Pour une démarche centrée sur l'utilisateur dans les ENT. Apport au Traitement Automatique des Langues.

Pierre Beust

HABILITATION À DIRIGER DES RECHERCHES

de l'Université de Caen Basse-Normandie

présentée et soutenue publiquement le 3 avril 2013 par

Pierre BEUST

Pour une démarche centrée sur l'utilisateur dans les Environnements Numériques de Travail : apport au Traitement Automatique des Langues

Composition du jury :

Prof. Jean-Claude BERTIN, Université du Havre, Rapporteur
Prof. Pierre DE LOOR, Ecole Nationale d'Ingénieurs de Brest, Rapporteur
Prof. Luigi LANCIERI, Université de Lille 1, Rapporteur
Prof. Pascale SÉBILLOT, IRISA / INSA de Rennes, Rapporteur
Prof. Ioannis KANELLOS, ENST de Bretagne, Examinateur
Prof. Mathieu VALETTE, INaLCO, Examinateur
Docteur Nadine LUCAS (HDR), GREYC CNRS UMR 6072, Marraine de recherche

à Delphine, mon moteur pour avancer …
à Cécile et Léna, mes deux jolies petites belles …

Remerciements

En premier lieu je tiens à remercier Nadine Lucas et à lui exprimer toute ma gratitude pour m'avoir accompagné, conseillé et soutenu (c'est rien de le dire !) dans ce parcours de préparation de l'habilitation qui, à l'inverse de ce que je pensais, était clairsemé d'embûches.

Je remercie vivement Pascale Sébillot, Pierre De Loor, Luigi Lancieri et Jean-Claude Bertin pour avoir accepté d'évaluer ce travail. Chacun d'eux apporte par sa thématique de recherche un éclairage particulier. J'adresse également un remerciement sincère aux autres membres de mon jury. Ioannis Kanellos a suivi dés le début les cheminements de mon travail d'HDR et il a toujours su être bienveillant et de bons conseils. Mathieu Valette a contribué directement au « coup d'envoi » de ce travail. Il m'avait fait l'honneur et le plaisir de m'inviter pour animer un séminaire à l'INaLCO afin de présenter un bilan des années de recherche passées. Dans l'assistance, il y avait entre autres François Rastier dont l'ampleur de la réflexion scientifique et son apport dans mon travail sont à l'égal de ses grandes qualités humaines. Tous deux m'ont convaincu qu'il y avait là toutes les pièces d'un puzzle qui préfiguraient une préparation de l'HDR. Merci de m'y avoir poussé quand je pensais que je n'arriverais jamais à trouver le temps de m'y mettre ...

En pensant aujourd'hui au chemin parcouru, je remercie spécialement Gaël Dias qui, tout fraîchement arrivé du Portugal pour prendre la direction de l'équipe de recherche à laquelle j'appartiens au GREYC (équipe HULTECH), a dû plaider la cause d'un nouveau collègue atypique en pleine difficulté.

Ce travail d'HDR doit beaucoup à tous ceux avec qui j'ai partagé mes idées, mes doutes, mes projets. Le présent mémoire retrace à sa façon toutes les interactions, tous les échanges de points de vue, tout le temps agréablement passé avec un bon nombre de collègues de disciplines souvent très variées. Je pense notamment à mes très chers collègues et amis du groupe « indiscipliné » *Nouveaux Usages* et plus particulièrement, et pour beaucoup plus encore que les aspects scientifiques, à Maryvonne Holzem, Jacques Labiche, Stéphane Ferrari et Denis Jacquet.

Je remercie spécialement ici mes deux amis et partenaires d'aventures extra-universitaires, Serge Mauger et Roger Cozien. Notamment, merci à Serge pour ses relectures en altitude …

Merci aussi à mes collègues du CEMU dont le directeur n'avait peut-être pas la meilleure disponibilité lorsqu'il était en pleine rédaction.

Enfin, et surtout, merci à mes proches, mon épouse et mes deux filles, qui m'auront vu tout le long de la préparation de cette HDR me battre contre le temps, contre moi-même, contre les obstacles en tout genre ...

Table des matières

Nous sommes condamnés ainsi au sens, dans la mesure où nous sommes condamnés à l'action, fût-elle méditative (F. Rastier)

1. Introduction

Ce mémoire d'habilitation à diriger des recherches se donne principalement deux objectifs. Le premier est de faire un bilan des projets de recherche et de développement réalisés depuis ma thèse de doctorat soutenue en 1998 et mon recrutement dans l'enseignement supérieur en 1999. Le second consiste, à la lumière du bilan, à identifier des spécificités fortes de l'approche centrée-utilisateur que je défends pour qu'en découlent des perspectives à venir.

Le lien principal qui unit mes travaux de recherche et mes implications dans la pédagogie numérique est incontestablement la place laissée à l'utilisateur de manière assumée. J'entends ici par utilisateur le sujet interprétant qui se trouve en interaction avec une machine (plus généralement un environnement numérique) dont il attend des fonctionnalités lui permettant de mettre à profit dans sa tâche des ressources, des corpus, des interactions humaines. Par la suite, en qualifiant ma démarche, j'utiliserai souvent la qualification condensée de « centrée-utilisateur » par facilité d'écriture pour indiquer en fait « centrée sur l'utilisateur ».

Tout au long de ce mémoire d'habilitation je chercherai à décrire les spécificités et les intérêts d'une approche centrée-utilisateur dans le domaine du traitement automatique des langues et dans le domaine des environnements numériques de travail. J'argumenterai quelques convictions que je pourrais très succinctement introduire par les quelques points suivants :

- Une approche centrée-utilisateur en informatique est importante et nécessaire dès que l'on cherche à instrumenter des tâches où une dimension sémiotique et langagière est centrale.

- Une approche centrée-utilisateur n'est absolument pas une approche de l'utilisateur isolé. Ses inter-relations avec les autres sont constitutives de ses points de vue sur la tâche qui fait l'objet

d'une instrumentation et l'approche centrée-utilisateur doit en rendre compte et inciter à plus d'interactions (notamment des interactions médiatisées par le numérique).

- L'approche centrée-utilisateur s'accompagne d'un changement de pratique dans les méthodes de conception informatique. Un environnement numérique ne s'évalue pas uniquement en terme de résultats de processus mais aussi en terme de couplage avec l'utilisateur. La recherche de ce couplage passe par une observation des usages en cours de conception. Il en découle souvent que les usages ne sont pas seulement le reflets des fonctionnalités prévues car une part importante du couplage tient au détournements et au contournements d'usages. Ce ne sont pas des épi-phénomènes et plus encore ce ne sont pas des « freins » au bon déroulement des fonctionnalités car ils sont souvent très vertueux. L'approche centrée-utilisateur doit chercher à les inciter plus qu'à les éviter.

- Une approche centrée-utilisateur ne doit pas chercher à remplacer l'utilisateur mais au contraire elle doit renforcer la place qui est la sienne. Il n'est donc pas question de chercher à extraire du sens en lieu et place de l'utilisateur mais au contraire à augmenter des conditions d'interprétation de l'utilisateur. L'approche centrée-utilisateur n'est pas une approche de l'utilisateur passif récipiendaire des résultats d'une application. C'est une approche de l'interprétation active où par ses capacités interprétatives et en couplage avec son environnement numérique l'utilisateur est « *créactif* ».

- L'approche centrée-utilisateur n'est pas un enjeu de l'informatique uniquement. C'est une approche incontestablement pluridisciplinaire.

1.1. Contours disciplinaires

Mes travaux de recherche s'intéressent aux environnements informatisés où les questions d'accès au sens (principalement des documents électroniques) sont au centre des tâches et des interactions entre l'environnement et l'utilisateur. Ces environnements informatisés sont principalement décrits par le terme générique d'Environnements Numériques de Travail (ENT). Cette habilitation se veut en premier lieu une contribution au domaine du Traitement Automatique des Langues (TAL). La problématique des ENT dépasse largement la problématique du TAL car dans bon nombre d'environnements de travail numériques (par exemple un environnement de GED[1]) on n'attend pas de l'environnement une assistance en terme d'accès au contenu.

Aborder en TAL la question de l'adaptation à l'utilisateur et des ENT a pour conséquence de rapprocher certains courants de l'informatique qui classiquement n'ont pas beaucoup d'interactions entre eux. Par exemple, il convient de rapprocher le champ du TAL et de celui des IHM (Interactions Homme-Machine) qui ont assez peu de recouvrements. D'autre part, un domaine de l'informatique qui de longue date étudie des ENT est celui des EIAH (Environnements Informatiques pour l'Apprentissage Humain). Il convient donc aussi de rapprocher le TAL et les EIAH.

Comme je l'expliquerai tout au long de ce mémoire, mettre en avant une approche centrée-utilisateur amène à repréciser des notions ou questions centrales dans le domaine du TAL comme par

1 Les systèmes de Gestion Électronique de Documents (GED) visent principalement l'assistance aux flux de création et de partages de documents numériques indépendamment de leur contenu en insistant notamment sur l'archivage, les différentes versions d'un documents, les vues possibles sur tels ou tels documents ...

exemple le sens, la signification, les ressources (principalement ressources lexicales) ou encore l'évaluation. Ces questions essentielles nous placent au sein d'un positionnement épistémologique par nature pluridisciplinaire. La problématique de l'interprétation est omniprésente et elle nous incite à tirer des ponts entre disciplines :

- ponts entre l'informatique et la linguistique, et plus précisément au sein de la linguistique la sémiotique et le courant de la sémantique interprétative de François Rastier ;

- ponts entre l'informatique et les sciences cognitives, et plus précisément au sein des sciences cognitives le courant de l'énaction ;

- ponts entre l'informatique et des sciences de l'éducation, et plus précisément les travaux sur les dispositifs numériques dans les sciences de l'éducation ;

- mais également, ponts entre l'informatique et l'histoire des sciences, et plus précisément le courant de la sérendipité qui étudie les découvertes fortuites où l'on fait le constat des richesses de l'interprétation. Notamment notre problématique des détournements et contournements d'usages a très probablement à être considérée de la même manière que la sérendipité.

L'approche centrée-utilisateur est donc une approche de décloisonnent disciplinaire à la fois au sein de l'informatique entre les domaines du TAL, de l'IHM et des EIAH et à la fois entre disciplines elles-mêmes (informatique, linguistique, sciences cognitives, sciences de l'éducation, ...). Plus qu'une complexité, ce décloisonnement doit être pris en compte comme une richesse en terme de pistes de recherches. Nous aurons donc à cœur de ne pas chercher à dénaturer l'approche centrée-utilisateur par une réduction de complexité trop simplificatrice.

1.2. Parcours professionnel

La problématique de l'utilisateur s'est imposée à moi de manière naturelle et presque insidieuse lors des différentes années de mon parcours professionnel, tant comme une évidence dans mes activités d'enseignant (chaque étudiant étant toujours très différent d'un autre dans sa façon de recevoir le message de l'enseignant) que dans mes missions en terme de TICE[2] à l'université et bien sûr comme un point commun dans tous les projets de recherche auxquels j'ai participé ou bien que j'ai encadrés. Il me semble donc utile en quelques lignes de rappeler les principales étapes de ce parcours (cf. le CV en annexe pour les détails) avant d'aborder plus en détail mes objectifs et ma problématique pour, au terme de cette introduction, annoncer les grandes parties de ce mémoire.

Je suis aujourd'hui maître de conférences à l'université de Caen Basse-Normandie. J'ai été recruté et 1999 (et titularisé en 2000). Je suis attaché à l'UFR des LVE pour mes activités d'enseignement et j'effectue mes activités de recherche au sein du laboratoire GREYC CNRS UMR 6072 où j'appartiens à l'équipe HULTECH (technologie du langage humain). Au sein de cette équipe, je participe principalement au thème « Interactions » et également au thème « Sémantique ».

J'ai suivi un parcours d'études universitaires de mathématiques et d'informatique classique à

2 Technologies de l'Information et de la Communication pour l'Enseignement

l'université de Caen Basse-Normandie jusqu'à un DEA en intelligence artificielle. Je suis alors parti faire un service national sous la forme d'un Volontariat à l'Aide Technique (VAT). J'étais affecté à l'université des Antilles et de la Guyane à Fort-de-France en Martinique. J'y ai enseigné l'informatique à des étudiants de Lettres et j'ai participé aux travaux de recherche du jeune laboratoire GIL (Groupe d'Informatique Linguistique) dirigé par Jacques Coursil. Ma rencontre avec Jacques Coursil est un élément qui a été déterminant pour la suite de mon parcours professionnel. Il m'a donné l'envie et le goût de la recherche en pluridisciplinarité, m'a initié aux théories linguistiques post-saussuriennes et m'a incité à me lancer dans une thèse de doctorat en me donnant les principales pistes de ce qui allait déterminer mon sujet de thèse. Pour tout cela et bien plus encore je voudrais ici l'en remercier très chaleureusement.

En 1995 de retour à Caen, j'ai commencé une thèse de doctorat (Beust 1998) en informatique sur la représentation lexicale en appliquant la méthode analytique de Jacques Coursil (méthode appelée Anadia, cf. chapitre 2). Cette thèse s'est faite sous une direction conjointe et pluridisciplinaire : Anne Nicolle, professeur en informatique et directrice de thèse et Laurent Gosselin, professeur en linguistique et co-directeur de thèse. Chacun d'eux m'a apporté des aspects essentiels qui ont façonné ma façon de travailler. Ainsi je remercie Laurent Gosselin pour m'avoir montré comment une théorie linguistique peut répondre à la question d'une description opératoire des variations de signification en contexte. Je remercie également Anne Nicolle pour m'avoir fait prendre conscience des richesses des interactions dans les développements informatiques. Elle m'a dès le début orienté vers des modèles interactionnistes qui aujourd'hui encore ne sont pas si fréquents que cela dans le domaine du TAL. Introduire l'interaction dans une thèse en informatique sur la sémantique des langues, c'était déjà se positionner dans une approche résolument centrée-utilisateur, ce qui m'est apparu avec plus d'évidence encore quelques années après la soutenance de ma thèse. A l'issue de mon travail de thèse, j'ai été qualifié dans les deux sections CNU informatique (27e) et sciences du langage (7e) ce qui représentait pour moi une double reconnaissance institutionnelle importante. Je mesure ainsi comment une problématique de doctorant peut s'enrichir d'une direction conjointe pluridisciplinaire et je pense que c'est un modèle à développer encore.

Depuis mon recrutement en tant que maitre de conférences, j'enseigne principalement à l'UFR des Langues Vivantes Etrangères (LVE) de l'université de Caen Basse-Normandie et plus précisément au sein du département Langue Étrangères Appliquées (LEA). J'assure des enseignements en bureautique (traitements de texte, tableurs, PréAO), en base de données (SQL et ACSI), en culture numérique Web et en veille documentaire. Au sein de l'UFR Sciences j'ai assuré des enseignements en systèmes à base de connaissances. Au sein de l'UFR des Sciences de l'Homme j'ai en charge un enseignement d'initiation au TAL. Enfin, à l'UFR de Droit et Sciences politiques j'ai en charge une unité d'enseignement de préparation au C2I.

Depuis 2006, l'université de Caen Basse-Normandie m'a chargé de mettre en œuvre sur l'ensemble de l'université le Certificat Informatique et Internet (C2I) niveau 1. Le C2I-1 est une certification nationale adossée à un référentiel de compétences qui vise à attester d'un niveau d'usage courant des outils informatiques et bureautiques et de l'internet. En tant que correspondant local à Caen pour le C2I-1, je participe à des réunions nationales regroupant les experts du domaine pour chacune des universités dans le but d'harmoniser les pratiques de formation et de certification et de mutualiser des contenus. D'un point de vue strictement universitaire, la mise en place du C2I-1 est un changement culturel. La question de la mutualisation nationale en est une première manifestation car souvent l'enseignant cherche à produire un enseignement dans lequel il fait passer un point de vue qui lui est propre. D'autre part la logique d'une certification et d'un référenciel de compétences est quelque chose qui est à l'opposé des pratiques classiques dans la gestion des diplômes universitaires courants. Par

exemple, pour réussir un diplôme il faut y obtenir au moins la moyenne alors que pour être certifié relativement à un référentiel il faut avoir validé toutes les compétences du référentiel. C'est bien sûr un changement culturel pour les enseignants mais également pour les étudiants. Par capitalisation des compétences validées les étudiants peuvent obtenir leur C2I sur plusieurs années lors de leur parcours. Là aussi, c'est un changement culturel car la logique annuelle et même semestrielle habituellement adoptée ne convient plus. Cela nous amène à changer nos idées sur la gestion des parcours de formation et notamment sur la quantification des échecs (par exemple, un étudiant auquel il ne manque que très peu de compétences à valider ne sera pas certifié cette année là mais pour autant il n'est pas en situation d'échec car il lui sera beaucoup plus facile d'être certifié l'année suivante en ne validant que les compétences qui lui manquent). En dehors des difficultés techniques et organisationnelles, le C2I est un défi intéressant relativement aux mutations qui attendent les milieux universitaires.

Le C2I Niveau 1 est un supplément au diplôme de licence, c'est-à-dire qu'il n'est pas obligatoire à la validation de la licence, mais il est vivement conseillé aux étudiants le plus tôt possible dans leur parcours pour leur donner des compétences numériques utiles à leur activité d'étudiant (par exemple, savoir maîtriser un environnement informatique, savoir produire et structurer professionnellement des documents électroniques, savoir effectuer des calculs dans un tableur, savoir mettre en œuvre une recherche d'information, savoir faire du travail collaboratif en ligne). Les effectifs de la formation C2I-1 sont très importants car par nature les étudiants de toutes les filières universitaires peuvent être concernés. En 2011-2012, nous comptions 1866 inscrits au C2I-1 à l'université de Caen Basse-Normandie. Pour former et certifier un tel volume d'étudiant, nous avons mis en place une plateforme de formation et de certification en ligne regroupant des vidéos de cours, des exercices, des tests de positionnement, des ressources multimédia ainsi que des moyens d'échanges tels que des forums, des séances de chat avec des tuteurs. Ce dispositif est basé sur le LMS (Learning Management System) libre Moodle[3] et donne entière satisfaction en permettant bien de former et de certifier un grand volume d'étudiants que nous ne pourrions pas réunir à un même endroit et à un même moment. En cela, l'expérience de la responsabilité du C2I-1 est intéressante car elle permet de faire tomber le paradoxe qui consisterait à énoncer que pour apprendre à utiliser des ENT il faudrait déjà savoir utiliser un ENT. En fait, on constate que l'apprentissage est concomitant à l'usage et qu'il résulte d'un couplage entre l'étudiant/utilisateur et le système qui lui est fourni. Ceci fait écho aux questions d'adaptation d'un utilisateur à son environnement et cette question est incontournable dans une approche centrée-utilisateur. Nous l'aborderons à nouveau plus tard dans ce mémoire quand nous examinerons comment la théorie de l'énaction nous apporte des éléments de réponse.

La problématique de l'usage des ENT dans les parcours étudiants relève de ce qu'on appelle aujourd'hui les TICE, Technologies de l'Information et de la Communication appliquées à l'Enseignement. Le développement des TICE permet de repenser l'organisation pédagogique et le suivi des étudiants en dehors des heures de présence en cours ou TD. Il en découle une palette de possibilités de parcours diplômants entre les formations intégralement présentielles, les formations mixtes avec présentiel et activités en ligne et enfin les formations intégralement à distance (la FOAD). Etant impliqué dans cette réflexion à l'université de Caen Basse-Normandie, il m'a été proposé de me porter candidat à direction du CEMU (Centre d'Enseignement Multimédia Universitaire), direction à laquelle j'ai été élu en février 2009. Le CEMU est une composante de l'université de Caen. C'est un service commun qui a une double mission :

- La Formation Ouverte À Distance (FOAD) :
  - Opérer et mettre en valeur l'offre de formation à distance de l'établissement
  - Mettre en œuvre des projets de formation continue à distance

3 Modular Object-Oriented Dynamic Learning Environment : http://www.moodle.org consultée le 17/11/12

  - Accompagner les composantes dans des projets de FOAD

- Les Technologies de l'Information et de Communication pour l'Enseignement (TICE) :
  - Faire bénéficier les étudiants de l'établissement de moyens numériques pour l'accompagnent en ligne en dehors du temps d'enseignement présentiel (plateforme Moodle par exemple)
  - Former les étudiants aux TICE via l'organisation du C2I Niveau 1
  - Offrir aux enseignants des outils et des compétences pour des activités pédagogiques en ligne : production et diffusion de contenus multimédia (podcast et ressources en compléments de cours), scénarisation pédagogique des enseignements.
  - Contribuer aux projets e-Learning de l'établissement (ENT, UNR[4], UNT[5] etc.)

Le CEMU est un service regroupant 22 personnes ayant des métiers divers : techniciens, ingénieurs, infographistes, enseignants, personnels administratifs. Il accompagne les enseignants et les décideurs en charge de la politique d'établissement dans les profondes mutations des métiers universitaires induites par l'essor du numérique et le développement de la formation continue (qui se fait le plus souvent dans des dispositifs où la FOAD est tout à fait adaptée).

Dans mes projets de recherche au sein du laboratoire GREYC, d'abord dans l'ex-équipe ISLanD (Interactions, Sémiotique, Langues, Diagrammes), puis dans l'équipe DLU (Documents, Langues, Usages) et maintenant dans l'équipe HULTECH, je cherche également à envisager l'essor du numérique dans les ENT et à proposer des solutions au service de l'utilisateur. Je détaillerai plus loin dans ce rapport plusieurs projets de développement logiciel qui ont été, ou continuent à aller, dans ce sens. Le travail de recherche en équipe est incontournable (en plus d'être agréable) et c'est même encore plus vrai lorsque le travail de recherche est la croisée de plusieurs disciplines. Je présenterai également dans ce rapport les différentes collaborations que j'ai eues dans ces dernières années. Au titre des plus importantes, j'ai participé à la directions de deux thèses de doctorats qui pour chacune d'elles ont apporté une contribution réelle et importante à ma problèmatique des approches centrées sur les utilisateurs :

- **Thibault Roy**, Doctorant allocataire moniteur (co-direction P. Beust à 80%, J. Vergne à 20%)
  - Titre de la Thèse : « Visualisation interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires »
  - Début de thèse : 1/10/2004
  - Date de soutenance : 17/10/2007,
  - thèse obtenue avec la mention très honorable
  - Jury : J. Vergne (directeur de thèse), P. Beust (co-directeur), B. Habert (rapporteur), P. Zweigenbaum (rapporteur), A. Nazarenko.
- **Vincent Perlerin**, Doctorant allocataire moniteur (co-direction P. Beust à 50%, A. Nicolle à 50%)
  - Titre de la Thèse : « Sémantique légère pour le document - Assistance personnalisée pour l'accès au document et l'exploration de son contenu»
  - Début de thèse : 1/10/2000
  - Date de soutenance: 7/12/2004, thèse obtenue avec la mention très honorable
  - Jury : A. Nicolle (directeur de thèse), P. Beust (co-directeur), B. Habert (rapporteur), I. Kanellos (rapporteur), F. Rastier, P. Sébillot.

4 Universités Numériques en Région

5 Universités Numériques Thématiques. Cf. http://www.universites-numeriques.fr consultée le 7/01/13.

1.3. Plan du mémoire

A la suite de cette première partie introductive du rapport, je structurerai le contenu en quatre chapitres avant de terminer sur un chapitre conclusif suivi d'une ultime partie sur les projets qui s'ouvrent dans la suite du travail.

Le deuxième chapitre vise un effet de « zoom réthorique » pour introduire la problématique qui est la mienne en la situant, bien sûr, au sein de l'informatique (quoique pas exclusivement), mais plus précisément du TAL, puis enfin des méthodes d'accès au contenu. Cet effet de zoom m'amènera à tracer les contours des rapports à l'utilisateur dans la question de l'accès au contenu.

Dans le troisième chapitre, je m'attacherai à dresser un bilan chronologique des différents projets de recherche et de développement auxquels j'ai participé ou que j'ai encadrés dans les dernières années. Je présenterai quelques expérimentations menées sur corpus lors de ces projets. Ce chapitre se concluera sur une description d'un projet en cours dans un contexte industriel avec la même approche centrée-utilisateur.

Le quatrième chapitre cherchera à rendre clairs les principes et méthodes au cœur de la démarche centrée-utilisateur. Nous y présenterons en quoi notre approche est diamétralement opposée à un courant dominant du TAL, celui des systèmes à base de connaissances ontologiques tels que ceux qui sont notamment envisagés dans les travaux se rapportant au Web Sémantique. Dans le domaine des innovations technologiques et sociologiques de l'internet dans les dernières années, un autre courant a émergé dernièrement, celui du Web 2.0, appelé ainsi par abus de langage. Je m'y arrêterai en expliquant que les principes du Web 2.0 sont très cohérents avec une problématique des ENT centrés-utilisateurs. En cela ils prouvent qu'il y a des alternatives crédibles aux courants du tout ontologique et tout automatique. Ce chapitre conclura son propos sur le choix de convoquer en TAL la théorie linguistique de la Sémantique Interprétative de François Rastier comme le cadre opératoire et épistémologique dans lequel l'approche centrée-utilisateur se positionne.

En face de l'ENT, il y a l'utilisateur. Entre les deux, il y a l'interaction et le couplage. Nous chercherons dans le cinquième chapitre à expliquer que la théorie cognitive de l'énaction est celle qui explique le mieux ce rapport de couplage entre un utilisateur et son environnement logiciel. Dans une approche énactive de l'interprétation des documents électroniques il convient de penser différemment les systèmes à concevoir. Plutôt que de viser un usage particulier, il semble préférable ne pas contraindre au préalable des besoins et des finalités des utilisateurs car justement, de manière énactive, ils se définissent dans l'interaction et c'est en partie cela qui créé du couplage et qu'il est particulièrement intéressant d'observer et d'étudier de manière pluridisciplinaire. Nous rejoignons ici l'idée de D. Dionisi et J. Labiche (Dionisi & Labiche 2006) qui consiste à caractériser des « processus logiciels » impliqués dans des « processus expérientiels », eux-mêmes impliquant des « processus cognitifs ». Nous chercherons à faciliter les différents usages vertueux des outils logiciels même quand ces usages sont en fait assimilables à des contournements ou des détournements d'usage (qui traditionnellement dans la conception informatique ne sont jamais entrevus et encore moins de manière vertueuse). Il en découlera qu'il convient de chercher à développer des ENT qui cherchent avant tout à rendre immédiate une appropriation par l'utilisateur et à faire émerger de manière énactive des usages observables. C'est ce que nous appellerons des ENT énactifs. Nous détaillerons le projet AIDé en cours de réalisation qui est un exemple d'ENT énactif.

~

Au terme de ce rapport il est fort probable que plus de questions auront été posées que de réponses apportées. Ce n'est pas un problème en soi car le domaine de recherche qui est le nôtre a certainement beaucoup plus à gagner des doutes et des interrogations que des certitudes. Dans le domaine de l'informatique et plus précisément du TAL, les certitudes se sont presque systématiquement effondrées d'elles mêmes le temps passant, comme en témoigne l'histoire de la traduction automatique ou encore du dialogue homme-machine. De ce point de vue, nous considérons donc qu'être à même de diriger des recherches ne revient surtout pas à orienter un chercheur dans telle ou telle direction mais plutôt d'être une sorte de contradicteur bienveillant qui cherche à œuvrer dans la prise de conscience des non-certitudes.

2. Problématique de recherche

Mes travaux prennent place institutionnellement en informatique (27e section du CNU) et en tant que tels au sein des sciences de l'ingénieur (département SPI du CNRS). Ce rattachement disciplinaire est essentiel car la conception d'environnements numériques est bien indiscutablement une problématique informatique. Mais ce positionnement disciplinaire n'est pas exclusif car dans une démarche centrée-utilisateur de l'accès au contenu des documents textuels, les apports de la linguistique et des sciences cognitives sont déterminants dans la mesure où en face de la machine il y a un utilisateur dont il faut comprendre les attentes et dont il faut assister les compétences interprétatives plutôt que de chercher à les reproduire. De plus, le champ théorique de la sémiotique est des plus importants là où la question est d'aborder les phénomènes de sens et de signification dans les applications informatiques. Enfin, le champ des interactions homme-machine (dont les rapports avec le TAL sont souvent assez éloignés) est lui aussi par nature un domaine d'investigation de premier plan. Loin de considérer que nous allons aborder une problématique intégralement cernée au sein de l'informatique par des apports des sciences humaines, nous préférons militer pour une vision scientifiquement vaste de la problématique informatique qui va largement jusqu'aux sciences humaines.

Parmi les courants de l'informatique qui sont nombreux, nous nous positionnons dans le champ communément appelé Traitement Automatique des Langues (TAL), acronyme traduisant en français le champ anglo-saxon du *Natural Language Processing* (NLP). Le TAL regroupe les travaux, principalement entre l'informatique et la linguistique, qui cherchent à développer des processus ou des ressources pour donner à une machine une compétence particulière dans la façon de traiter des données linguistiques (textes, paroles, énoncés, documents …).

Le TAL est lui aussi (à l'instar de l'informatique dont il est un sous-domaine) un champ d'activités

scientifiques très étendu. Premièrement, les travaux en TAL peuvent être aussi bien des travaux de recherche fondamentaux théoriques (en témoigne par exemple des recherches sur les ontologies) que des réalisations industrielles apportant des solutions à des problèmes applicatifs très concrets (la correction orthographique et grammaticale dans les applications bureautiques par exemple). Deuxièmement, les objets d'études en TAL sont très diversifiés et vont quasiment du traitement du signal et de la phonétique jusqu'aux caractérisations du contexte dans des approches de la pragmatique, et de l'analyse de requêtes faites de mots-clés jusqu'au dialogue homme-machine.

Au sein du TAL, les questions qui m'intéressent sont celles qui abordent les notions de sens, de signification, d'interprétation et d'inter-compréhension. Ces questions d'ordre sémantique et pragmatique relèvent d'une branche du TAL qu'on identifie souvent par l'appellation « accès au contenu ». Je vais détailler les grandes directions prises dans les méthodes d'accès au contenu en TAL et je décrirai ensuite la question, pour moi incontournable, du rapport entre le contenu et l'utilisateur/interprétant.

2.1. L'accès au contenu

Les technologies de l'information (notamment sur l'Internet) forment un domaine d'application direct du TAL et plus précisément de l'accès au contenu des documents. La taille des données textuelles à traiter ainsi que le nombre et la variété des traitements à réaliser rendent incontournable le développement de méthodes d'analyses automatiques les plus fiables possibles et rapides. A titre d'exemple, on peut se rappeler que le fameux moteur de recherche Google indexe aujourd'hui environ 8 milliards de documents et estimait déjà en février 2005 traiter environ 250 millions de requêtes par jour (dernièrement on estime à 34 000 par seconde le flux de requêtes traitées par Google[6]).

Plusieurs types d'outils de TAL sont spécifiquement dédiés à la problématique de l'accès au contenu du document. Ils constituent une évolution majeure du TAL aujourd'hui. Dans certains cas y sont réinvestis des travaux sur la compréhension des textes provenant de la tradition logico-grammaticale (c'est par exemple le cas des systèmes mis en compétition dans le cadre des conférences MUC). Dans d'autres cas, on observe des démarches plus pragmatiques qui tentent de tirer profit de larges corpus et de méthodes d'apprentissage automatiques (Claveau, 2003). Nous reviendrons sur ces différences d'approches qui marquent différents courants du TAL dans le chapitre n°4.

Adeline Nazarenko dans (Condamines & al., 2005, Chap. 6) établit quatre familles de méthodes automatiques d'accès au contenu des documents :

- l'extraction d'information,
- les méthodes de Question/Réponse,
- le résumé automatique,
- l'aide à la navigation.

On entend par extraction d'information les méthodes qui consistent à rechercher dans un corpus très homogène en genre et en langue (par exemple des dépêches d'actualité ou encore des articles scientifiques, souvent en anglais) des informations dont on sait qu'elles s'y trouvent. Ainsi on cherche par exemple dans un corpus d'actualité boursière à extraire les transactions de rachat et de fusion de

6 Blog Média et technologies (25 janvier 2010) : http://media-tech.blogspot.com

sociétés ce qui revient à chercher à remplir des sortes de formulaires électroniques indiquant notamment qui a acheté qui, à quel prix et quand. Il s'agit donc ici d'alimenter de manière automatique des bases de données préexistantes à partir de corpus soigneusement sélectionnés. Les méthodes dites de Questions/Réponses n'ont pas le même objectif. Elles consistent à chercher un fragment de texte extrait d'un corpus volontairement assez généraliste dans lequel un sujet interprétant a de bonnes chances de trouver la réponse à une question qu'il aura formulée en langue naturelle. Par exemple extraire une séquence du style « (…) la vie de Baudelaire, auteur des *Fleurs du mal*, fut (…) » à la question « Qui a écrit les *Fleurs du mal* ? ». La bonne construction linguistique de la réponse n'est pas ici visée car il ne s'agit que de fournir une « fenêtre » appartenant à un texte censée répondre à la question posée, éventuellement en essayant tout de même de ne pas couper des mots en leur milieu. Lors des conférences d'évaluation TREC9, les systèmes de Questions/Réponses avaient pour consigne de rendre des réponses de moins de 250 caractères à partir de 980 000 documents et de 700 questions. A la différence des méthodes d'extraction d'information, on s'en remet à l'interprétation d'un sujet humain quant à la qualité des réponses trouvées. Les méthodes de résumé automatique s'appuient aussi largement sur l'interprétation de celui à qui est destiné le résumé. Bien souvent il est plus juste de parler de condensation ou de réduction de textes plutôt que de résumé (dans le sens de ce qu'est un résumé quand il est rédigé par un sujet humain). L'enjeu technique est de rechercher des phrases dont on pense qu'elles ont un statut assez significatif (par exemple une phrase qui commencerait par « en somme, on constate que (…) » a de bonnes chances de synthétiser ce qui est dit avant) et de les juxtaposer dans un « résumé » dont on fait l'hypothèse que celui qui le lira pourra rétablir une certaine cohérence textuelle, par exemple relativement aux rattachements anaphoriques.

Les méthodes d'extraction d'information, de Question/Réponse et de résumé automatique s'adressent principalement à la dimension rhématique des documents en cherchant d'une certaine façon à savoir ce qui est dit, où et comment. En général, les méthodes d'aide à la navigation s'adressent plus spécifiquement à la dimension thématique des documents (dans le sens où l'on cherche de manière plus globale à savoir de quoi traite un document ou un ensemble de documents). Les applications les plus courantes de ces méthodes sont l'indexation de document, l'extraction de terminologies, l'aide à la lecture (visualisation de documents ou encore création d'index par exemple), le groupement en classes de documents, la cartographie de corpus.

Les quatre familles de méthodes d'accès au contenu présentées ci-dessus regroupent des projets de recherche où sont mis en œuvre beaucoup d'intelligence du point de vue des collaborations interdisciplinaires, notamment entre la linguistique et l'informatique. Cependant force est de constater que peu d'entre eux sont mis en application et évalués dans des outils sur Internet à destination du plus grand nombre. Cela a des conséquences comme le montrent Lavenus & al. (2002) à propos des méthodes de Question/Réponse en mettant en évidence la différence entre les corpus de référence utilisés dans les conférences TREC par rapport à des vraies questions d'utilisateurs en recherche documentaire. Les auteurs notent que les questions du corpus de référence sont toutes des interrogatives canoniques courtes (par exemple « What does a defibrillator do ? ») alors que la majorité des demandes de « vrais » utilisateurs sont couramment des affirmatives complexes du style « je voudrais savoir (…) ».

Paradoxalement, si d'un point de vue informatique et algorithmique les méthodes couramment utilisées notamment par les moteurs de recherche sont très fines et fiables, on constate effectivement qu'elles restent linguistiquement relativement pauvres, à la fois du point de vue de leur fonctionnement propre mais également du point de vue de l'interaction avec leurs utilisateurs. Le recours à l'usage de mots clés, éventuellement agencés dans des requêtes booléennes, reste souvent la seule façon de voir la recherche d'information. Nous verrons (au chapitre n°3) que nous pouvons

proposer d'autres méthodes pour cela.

Les méthodes d'indexation utilisées par les moteurs de recherche pour associer des documents à des mots clés potentiels sont un exemple de la « pauvreté » linguistique des méthodes. Cette indexation est dite « *Full Text* » dans le sens où tous les mots figurant dans un document sont gardés comme entrée d'index pour ce document. Pas étonnant dans ces conditions que les mots grammaticaux indexent potentiellement une multitude de documents (expérience faite sur Google.fr le 17/11/2012 : une recherche stupide avec le mot clé unique « de » donne 25 270 000 000 réponses, ce qui par ailleurs est une estimation évidemment fausse quand on sait que la base d'index est de 8 milliards de documents, soit 3 fois moins). Ceci a des inconvénients, notamment la taille énorme des bases d'index que le moteur de recherche doit archiver et doit être capable d'interroger rapidement. En fait, l'intérêt de cette indexation un peu brutale réside dans le fait de pouvoir garder facilement comme index tout ce qui dans un texte ne peut être retrouvé dans un dictionnaire. C'est surtout le cas des entités nommées telles que des noms propres, des expressions temporelles par exemple, qui restent importants par rapport aux documents (on imagine mal par exemple que le nom d'une société ne puisse pas être gardé comme entrée d'index pour son site web) mais dont il est difficile de dresser un catalogue fiable et durable. La question du repérage et même de l'étiquetage (en tant que nom d'organisation ou de nom de lieu par exemple) des entités nommées est un enjeu important[7] du TAL aujourd'hui et de nombreux projets de recherche abordent cette question avec des résultats intéressants mais leurs avancées n'ont pas encore eu de retombées pour affiner les méthodes d'indexation utilisées sur Internet, notamment parce que certaines graphies d'entités nommées ne sont pas prédictibles (elles varient avec l'actualité).

Dans leurs interactions avec les utilisateurs, les moteurs de recherche sont souvent assez rudimentaires d'un point de vue linguistique. Il faut bien souligner que l'utilisateur et son objectif de recherche sont uniquement considérés sous la forme d'une liste de mots clés (dont la casse et l'accentuation et même l'ordre sont d'ailleurs rarement pris en compte) considérés pour une seule recherche dans la mesure où toutes les requêtes sont traitées indépendamment les unes des autres. Dans la pratique on s'aperçoit que pour mener à bien une recherche sur le web, il convient en fait d'interroger successivement plusieurs fois le (ou les) moteur(s) en ajoutant ou en précisant certains mots clés en fonction des résultats rendus à chaque étape. C'est donc le plus souvent à l'utilisateur seul qu'il revient de développer des stratégies efficaces pour trouver des mots clés adaptés à sa recherche. Certaines tentatives sont mises en place par certains moteurs pour aller un peu plus loin que la simple prise en compte de mots clés. Par exemple, Google permet de rechercher un mots clé ou un de ses synonymes avec l'opérateur tilde ~ (par exemple une recherche sur *powerpoint ~help* effectuera une recherche sur *powerpoint* ET *help* ou *tips*, *faq*, *tutorial*). Cependant, c'est le moteur lui-même qui établit ses listes de synonymes et il serait peut être plus judicieux que celles-ci soit validées par les utilisateurs quand ils les utilisent. Il convient donc, en tant qu'utilisateur, de rester très prudent quant aux compétences linguistiques des moteurs. Toujours à propos de Google, on trouve un exemple de résultat assez malheureux de l'opérateur define sur le blog de Jean Véronis[8]. L'opérateur *define* (disponible pour les pages en français depuis avril 2005) sert à rechercher à propos d'un mot des pages Web où ce mot ferait visiblement l'objet d'une définition. L'expérience relatée consiste à rechercher ainsi sur Google une définition du mot femme avec la requête *define:femme*. Les résultats donnés sont très contestables. On aurait donc bien tort de croire à la fiabilité de l'opérateur *define* (qui pourtant est présenté par Google comme un outil de recherche de définition sans plus de détails) comme on aurait tort aussi de considérer le Web dans son ensemble comme une encyclopédie dans lequel on puisse rechercher des définitions attestées, notamment d'un point de vue moral. En matière d'ingénierie

7 cf. http://www.slideshare.net/aixtal/jean-vronis-2010-seo-campus-3176089 consultée le 17/11/12

8 cf. http://blog.veronis.fr/ consultée le 17/11/12.

documentaire la tendance actuelle est pourtant de renforcer ce genre d'utilisations du Web en cherchant à en faire une vaste base de connaissances, ce qu'évidemment il n'est pas. C'est la démarche considérée dans le projet du Web Sémantique dont nous reparlerons au chapitre n°4.

2.2. Le rapport à l'utilisateur

Les méthodes d'accès au contenu que nous avons évoquées précédemment ont pour point commun de tenter de vouloir limiter le plus possible l'intervention de l'utilisateur. Dans la démarche centrée utilisateur, on part d'une position radicalement opposée où l'on considère que les traitements sémantiques appliqués à l'accès au contenu des documents ont tout à gagner à être le plus possible subjectivés, tant du point de vue des ressources que du point de vue des résultats opératoires. Cette démarche nous paraît notamment être une réponse au constat que dressent Didier Bourigault et Nathalie Aussenac-Gilles à propos de la variabilité des terminologies :

> *(...) le constat de la variabilité des terminologies s'impose : étant donné un domaine d'activité, il n'y a pas une terminologie, qui représenterait le savoir du domaine, mais autant de ressources termino-ontologiques que d'applications dans lesquelles ces ressources sont utilisées (Bourigault & al., 2003, p. 24).*

Notre objectif scientifique est d'apporter une contribution aux développements des ENT où les problématiques sémiotiques et linguistiques sont des points clés de l'interaction homme-machine. Dans ce type d'ENT (par exemple, des environnements pour la veille documentaire, pour le *e-learning* ...), il convient de modéliser et d'expérimenter des ressources (principalement terminologiques) et des modèles d'analyse informatique des textes et du sens. Notre contribution réside principalement dans une approche pluridisciplinaire et résolument centrée sur l'utilisateur. Viser des formes d'instrumentations dites centrées utilisateurs consiste à construire des applications et des ressources manipulées avant tout autour des spécificités socio-linguistiques des utilisateurs. La priorité en terme de description concerne leurs centres d'intérêt, leurs habitudes terminologiques, leurs parcours interprétatifs dans les textes.

La question du sens et de l'accès au contenu des documents électroniques est de toute évidence très liée aux rapports entre ces documents (majoritairement textuels) et des utilisateurs travaillant avec ces documents et en produisant par ailleurs. L'idée au centre de notre travail est une certaine façon de considérer ce qu'est le sens dans des interactions homme-machine (que ce soit le sens d'un énoncé, d'une phrase, d'un texte, d'une collection de documents, une image, …) et plus généralement dans les interactions humaines médiatisées (ce qui est une autre extension du sigle IHM, cf. Lancieri 2009) : le sens est avant tout le fait d'une **interprétation**.

On se place ici dans la suite des travaux de Jacques Coursil qui montre que le sens est au moins autant du côté de l'interprétant que du côté du sujet parlant. Ainsi c'est parce que le sujet parlant est aussi le premier interprétant de ce qu'il dit, au moment où il le dit, qu'il peut donner forme à sa parole en continu. C'est le principe de non préméditation de la chaîne parlée (Coursil 2000). Il n'est bien sûr pas question ici de comprendre qu'un sujet parlant ne sait pas ce qu'il veut dire au moment où il commence à le dire mais plutôt qu'il ne sait pas exactement comment il va le dire avant de proférer une parole. Ce principe de non préméditation et une mise en évidence d'un couplage intéressant entre le sujet parlant/interprétant et l'environnement dans lequel il produit sa parole.

Dans ce rapport à l'interprétation, notre travail trouve naturellement un positionnement dans la

lignée des travaux en Sémantique Interprétative (SI) de François Rastier (Rastier 1987), elle-même en filiation avec les travaux en sémantique structurale dont l'origine remonte aux travaux de Saussure. Dans la SI le sens est vu comme une perception sémantique, perception forcément individuelle, dont toute tentative d'objectivation est une sommation incomplète de points de vue. Comme nous le verrons au chapitre n°4 de ce mémoire, la sémantique interprétative est un ancrage fondamental dans notre démarche. Elle apporte un modèle de description sémique des opérations interprétatives d'une grande richesse conceptuelle ainsi qu'un positionnement épistémologique très fort dans le champ des sciences de la culture.

Pour nous, la priorité en terme de description et de représentation est donnée aux spécificités socio-linguistiques des utilisateurs. C'est d'une personnalisation de l'environnement de travail que peut découler une meilleure compréhension par l'utilisateur de ce qui est effectif dans l'interaction et donc une meilleur appropriation, un meilleur couplage. Le couplage entre l'utilisateur et son environnement est un point clé de la problématique. Un ENT n'est pas un programme qu'on lance et dont on attend un résultat. A la manière d'un système d'exploitation ou encore d'une interface graphique, l'interaction dans un ENT n'est pas finalisée en soi, c'est-à-dire que le but de l'interaction est de maintenir l'interaction. Pour que l'utilisateur éprouve l'envie ou le besoin de prolonger à chaque instant l'interaction il faut qu'un couplage personne/système soit effectif et productif. Tel est le couplage s'il occasionne une émergence de sens. Une interaction qui relativement à son utilisateur ne produirait pas de sens deviendrait inutile à prolonger. Dans cette interaction, le sens ne peut pas être réduit à une représentation formelle calculée à un moment donné car il résulte du déroulement de l'interaction autant qu'il la conditionne.

Dans notre approche, le sens n'est pas le résultat d'un calcul, c'est une activité au centre d'une interaction (activité qui de plus n'est pas forcément préalablement finalisée dans le temps). Ainsi, nous remettons en cause l'idée que quelque chose ait du sens (ou non) pour défendre plutôt l'idée qu'il y a des choses qui font sens (ou pas) pour quelqu'un. On se rapproche en cela des principes de la sémiotique triadique de Peirce (Peirce 1978). L'approche adoptée nous amène donc à préférer une instrumentation interactive du sens à une construction compositionnelle (ou à un calcul) du sens (comme c'est le cas dans beaucoup de travaux de sémantique en TAL, comme en témoignent par exemple les thèses de Guillaume Jacquet[9] ou de Fabienne Venant[10]). Nous y reviendrons lors du chapitre n°4. A la manière de Coursil qui définit les principes de non consignation et de non préméditation de la chaîne parlée (Coursil 2000), nous défendons un principe de « non transformation » du sens dans la mesure où l'on estime qu'il n'y a pas de forme extralinguistique du sens. Il n'y a pas non plus de pensée construite possible qui ne soit pas déjà sous forme langagière. Le sens est par nature intralinguistique et donc non représentable dans les registres d'une machine. Il n'y a alors pas à chercher une représentation formelle extralinguistique du sens (comme c'est le cas par exemple en logique). On rejoint ici l'avis de (Nicolle 2005) pour qui le sens n'est jamais capté par ses représentations. Toute représentation du sens est forcément incomplète et il n'y a donc pas de langage formel qui puisse reproduire fidèlement le sens d'un énoncé en langue naturelle[11]. La compréhension, l'interprétation langagière, est en cela bien distincte de l'interprétation logique se résumant à la traduction dans un autre langage (une forme de compilation) car pour nous, il n'y a pas d'autre langage. D'autre part, et c'est aussi une différence importante par rapport à une formalisation qui se veut objective, le sens est subjectif dans la mesure où il n'y a pas forcément de consensus d'un

9 Thèse intitulée « Polysémie verbale et calcul du sens » (Jacquet 2005)

10 Thèse intitulée « Représentation et calcul dynamique du sens » (Venant 2006)

11 Anne Nicolle en tire la conséquence que la langue est un langage « terminal ». Mais c'est aussi le langage « premier ».

interprétant à un autre sur une explicitation (partielle ou complète) du sens d'un texte, si court soit-il.

Ainsi nous préférons de loin l'idée d'une instrumentation du sens à celle de la construction du sens. Nous chercherons à assister les compétences interprétatives mais non, et surtout pas, à les remplacer. D'une manière métaphorique, il nous semble qu'un traitement sémantique informatisé centré-utilisateur a plus de points communs avec un outil tel un microscope ou un miroir par exemple, c'est-à-dire quelque chose qui nous montre ce qui est déjà là mais qu'on ne voyait pas de cette façon, qu'avec un outil qui produirait à partir de quelque chose existant une autre chose qui n'existait pas avant. Il s'agit donc de considérer que le sens tel que le produit une interprétation humaine n'est pas à la portée d'un seul traitement informatique. Cela ne veut pas dire pour autant que les machines ne puissent pas avoir une activité d'analyse des textes et être d'une utilité certaine dans une problématique d'accès au contenu. Les machines ont une approche calculatoire de l'accès au contenu là où les sujets humains ont une approche interprétative. Ces deux formes de rapport aux textes ne sont pas en concurrence car l'activité de la machine n'a en aucun cas, dans nos travaux, le but de supplanter celle de l'utilisateur. Au contraire, dans l'environnement numérique de travail elle doivent être complémentaires. L'environnement doit avoir comme objectif de produire dans l'interaction des signes (notamment, par exemple, au moyen de techniques de visualisation adéquates) qui vont participer aux interprétations du ou des utilisateurs et ainsi prolonger le couplage.

~

L'approche interprétative centrée-utilisateur de l'accès au contenu pourrait sembler porteuse d'un paradoxe. Si le sens est le résultat d'une interprétation qui en retour conditionne le sens, alors il n'y a jamais de véritable accès au contenu dans la mesure où ce contenu se trouve toujours recréé. L'idée d'un accès au contenu part du principe que le contenu existe indépendamment des interprétations qu'on pourra en faire. Par opposition, voir les choses de manière centrée-utilisateur consiste à affirmer que l'interprétation n'est pas une sorte de décodage et que ce sont les spécificités du ou des utilisateurs qui se projettent dans une rencontre avec le texte et que le sens résulte de cette projection.

Une vision stricte de l'accès au contenu devrait logiquement considérer que les signes portent leurs significations et qu'ils se combinent compositionnellement dans les textes de telle sorte que ces textes portent leur sens. C'est une vision inexacte et simpliste qui au final s'apparente à une métaphore, tout aussi contestable, par exemple, que le modèle gravitationnel de l'atome en chimie inspiré par les systèmes planétaires. Dans une vision de l'accès au contenu aux contours beaucoup moins tranchés et nettement plus réaliste, nous défendons l'idée que le sens n'est pas dans les textes mais dans les interprétations qu'en font les utilisateurs/lecteurs/interprétants. De même pour la signification des signes qui est déterminée par la mise en contexte plus qu'elle ne la conditionne. Le contenu n'est donc pas « contenu » et encore moins de manière extrêmement localisée (comme c'est envisagé dans les tâches de Questions/Réponses par exemple).

Nous dépassons ainsi le paradoxe en ramenant la question de l'accès au contenu à la celle de l'instrumentation des facultés interprétatives de l'utilisateur. Instumenter l'interprétation c'est permettre une rencontre entre l'utilisateur et des textes d'où du sens pourra émerger dans une boucle vertueuse où seront créées et convoquées des ressources (par exemple lexicales) propres à l'utilisateur.

Ce repositionnement de la problématique de l'accès au contenu qui remet en cause très vivement la question de la compositionnalité n'est pas pour autant l'aveu d'une impossibilité à mettre en place des traitements automatiques. L'approche centrée-utilisateur est tout à fait conciliable avec l'idée de

traitements opérationnels tant que demeure la nature interprétative du sens et de la signification. C'est là que se situe la présente contribution au TAL. Le chapitre suivant montre comment différents projets menés depuis une douzaine d'années donnent des réalisations concrètes de développement et d'expérimentations centrés sur l'utilisateur en TAL.

3. Développements logiciels et expérimentations

Résumer plusieurs années de recherche est un défi. Les intuitions, les hypothèses, les changements de contextes, par exemple, ne s'exposent pas toujours facilement et de manière linéaire. Dans ce chapitre, nous prenons le parti de retracer chronologiquement ces années sous le prisme des projets de développement logiciels et des expérimentations qu'ils ont permis. C'est donc forcément réducteur mais plus qu'une liste de réalisations il faut y voir la partie émergée de l'iceberg qu'est l'activité scientifique.

Je détaillerai premièrement les différents projets de logiciels d'étude réalisés en laboratoire depuis ma thèse de doctorat jusqu'à dernièrement. Puis de manière non exhaustive je présenterai quelques expérimentations sur corpus réalisées avec ces logiciels d'étude. Enfin je finirai le présent chapitre sur un projet en cours de développement logiciel d'un produit logiciel en partenariat avec une société de développement informatique parisienne.

3.1. Logiciels d'étude

On entend ici par logiciel d'étude un développement informatique conçu pour expérimenter et valider de manière opératoire une approche dans le champ d'action visé par le logiciel (Nicolle 1996). Les logiciels d'étude permettent de s'intéresser à un double objectif. D'une part, savoir comment s'y prendre conceptuellement et techniquement pour mettre en œuvre un processus automatique et/ou interactif qui opère sur le champ d'étude visé. D'autre part, observer ce qu'un outil logiciel opératoire change dans la façon d'appréhender le champ d'étude. Dans ce double objectif, les logiciels d'études sont presque toujours des développements réalisés sous forme de prototypes de laboratoire. En général dans un prototype, on ne se pose pas vraiment prioritairement les questions d'optimisation de traitement, de montée en charge du logiciel, d'ergonomie de l'interface, de portabilité ou encore de

banc d'essai en situation réelle. Le plus souvent, l'utilisateur se trouve être de plus le (ou un des) concepteurs(s) du logiciel ce qui occasionne une moindre exigence en terme de documentation, d'aide en ligne et plus globalement de prise en compte de l'appropriation du logiciel.

On comprend bien que ceci se justifie car l'objectif du logiciel d'étude est avant tout d'alimenter l'activité du chercheur mais, par ailleurs, il faut être conscient que cela peut poser certains problèmes. Tout d'abord, dans une démarche centrée-utilisateur, le chercheur s'observe lui même jusqu'à ce qu'il éprouve le besoin de se confronter à d'autres utilisateurs. Ensuite, le fait que le développement soit peu documenté rend les solutions logicielles peu évidentes à maintenir dans le temps, surtout quand il arrive que le principal auteur ait quitté le laboratoire. La pérennité d'un logiciel d'étude n'est pas garantie et l'on constate que certains évoluent inévitablement vers des sortes d'*abandonware*[12].

Dans la suite du chapitre, je vais décrire les différents logiciels d'études sur lesquels j'ai travaillé :

- Anadia (1995-1999) : Représentation lexicale et analyse par calcul d'isotopies
- ThèmeEditor (2000-2002) : Coloriage d'isotopies génériques
- Lucia (2000-2004) : Structuration du contenu lexical et application à la recherche d'information
- ProxiDocs (2004-2007) : Cartographie thématique de corpus
- CartoMail (2009-2010) : Cartographie de boites mails

3.1.1. Anadia

Anadia est le nom donné au développement logiciel que j'ai effectué lors de ma thèse de doctorat (Beust 1998). L'objectif du modèle, et donc par extension du logiciel d'étude réalisé, était de fournir un cadre de représentation pour la structuration du contenu lexical et d'en déduire des procédures d'analyse textuelle. Les expérimentations réalisées ont concerné l'étude de tours de paroles dans des énoncés de dialogue où l'explicitation lexicale de certains termes était l'objet du dialogue.

Anadia a pour objectif une représentation différentielle du contenu lexical. Nous en faisons ici une présentation volontairement très succincte et renvoyons vers (Beust 1998) et (Nicolle & al. 2002) pour plus de détails.

Le principe central du modèle (et donc du logiciel d'étude) Anadia est la notion de valeur, notion introduite par Saussure (Saussure 1915) pour rendre compte de la signification dans les langues et reprise dans la sémiotique de Hjelmslev, pour qui le premier principe à considérer dans l'analyse d'une structure est que "une totalité ne se compose pas d'objets mais de dépendances" (Hjelmslev 1943, p. 37). Vu comme une valeur, le signifié, contrairement à la notion de concept ontologique, n'entretient aucun lien avec une classification des choses ou des espèces animales. Il est défini de façon purement différentielle, c'est-à-dire, non pas positivement par son contenu mais négativement par ses rapports avec les autres signes du système.

"Leur plus exacte caractéristique est d'être ce que les autres ne sont pas." (Saussure

12 Le terme *abandonware* provient du monde des jeux vidéo. Un abandonware (ou "logiciel abandonné" en français) est un jeu dont les éditeurs ou producteurs légitimes n'assurent plus la vente, ni le service après vente depuis longtemps. Ces jeux étant laissés à l'abandon, ils ne sont plus une source lucrative pour leurs auteurs. cf. http://www.abandonware-definition.org/def_actuelle.php consultée le 7/01/13

1915, p. 162)

Anadia procède à la représentation par une méthode de discrimination des valeurs en tables. Cette méthode à été en premier formulée (et de façon opératoire) par Jacques Coursil pour rendre compte de ce qu'il appelle la dimension défective de la langue (Coursil 1992), c'est-à-dire le fait que la langue se constitue essentiellement par des coupures, des différences. Notamment, il a pu montrer clairement avec cette méthode analytique la structure systémique et signifiante de la phonologie du français contemporain telle que l'a étudiée Jakobson (Jakobson 1963), ainsi que la structure des morphèmes dans la constitution syntagmatique des signes.

Le principe de représentation lexical d'Anadia est un principe componentiel, c'est-à-dire que le signifié est décrit en terme de traits sémantiques appelés *sèmes*. La spécificité du modèle est de considérer le sème comme une entité qui établit une différence dans un domaine d'interprétation. Ainsi plutôt que de considérer 3 traits indépendants /solide/, /liquide/, /gazeux/ on représente un sème opposant les 3 valeurs solide, liquide et gazeux dans le domaine d'interprétation de l'état physique. Par la suite, on adopte la notation [Etat physique : /solide/ vs. /liquide/ vs. /gazeux/]. Par combinaison de sèmes le modèle permet d'identifier des valeurs en leur attribuant des signifiants dans des structures de tables et de topiques (cf. illustrations 1 et 2).

Les tables du modèle Anadia peuvent être liées entre elles par des mécanismes de sous catégorisation. La représentation lexicale amène à procéder par paliers : tous les sèmes ne sont pas pertinents pour toutes les lexies. Faire une seule et unique grande table avec un maximun de sèmes serait une très mauvaise représentation avec, pour le coup, un nombre important de valeurs potentielles non réalisées. Il faut donc créer des tables différentes pour chaque classe terminologique où des lexies s'opposent, avec des sèmes pertinents seulement. Certains sèmes deviennent pertinents dans une sous-classe du fait que d'autres attributs ont une certaine valeur (par exemple, seuls les objets matériels peuvent avoir un poids et un volume), c'est ce qu'Aristote appelait les attributs propres de la catégorie (Aristote , Trad. J. Tricot 1969).

	Type de medium	Mode d'utilisation
clavier, souris, appareil photo numérique	image et texte	entrée
caméra vidéo	image et texte et son	entrée
microphone, micro	son	entrée
écran, imprimante	image et texte	sortie
écran avec haut-parleur intégré, écran avec hauts-parleurs intégrés	image et texte et son	sortie
haut-parleurs, HP, enceintes, enceinte, haut-parleur	son	sortie
	image et texte	entrée / sortie
support de données	image et texte et son	entrée / sortie
clavier midi avec haut-parleur intégré, clavier midi avec hauts-parleurs intégrés	son	entrée / sortie

Illustration 1 : Table Anadia
Cette table est construite par combinatoire des valeurs de sèmes [Type de médium : /image et texte/ vs. /image et texte et son/ vs. /son/] et [Mode d'utilisation : /entrée/ vs. /sortie/ vs. /entrée/sortie/]. La table identifie des signifiés pour les signifiants indiqués dans la colonne de gauche. Deux termes d'une même ligne sont localement identifiés par la même valeur.

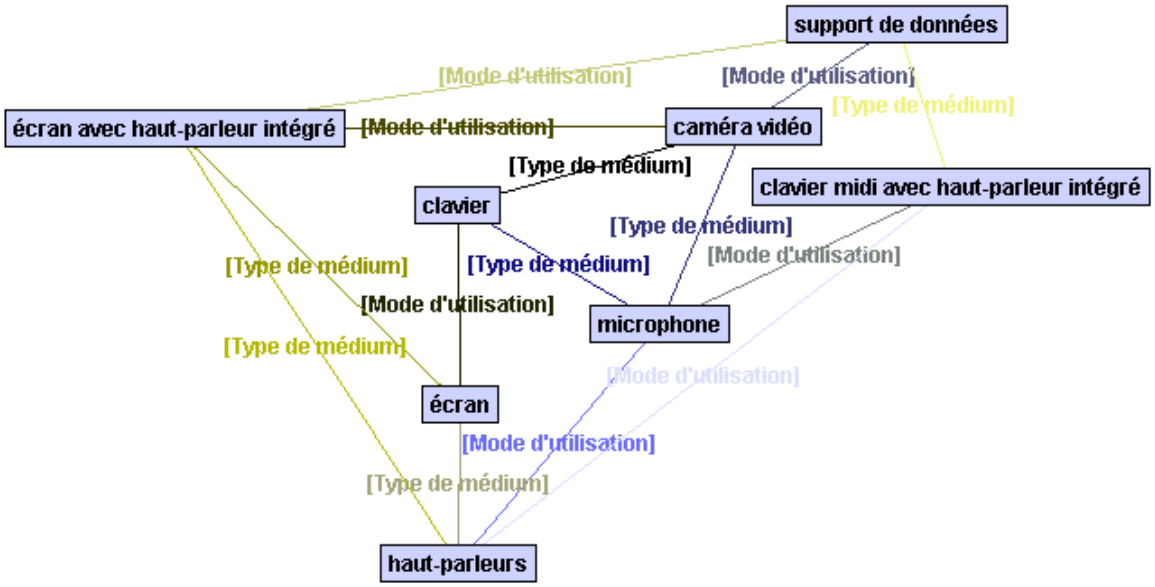


Illustration 2 : Topique Anadia.
Cette topique est construite à partir de la table précédente. Elle représente le tissu différentiel des valeurs. Les sommets du graphe représentent des termes définis dans la table. Les arcs représentent une relation de différence à un sème près, celui indiqué sur le lien. Ainsi, par exemple, « haut parleurs » est une valeur à un sème près de celle de « microphone », la différence portant sur le sème [mode d'utilisation].

La représentation lexicale d'un terme est la collection des sèmes (et des valeurs de sèmes correspondants) extraite de la hiérarchie des tables où le terme est définit. C'est ce qu'on appelle son sémème (Rastier 1987). Le sémème (cf. illustration 3) est composé des sèmes dits génériques qui sont ceux hérités des tables de niveaux supérieur (l'ensemble des sèmes génériques est appelé classème) et des sèmes dit spécifiques (l'ensemble des sèmes spécifiques est appelé sémantème) qui différencient la lexie des autres lexies de sa classe de plus bas niveau dans la hiérarchie des tables (classe qu'on appelle le taxème).

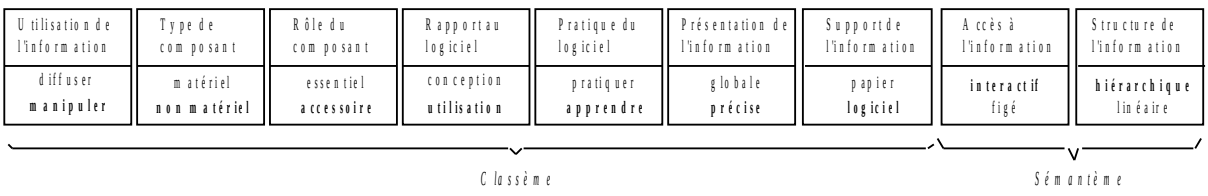


Illustration 3 : Sémème de la lexie *assistant*

Une fois un ensemble de lexies représenté en terme de sèmes, l'analyse syntagmatique menée dans

Anadia consistait à chercher localement des redondances de sèmes. Ces redondances de sèmes en contexte sont importantes dans la théorie de la sémantique interprétative (Rastier 1987) car elles expliquent la dynamique des sèmes à l'œuvre dans les phénomènes interprétatifs. Ces redondances sont appelées *isotopies*. Plus un énoncé, et plus généralement un texte, induit des isotopies plus il est porteur d'une cohésion. L'isotopie explique la cohésion interprétative mais l'inverse est également vrai car interpréter (et donc trouver une cohésion) c'est chercher des isotopies, à tel point que l'interprétation des messages énigmatiques s'explique notamment par des afférences de sèmes indiquant des isotopies (Mauger 1999). En d'autres termes, l'interprétation ne s'appuie pas sur des signes déjà donnés, elle reconstitue les signes en identifiant leurs signifiants et en les associant à des signifiés.

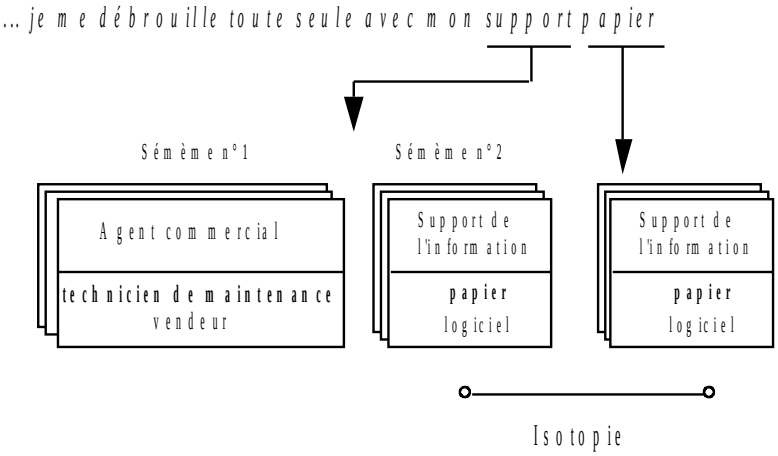


Illustration 4: Désambiguisation lexicale contextuelle.
Dans cet énoncé, l'identification de l'isotopie du sème [Support de l'information : papier vs. Logiciel] permettait de sélectionner le bon sémème en contexte de la lexie support.

Lorsqu'un terme est polysémique, sa représentation componentielle est faite de plusieurs sémèmes, chaqu'un d'eux représentant une valeur. En utilisant Anadia, nous avons pu constater notamment que l'isotopie est un concept opératoire très intéressant pour rendre compte en contexte de la désambiguisation de la polysémie (cf. illustration 4). En sélectionnant les sémèmes renforçant une ou plusieurs isotopies, on virtualise les sémèmes non pertinents.

Anadia était un logiciel d'étude intéressant dans plusieurs raisons. Premièrement il a montré une réalisation opératoire des principes de description linguistique de la Sémantique Interprétative (SI). D'autre travaux l'ont également fait, un peu près au même moment (Tanguy 1997, Thivilitis 1998) mais sans ce modèle oppositionnel du sème. Depuis la SI s'est conforté comme un ancrage théorique majeur pour d'autres travaux en TAL (Bommier-Pincemin 1999, Perlerin 2004, Rossignol 2005, Roy 2007 par exemple) ainsi que dans d'autres champs, notamment les domaines de la linguistique de corpus (Bourion 2001, Loiseau 2006, ou encore Poudat 2006) et de la recherche d'information (Valette 2004, Mauceri 2007, Kanellos & Mauceri 2008 ou encore Valette & Slodzian 2008). Deuxièmement, Anadia était intéressant pour sa démarche interactionniste où le processus de description lexicale était indissociable de l'analyse syntagmatique et donc de l'interprétation. Classiquement en TAL, la constitution de ressources est un travail indépendant du développement des analyseurs. Avec Anadia,

nous montrions un modèle ne visant pas l'exhaustivité en terme de couverture de ressources mais motivant les ressources par les analyses qu'on pouvait faire avec. On rejoignait ainsi le propos ci-dessous de Umberto Eco et on était ainsi déjà dans un processus d'ordre énactif (cf. chapitre n°6).

> *"Donc, le problème sémiotique de la construction du contenu comme signifié est étroitement solidaire du problème de la perception et de la connaissance comme attribution de sens à l'expérience" (Eco 1988, p. 81)*

3.1.2. ThèmeEditor

L'outil ThèmeEditor est un logiciel d'étude disponible gratuitement en *open source* via Internet[13]. Il a été réalisé en Java par deux étudiants en maîtrise d'informatique à l'Université de Caen en 2000.

ThèmeEditor est une suite du projet Anadia dans laquelle on a cherché à prolonger l'idée de la création de ressources lexicales intégrée dans l'activité d'analyse interprétative de corpus. Le but est de mettre en place une boucle interactive où l'on constitue des classes thématiques de termes interactivement à partir de la visualisation d'isotopies génériques sur corpus. A la différence d'Anadia, nous ne cherchons pas ici à représenter du contenu lexical. L'objectif terminologique est simplement de constituer des classes relevant d'un même thème, ce qu'on a (par abus de langage) appelé des sacs de mots. Du coup, l'analyse en terme de recherche d'isotopies est réduite aux isotopies génériques que représentent les redondances d'un même thème.

De même que Pichon & Sébillot (1999), nous entendrons ici par « thèmes » les sujets abordés dans un texte ou dans un corpus et nous les mettons en évidence par un principe de coloriage. Par exemple on pourrait définir le thème *Politique* (et lui affecter une couleur) par la liste de graphies : *élus, élu, assemblée, assemblées, premier ministre, ministre, ministres, président, présidents, député, députés, référendum local, ministère, commune, communes, vote, gouvernement* … Les thèmes ne visent pas une représentation lexicale exhaustive et objective mais donnent la possibilité à un utilisateur d'identifier certaines lexies d'un domaine (indiquant un ou plusieurs vocabulaires de spécialité) qui, pour lui, ont une pertinence dans certains textes.

Le coloriage thématique est une façon très efficace d'un point de vue ergonomique d'identifier ces sujets dans les textes. Le principe du coloriage thématique consiste à affecter une couleur à chaque isotopie et à « surligner » les mots du texte sur lesquels elles s'appuient. Le coloriage de textes permet ainsi de faire apparaître les différentes isotopies qui recouvrent un texte. On peut alors en examiner la répartition 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. En cela, on se situe dans le même courant d'étude que le logiciel PASTEL développé par Ludovic Tanguy (Tanguy 1997).

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 classes sémantiques 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 indiquant le champ lexical des aliments qu'à une classe sémantique révélant un champ lexical juridique. Du point de vue de la méthode de coloriage, la question qu'il convient alors de se poser consiste à savoir quelle est la couleur à attribuer

13 http://users.info.unicaen.fr/~beust/ThemeEd/ consultée le 7/01/13.

à un mot rencontré dans un texte si ce mot appartient à plusieurs classes. Dans un tel cas, l'heuristique considérée est celle que nous avions déjà expérimentée avec Anadia et qui tend à prolonger le plus possible les isotopies du texte (c'est-à-dire à favoriser la redondance). Ainsi, parmi ses couleurs possibles, on attribuera au mot la couleur la plus représentée dans le texte, comme on le montrent les deux exemples suivants :

En faisant mon marché, j'ai vu des poireaux, des concombres et des avocats

En sortant du tribunal, j'ai vu mon avocat.

La méthode de construction des classes sémantiques par le coloriage thématique est essentiellement manuelle. C'est la différence avec des systèmes qui proposent automatiquement des thèmes par analyse des voisinages de mots (par exemple Pichon & 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 (Tanguy 1997), notre méthode est basée sur une analyse interprétative de textes électroniques. Cette analyse est éclairée par des calculs statistiques réalisés par le logiciel. Partant d'un corpus de textes, le système permet de dresser des listes d'occurrences de mots (listes de type Zipf[14]) à partir desquelles on enrichit les classes sémantiques qui en retour permettent le coloriage du corpus. L'utilisateur peut réitérer ce processus autant fois que nécessaire à la lumière des statistiques issues du coloriage (cf. illustration 5).

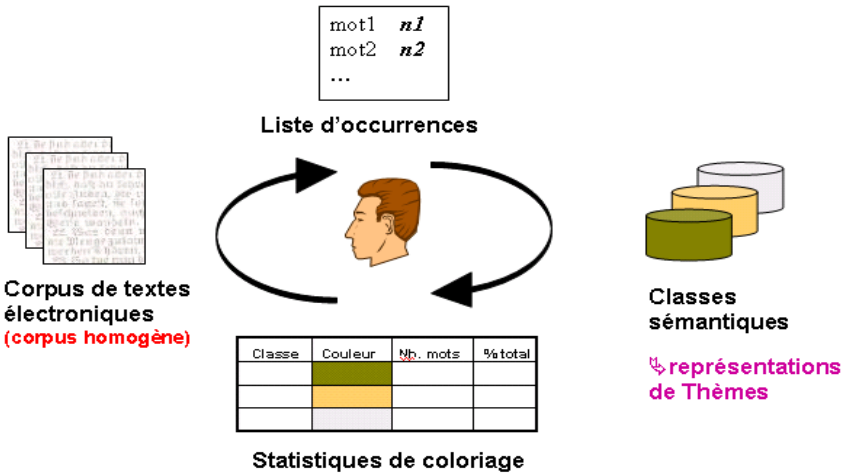


Illustration 5 : Coloriage et construction de classes terminologiques

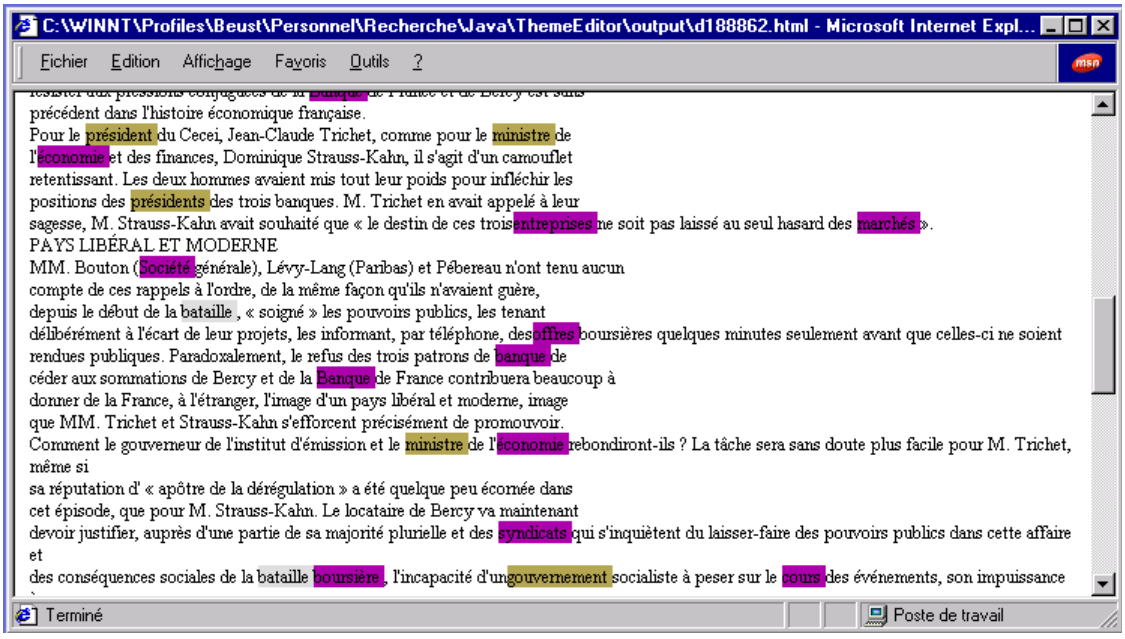
L'interface de l'application présente plusieurs zones :

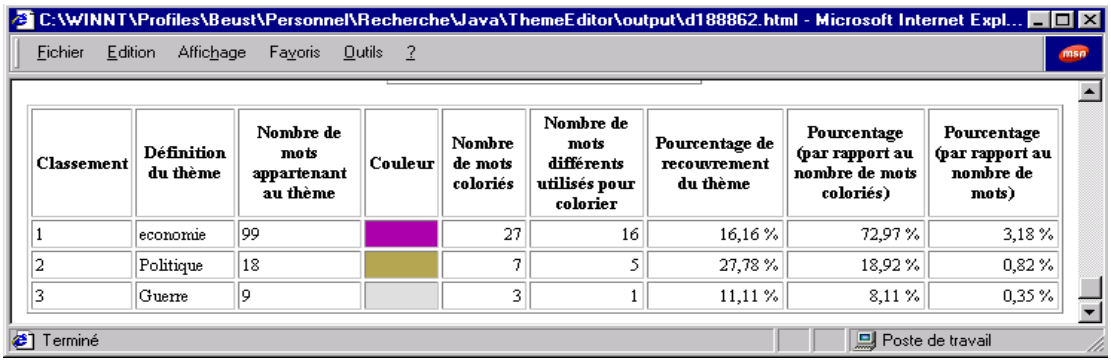
- la liste des thèmes chargés. Si l'on ouvre l'un des thèmes on peut en consulter le contenu, c'est-à-dire la liste des termes retenus pour représenter le domaine sémantique visé,
- les noms des fichiers (fichiers textes) contenus dans le dossier qui contient le corpus d'étude,
- le résultat du calcul du nombre d'occurrences des différents mots du (ou des) document(s)

14 (Zipf 1949)

sélectionné(s) dans la deuxième zone. La liste des mots est présentée par ordre décroissant de nombre d'occurrences. On peut ainsi extraire les mots les plus fréquents d'un texte, de plusieurs textes ou même de tout le corpus d'étude en fonction de la sélection de documents considérée (simple ou multiple).

- les résultats du coloriage sur le ou les textes sélectionnés. On trouve à la suite des textes où figurent colorés les mots des différents thèmes, un jeu de calculs statistiques sur le coloriage réalisé, par exemple le classement des thèmes colorés dans la sélection avec pour chacun des thèmes le nombre de mots utilisés pour colorier ou encore le pourcentage que ces mots représentent par rapport à l'ensemble de la sélection (cf. illustration 6).





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 %

Illustration 6: Coloriage du texte d198862.txt
Le coloriage fait apparaître par ordre d'importance les thèmes *Economie, Politique* et *Guerre*

Les classes sémantiques sont construites de manière incrémentale, soit en y entrant directement des lexies au clavier, soit en cliquant sur les mots présentés dans la liste des occurrences pour les ajouter au thème préalablement sélectionné.

Les textes automatiquement coloriés grâce aux classes sémantiques construites peuvent être visualisés dans la zone de l'application qui indique les statistiques, mais ils peuvent aussi être enregistrés au format HTML et ainsi être visualisés dans un navigateur. Ils peuvent également être enregistrés dans un document structuré par une DTD XML pour d'éventuels traitements automatiques ultérieurs. En plus du coloriage de fichiers textes, il est également possible de réaliser un coloriage en surchargeant la structure d'un document HTML.

Une extension de ThèmeEditor a été réalisée (sous la forme d'un deuxième projet d'étudiants d'informatique) dans le but de caractériser l'homogénéité thématique des flux documentaires. A la différence d'un corpus de textes, la spécificité d'un flux documentaire tient à son ancrage temporel. En fonction de l'empan temporel considéré, le flux change par l'apparition et la disparition de documents. Cette dimension temporelle est à prendre en compte dans l'interprétation des documents du flux. Par exemple, on veut pouvoir rendre compte de l'apparition, des modifications et de la disparition de classes sémantiques en fonction de l'empan considéré (cf. illustrations 7 et 8). Dans ce but, le logiciel présente différentes statistiques sur une période choisie : la densité du flux en terme de documents, l'évolution des thèmes dans la période, la co-présence de certains thèmes dans cette période ou encore les proportions relatives des thèmes sur les documents de la période.

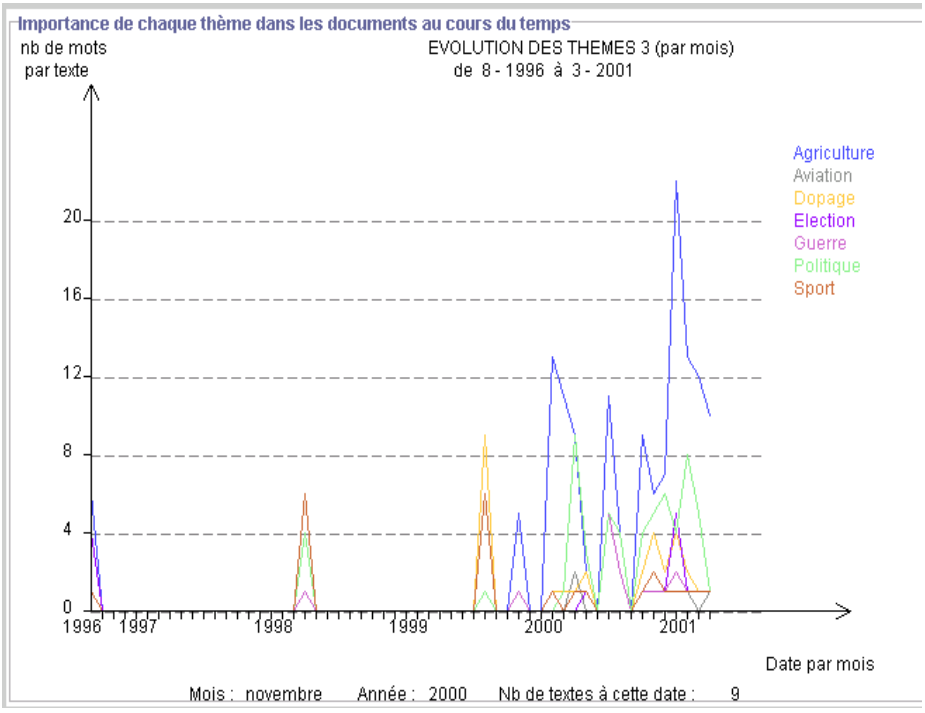


Illustration 7: Analyse de l'évolution des thèmes dans un flux documentaire

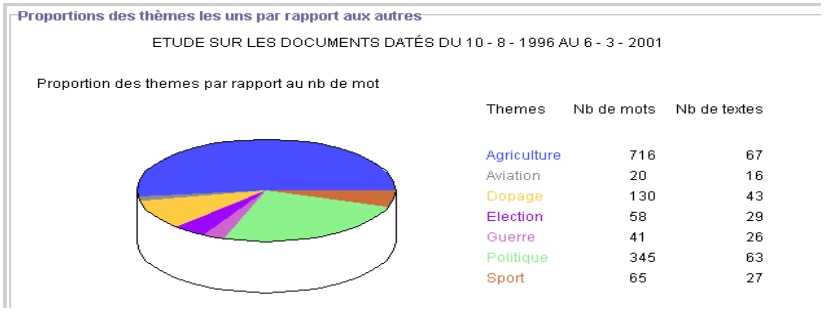


Illustration 8: Analyse de l'évolution et de la proportion des thèmes dans un flux documentaire

ThèmeEditor est un outil pour l'extraction rapide (une sorte de prototypage) de termes préférentiels de chaque utilisateur qui sont le reflet d'une analyse interprétative de leurs textes. Les résultats statistiques et les coloriages sont présentés pour enrichir la perception qu'a l'utilisateur de son corpus et de ses ressources. Les classes terminologiques sont principalement construites manuellement et de façon incrémentale d'une session de travail à une autre et révèlent les interprétations qui ont été faites. L'outil est volontairement très simple d'utilisation et de prise en main et c'est la raison pour laquelle nous avons voulu y mettre en œuvre, à la différence d'Anadia, un principe de structuration sémantique des classes qui est très rudimentaire (ce ne sont que les listes de termes et de graphies). De cette simplicité d'usage dépend un couplage avec l'utilisateur très naturel et en conséquence un apport assez significatif à la démarche interprétative. Un intérêt de ThèmeEditor était donc de mettre l'accent sur les aspects vertueux de la boucle personne-système.

3.1.3. LUCIA

Le projet LUCIA (Located User-Centered Interpretative Analyser) a été réalisé entre 2001 et 2004 durant la thèse de Vincent Perlerin (Perlerin, 2004) que j'ai co-dirigée. L'objectif de départ de cette thèse était de reprendre les concepts de base d'Anadia et de les appliquer dans plusieurs situations d'assistance interprétative personnalisée (analyse de contenu, recherche d'information) en affirmant au maximum une démarche centrée-utilisateur que le projet Anadia n'a que laissé entrevoir.

Faisant le constat que les ressources terminologiques généralistes sont peu utilisables et que les méthodes d'analyses compositionnelles sont peu fructueuses, la thèse de Vincent Perlerin argumente pour une approche centrée-utilisateur où la non exhaustivité des ressources est plus un atout qu'un problème, y compris d'un point de vue strictement algorithmique dans les analyses qui les exploitent. Il en découle les principes d'une *sémantique légère* pour les approches centrées sur l'utilisateur.

La thèse de Vincent Perlerin a marqué une avancée significative dans la maturation des idées de fond sur la façon d'utiliser la sémantique interprétative de François Rastier dans une approche informatique centrée-utilisateur. De plus, nous devons à Vincent Perlerin une très grande qualité de développement d'outils logiciel. Le projet LUCIA a été articulé en deux outils distincts mais complémentaires :

- VisualLuciaBuilder : Outil qui permet à un utilisateur (voir un groupe d'utilisateurs) de mettre au point des Ressources Termino-Ontologies (RTO) appelées dispositifs Lucia, c'est-à-dire des représentations lexicales différentielles basées sur les mêmes principes componentiels

qu'Anadia (cf. illustration 9).

- LuciaAnalyser : Outils qui permet de visualiser sur le corpus de l'utilisateur la projection des ses ressources en mettant en évidence des zones de localité textuelle portant des faisceaux d'isotopies.

VisualLuciaBuiler et LuciaAnalyser sont deux outils qui s'alimentent mutuellement d'un point de vue opératoire dans la mesure où les dispositifs de représentation différentielle mis au point avec VisualLuciaBuilder servent de bases aux analyses que permet LuciaAnalyser, analyses qui en retour conditionnent les évolutions des dispositifs. D'un point de vue technique cette interopérabilité à été réalisée en exploitant les richesses de la représentation XML ainsi que les technologies de transformation d'arbres XSL et XSLT.

Plutôt que de recourir à des ressources génériques et les plus exhaustives possibles, nous cherchons à exploiter au maximum des ressources légères et non exhaustives car spécifiques au besoin de l'utilisateur qui les crée. Une question qui en découle est celle de la stabilité dans le temps de ces ressources du point de vue de son (ou de ses) auteur(s). Le rapport au corpus est prépondérant. C'est en utilisant ses ressources qu'un utilisateur est amené à les mettre à l'épreuve et peut être incité à les rectifier et/ou les compléter. Par exemple, si on s'aperçoit qu'une lexie n'est jamais trouvée en corpus, on peut la supprimer de la ressource. A l'inverse, si l'une est très présente avec des significations différentes, on peut vouloir en rendre compte dans la ressource (notamment, en créant une nouvelle table dans le dispositif). La taille raisonnable des ressources (environ une centaine de termes en moyenne) rend possible ce genre de révision incrémentale au cours de l'utilisation de la ressource.

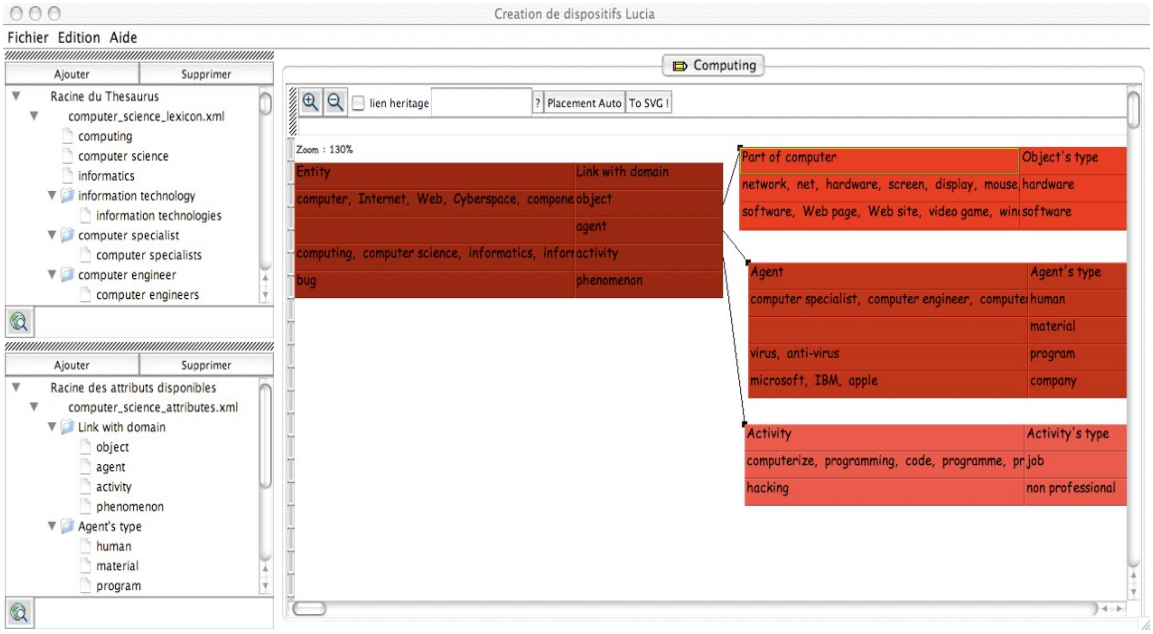


Illustration 9: Interface utilisateur de VisualLuciaBuilder.
Création de dispositifs de représentation lexicale différentielle.

L'atelier formation du CNRS « Variation, construction et instrumentation du sens » qui s'est tenu en juillet 2002 (île de Tatihou, Manche) a été pour nous l'occasion de mettre en place une première expérimentation sur LUCIA (Perlerin & al. 2003). Nous souhaitions tester la capacité d'utilisateurs

novices à s'approprier les principes généraux du modèle (attributs, tables, dispositifs) en leur demandant de construire dans un temps imparti, un dispositif sur un sujet précis (en l'occurrence 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 différents (linguistique, psychologie, ergonomie, informatique, microbiologie, etc.). Après un exposé introductif sur les principes du modèle, nous avons fourni aux participants une liste de 216 lexies issues du corpus Le Monde sur CD-ROM. Cette liste avait été obtenue à partir d'un calcul de fréquences de mots 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 (cette liste contenait par exemple des lexies comme action, back office, dévaluation, OPA ou encore palais Brongniart). Les consignes données aux participants se bornaient à leur demander de construire sur papier un dispositif selon leur façon propre de parler du domaine (la consigne n'imposait pas nécessairement d'intégrer les 216 lexies dans le dispositif).

A l'issue des deux séances d'expérience, tous les participants ont au moins proposé des groupes de lexies, précisé les différences qu'ils considéraient effectives au sein de ces groupes et créé des tables avec un ou plusieurs attributs. Cependant aucun participant n'a estimé au bout de l'expérience être parvenu à un résultat finalisé. Après entretien avec eux, nous avons pu estimer tout d'abord que l'expérience présentait un certain nombre de biais. 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 lexies proposées a également été ressentie comme un handicap par les participants. Ceci nous montre bien que le retour en permanence aux textes est fondamental. C'est un élément important favorisant le couplage avec l'utilisateur.

Cette expérience sans corpus, ni outil logiciel à disposition, permettait simplement de tester la faisabilité de la construction de RTO différentielles personnelles et donc d'apprécier la capacité des participants à amorcer un processus de construction cyclique. En l'occurrence, nous avons constaté que la méthode de construction des RTO LUCIA 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 montre que les utilisateurs intègrent leur propre sensibilité par rapport à un domaine (cette sensibilité pouvant relever d'une méconnaissance totale de ce domaine) tout en se conformant aux usages qu'ils ont pu rencontrer des lexies proposées.

Bien que la tâche de description des significations ne soit pas triviale (c'est un métier à part entière), nous avons pu constater que des utilisateurs arrivent à formuler dans un temps raisonnable des représentations terminologiques reflétant leur point de vue sur le domaine proposé. Ceci constituait l'une des hypothèses que nous cherchions à estimer. Durant cette expérimentation et, à travers les différences et les points communs entre les dispositifs et les groupes de mots construits 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 qui ne peut pas être absente de ce type de construction (Nicolle & al., 2002, p.61). Ceci a renforcé notre point de vue résolument tourné vers l'utilisateur, mais pas l'utilisateur seul.

Durant sa thèse, Vincent Perlerin a expérimenté la création et l'utilisation de RTO LUCIA dans plusieurs champs lexicaux allant du vocabulaire boursier à la terminologie métier issue d'une collaboration avec un partenaire industriel (la société CORRODYS) proposant des services dans le traitement de la biocorrosion (corrosion de matériaux par des organismes vivants).

Une utilisation très convaincante des dispositifs LUCIA a été réalisée dans le champ applicatif de la recherche d'information sur Internet (Perlerin 2004, p. 228 et suivantes). L'idée était de coupler les

outils LUCIA à une démarche de veille technologique sur Internet dans des usages complémentaires aux moteurs de recherche utilisés. Ceci s'est traduit par une déclinaison de LuciaAnalyser spécifique à la veille documentaire, LuciaSearch. Ainsi suite à une requête sur Google, LuciaSearch permettait de traiter les premiers documents fournis en construisant notamment pour chaque document une miniature du rendu visuel (un peu à la façon des TileBars de Hearst[15]) enrichie par des zones de localité colorées par le faisceau d'isotopie dominant. Il s'en suivait des propositions de filtrages et de reclassements personnalisés sur les réponses de requêtes sur le Web.

La création des miniatures symboliques (cf. illustration 10) est un résultat opératoire dont nous nous sommes aperçu qu'il était en fait beaucoup plus intéressant qu'on le pensait initialement. Comme on le voulait, cela permettait bien de produire des repérages rapides de documents intéressants mais, plus encore, la représentation sémiotique personnalisée de documents est fortement vecteur de couplage personne-système. On constate assez vite que la forme visuelle et matérielle des documents sur le Web est un signe tout à fait intéressant dans une perspective interprétative. Ainsi, par exemple, l'aspect visuel d'un article du journal *Le Monde* est une signature du genre textuel en question car rien ne ressemble plus à un article du *Monde* qu'un autre article du *Monde* (ce serait vrai également sur d'autres sources de publication). Dans la globalité d'une interprétation, la textualité d'un document ne se limite pas à une chaîne de caractères, comme d'ailleurs une image ne peut non plus se réduire à une simple suite de pixels.

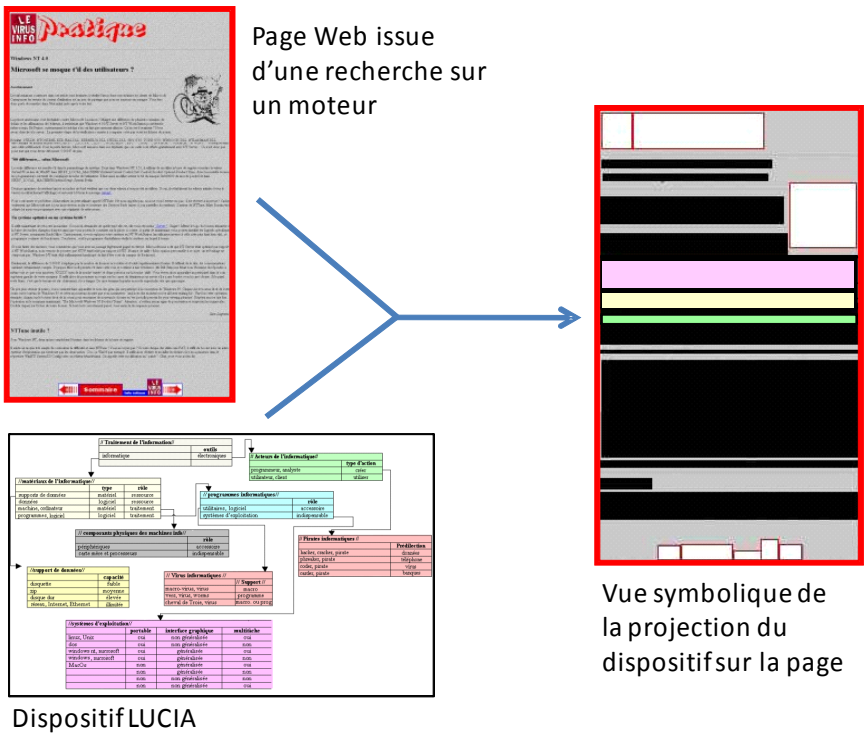


Illustration 10: LuciaSearch
Sur la visualisation à droite de la page de Web de gauche on remarque des zones colorées en rose, jaune et vert indiquant des occurrences significatives de termes définis dans les tables de couleurs correspondantes dans le dispositif de gauche.

15 (Hearst 1995)

D'une part, nous avons réaffirmé le constat que toute la sémiotique des documents est importante dans une approche interprétative. Ce même constat a également pu être constaté dans les travaux du projet PRINCIP (Valette 2004) sur la détection de sites à contenu raciste ou antiraciste. D'autre part, la thèse de Vincent Perlerin nous incitait à développer des outils qui allaient chercher encore plus à investir des stratégies de visualisation en tant que celles-ci produisent du couplage. C'est la réflexion qui a été à l'origine du sujet de thèse de Thibault Roy et du projet ProxiDocs.

3.1.4. ProxiDocs

Nous rejetons un vision simpliste compositionnelle du sens dans les phrases ou énoncés, puis dans les textes et enfin dans les collections documentaires. Le sens d'un texte n'est pas l'unique résultant des mots et des phrases du texte. Suivant la Sémantique Interprétative, nous adoptons le principe de détermination du local par le global. C'est-à-dire qu'une zone de localité textuelle a un sens qui dépend pour une bonne partie de son entour textuel. De la même façon le sens d'un document au sein d'une collection dépend autant de la collection que la collection dépend du texte. C'est un principe dit holiste (Gosselin 1996). L'approche holiste prend sa source dans la sémiotique de Saussure et, plus précisément, dans la valeur syntagmatique des signes :

> *"Le tout vaut par ses parties, les parties valent aussi en vertu de leur place dans le tout, et voilà pourquoi le rapport syntagmatique de la partie au tout est aussi important que celui des parties entre elles" (Saussure 1915,p. 177)*

Si les approches compositionnelles se prêtent bien par nature à des implémentations, c'est moins immédiat pour les approches holistes. Nous avons cherché avec le projet ProxiDocs à produire une instrumentalisation informatique du principe de détermination du local par le global en essayant de donner à un utilisateur une visualisation globale d'une collection documentaire.

Dans bon nombre de situations (analyse de flux documentaires, recherche d'informations, extraction d'informations …) l'appréhension de la thématique globale des documents ainsi que l'homogénéité thématique d'une collection de documents est une première analyse importante. ProxiDocs a pour objectif de fournir à son utilisateur une aide dans ce type d'analyse en lui permettant de construire, dans une approche centrée-utilisateur, des cartes thématiques de collections de documents.

ProxiDocs a été développé et expérimenté entre 2004 et 2007 durant la thèse de doctorat de Thibault Roy que j'ai co-dirigée avec Jacques Vergne. ProxiDocs est un outil logiciel gratuit *open source*, disponible avec sa documentation sur le Web[16].

ProxiDocs aide son utilisateur à extraire les tendances thématiques d'un corpus via des représentations graphiques que nous avons appelées *cartes*. Ces cartes permettent de mettre en évidence les thèmes abordés dans le corpus, mais également leur répartition et leur cohésion au sein des documents. Chaque document qui est représenté dans une carte indique sa position relative par rapport aux autres documents selon la propriété suivante : si deux documents abordent des thèmes similaires, alors nous supposons que les points les représentant seront proches sur la carte. A l'inverse, si deux documents n'abordent pas les mêmes thèmes, alors nous pouvons faire l'hypothèse que les points les représentant seront éloignés sur la carte. Les intérêts de la métrique de ProxiDocs sont

16 cf. http://roythibault.free.fr/ consultée le 7/01/13.

multiples, par exemple le regroupement des documents thématiquement similaires ou encore l'extraction d'un document moyen représentant un groupe de documents de la carte.

ProxiDocs prend en entrée un corpus de textes recueilli par l'utilisateur et un jeu de thèmes qu'il a construit. Ces thèmes peuvent être de 2 formes différentes : soit des classes thématiques constituées avec ThèmeEditor, soit des ressources termino-ontologiques construites sous forme de dispositif LUCIA. Les différentes étapes de calcul réalisées par ProxiDocs s'effectuent d'une manière chaînée (cf. illustration 11). Il y a des étapes de comptage d'occurrences de lexies, de représentation vectorielle de l'espace documentaire, de visualisation par projection et de regroupements. A travers ces différentes étapes de calcul, ProxiDocs met en oeuvre un processus de traitement statistique qui s'apparente à une technique d'analyse de *Sémantique Latente* décrite dans (Dumais 1988). Nous renvoyons à (Roy 2007) ou encore (Roy & Beust 2004) pour plus de détails. Les cartes thématiques produites en sortie de l'application sont représentées dans le format SVG (W3C 2001) afin de permettre à l'utilisateur d'effectuer des zooms et des déplacements sur ses cartes, et de garantir la portabilité des cartes.

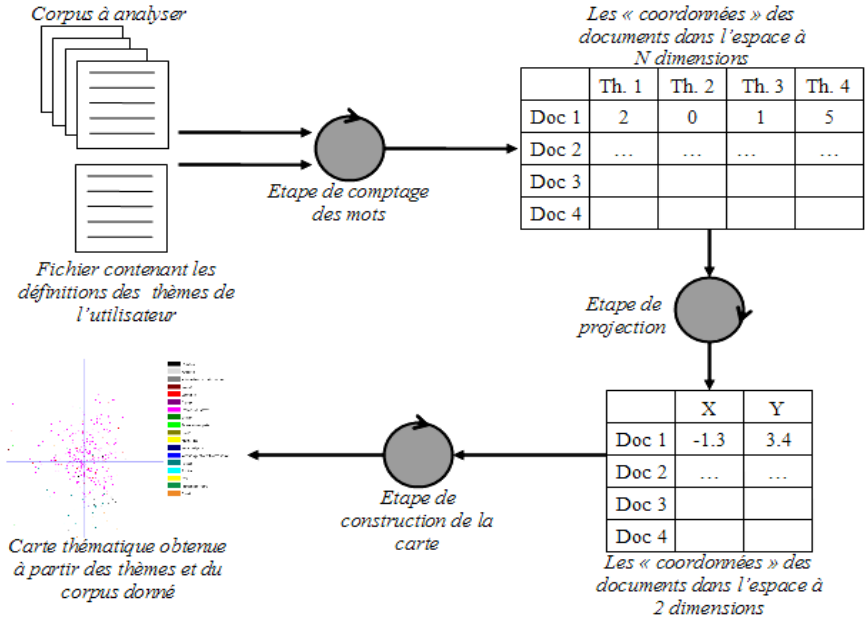


Illustration 11: La chaîne de traitement de ProxiDocs

L'étape de comptage est nécessaire pour établir une représentation vectorielle des documents du corpus étudié. Il faut compter, pour chaque document, le nombre des occurrences de mots de chaque thème. Chaque document se voit donc attribuer un vecteur contenant un nombre de données égal au nombre de thèmes utilisés. Soit N, un tel nombre, le vecteur décrira donc un point possédant N coordonnées. Soit D un document où figure 3 mots du thème « Bourse », 0 du thème « Météo » et 2 du thème « Sport », le vecteur obtenu sera alors : vecteur_D = (3, 0, 2).

Après comptage et création de la matrice de dimension N (nb de thèmes utilisés), les documents sont représentés par des points possédant N coordonnées. La carte étant un support en deux ou trois dimensions, il faut établir une projection de l'espace à N dimensions. La méthode la plus probante que nous avons implémentée est l'analyse en composantes principales (ACP) (Bouroche & Saporta 1980). Cette méthode (de même que les analyses factorielles des correspondances) est bien connue dans les domaines de l'analyse de données et de la lexicométrie (Salem 1993). Nous ne la détaillerons pas ici

mais nous nous limiterons à rappeler une idée de son application à notre problématique. L'idée majeure de l'ACP est que si les thèmes sont corrélés entre eux alors ils sont partiellement redondants. Prenons un exemple simple pour illustrer cette idée. Supposons que l'utilisateur ait défini les thèmes « Equitation » et « Jeux » parmi l'ensemble de ses thèmes. Supposons de plus, que son corpus soit principalement composé d'articles concernant les courses hippiques. Dans ce cas, les deux thèmes sont fortement corrélés vis-à-vis du corpus. Ils peuvent donc être regroupés en un seul groupe de thèmes. L'ACP va généraliser ce processus à l'ensemble des thèmes, pour ne retenir que les deux ou trois principaux groupes de thèmes. Ces deux ou trois groupes sont appelés les composantes principales, la projection des documents se fera alors sur ces deux ou trois axes (cf illustrations 12 et 13).

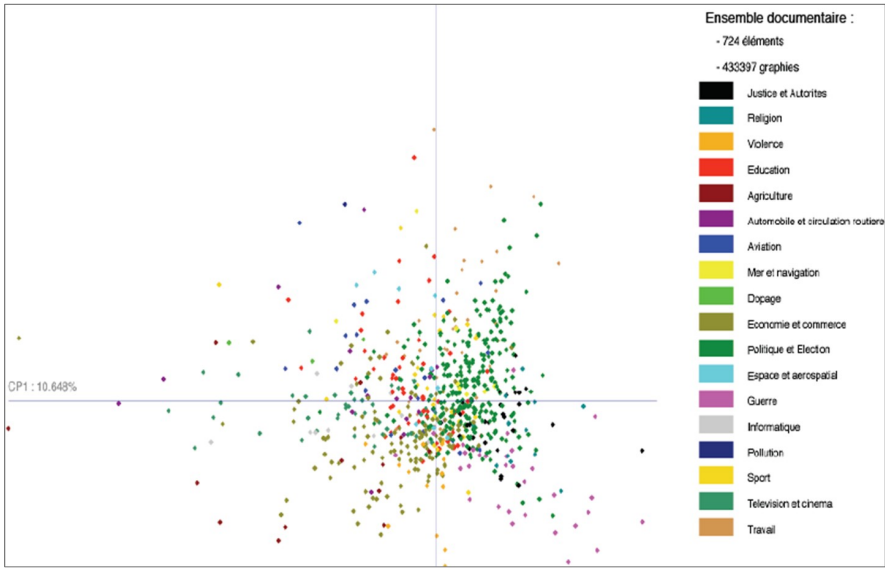


Illustration 12: Carte thématique de documents 2D.
Chaque point représente un document qui est de la couleur qui correspond au thème majoritaire du document.

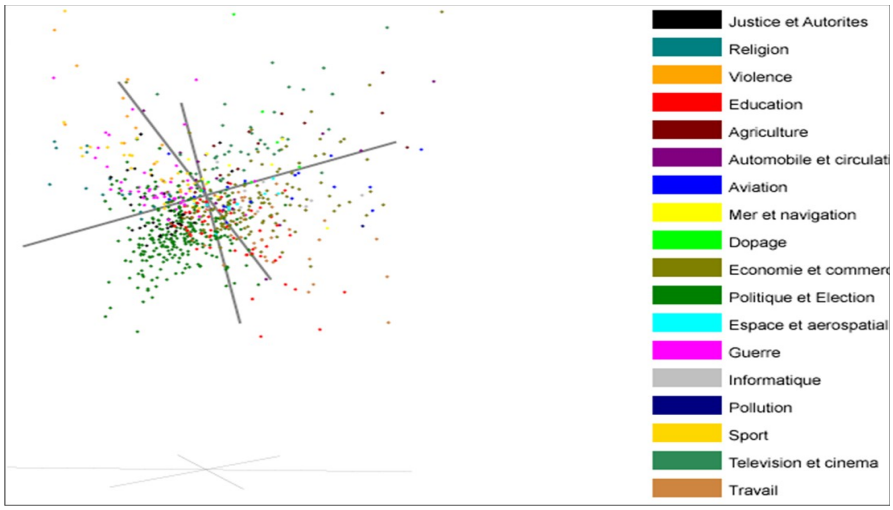


Illustration 13: Carte thématique de documents 3D.

Les cartes thématiques ProxiDocs sont des visualisations interactives. L'utilisateur peut y réaliser

diverses manipulations à l'aide de sa souris : zooms, déplacements, accès au coloriage thématique d'un document en cliquant sur le point représentant le document, mise en évidence des composantes thématiques du corpus en cliquant dans la légende sur le thème voulu (cf. illustration 14).

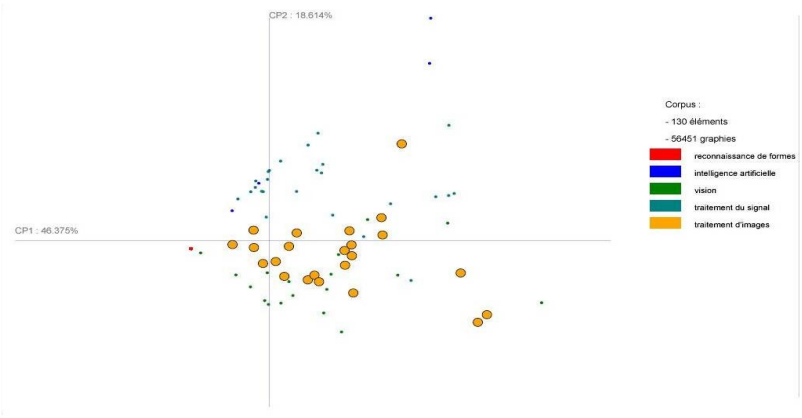


Illustration 14: Mise en évidence d'une composante thématique. Les points en jaune apparaissent lorsqu'on passe la souris sur le thème jaune.

L'étape suivant la présentation de la carte est une étape de *clustering* où le logiciel va proposer à son utilisateur de calculer des regroupements de documents. Ces regroupements sont motivés par la distance euclidienne qui sépare les différents points de la carte. On va chercher à regrouper les points proches dans la carte 2D ou 3D pour former des groupes indiquant des tendances thématiques du corpus induites par la répartition globale du corpus selon les thèmes de l'utilisateur. Des cartes de regroupements sont produites et indiquent chaque groupe par des disques (ou des sphères en 3D) dont la taille est proportionnelle au nombre de documents agrégés et dont la couleur indique le thème majoritaire dans le groupe (cf. illustrations 15 et 16).

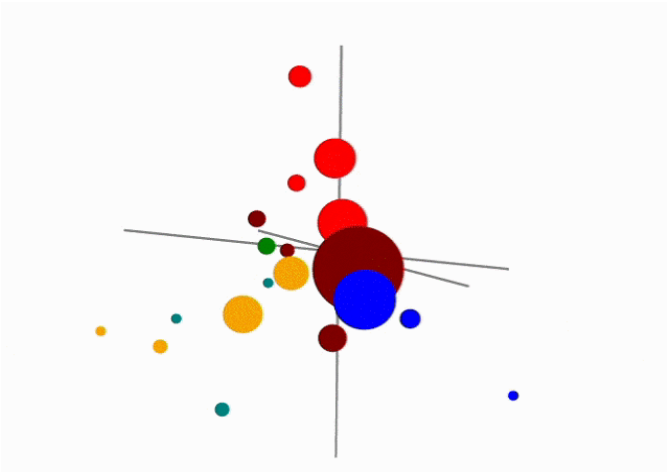


Illustration 15: Carte de groupes de documents 3D

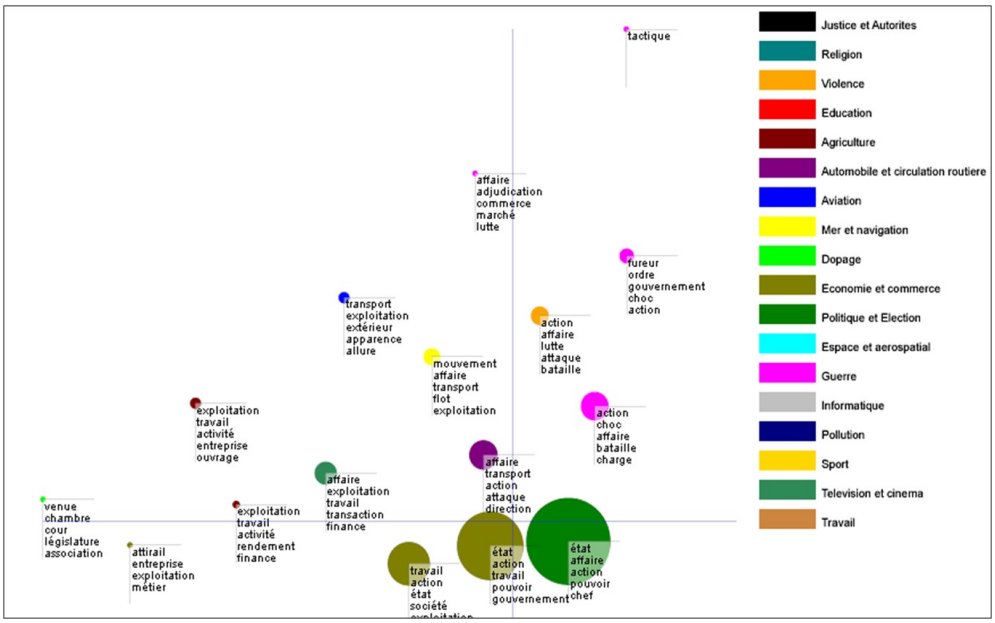


Illustration 16: Carte de groupes de documents en 2D produite à partir de la carte de l'illustration n°11.
Les étiquettes textuelles portées par les groupes indiquent les lexies les plus occurrentes dans le groupe.

Les cartes de groupes de documents sont également des cartes interactives qui permettent à l'utilisateur d'opérer des zooms et des déplacements. Elles permettent également en cliquant sur un groupe d'éditer un rapport d'analyse du groupe. Ce rapport présente :

- le nombre de documents du groupe,
- un lien vers le document moyen du groupe (une sorte de barycentre thématique),
- le nombre de graphies comptabilisées dans le groupe,
- le classement des thèmes du groupe,
- les occurrences de lexies définies dans les ressources de l'utilisateur qui ont été comptabilisées dans le groupe,
- une comparaison des tendances thématiques propres du groupe relativement aux tendances thématiques du corpus (l'idée étant par exemple de savoir si le groupe est atypique par rapport au corpus)

Une fonctionnalité proposée de manière complémentaire sur les groupes et de leur appliquer récursivement le même processus de cartographie que celui appliqué au corpus dans son ensemble. Il peut ainsi en découler une représentation des tendances du corpus globale de manière arborescente avec des tendances de premier niveau, de second niveau etc …

Lorsqu'un corpus que l'on cherche à analyser est un flux documentaire, la dimension temporelle est essentielle et cette dimension le différencie d'une collection où les documents ne sont pas forcément datés. Pour répondre à cette observation, nous construisons une frise temporelle du corpus ne

représentant simultanément que les documents situés dans une certaine fenêtre de temps et mettant dynamiquement en évidence le « glissement » de cette fenêtre sur l'axe du temps. Les traitements statistiques réalisés pour la construction de cette carte restent identiques à ceux présentés précédemment, seul le mode l'affichage change. Nous pouvons mettre ainsi en évidence une diachronie du corpus en montrant que certains groupes peuvent apparaitre, grossir, se réduire ou disparaître (cf. illustration 17).

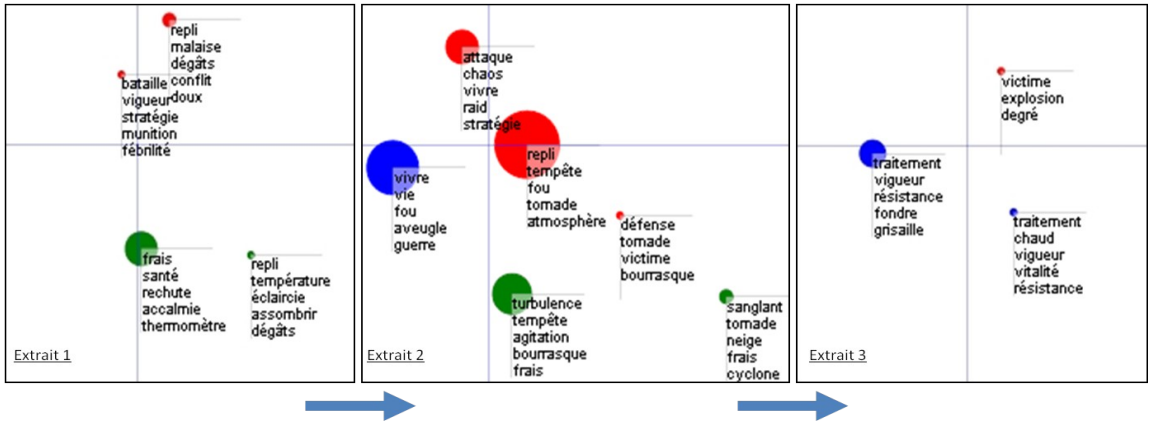


Illustration 17: Frise temporelle de flux documentaire.
La frise est construite avec une fenêtre temporelle d'un mois et une unité de déplacement de cette fenêtre d'un jour.

Le développement de l'outil ProxiDocs a donné lieu à de multiples pistes de recherche très différentes. Nous renvoyons pour plus de détails à la thèse de Thibaut Roy (Roy 2007). Nous avons par exemple travaillé comparativement sur différents types de méthodes statistiques à employer pour les étapes de projection et de classification automatique (*clustering*). Nous avons aussi travaillé sur différents genres textuels sur lesquels on a mis en œuvre un assistance interprétative par cartographie thématique (corpus de textes, forums de discussion, réponses à des requêtes de recherche …).

Une autre expérimentation sur ProxiDocs concernait une étude sur l'ergonomie des cartes thématiques. En collaboration avec des collègues biologiste-éthologue et psychologue (Henri Roussel et Stéphane Breux), nous avons monté une expérimentation avec plusieurs sujets (23 sujets) pour analyser leurs tracés oculaires sur ce type d'objets graphiques relativement à des questions permettant de vérifier la lisibilité des cartes. Nous renvoyons aux articles publiés sur cette expérimentation pour plus de détails (Beust & al. 2008, Roussel & al 2011). Cette étude a permis de mettre en évidence certains résultats intéressants.

Premièrement, nous avons pu prendre conscience de la grande variabilité des profils d'utilisateurs relativement à leur stratégies de parcours oculaires (cf. illustration 18) en dégageant 5 classes d'utilisateurs :

- La classe des sujets dits "mobiles longs" - 3 sujets : ce sont des sujets qui réalisent beaucoup de fixations (20 fixations par image), en moyenne assez longues (0,5 seconde) ;
- La classe des sujets dits "mobiles courts" - 3 sujets : comme pour la catégorie précédente, ce sont des sujets qui réalisent beaucoup de fixations (22 fixations), la différence se situe au niveau de la durée moyenne de fixation un peu plus courte (0,4 seconde) ;

- La classe des sujets dits "sédentaires" - 5 sujets : cette catégorie caractérisent des sujets peu mobiles (assez faible nombre de fixations, environ 15) et dont la durée moyenne de fixations est assez longue (0,6 seconde) ;
- La classe des sujets dits "saccades courtes" - 4 sujets : les sujets de cette catégorie ont réalisé beaucoup de fixations (22 en moyenne) sur des durées moyennes très courtes (0,3 seconde) ;
- La classe des sujets dits "grandes saccades" - 3 sujets : cette dernière catégorie regroupe des sujets réalisant peu de fixations (14 en moyenne) sur des durées très courtes (0,3 seconde).

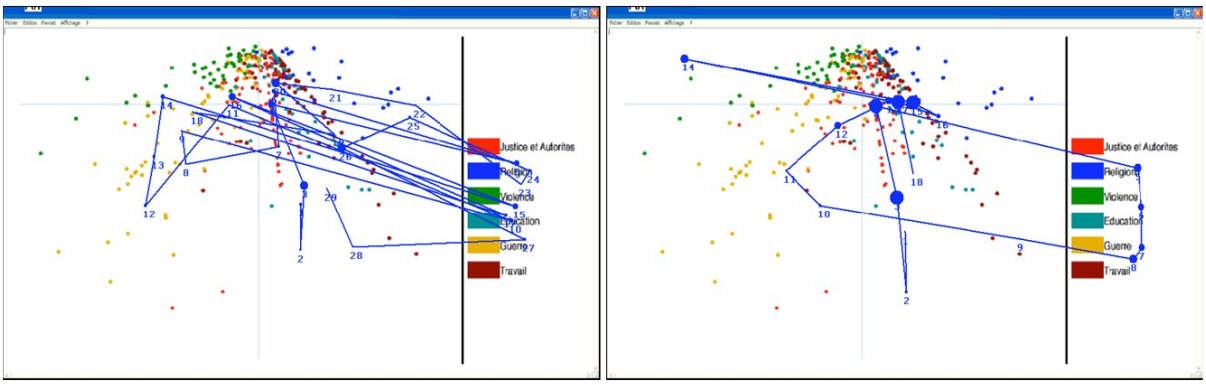


Illustration 18: Parcours du regard de deux sujets différents sur la carte des documents. Les points bleus sur les parcours oculaires représentent les fixations du regard. La taille du point est proportionnel au temps de la fixation.

Deuxièmement, contrairement aux idées préconçues, les temps des durées de fixation ne permettent pas d'expliquer la réussite aux questionnaires vérifiant la lisibilité des cartes. Nous avons pu statistiquement montrer (Roussel & al. 2011) que les transitions entre différentes parties des cartes (légende, cadrans haut droite, haut gauche, bas droite, bas gauche) sont bien plus pertinentes. Nous en tirons la conséquence que la compréhension d'une carte ProxiDoc n'est pas une « captation d'informations » qui du coup serait proportionnelle au temps de fixation de regard sur la carte. C'est en fait bien plus une construction d'information globale.

Les usages de ProxiDocs qui ont été expérimentés ont été nombreux et très riches d'enseignements, aussi bien, par exemple, pour l'analyse de corpus médicaux (Roy & Névéol 2006), que pour l'analyse de discussions entre professionnels de l'éducation (Roy 2007, p. 154-167). Il en est ressorti presque à chaque expérimentation que l'environnement interactif et visuel de ProxiDoc était d'un usage complémentaire aux outils que les utilisateurs manipulent dans leurs tâches quotidiennes. C'est ainsi que s'est imposée l'idée de considérer d'une manière opérationnelle une assistance personnalisée à l'interprétation comme une composante à part entière de l'environnement de travail de l'utilisateur. Pour aller dans ce sens nous avons chercher à appliquer les principes de ProxiDocs à une tâche quotidienne des utilisateurs à forte dimension textuelle : la gestion du courrier électronique. C'est le projet CartoMail.

3.1.5. CartoMail

Les utilisateurs de messageries électroniques n'ont jamais été aussi nombreux et il est fréquent qu'ils utilisent conjointement plusieurs comptes de messagerie électronique. Du point de vue des usages, les applications de client de messagerie (Outlook, Thunderbird pour les plus répandues) ainsi

que les services WebMails (Gmail, Hotmail …) sont de moins en moins vus comme des applications et services parmi tant d'autres. De même que les périphériques de connexion réseau, ils tendent à devenir des éléments de base incontournables des ENT au même titre, par exemple, que les gestionnaires de fichiers. Dans bon nombre de cas, on constate même que la différence entre architecture de dossiers/fichiers et boites de courriers électroniques est une distinction de plus en plus artificielle relativement aux usages et tâches des utilisateurs.

Les interactions offertes par les clients de messagerie et les architectures WebMails sont relativement pauvres du point de vue des fonctionnalités offertes aux usagers. En dehors de l'architecture hiérarchique des boites et des fonctions de recherche, l'utilisateur ne dispose pas vraiment d'outils pour manipuler et appréhender ses courriers électroniques globalement alors que, paradoxalement, la messagerie est à la fois le contexte de travail et l'environnement d'archivage du travail en cours.

Le but du projet CartoMail est de réaliser une application qui permette de visualiser de manière interactive et personnelle un compte de messagerie électronique. L'application CartoMail est une extension de ProxiDocs réalisée au stade de maquette dans le cadre d'un projet d'étudiants de Master en informatique (au département Informatique de l'Université de Caen Basse-Normandie). Cette maquette devra encore être améliorée car l'application n'a pas été assez finalisée dans la durée du stage pour la diffuser. L'application CartoMail analyse le compte de messagerie et produit en temps réel des visualisations interactives et personnalisables présentées à travers un navigateur HTML. Cette application ne doit pas être vue comme une alternative aux clients de messagerie et aux WebMails mais, au contraire, dans l'optique de son intégration à un ENT, elle doit être pensée dans un usage complémentaire aux outils de l'utilisateur et à son architecture hiérarchique de boites mail.

CartoMail propose à son utilisateur plusieurs types de visualisations interactives :

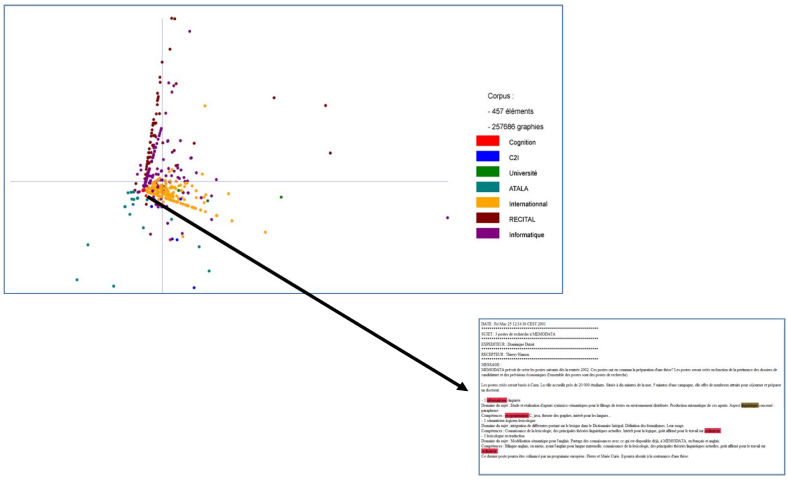


Illustration 19: Carte thématique de mails.
Chaque point représente un mail colorié de la couleur du thème majoritaire. Un clic sur un point affiche un coloriage thématique du mail.

- une carte thématique des messages (avec possibilité en cliquant sur un message d'en visualiser un coloriage thématique) qui donne une vision globale de l'espace thématique de la boite de courriers électroniques dans son ensemble (cf. illustration 19) :
- une carte de regroupement en classes thématiques des messages proches (avec possibilité d'établir un rapport statistique de la classe et un affichage des termes les plus occurrents de la classe) (cf. illustration 20) :

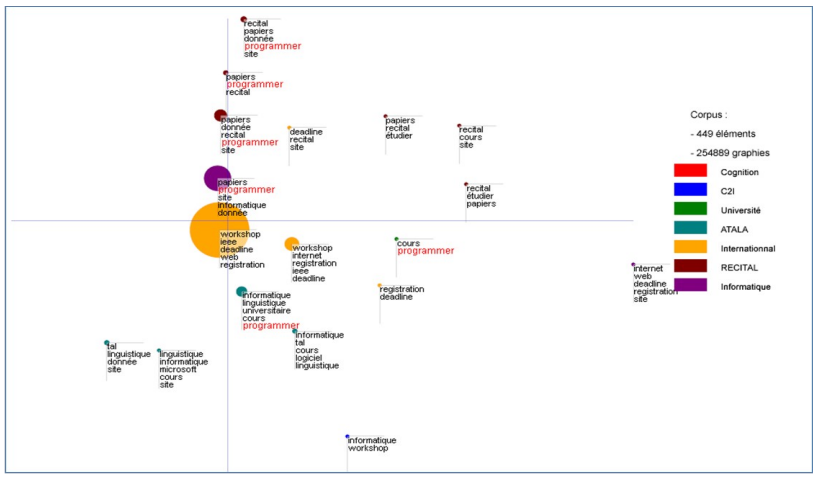


Illustration 20: Carte de groupes de mails.
Chaque disque représente un groupe de mail colorié de la couleur du thème majoritaire. Un passage de souris sur un terme affiche en rouge les autres occurrences du terme dans les autres groupes.

- une carte des messages dans laquelle apparaissent de manière singulière les nouveaux messages arrivés dans la boite de messagerie (cette visualisation est proposée selon 2 méthodes au choix de l'utilisateur : soit projeter les nouveaux messages dans l'espace des anciens messages, soit recalculer un nouvel espace de projection en tenant compte des nouveaux messages arrivés)(cf. Illustration 21) :

Illustration 21: Carte de mails.
Les points qui apparaissent en plus grosse taille que les autres indiquent la place des nouveaux messages dans l'espace thématique de la messagerie.

3.2. Expérimentations sur corpus

Les différents logiciels d'étude présentés précédemment ont fait l'objet de nombreuses expérimentations sur corpus et l'objectif ici n'est pas d'une faire une description exhaustive. Ces expérimentations ont surtout été menées dans le cadre des thèses de Vincent Perlerin et Thibault Roy (auxquelles nous renvoyons pour des détails). Monter des protocoles expérimentaux avec des utilisateurs en situation est un travail qui demande un certain temps de préparation, de réalisation et d'analyse, temps qui n'est jamais rendu possible dans un travail plus court (par exemple un projet de Master). Les principaux champs d'expérimentation ont été les suivants et nous indiquons pour chaque champ d'étude le corpus étudié, le ou les outils mobilisés et les références bibliographiques à consulter pour plus de détails:

- Observations d'énoncés de dialogue :
  - corpus : 2 heures de dialogue à trois sur la conception d'une documentation utilisateurs
  - outil : Anadia
  - références : (Beust 1998), (Nicolle & al. 2002)
- Analyse de corpus journalistiques :
  - corpus : Le Monde sur CD-ROM – 1987/1989 (plus de 500 articles : 500 000 mots).
  - outils : ThèmeEditor, LUCIA, ProxiDocs
  - références : (Beust 2002), (Perlerin & Beust 2003), (Roy & Beust 2004), (Roy & Beust 2005)
- Analyse de métaphores conceptuelles
  - corpus : Le Monde sur CD-ROM – 1987/1989
  - outils : LUCIA, ProxiDocs
  - références : (Perlerin & al. 2003), (Roy & al. 2007)
- Veille documentaire et stratégique sur Internet
  - corpus : corpus sur la Biocorrosion, pages du site de la société CORRODYS : 3445 mots.
  - outils : LUCIA
  - références : (Perlerin 2004)
- Assistance dans une recherche d'information médicale
  - corpus : 70 documents en français extraits du catalogue CISMeF
  - outils : ProxiDocs
  - références : (Roy 2007), (Névéol 2005)
- Observation des usages d'une terminologie professionnelle dans des forums
  - corpus : CALICO[17] – 6 forums d'échanges entre professeurs des écoles stagiaires entre 2002 et 2004: environ 145000 mots
  - outils : ProxiDocs
  - références : (Roy 2007)
- Analyse de terminologie pour l'indexation de ressources documentaires

17 CALICO (Communautés d'Apprentissage en Ligne, Instrumentation, COllaboration) est une Équipe de Recherche Technologique éducation (ERTé) qui s'est réunie de 2006 à 2009 qui avait un objectif de recherche sur les formations à caractère professionnalisant se déroulant partiellement ou totalement à distance et qui intègrent des modalités de travail collaboratif. La plateforme Calico est toujours accessible à l'URL http://www.stef.ens-cachan.fr/calico/calico.html ou bien http://woops.crashdump.net/calicorss2/. Pour consulter le rapport de l'ERTé : http://www.stef.ens-cachan.fr/calico/calico_rapport_final_fev_2010.

- corpus : 130 résumés de thèses : 56 451 mots
- outils : ProxiDocs
- références : (Roy 2007), (Beaudouin 2008)

La confrontation à des corpus et à des genres est une démarche importante qui est riche d'enseignement mais plus encore c'est la confrontation à de « vrais » utilisateurs qui est ce que nous recherchions. En tant que concepteurs d'outils logiciels centrés sur l'utilisateur nous étions dans une situation inconfortable dès lors qu'il fallait appliquer nos outils à nos propres perspectives d'analyse car la « double casquette » concepteur/utilisateur est créatrice de biais potentiels dans l'analyse des usages. Il est assez probable que la capacité à contourner un usage prévu de manière fortuite est accrue lorsque l'utilisateur n'a qu'une idée fonctionnelle mais pas conceptuelle de l'outil. Et justement, ces contournements font partie de ce qu'on veut pouvoir observer.

Dans la suite de ce chapitre nous allons nous arrêter rapidement sur deux de ces expérimentations : l'analyse de métaphores conceptuelles et l'indexation de ressources documentaires. Toutes deux à leur manière valident l'approche expérimentale déployée et font émerger des questions intéressantes.

3.2.1. Analyse de métaphores conceptuelles

L'analyse de métaphores conceptuelles est un projet mené (entre 2002 et 2007) en collaboration avec Stéphane Ferrari et successivement Vincent Perlerin dans un premier temps puis Thibault Roy dans un second temps (projets appelés ISOMETA 1 puis ISOMETA 2). Stéphane Ferrari travaille depuis sa thèse (Ferrari 1997) sur les métaphores conceptuelles et conventionnelles au sens de (Lakoff & Johnson 1985). L'objectif du projet ISOMETA (dans ses deux phases) était de se livrer à une analyse interprétative de ces métaphores avec une approche componentielle différentielle des domaines lexicaux sources et cibles. Le corpus étudié était constitué de l'actualité boursière du journal *Le Monde* de 1987 à 1989 totalisant environ 500 000 mots et les métaphores observées ont concerné l'usage des vocabulaires météorologique, guerrier et de la santé dans le domaine boursier. Les résultats et analyses menées sont nombreuses. Des méthodes de visualisations spécifiques ont été développées en utilisant des technologies XML, XSLT. Des études ont été menées en synchronie et en diachronie notamment au moyen de visualisations dynamiques ProxiDocs. Je ne les détaille pas toutes ici mais je renvoie à (Roy 2008) et plus récemment (Ferrari 2010) pour un exposé beaucoup plus exhaustif. Un des résultats de ISOMETA que je voudrais reprendre ici concerne un questionnement sur une typologie des sèmes utilisés.

Pour amorcer notre étude des usages métaphoriques des thèmes de la météorologie, la guerre et la santé, dans le champ de l'actualité boursière nous avons développé 3 dispositifs[18] LUCIA afin de structurer en tables et en sèmes les termes retenus pour les trois thèmes. Dans un premier temps, nous avions structuré les trois thèmes avec des sèmes tout à fait indépendants d'un domaine à l'autre. Nous nous sommes alors aperçus de manière assez logique que les isotopies observées ne permettaient que de rendre compte en contexte de l'alternance des thèmes mais pas d'un apport interprétatif de la métaphore. Nous nous sommes donc remis au travail de structuration lexicale en cherchant le plus possible à utiliser certains sèmes dans plusieurs dispositifs de sorte à s'attendre à analyser les métaphores en terme d'isotopies trans-domaines (par exemple une isotopie d'un même sème dans un emploi métaphorique porté par une lexie d'un domaine et par une autre lexie d'un autre domaine). Cela donne effectivement d'autres résultats plus intéressants du point de vue des métaphores (on constate ici notamment que la création de ressources lexicales est bien dépendante d'un objectif interprétatif). Les

18 Les ressources termino-ontologiques produites avec LUCIABuilder sont appelées *dispositifs* (Perlerin 2004).

ressources lexicales finalement produites et utilisées relativement à l'interprétation des métaphores comptaient les nombres de lexies suivants :

Domaine	Nb de lexies de la ressource	Nb total de lexies et de leurs flexions dans les articles du corpus
Guerre	64	831 occurrences
Météo	112	569 occurrences
Santé	111	485 occurrences

La dynamique des sèmes observée dans les usages métaphoriques nous ont permis de rendre compte de deux formes d'apports interprétatifs de la métaphore déjà décrits notamment chez (Indurkhya 1987) : l'interprétation par analogie et l'interprétation par nouveauté.

L'interprétation par analogie de la métaphore est le résultat d'une isotopie d'un sème commun entre des lexies des domaines sources et cibles. C'est par exemple le cas dans l'extrait ci-dessous (illustration 22), provenant de l'article n°126 du corpus, où nous avons observé une isotopie du sème [Rôle : /intervient/ vs. /étudie, analyse/] portée par les lexies *Dow Jones* et *thermomètre* et actualisée de part et d'autre sur la valeur de sème /étudie, analyse/. Ici l'analogie peut s'interpréter alors de la façon suivante : *à la façon d'un thermomètre dans le domaine de la météorologie, le Dow Jones est un outil d'étude et d'analyse du domaine de la bourse.*

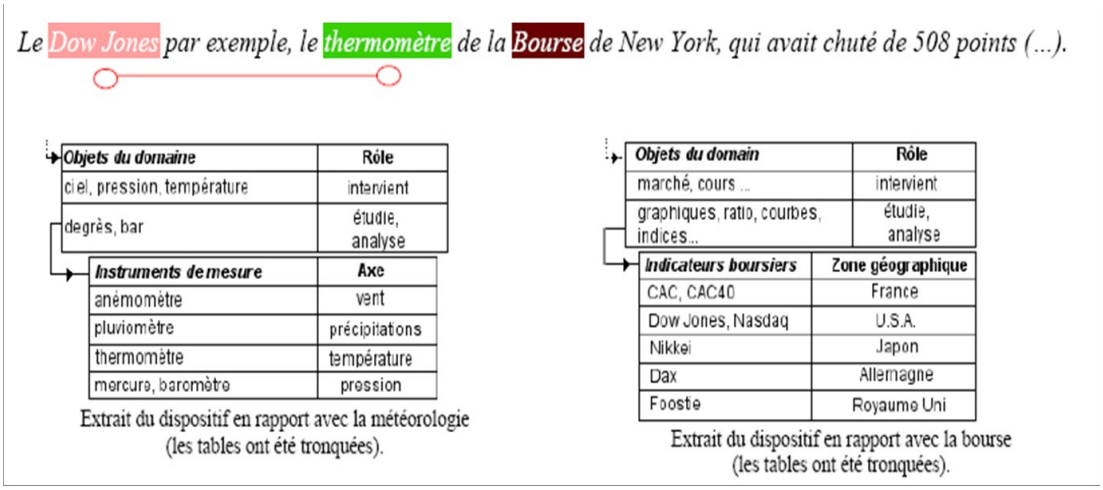


Illustration 22: Interprétation par analogie d'un emploi métaphorique. En dessous de l'exemple, nous avons reproduit les extraits de dispositifs LUCIA où sont définies les lexies thermomètre et Dow Jones. On constate que ces deux lexies portent de manière générique le sème [Rôle : /intervient/ vs. /étudie, analyse/].

L'interprétation de la métaphore en terme de nouveauté exprime l'actualisation en contexte d'un sème du domaine source. Considérons par exemple l'extrait ci-dessous (illustration 23), provenant de l'article n°153 du corpus. Un faisceau de 2 isotopies est remarquable entre les lexies *krach* et *tempête* (lexie marquant un emploi métaphorique) : l'isotopie du sème [Direction : /descend/ vs. /Monte/] et [Évaluation : /bon/ vs. /mauvais/ vs. /pas connoté/]. Les deux lexies actualisent des valeurs différentes du sème Direction (/descend/ pour *krach* et /monte/ pour *tempête*). On considère de ce fait que le sème est ici virtualisé. Ce n'est pas le cas du sème Evaluation qui est bien redondant avec la même valeur de

sème. Comme dans l'exemple précédent, cela caractérise un apport interprétatif par analogie de la métaphore. Le sémème de *tempête* actualise le sème [Force : /violent/ vs. /très violent/] avec sa valeur actualisée /violent/. Ce sème n'est pas un sème propre du domaine de la météorologie (à la différence par exemple dans l'illustration 22 de [Axe : /vent/ vs. /précipitations/ vs. /température/ vs. /pression/]). C'est un sème partageable. En tant que tel, il nous paraît ici être actualisable marquant un apport par nouveauté de la métaphore que nous pourrions interpréter de la sorte : *le krach boursier est un phénomène évalué négativement qui, à l'identique d'une tempête, est en plus un phénomène violent.*

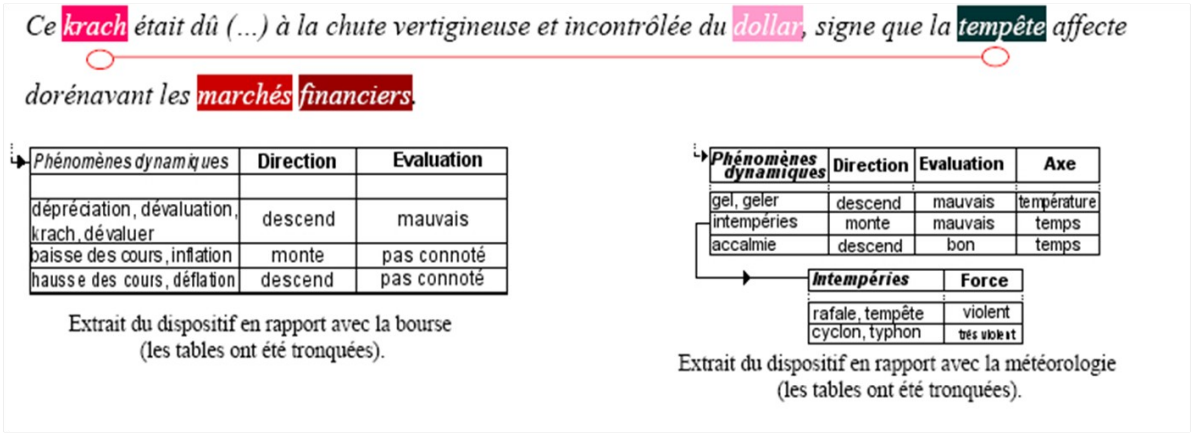


Illustration 23: Interprétation par nouveauté d'un emploi métaphorique. En dessous de l'exemple, nous avons reproduit les extraits de dispositifs LUCIA où sont définies les lexies *krach* et *tempête*. On constate que *tempête* porte un sème spécifique partageable [Force : /violent/ vs. /très violent/].

Ces deux exemples d'analyse d'emplois métaphoriques sont particulièrement intéressants du fait de la dynamique sémique qu'ils mettent en évidence. Plus largement que la métaphore, ce qui nous intéresse ici est de comprendre et de décrire de manière opératoire des phénomènes de base de toutes les interprétations. De ce point de vue, nous considérons que les tropes ne sont pas par principe des formes particulières de chaînes syntagmatiques mais, au contraire, qu'ils mettent en évidence (presque plus clairement encore) des principes interprétatifs généraux. L'étude des interprétations en terme d'actualisations et de virtualisation de sèmes est un enjeu de la sémantique interprétative. Nous y contribuons ici en cherchant à apporter des éléments de description en complément. La sémantique interprétative prévoit une typologie des sèmes en terme de sèmes génériques et spécifiques relativement à leur fonction dans le sémème. Nous considérons qu'on peut ajouter à cette typologie des sèmes une autre distinction entre des sèmes propres et des sèmes partageables. Cette distinction n'est pas motivée par une fonction au sein du sémème mais par une nature du sème relativement aux champs lexicaux décrits. Ces distinctions nous semblent complémentaires dans un modèle analytique de la dynamique sémique. De plus, nos exemples à propos des métaphores nous montrent que le modèle différentiel du sème (introduit avec Anadia et utilisé dans les descriptions componentielle différentielle LUCIA) permet de décrire finement des phénomènes de virtualisation ou d'actualisation grâce au jeu des valeurs de sèmes actualisées ou pas dans les isotopies. Ceci nous incite à poursuivre dans la direction pour chercher à décrire de la même façon d'autres formes d'opérations interprétatives

basées sur l'afférence de sème, par exemple les cas d'assimilation[19] et de dissimilation[20]. C'est notamment un objectif d'un projet qui démarre dans la suite de ISOMETA : le projet SemComp (projet sur les applications de la Sémantique Componentielle) dont il sera à nouveau question dans le chapitre n°7 « Projets et perspectives ».

Cette étude sur la métaphore est intéressante également d'un autre point de vue. Nous nous sommes confrontés à la description componentielle de plusieurs champs lexicaux simultanément et nous avons bien ressentis que deux aspects sont absolument déterminants dans cette tâche : l'objectif interprétatif (le fait de vouloir expliquer un lien métaphorique) et le rapport au corpus. Le temps de la représentation lexicale n'est pas une phase préalable au temps du traitement et de l'analyse syntagmatique. Les corpus et les ressources se construisent et s'analysent en même temps au sein d'une boucle où le projet de l'utilisateur est au centre. C'est pour nous un argument pour affirmer encore une fois la pertinence d'une approche située et centrée-utilisateur. De manière opératoire, il convient donc de mettre en place des outils installant des interactions entre corpus et ressources afin de permettre à un utilisateur d'amorcer un cycle de représentation terminologique relativement à sa tâche.

3.2.2. Indexation de ressources documentaires

Une autre expérimentation sur corpus intéressante de nos développements logiciels est celle réalisée par Nathalie Beaudouin en 2007. Dans une perspective de caractérisation de contenu pour une indexation de ressources documentaires, elle a utilisé les outils ThèmeEditor et ProxiDocs. En comparaison notamment avec notre travail décrit précédemment sur les métaphores, cette expérimentation présentait un intérêt de taille : nous n'étions pas l'utilisateur ce qui nous permettait de pouvoir recueillir un avis objectif d'un utilisateur (en l'occurrence une utilisatrice) qui n'avait pas participé au développement des outils.

Nathalie Baudouin a réalisé une thèse en sciences du langage à l'université de Rouen intitulée *Problèmes d'ergonomie linguistique en traitement d'images : une approche socioterminologique* (Beaudouin 2007). Elle s'y penche sur les problématiques linguistiques que soulève la conception d'une plateforme logicielle dédiée au traitement d'images. S'attachant à résoudre les problèmes d'ergonomie linguistique d'une interface homme-machine, elle fonde son approche dans le courant de la socioterminologie. Il s'agit, en l'occurrence, de veiller à ce que la terminologie utilisée (par exemple dans le cadre de l'aide en ligne) rende compte des usages d'une communauté d'utilisateurs issus d'horizons différents.

La méthodologie mise en place vise à constituer un corpus oral fondé sur l'interview d'experts du domaine pour réaliser une « base de connaissances socioterminologique » (appelée Termedit) dont la

19 L'assimilation est une afférence co-textuelle (i.e. le sème afférent est inhérent dans d'autres sémèmes du co-texte) d'un sème générique par présomption d'isotopie. Par exemple dans *Voici des choux, des concombres et des scoubidous* on peut décrire un enrichissement du contenu sémique de 'scoubidou' avec le sème afférent générique /légume/ (on pourrait dire que le sème /vert/ est également inhérent à 'choux' et 'concombre' mais il est spécifique donc il n'est pas objet d'assimilation, ce que l'on constate par exemple dans *Voici des choux, des concombres et des carottes*)

20 La dissimilation est l'opération interprétative inverse de l'assimilation. Alors que l'assimilation diminue, par afférence, les contrastes forts, la dissimilation, quant à elle, augmente les contrastes faibles. La dynamique des sèmes en cause dans la dissimilation n'est plus une afférence de sème générique, comme c'est le cas pour l'assimilation, mais une afférence de sèmes spécifiques pour différencier, dans un co-texte, les sémèmes appartenant au même taxème. Exemple *Il y a musique et musique*.

structure réponde aux besoins des utilisateurs de la plateforme et tienne compte de son évolution. Cette base terminologique offre des champs d'entrée particuliers introduits par Nathalie Beaudouin et extraits des entretiens : la fonction typique et l'objet typique. Ces champs sont au centre de la problématique ergonomique. La fonction typique vise les différentes dénominations des opérations algorithmiques réalisées et enchaînées par le logiciel dans le traitement de l'image (par exemple *binariser l'image, seuillage de l'image, reconnaissance de forme ...*). L'objet typique est lié aux différents termes qui indiquent ce sur quoi portent les opérations du logiciel (par exemple *pixel, contour ...*). La question de l'évolution de Termedit est essentielle afin qu'elle puisse continuer à avoir une pertinence socioterminologique et donc une utilisation en ergonomie linguistique. Afin que Termedit suive en temps réel l'avancée des connaissances en traitement d'image (néologie et nouveaux usages), Nathalie Beaudouin cherche à l'enrichir en extrayant des patrons sur un corpus dynamique car en prise sur l'avancée des connaissances : un corpus de résumés de thèses et d'HDR principalement en traitement d'image (ce corpus est disponible et mis à jour régulièrement sur la base de données en lignes Pascal produite à l'INIST).

Pour mieux prendre en compte ce corpus de ressources documentaires dans l'alimentation de la base Termedit, un préalable était d'indexer les ressources par des termes fréquemment utilisés dans les domaines scientifiques connexes au traitement d'image. Nathalie Beaudouin connaissait déjà bien les termes qu'elle voulait utiliser pour les avoir déjà recensés dans 5 thèmes différents :

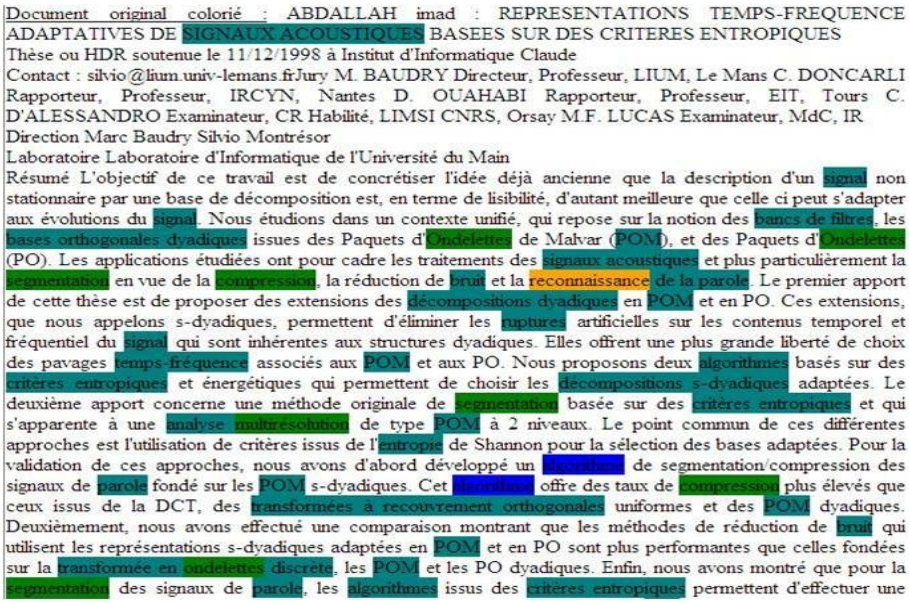
- Reconnaissance de formes : 21 termes[21]
- Intelligence artificielle : 41 termes
- Vision : 193 termes
- Traitement du signal : 222 termes
- Image : 183 termes

Elle connaissait bien par ailleurs son corpus de résumés (130 résumés de thèses et d'HDR, 56 451 graphies) et cherchait dans l'usage de ProxiDocs et ThèmeEditor une mise en relation automatique des termes et des résumés.

L'interêt de l'usage conjoint de ProxiDocs et ThèmeEditor dans une perspective telle que celle de Nathalie Beaudouin était de pouvoir offrir une approche du corpus d'une part très locale grâce au coloriage thématique des textes (cf. illustration 24) et d'autre part très globale grâce aux cartes de corpus et de groupes (cf. illustration 25). De plus les rapports de regroupement de ProxiDocs (document moyens, statistiques du groupe relatives au corpus) permettent de mettre en rapport le local et le global dans une appréhension intertextuelle du corpus, appréhension guidée par des ressources terminologiques propres. Bien que connaissant parfaitement son corpus et ses ressources lexicales, c'est ce rapport complexe entre local et global qui manquait à Nathalie Baudouin dans sa tâche d'indexation. C'est aussi un constat fait par Aurèlie Névéol dans son utilisation de ProxiDocs sur l'indexation d'extraits du catalogue CISMeF (Névéol 2005).

Cette expérimentation est intéressante pour plusieurs raisons. Avant toute chose, elle montre comment enrichir une tâche interprétative avec une approche globale inter-textuelle et personnelle du corpus. L'approche centrée utilisateur n'est pas ici une alternative à une démarche guidée par des ressources terminologique généralistes, au contraire c'est une nécessité. En effet, il est fort peu

21 Dans chacun des 5 thèmes les termes identifiés allaient de lexies simples telles que *classification* à des lexies beaucoup plus complexes telles que *fonction de densité de probabilité*



Document original colorié : ABDALLAH imad : REPRESENTATIONS TEMPS-FREQUENCE ADAPTATIVES DE SIGNAUX ACOUSTIQUES BASEES SUR DES CRITERES ENTROPIQUES
Thèse ou HDR soutenue le 11/12/1998 à Institut d'Informatique Claude
Contact : silvio@lium.univ-lemans.frJury M. BAUDRY Directeur, Professeur, LIUM, Le Mans C. DONCARLI Rapporteur, Professeur, IRCYN, Nantes D. OUAHABI Rapporteur, Professeur, EIT, Tours C. D'ALESSANDRO Examinateur, CR Habilité, LIMSI CNRS, Orsay M.F. LUCAS Examinateur, MdC, IR
Direction Marc Baudry Silvio Montrésor
Laboratoire Laboratoire d'Informatique de l'Université du Main
Résumé L'objectif de ce travail est de concrétiser l'idée déjà ancienne que la description d'un signal non stationnaire par une base de décomposition est, en terme de lisibilité, d'autant meilleure que celle ci peut s'adapter aux évolutions du signal. Nous étudions dans un contexte unifié, qui repose sur la notion des bancs de filtres, les bases orthogonales dyadiques issues des Paquets d'Ondelettes de Malvar (POM), et des Paquets d'Ondelettes (PO). Les applications étudiées ont pour cadre les traitements des signaux acoustiques et plus particulièrement la segmentation en vue de la compression, la réduction de bruit et la reconnaissance de la parole. Le premier apport de cette thèse est de proposer des extensions des décompositions dyadiques en POM et en PO. Ces extensions, que nous appelons s-dyadiques, permettent d'éliminer les ruptures artificielles sur les contenus temporel et fréquentiel du signal qui sont inhérentes aux structures dyadiques. Elles offrent une plus grande liberté de choix des pavages temps-fréquence associés aux POM et aux PO. Nous proposons deux algorithmes basés sur des critères entropiques et énergétiques qui permettent de choisir les décompositions s-dyadiques adaptées. Le deuxième apport concerne une méthode originale de segmentation basée sur des critères entropiques et qui s'apparente à une analyse multirésolution de type POM à 2 niveaux. Le point commun de ces différentes approches est l'utilisation de critères issus de l'entropie de Shannon pour la sélection des bases adaptées. Pour la validation de ces approches, nous avons d'abord développé un algorithme de segmentation/compression des signaux de parole fondé sur les POM s-dyadiques. Cet algorithme offre des taux de compression plus élevés que ceux issus de la DCT, des transformées à recouvrement orthogonales uniformes et des POM dyadiques. Deuxièmement, nous avons effectué une comparaison montrant que les méthodes de réduction de bruit qui utilisent les représentations s-dyadiques adaptées en POM et en PO sont plus performantes que celles fondées sur la transformée en ondelettes discrète, les POM et les PO dyadiques. Enfin, nous avons montré que pour la segmentation des signaux de parole, les algorithmes issus des critères entropiques permettent d'effectuer une

Illustration 24: Coloriage thématique ThèmeEditor d'un résumé de thèse. Le thème vert foncé majoritaire est celui du traitement du signal.

probable qu'on puisse trouver une ressource terminologique comprenant des lexies aussi complexes que *transformées à recouvrement orthogonales* ou encore *réseaux de Petri partiellement stochastiques*.

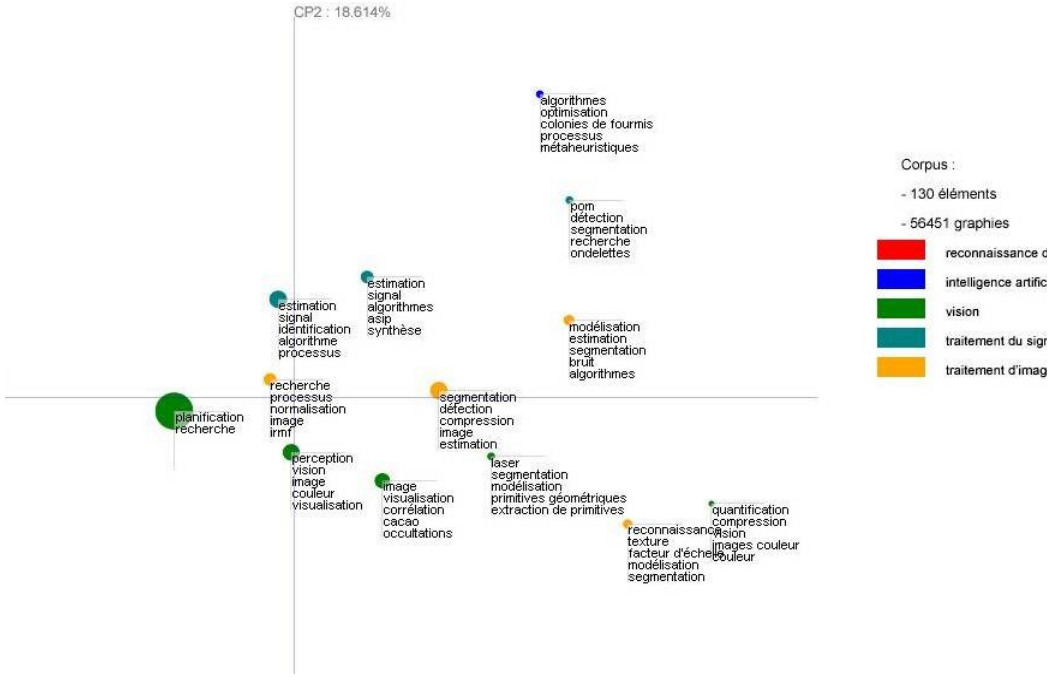


Illustration 25: Carte de regroupements ProxiDocs

De plus cette expérimentation remet en cause une certaine vision naïve de ce qu'est l'indexation. Une tâche d'indexation ne se réduit pas à une extraction de contenu car elle donne manifestement une place importante à l'influence du corpus sur les index choisis pour les textes. De ce point de vue, indexer est une tâche interprétative à part entière et en tant que telle elle est fortement conditionnée par les rapports inter-textuels.

Enfin, le travail de Nathalie Beaudouin et l'usage des outils que nous lui avons mis à disposition montrent une collaboration pluridisciplinaire fructueuse. Pour alimenter un développement informatique, on a recours à un travail d'ergonomie linguistique qui lui même fait intervenir des outils informatique d'étude de corpus. Il en ressort que la problématique linguistique n'est pas ici simplement appliquée à un développement informatique mais en constitue une partie intégrante motivée par la prise en compte de l'utilisateur. Pour reprendre une formulation de François Rastier, ceci montre bien plus qu'une linguistique *appliquée*, mais plutôt une linguistique *impliquée*. ProxiDocs et ThèmeEditor ont aidée à cette implication et c'est de notre point de vue très positif.

3.3. Développements industriels

Comme on l'a vu précédemment, quand un logiciel que l'on développe avec une certaine approche et des idées sur une tâche est repris et utilisé par quelqu'un d'autre que les développeurs, il en ressort des retours d'expériences très riches. Dans une approche centrée-utilisateur c'est encore plus vrai, car il est toujours difficile, délicat et pas très satisfaisant pour le concepteur de se mettre à la place de l'utilisateur. Les collègues universitaires qui ont utilisé les outils et nous ont fait des retours de qualité comprenaient bien que l'outil qui leur étaient proposé est un logiciel d'étude de laboratoire. Ils avaient une certaine indulgence (dont nous les remercions) sur les aspects fonctionnels qui restent à optimiser (temps de traitement, passage à l'échelle). Une spécificité d'une démarche industrielle serait de travailler à ce que l'outil soit de ce point de vue complètement finalisé, acquérant ainsi un statut de produit là où il n'était avant qu'un logiciel d'étude de laboratoire. C'est nécessaire parce que les utilisateurs ne sont plus des universitaires bienveillants mais des clients exigeants. Par abus de langage, on pourrait presque dire des « utilisateurs normaux ». C'est naturellement assez motivant dans une démarche centrée-utilisateur de se tourner vers eux.

Encore faut-il que le produit soit à la hauteur de ce qu'un client attend. Pour cela, il nous semble qu'il faut veiller à deux aspects. Une maturité des idées, de l'approche, des concepts mis en œuvre et une qualité de développement technique sans faille. Même en ayant la chance de travailler avec des collègues qui maîtrisaient très bien les aspects techniques du développement logiciel (c'était le cas de Vincent Perlerin et Thibault Roy) il s'avère très rare qu'un prototype de laboratoire puisse être décliné en produit industriel. Ce n'est le cas ni de ThèmeEditor, ni de LUCIA, ni de ProxiDocs car il faudrait quasiment envisager un redéveloppement intégral (éventuellement même dans d'autres langages que le langage Java utilisé). Au tout début du travail de développement, un prototype de laboratoire cherche à prouver le bien-fondé scientifique des objets impliqués, alors qu'au tout début du développement logiciel d'un produit on anticipe les aspects techniques qui se poseront par la suite. L'un n'est pas meilleur que l'autre. Les deux sont différents. Dans un cas on est dans un travail de recherche. Dans l'autre, on est dans un travail d'ingénierie.

Dans notre cas, la question n'est donc pas de valoriser un prototype de laboratoire en produit industriel, ce qui traditionnellement se fait par des contrats entre laboratoires de recherche et sociétés de développement, contrats justement appelés *contrats de valorisation*. La question est de s'entendre avec un partenaire industriel pour qu'il développe (avec son savoir faire) des produits qui mettent à

profit les idées et approches qu'on lui propose. On donne ainsi un autre devenir à des idées de recherche et c'est tout à fait motivant. La société eXo maKina (présentée ci-dessous) s'est signalée intéressée auprès de moi pour développer un outil centré utilisateur permettant de donner un point de vue global à un utilisateur sur une collection documentaire. Cet outil est actuellement en cours de développement et est appelé Canopée©.

3.3.1. La société eXo maKina

La société eXo maKina[22] est une petite SARL parisienne née de la rencontre de Roger Cozien, spécialiste de la sécurité et des technologies numériques, et de Dominique Haglon, développeuse informatique. J'ai rencontré Roger Cozien lors des journées de Rochebrune de 2005 et nous sommes depuis restés en contact régulièrement. J'ai suivi ses évolutions professionnelles et notamment la création de sa société en 2009.

La société eXo maKina est spécialisée dans la conception et la production de logiciels spécifiques à destination d'usages professionnels. Ses principaux clients sont les agences de presse et les services de défense et de renseignement. Elle développe notamment des solutions logicielles innovantes dans les technologies d'analyse d'images :

- marquage (propriété, source …) implicite et invisible d'images résistant aux altérations (logiciel Tantale©)
- analyse des structures mathématiques et statistiques des images contrefaites destiné à la photo-interprétation avancée[23] (logiciel Tungstène©).

Les photo-montages et les contrefaçons d'images ne sont pas récents. Dès que l'image a servi un but politique ou commercial, s'est posée la question de l'existence et de l'authenticité de la scène représentée. L'avènement du numérique n'a que peu changé de choses dans les buts poursuivis. En revanche, l'informatique de masse a diffusé à la fois les outils de création de retouches et contrefaçons ainsi que les vecteurs de diffusion de ces images aux contenus altérés.

eXo maKina développe la première plateforme logicielle spécialisée dans l'assistance à la photo-interprétation avancée pour l'identification des images numériques altérées. Les altérations dont il est ici question n'ont pas de visées artistiques mais poursuivent un but unique : provoquer une version alternative du sens de l'image/photographie originale et par conséquent, induire, chez l'observateur, une réalité perçue différente, voire mensongère. Tungstène intègre une palette de filtres et de fonctions d'analyse qui autorisent l'opérateur à entrer au plus profond de l'image et à y rechercher les incohérences et altérations qui se répartissent en deux grandes familles. Premièrement, les ruptures dans les statistiques profondes de l'image numérique. Deuxièmement, les incohérences dans les lois physiques qui régissent la diffusion des rayons de lumière ainsi que la chrominance[24]. Parallèlement à la stricte technique informatique, la société développe une méthodologie complète d'interprétation des

22 http://www.exomakina.com/ consultée le 11/12/12.

23 Le 2/5/11, le logiciel Tungstène a été utilisé par des organismes de presse pour démontrer quelques heures seulement après son apparition dans les dépêches de presse que la photographie du visage de O. Ben Laden le jour de sa mort était un faux composé à partir de clichés antérieurs à la mort (cf. Nouvel Observateur, numéro 2427 du 12/05/11, p. 106-107).

24 La chrominance désigne la partie de l'image correspondant à l'information de couleur. Elle est complémentaire de la luminance correspondant à l'intensité de lumière.

résultats du logiciel afin d'aider l'opérateur à identifier les réelles tentatives de désinformation et d'ingérence par l'image. Cette méthodologie opérationnelle est basée sur la sémiotique de l'image que la société a élaborée avec Serge Mauger (GREYC CNRS UMR 6072).

Tungstène ambitionne une extension de la vision et de l'intuition de l'expert. Le logiciel favorise la boucle de rétro-action positive entre cet expert et l'image. C'est un logiciel qui, comme les outils que nous développons en TAL, ne se substitue pas à l'interprétation de son expert-utilisateur mais, au contraire, enrichit l'interprétation humaine dans (et par) la boucle d'interaction. Les conditions d'interprétation des images et des textes sont manifestement très comparables, notamment en ce qui concerne l'apport du contexte, de la culture et des autres documents (images ou textes) au sein d'un corpus. En suivant le même type d'approche eXo maKina cherche à développer un produit qui assiste son utilisateur dans une tâche d'assistance à l'interprétation de flux documentaires. C'est le produit Canopée.

3.3.2. Le produit Canopée

eXo maKina s'intéresse à l'extraction des éléments linguistiques qui font sens dans les documents (numériques) textuels (et plus particulièrement ceux qui s'accompagnent d'images, par exemple de dépêches d'agence de presse). L'approche vise à offrir à chaque utilisateur/analyste la possibilité de définir son propre lexique et ses centres d'intérêt pour explorer des collections documentaires et en mettre en évidence de manière personnalisée les grandes tendances sémiotiques. Dans les tâches numériques quotidiennes de la majeure partie des utilisateurs, l'appréhension de collections documentaires revêt une importance croissante. Ainsi, l'accès à un dossier ou à un répertoire est déjà une confrontation à une collection potentiellement assez grande. La gestion régulière d'une boite de courrier électronique est aussi une façon d'appréhender une collection qui par nature est de taille croissante. La réponse même d'une requête à un moteur de recherche est aussi une collection assez vaste (en témoigne les estimations de nombres de réponses fournies) dont l'utilisateur ne regarde dans presque tous les cas que les premiers documents. Enfin, les flux d'informations (dépêches d'actualités, flux RSS, etc.) sont aussi des collections à part entière avec, en plus, la spécificité chronologique. Le projet Canopée cherche à produire une assistance logicielle dans le rapport entre un utilisateur particulier (voir un groupe d'utilisateurs) et une collection, rapport sous-tendu par les capacités interprétatives de l'utilisateur.

L'environnement documentaire des utilisateurs (qui plus est quand il est numérique) est un milieu complexe où les textes en tout genre (documents numériques, mail, flux RSS, *tweets* …) foisonnent, apparaissent, passent, disparaissent … L'utilisateur est comme désorienté et souhaite avoir une vision globale de sa tâche et de son activité interprétative en visualisant les grandes tendances de son environnement textuel numérique. Canopée répond à ce besoin. Il reprend des idées, notamment une approche centrée-utilisateur, que nous avons préalablement expérimentées dans ProxiDocs et ThèmeEditor. A la manière dont un système de réalité augmentée (des systèmes les plus évolués jusqu'au tableau de bord d'un véhicule) fonctionne, assister un utilisateur ce n'est pas chercher à décider des choses à sa place mais au contraire c'est lui donner « toutes les cartes en main » pour l'aider dans sa propre décision. Canopée vise un fonctionnement en synergie avec son utilisateur sur le mode de la réalité augmentée. Le système cherche à éclairer les compétences interprétatives de l'utilisateur en lui suggérant des rapprochements de textes, en lui indiquant en quoi des documents s'opposent, en lui montrant comment un ensemble documentaire donne à voir ses propres centres d'intérêt. D'une certaine façon, on peut dire que Canopée se veut un système de ***sémantique augmentée personnalisée***.

Canopée est un produit qui est encore en phase de développement. Un prototype avec une première version de l'interface est réalisé et sert de démonstrateur dans des rendez-vous avec des clients potentiels. Le logiciel est un développement complet qui bien que reprenant des fonctionnalités communes avec ProxiDocs et ThèmeEditor ne peut en partager aucune ligne de code. Pour des questions d'optimisation de traitement, de passage à l'échelle et de portabilité de librairies graphiques, le développement sous Java (à l'identique de ProxiDocs et ThèmeEditor) a du être abandonné au profit de C++.

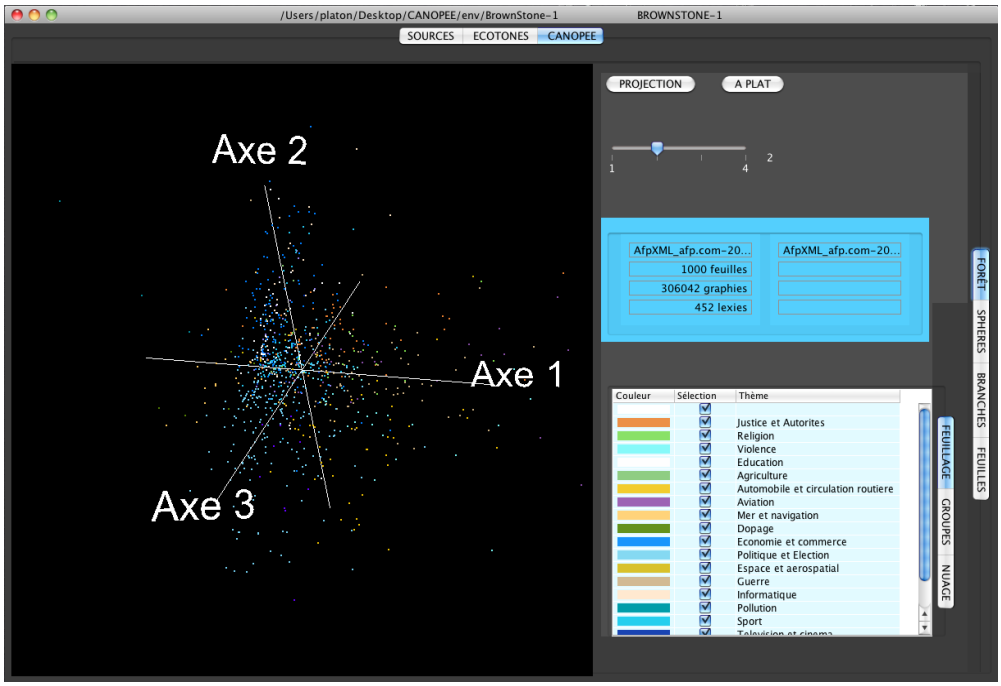


Illustration 26: Visualisation d'un flux documentaire avec Canopée

Les fonctionnalités de Canopée qui sont déjà opérationnelles concernent la cartographie de corpus (cf. illustration 26), le coloriage thématique (cf. illustration 27), l'extraction de tendances thématiques (cf. illustrations 28 et 29) et la saisie des ressources termino-ontologiques personnalisées. Le format des ressources termino-ontologiques de Canopée n'est pas de même nature que les ressources ThèmeEditor ou les ressources LUCIA. Les représentations de ThèmeEditor sous formes de listes de termes avaient l'intérêt d'être conceptuellement très simples d'où une bonne possibilité d'appropriation par un utilisateur lambda. Par contre, la représentation sémique du contenu leur fait défaut. A l'opposé le modèle différentiel du sème de LUCIA est intéressant pour en déduire une richesse d'analyse syntagmatique, mais l'élaboration de dispositifs LUCIA reste une tâche complexe qui induit peu de couplage avec un utilisateur non rompu à la description lexicale. Dans le cadre d'un produit industriel il fallait trouver une voie moyenne. La représentation lexicale de Canopée est une description thématique. Dans chaque thème on y décrit les lexies du thème et on étiquette librement les lexies par des sèmes représentés par une simple description textuelle. Le tout se fait dans une partie de l'application qui gère l'encodage XML des ressources, le partage des sèmes utilisés, les flexions des lexies (module appelé Kotoba, cf. illustration 30).

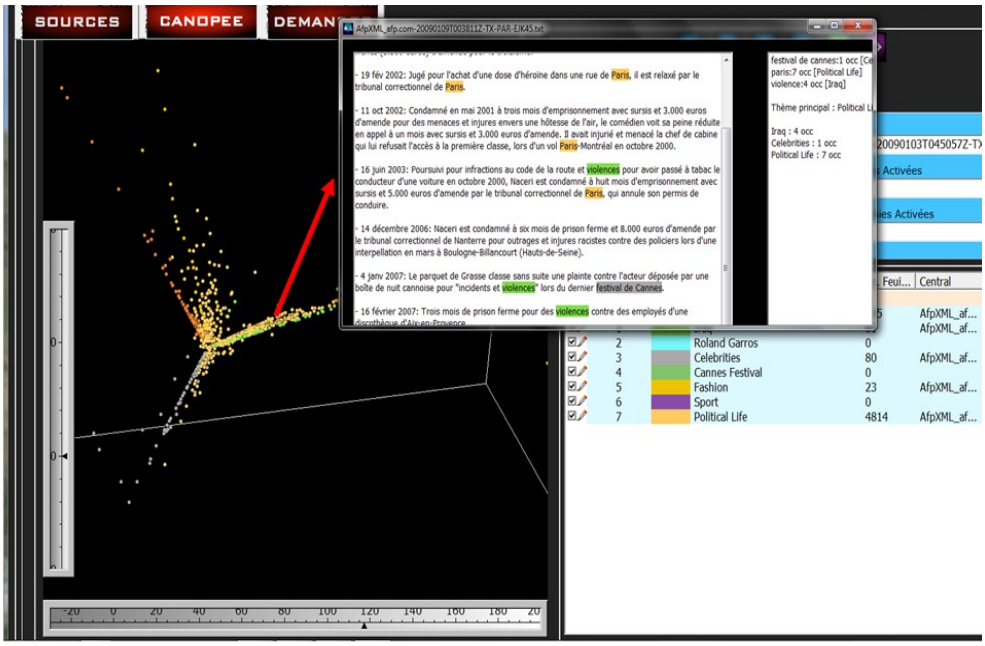


Illustration 27: Coloriage thématique

Un clic sur un point de la cartographie permet d'afficher le coloriage thématique du texte.

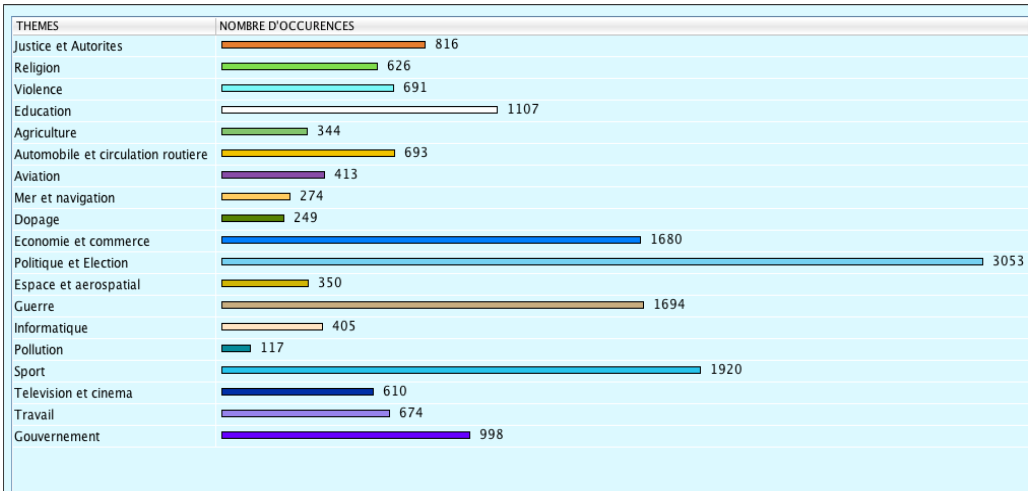


Illustration 29: Importances relatives des thèmes dans un flux documentaire

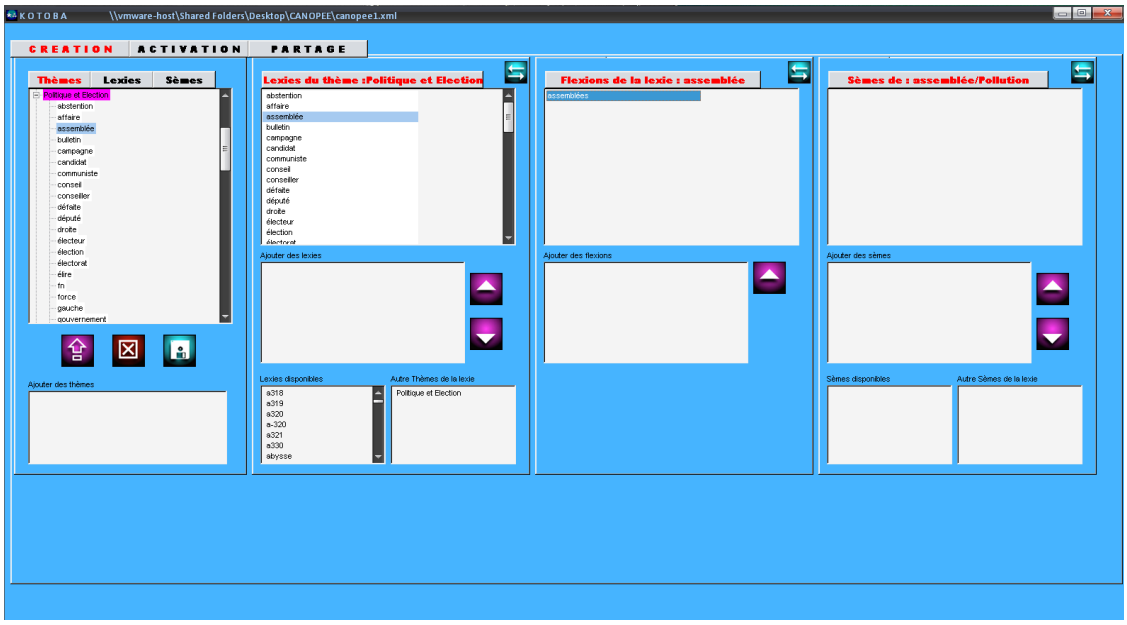


Illustration 30: Interface de description du contenu lexical (Kotoba)

En tant qu'outil pour la veille documentaire, Canopée se doit d'intégrer une fonctionnalité de recherche de documents dans les collections parce que c'est une attente forte des utilisateurs/clients et qu'en tant que tel c'est un vecteur de couplage. La majeure partie des moteurs de recherche dans des collections sont construits sur le même modèle opérationnel de la base de données d'index de documents. Dans l'approche centrée-utilisateur et cartographique qui est la nôtre nous avons cherché à proposer une fonctionnalité de recherche différente basée sur la topologie de l'espace documentaire personnel plus que sur des entrées d'indexation indépendantes. Du point de vue de l'utilisateur, l'originalité réside dans la nature de la requête. Dans les moteurs de recherche classiques à base d'indexation *full-text*, une requête s'exprime comme une combinatoire (plus ou moins structurée) de mots clés. Nous avons cherché ici à proposer une fonctionnalité de recherche où la requête et les documents recherchés sont de même nature, c'est-à-dire textuels. L'utilisateur peut écrire en quelques lignes de texte la thématique et les spécificités du document qu'il cherche (il peut également en contournant l'usage soumettre dans une requête un copier/coller d'un texte pour chercher ceux de son corpus qui s'en rapprochent le plus). La formulation de la requête est considérée comme un texte à part entière au même titre que les documents du corpus. On peut donc associer à la requête un vecteur dans l'espace vectoriel et rendre comme réponses à la requête les vecteurs de l'espace qui sont les plus proches en terme de distance du vecteur de la requête. Cela s'apparente à une sorte de recherche documentaire par l'exemple. C'est un point de vue considéré en recherche d'information que l'on trouve dans le modèle vectoriel de Salton (Salton et McGill, 1983) dont, d'une certaine façon, Canopée ou encore la sémantique latente (Landauer et al., 1998) sont des déclinaisons.

Dans l'interface de Canopée la réponse à une recherche est un point mis en évidence dans l'espace 3D. Ce point est mis en exergue sous la forme d'un petit cube rouge pour le différencier des autres points qui représentent les documents. L'ensemble des documents formant les réponses trouvées dans le corpus se visualise comme une sphère (dont on peut paramétrer le rayon) autour du cube rouge et incluant les documents recherchés (cf. illustration n°31).

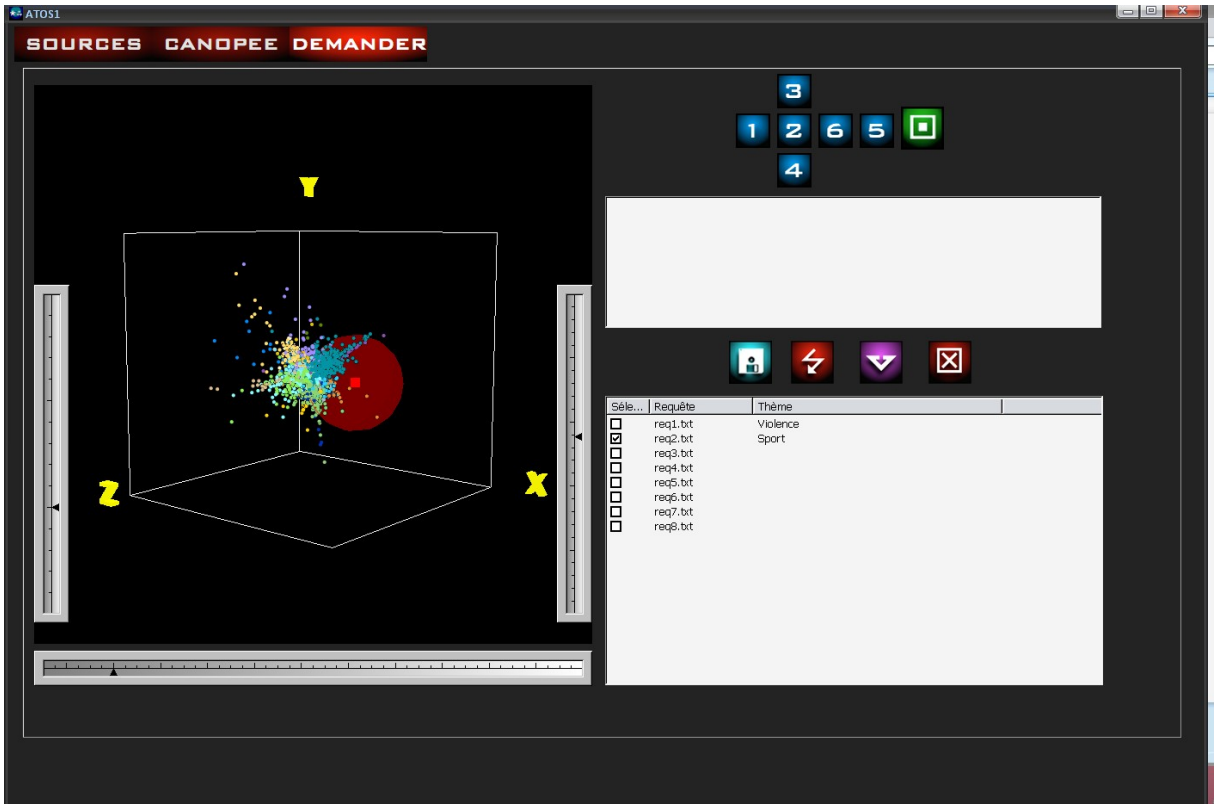


Illustration 31: Interface graphique pour le requêtage en recherche documentaire La requête req2.txt est ici considérée comme un texte dans l'espace vectoriel. Elle est visualisée par le carré rouge au centre de la sphère rouge. Les documents de l'espace inclus dans la sphère sont des réponses potentielles à la requête.

Comme on l'a vu, poser une requête non nécessairement réduite à une liste de mots clé est une différence entre les moteurs de recherches classiques et les approches vectorielles du type Canopée. Une autre différence est également à souligner. Les moteurs de recherche ne permettent de traiter qu'une seule requête à la fois (alors que la majeur partie des recherches s'organisent souvent en plusieurs requêtes exprimant tel ou tel point de vue ou affinant la stratégie de recherche). L'espace vectoriel que visualise Canopée permet de situer de manière topologique plusieurs requêtes simultanément (cf. Illustration 32). Ceci est générateur d'une affordance[25] dans le sens où l'utilisateur peut en déduire la connaissance d'un sous-espace au sein de son corpus délimité par son jeu de requêtes.

25 les affordances sont caractérisées par la perception directe d'objets signifiants, notamment d'une signification immédiate pour l'action, d'une possibilité d'action (Gibson 1977).

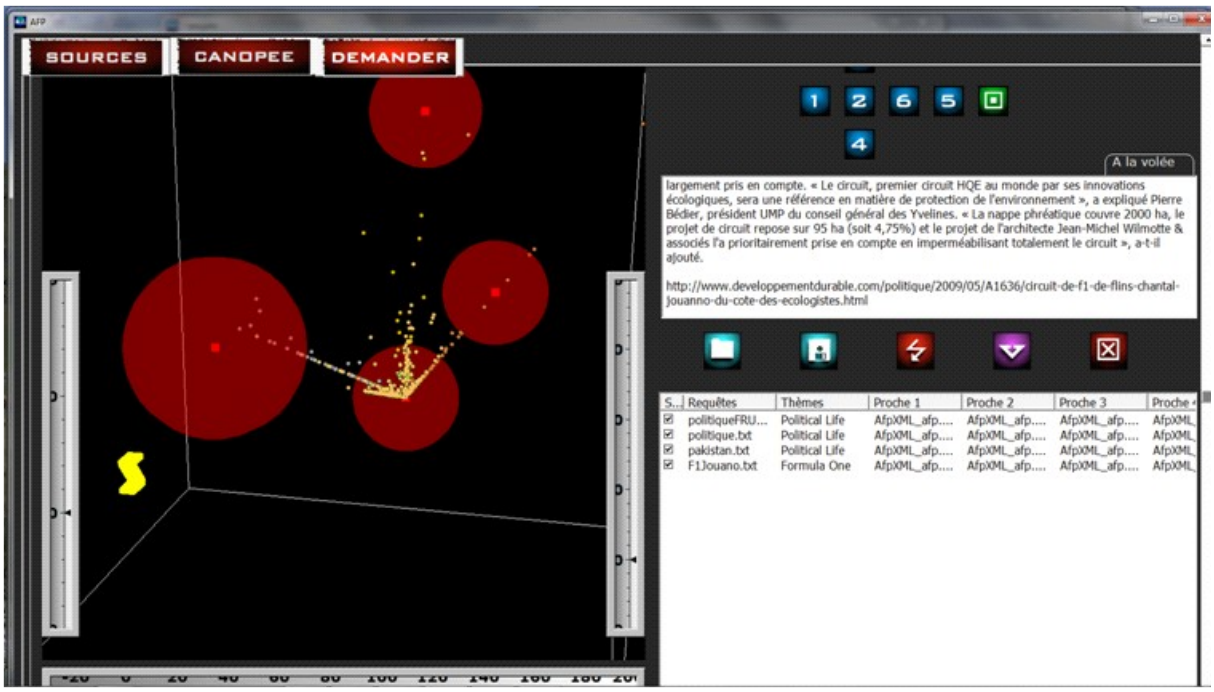


Illustration 32: Requêtage multiple
L'utilisateur a soumis ici 4 requêtes qui sont relativement éloignées les unes des autres (elles n'ont d'ailleurs pas d'intersections en terme de sphères de réponses) et délimite un espace assez étendu couvrant une grande part de la collection documentaire.

En complément du requêtage simple (ou même multiple), on pourrait également penser qu'avoir une connaissance des documents de la collection les plus éloignés de la requête (voir de l'espace formé des différentes requêtes en cas de requêtage multiple) pourrait être aussi une information utile. C'est aussi une forme d'information à propos de la requête, forme à laquelle les utilisateurs ne sont pas du tout habitués étant donné qu'un moteur de recherche classique ne peut pas, de manière similaire, répondre « en creux » à une requête en donnant les documents qui a priori n'auraient rien à voir. C'est une fonctionnalité à expérimenter qui nous semble intéressante et potentiellement génératrice d'autres affordances et de couplage avec l'utilisateur.

D'autres fonctionnalités de Canopée sont encore en cours de développement comme celle de cartographie dynamique de la collection documentaire d'une certaine date à une certaine date en précisant une taille de fenêtre de déplacement. C'est une fonctionnalité qu'on avait déjà expérimentée dans le projet ProxiDocs. D'autres fonctionnalités qui n'ont pas déjà été expérimentées dans ProxiDocs sont également en cours de développement.

C'est le cas de la cartographie des ressources. La boucle interactive de l'outil permet et incite à des aller-retours entre le corpus et les ressources. Cette fonctionnalité supplémentaire vise à accroître encore ces aller-retours d'où peut émerger une meilleure connaissance de la collection documentaire. Quand on fait une cartographie du corpus en fonction des ressources, on projette les ressources sur le corpus. C'est à dire qu'on crée un espace vectoriel de documents où chaque vecteur de dimension N (avec N le nombre de thèmes) indique la façon dont chaque thème est occurrent dans le document (moyennant une normalisation en fonction de la taille du document).

	Th 1	**Th 2**	…	…	…	…	**Th n-1**	**Th n**
Doc x	x_1	x_2					x_{n-1}	x_n

La cartographie produite à partir de cet espace vectoriel indique donc un « mappage » de la ressource sur le corpus. L'idée serait de faire une sorte de cartographie duale en proposant un « mappage » du corpus sur la ressource. C'est-à-dire projeter le corpus sur les ressources en créant un espace vectoriel de lexies où chaque vecteur de dimension P (avec P le nombre de documents du corpus) indique la façon dont chaque document porte un nombre d'occurrence de la lexie en question.

	Doc 1	**Doc 2**	…	…	…	…	**Doc p-1**	**Doc p**
Th y	y_1	y_2					y_{p-1}	y_p

Il suffit d'appliquer exactement les mêmes traitements ACP que dans la carte initiale. On le fait juste sur la matrice transposée de la matrice initiale : $M^{-1}[i;j] = M[j;i]$

Cet espace vectoriel transposé de l'espace initial est une matrice d'indexation au sens de (Salton & Yang 1975). Cette cartographie issue de la matrice d'indexation présenterait l'intérêt d'amener l'utilisateur à une réflexion supplémentaire sur ses ressources. Certains thèmes ou certaines lexies sont peut-être peu occurrentes là ou d'autres le sont beaucoup plus et peut-être de manière corrélée avec d'autres thèmes ou lexies. La cartographie des documents visualise des rapprochements entre textes. Ici on vise plutôt des rapprochements entre lexies et entre thèmes. Ce serait complémentaire de l'histogramme d'importance des thèmes (cf. illustration 29). Une cartographie de la ressource permettrait de mettre les thèmes de l'utilisateur dans un espace topologique d'où on peut tirer des informations utiles sur l'adéquation globale corpus-ressource.

- Quelle est le thème le plus central dans le corpus ?
- Quels sont les thèmes corrélés par le corpus (proches dans la cartographie) ? ce qu'on pourrait interpréter par « de quoi parle-t-on en général de manière conjointe ou dans une répartition statistique peu différente ? »
- Quels sont les thèmes qui formellement sur le corpus semblent s'exclure ?
- Y-a-t-il une singularité dans la répartition topologique des thèmes ? si oui ou non, qu'est-ce qui pourrait l'expliquer ? nombre de termes dans les thèmes ? thèmes liés à des registres de langues (discours familier, scientifique …etc) ?

Canopée est clairement un produit ouvert en terme de fonctionnalités additionnelles. La version en cours de finalisation est la n°1. Elle devrait s'enrichir de ces fonctionnalités dans les versions futures.

3.3.2.1. Expérimentations de Canopée sur corpus

Nous allons dans cette sous-partie rendre compte de certaines analyses que nous avons menées avec Canopée sur différents types de corpus avec différentes ressources lexicales. Ces analyses proviennent essentiellement de démonstrations que nous avons faite du logiciel auprès de clients potentiels pour eXo maKina, essentiellement des agences de presse. L'objectif ici est de montrer que l'usage d'un environnement interactif de médiatisation des interprétations produit une valeur ajoutée émergente en terme de compréhension globale et que ce qui émerge dans ces analyses n'était pas forcément des résultats prévisibles et attendus.

Il ressort de la quasi-totalité des analyses menées avec Canopée que la cartographie de corpus n'est réellement informative que si elle est un espace d'interactions et non simplement un résultat sous forme d'image fixe. C'est particulièrement vrai dans une visualisation 3D comme celles que Canopée produit car la vue sous un certain angle d'un nuage de points ne permet de déduire une réalité topologique que si elle est corroborée par un décalage d'angle. Là où deux points peuvent sembler proches sous une vue particulière ils peuvent être très éloignés, ce dont on ne s'aperçoit que si l'on opère une rotation de vue. Canopée permet à son utilisateur de zoomer et de « tourner » à la souris dans la visualisation 3D assurant ainsi la possibilité de comparer plusieurs vues complémentaires sur le même objet. C'est ainsi qu'on peut mettre en place une « expérience topologique » de l'utilisateur sur sa collection documentaire qui lui permet d'inférer et de vérifier des propriétés de distance, de compacité ou encore d'homogénéité de son nuage de points et de se forger ainsi une appréhension globale topologique et thématique de la collection.

Les effets de projection inhérents au mode de calcul de la visualisation de l'espace vectoriel amplifient le besoin d'interaction avec la cartographie. Ces effets sont encore plus sensibles quand l'espace vectoriel projeté dans un espace 3D compte initialement peu de dimensions. Nous l'avons constaté lors d'une analyse pour une entreprise de veille où l'on faisait une démonstration de Canopée sur un petit corpus d'articles du journal *Le Monde* avec 4 thèmes (politique intérieure, politique étrangère, économie, sport).

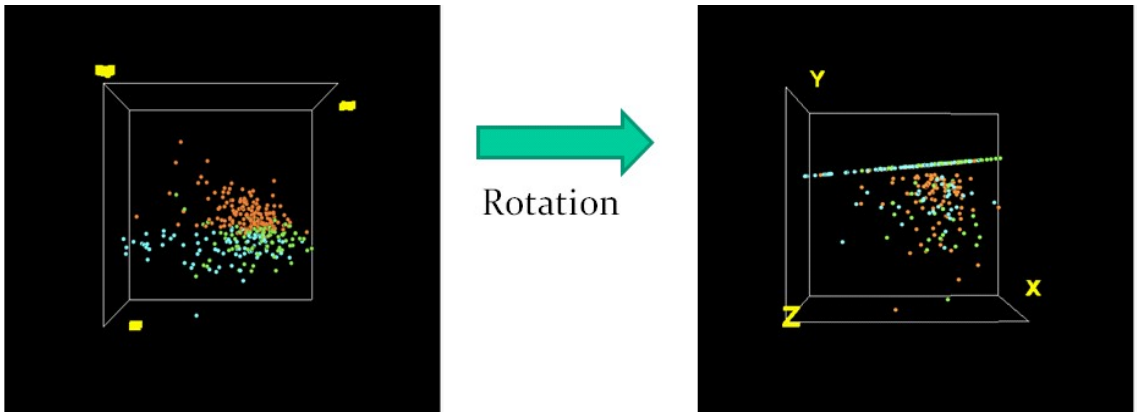


Illustration 33: Effets topologiques singuliers
Les deux cartographies sont le même espace vectoriel cartographié. L'image de droite résulte d'une rotation de 90° vers l'arrière de l'image de gauche.

Nous avons obtenu une carte 3D présentant un plan assez caractéristique (cf. Illustration 33, carte de droite). En examinant les points constitutifs de ce plan relativement aux autres on s'est aperçu qu'il provient d'un effet de projection regroupant des documents où les 4 thèmes utilisés ne sont pas tous simultanément occurrents. Sous un autre angle de vue ce plan n'aurait pas été visible. D'autre part, sur la même carte on a observé que sous une certaine vue l'espace semble marquer une frontière assez nette entre les documents traitant majoritairement de politique étrangère (documents de couleur marron sur la carte de gauche de l'Illustration 33) et les autres. Aucune de ces deux observations n'était a priori prédictible et elles émergent uniquement de l'interaction personne-système.

D'autre part, il est également une autre forme d'interaction qui est particulièrement importante et génératrice de couplage avec l'utilisateur : celle qui consiste à passer alternativement des cartes aux coloriages de textes. Les phénomènes sémantiques locaux dans les textes et plus généralement leurs

interprétations sont fortement liés aux spécificités globales des corpus dont font partie ces textes. C'est le principe d'architextualité décrit par François Rastier (Rastier 2001) que nous détaillerons au cours du prochain chapitre. Par exemple, des usages locaux et co-occurentiels (dans un texte) de plusieurs lexies doivent être mis en relation avec l'entourage intertextuel du document pour ainsi caractériser si ces usages sont une spécificité du document ou un « signal fort » du corpus pour le lequel le document ne se distingue pas. De manière duale, il y a aussi souvent besoin d'expliciter les quantifications globales par des observations locales contextuelles. Par exemple, nous avons travaillé a des fins de démonstration sur un corpus qui nous a été soumis. C'est un corpus de *tweets*, plus précisément les *tweets* émis dans le *hashtag* DSK entre le 15 et le 21 mai 2011, c'est à dire en pleine fameuse affaire Strauss-Kahn new-yorkaise[26]. Le corpus rassemble environ 580 000 *tweets* segmentés en 6 sous-corpus regroupant chacun les *tweets* d'une journée. Un premier « réflexe » lorsque l'on aborde un nouveau corpus et que l'on cherche à mettre au point des ressources Kotoba pour commencer la cartographie est de faire une analyse fréquencielle en courbe de Zipf et d'en extraire des bribes de familles thématiques (c'est tout à fait la démarche que nous avions également mis en œuvre dans l'outil ThèmeEditor). Sur le corpus du 17 mai 2011 on a observé une fréquence d'occurrences surprenante (et singulière par rapport aux autres jours) de la graphie « Bernard ». Avec 3138 occurrences, la graphie arrivait dans les 30 graphies de substantifs les plus occurrentes. Autant, à cette date précise, on comprenait bien les fortes occurrences de « Rikers » et « Island » (prison où a été incarcéré D. Strauss-Kahn le 17/5/11), autant « Bernard » ne ne semblait pas facilement justifiable. Nous avons effectué plusieurs retours aux textes par des extractions de concordances et avons constaté un effet simultané de 3 origines lexicales sur des *tweets* qui font l'objet de nombreux *retweets* : « Bernard Debré », « Bernard-Henri Levy », « Bernard Madoff ». En témoignent les extraits suivants :

```
<ami:content>RT @DSKGate: Mais pourquoi Bernard Debré en veut-il à DSK ? - Rue89: Mais pourquoi Bernard Debré en veut-il (cont) http://tl.gd/ah928h</ami:content>
<twitter:id>70579338319364100</twitter:id>
<date>2011-05-17 22:00:00</date>
<ami:author>Suite_2806</ami:author>

<ami:content>French philosopher Bernard-Henry Levy defends accused IMF director: http://bit.ly/ikbC8Z</ami:content>
<twitter:id>70584888025284610</twitter:id>
<date>2011-05-17 22:22:03</date>
<ami:author>Clauni</ami:author>

<ami:content>Robert de Niro bientôt à l'affiche d'un téléfilm sur l'histoire de Bernard Madoff, pour la chaîne HBO.... et sur celle de DSK ? #bistougate</ami:content>
<twitter:id>70439858031689730</twitter:id>
<date>2011-05-17 12:45:45</date>
<ami:author>tabounet</ami:author>
```

Les analyses que nous avons menées sur le « corpus Twitter DSK » comparativement à d'autres analyses menées sur d'autres corpus nous ont également permis appréhender un autre intérêt issu de l'interaction avec la cartographie. Il s'agit de pouvoir rendre compte et d'objectiver certaines propriétés topologiques de certains genres textuels, notamment en terme de compacité de l'espace cartographié. Nous avons vu précédemment que le nombre de thèmes utilisés pour dresser la cartographie induit certaines propriétés spatiales. La quantité et la qualité de l'information dans les documents constituant

26 Agression sexuelle sur une femme de chambre de l'hotel Sofitel à New-York le 14 mai 2011.

le corpus est également un facteur déterminant de la topologie de l'espace. Comparons par exemple le corpus de *tweets* DSK de la journée du 18 mai 2011 et un corpus de dépêche de l'AFP. En terme de qualité de l'information les deux corpus sont très différents. Les dépêches, même quand elles sont courtes comme dans l'exemple ci-dessous, montrent un travail rédactionnel de la part de l'auteur avec des constructions syntaxiques correctes et un message clair :

AfpXML_afp.com-20090101T052623Z-TX-PAR-DFV66.txt
1er Janvier 2009
Thaïlande-incendie
BANGKOK
Au moins 58 personnes sont mortes, et 200, dont des étrangers, ont été blessées, dans la nuit de mercredi à jeudi à Bangkok dans l'incendie d'une boîte de nuit où les clients fêtaient le Nouvel An, ont annoncé les services de secours.
(PAPIER GENERAL) 5/600 mots =(PHOTO)= 05h00 GMT
Deskinter/jr

Il est loin d'en être de même dans le corpus de *tweets* où comme le montrent les exemples suivants la qualité et la clarté de l'information sont très variables :

<ami:content>j'ai parlé avec un pote receptionniste ds un hotel 3* il m'a dit que la femme de ménage n'avait rien à foutre ds la chambre ! #dsk</ami:content>
<twitter:id>70920185418219520</twitter:id>
<date>2011-05-18 20:34:24</date>
<ami:author>Lunack</ami:author>

<ami:content>#DSK ??????? http://uk.twend.it/DSK</ami:content>
<twitter:id>70869575943729150</twitter:id>
<date>2011-05-18 17:13:18</date>
<ami:author>kadido02</ami:author>

D'un point de vue quantitatif, les deux corpus diffèrent beaucoup et les analyses menées avec Canopée aussi, notamment en terme de couverture de ressources lexicales :

- Corpus AFP : 5288 dépêches (1 691 204 tokens ; environ 320 lexies en moyenne par dépêche), 17 thèmes généralistes couvant 552 lexies (98 156 occurrences dans le corpus).

- Corpus DSK : journée du 18 mai 2011 : 75 521 *tweets* multilingues (3 389 602 tokens ; environ 44 lexies en moyenne par tweet), 4 thèmes très spécifiques à l'affaire DSK (agression, politique, justice, hôtel) couvrant 43 lexies (222 353 occurrences dans le corpus)

Ces différences très nettes se retrouvent de manière très sensible sur les cartes produites en montrant des espaces topologiquement très différents en terme de compacité notamment (cf. Illustration 34). Afin de quantifier cette différence de compacité un calcul d'entropie du nuage de point pourrait être mené mais c'est une fonctionnalité qui n'est pas encore mise en œuvre dans Canopée.

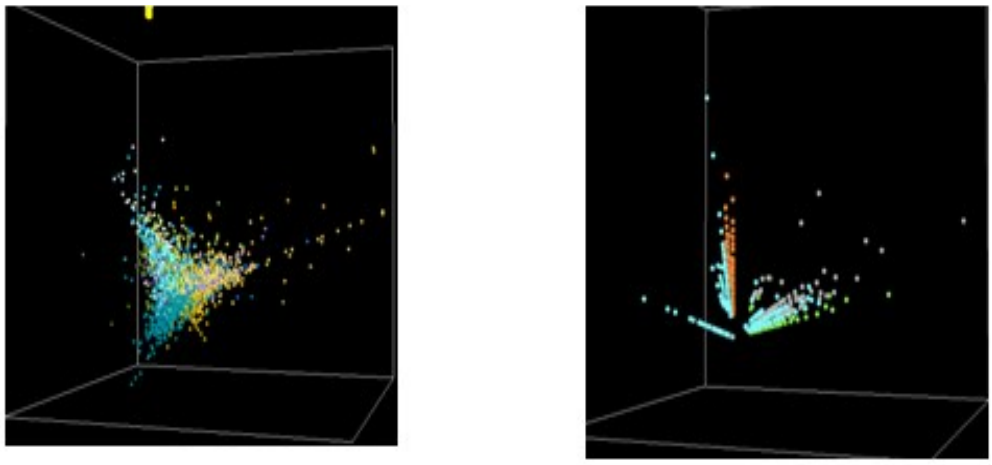


Illustration 34: Deux cartes Canopée différentes
A gauche une carte du corpus AFP, à droite une carte du corpus de *tweets* DSK.

Ce type d'observations nous invite à approfondir dans les usages de Canopée une tentative de caractérisation topologique des genres textuels corrélés aux ressources lexicales que l'on projette dessus. La projection dans l'espace vectoriel des documents révèle des similitudes et des différences. Les similitudes (et donc la proximité) entre deux documents s'apparentent à des profils d'occurences en corrélation dans les thèmes retenus pour établir la cartographie. Si deux documents différents ont les mêmes nombres d'occurrences pour les mêmes thèmes et des longueurs proches, alors les deux points qui les représenteront dans l'espace seront peu éloignés. C'est ce que l'on peut observer dans la cartographie de l'Illustration 35 où l'ovale rouge que nous avons ajouté sur la carte montre les *tweets* ouverts sur le côté qui évoquent tous la question de la « femme de ménage », lexie qui n'était pas définie en tant que telle (à la différences de « femme de chambre » mais dont on s'aperçoit qu'il faudrait la rajouter à la ressource) et qui produit des profils d'occurrences similaires (1 occurrence pour « femme », 1 occurrence pour « ménage »).

Dans les corpus de *tweets* que nous avons analysés on peut constater que les *tweets* très similaires en terme de profils thématiques sont très nombreux. La taille nécessairement réduite du contenu textuel amplifie ce phénomène. On remarque aussi que les pratiques des abonnés de Twitter abondent dans ce sens et que le genre textuel est bien une résultante des pratiques. Lorsqu'une information apparaît dans une actualité suivie en temps réel dans un *hashtag* Twitter, cette actualité se retrouve simultanément diffusée le plus rapidement possible par les abonnés du réseau social et le peu de latitude en terme de rédaction fait que cette actualité est souvent diffusée de manière multiple au mot près. Un usage très fréquent des abonnés de Twitter dans cette diffusion d'une actualité est la fonctionnalité de *retweet*. Si un abonné A reçoit de B dont il est « *follower* » un *tweet*, il peut « *retweeter* » le *tweet* pour à son tour communiquer l'information à l'identique à ses propres « *followers* ». Il en résulte au sein du *hashtag* une part importante de *tweets* et de *retweets* identiques qui ne se différencient quasiment que par le signe « RT » (*retweet*) en début de message. Nous nous sommes aperçu que ces phénomènes de *tweets* ayant le même contenu et de *retweets* donnent lieu à des formes topologiques singulières dans les cartes Canopée se matérialisant (par effet de normalisation) comme des alignements de points (cf. Illustrations 35 et 36)

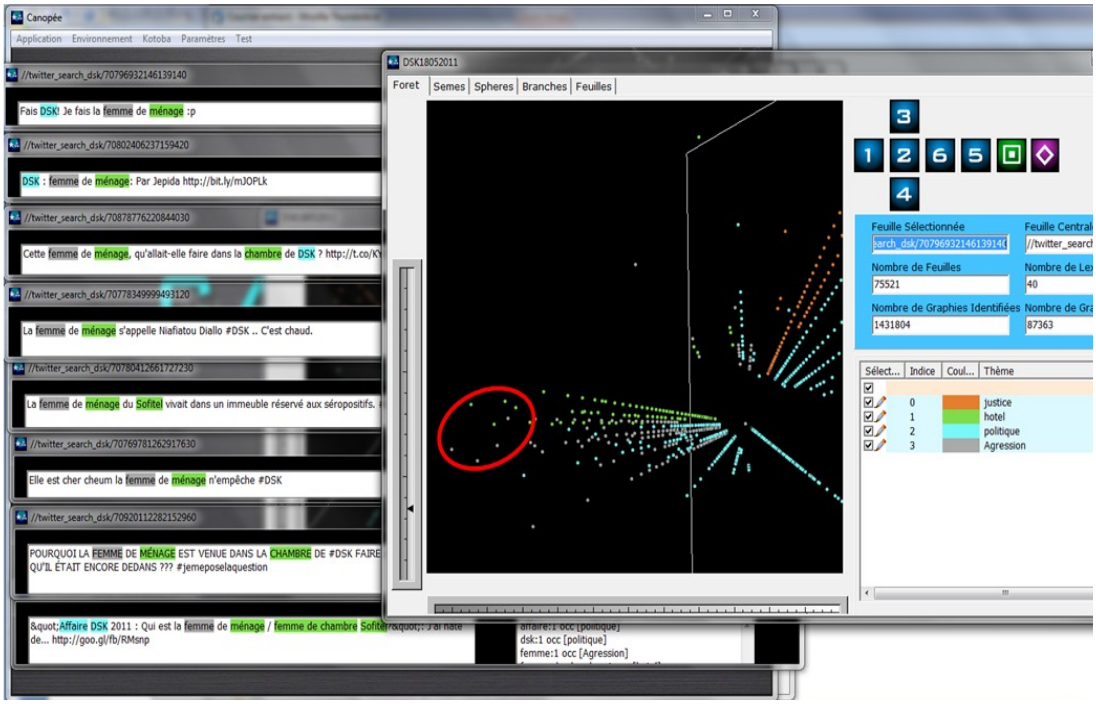


Illustration 35: *Tweets* similaires
Le *tweets* ouverts sur la gauche sont thèmatiquement proches (il évoque un sujet commun, ce lui de la « femme de ménage »). On les retrouve dans le cercle rouge.

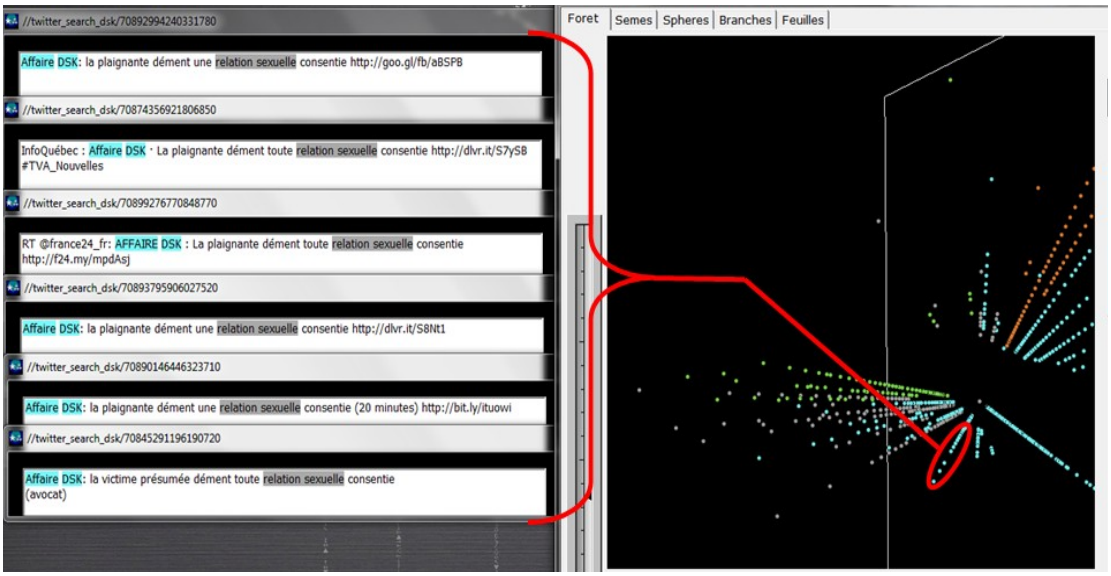


Illustration 36: *Tweets* recopiés et *retweets*

Il ressort de cette observation sur les alignements de *tweets* et de *retweets* que certaines traces topologiques pourraient devenir des types de signatures de pratiques et donc de genres textuels. On caractériserait ainsi les corpus de *tweets* comme des espaces peu denses et présentant significativement des alignements nombreux. A la différence, un corpus de dépêches ou d'articles de presse se

présenterait comme un espace en nuage de points dense. Bien évidemment, nous n'en sommes là qu'au stade de l'intuition de recherche et cela mériterait d'être étudié comparativement en faisant varier les genres textuels et les ressources utilisées pour les cartes.

En complément de la perspective purement industrielle, nous continuons donc aussi à travailler au projet Canopée dans l'optique d'un « laboratoire » d'analyse des pratiques en veille de presse et d'étude des attentes des veilleurs.

~

A travers les différents projets de développement et les différentes expérimentations que nous avons menés (et que nous poursuivons), il apparaît que la problématique du document numérique nécessite d'être abordée de manière indissociable à l'activité du ou des humains qui les produisent, recueillent, indexent, classent et recherchent. Les pratiques conditionnent les genres qui fondent les textes. Il en va logiquement de même des instrumentations logicielles qui permettent l'accès aux informations numérisées. Elles aussi doivent être abordées dans un rapport nécessaire à l'utilisateur. Les conditions de l'appropriation des outils logiciels par leurs utilisateurs sont cruciales. Chaque action dans le logiciel s'accompagne d'un engagement (cognitif et physique) de l'utilisateur, et induit une évolution dynamique des connaissances et informations en émergence, d'où une valeur ajoutée en terme de résultat. C'est ce qu'on identifie comme un couplage personne-système. Dans l'installation de ce couplage, on observe très régulièrement une tendance qu'ont naturellement les utilisateurs qui consiste à détourner ou contourner l'usage prévu des logiciels et ainsi en faire émerger certains qui n'étaient pas initialement prévus. Ceci est largement positif et il convient de chercher plus à l'encourager qu'à l'empêcher.

La question du couplage personne-système est la question centrale des approches centrées-utilisateurs et c'est un aspect qu'il faut pouvoir concilier avec les exigences techniques pour qu'au final le système soit bien centré-utilisateur et pas techno-centré. Le TAL (dans le prolongement de l'intelligence artificielle) est un domaine qui se prête souvent à des approches très techniques et calculatoires ce qui ne rend pas souvent simple d'y affirmer une approche centrée-utilisateur. C'est ce que nous allons développer dans le chapitre suivant.

4. Approche centrée-utilisateur et TAL

Le « A » de TAL signifie *Automatique*. Implicitement cela pose un objectif qui n'est pas anodin : rendre automatiques (ou du moins automatisables) des compétences linguistiques humaines. La question serait donc de chercher à rendre effectives ces compétences en s'abstrayant complètement de l'humain. On sent bien dès lors qu'une démarche centrée-utilisateur en TAL apparaît comme un positionnement qui ne coule pas de source, pour ne pas dire un antagonisme. Nous verrons dans ce chapitre que, de fait, l'approche centrée-utilisateur que nous défendons ne trouve pas naturellement un espace bien balisé au sein du TAL. Nous expliquerons que cela provient essentiellement des courants historiques du TAL qui ne laissent que très peu de place à la question de l'interaction (elle même, introduisant le rapport à l'utilisateur). Pourtant certaines évolutions de l'informatique ont récemment montré que les approches centrées-utilisateurs en changeant complètement de point de vue demeurent pour autant plausibles et même efficaces. C'est ce que nous montrerons en détaillant le cas de ce que l'on appelle le « Web 2.0 ».

Le TAL est majoritairement dans une certaine tradition épistémologique de l'ontologie et de la compositionnalité, comme en témoigne (nous le verrons) le champ actuel du Web Sémantique. C'est un constat. Nous adopterons une démarche alternative à ce constat, celle de la Sémantique Interprétative (SI) de François Rastier. Pour autant, adopter la SI dans une démarche de TAL est tout à fait envisageable. Certains autres travaux que les nôtres l'ont déjà fait[27] et nous expliquerons que c'est bien là que nous nous positionnons également. La SI est une théorie linguistique qui fournit principalement avec les notions de sèmes et d'opérations interprétatives un bagage théorique très fin pour l'analyse des phénomènes sémantiques. En tant que telle, elle permet d'éclairer une modélisation opératoire en TAL. Nous verrons que son ancrage épistémologique est un apport au TAL tout aussi important. Nous expliquerons que le sens n'est pas à chercher dans une matérialité des textes mais

27 (Tanguy 1997), (Thlivitis 1998), (Bommier-Pincemin 1999), (Rossignol 2005) par exemple.

dans une médiatisation de l'espace interprétatif de l'utilisateur. Ainsi, loin de vouloir « tourner le dos » au TAL, nous y reviendrons en ramenant de la SI la problématique de l'utilisateur qui fait crucialement défaut au courant dominant du TAL.

4.1. Courants du TAL

Comme toutes les autres disciplines, l'informatique et la linguistique n'échappent pas à des effets d'école autour de certains travaux comme à certaines ruptures au cours de leurs évolutions. Par exemple, F. de Saussure a marqué une véritable discontinuité en linguistique en montrant qu'on peut étudier scientifiquement la langue comme un système de valeurs, à tel point qu'on dit aujourd'hui de lui qu'il est le père de la linguistique dite moderne. Dans le champ du TAL, un exemple de rupture brutale est le rapport ALPAC[28] en 1964 dont les conclusions trop négatives sur la traduction automatique ont conduit le gouvernement américain à ne plus financer d'études à ce sujet pendant plus de 10 ans. En tant qu'activité humaine et sociale, la science n'avance finalement que par le fait d'allers et retours, de discontinuités et de ruptures. Certaines sont immédiatement sensibles (c'est le cas du rapport ALPAC) et d'autres demandent plus de temps car les positionnements épistémologiques sont difficiles à abandonner. Nous verrons ainsi que le paradigme dominant du TAL que constituent les approches calculatoires et représentationnelles (notamment en ce qui concerne le sens) est bien toujours actif depuis les années 1950. C'est ce que nous appellerons les approches théoriques. Un contrepied notable à ces approches est apparu dans les 20 dernières années avec l'émergence des corpus et des données réelles accompagnés de l'essor des méthodes statistiques. C'est ce que nous appelons les approches empiriques. Si incontestablement, l'observation de données attestées est scientifiquement une approche plus défendable que de créer ses propres données, nous expliquerons que l'utilisateur, quant à lui, n'est toujours pas vraiment envisagé et problématisé.

4.1.1. Les approches théoriques

Les premiers pas du TAL sont ceux de la traduction automatique dans le contexte particulier de la guerre froide. En 1954, IBM fait une première expérience de traduction réalisée sur un ordinateur (on cherchait à traduire du russe en anglais !). Les résultats sont peu probants mais tout le monde s'accorde alors à penser que l'évolution fulgurante réelle des ordinateurs[29] permettrait de rendre rapidement de bien meilleurs résultats. C'est avec le rapport ALPAC, 10 ans plus tard, qu'on commençait à revenir sur cette idée. Pourtant, dans les autres champs d'applications du TAL (par exemple dans le dialogue homme-machine[30]) les mêmes types d'approches ont perduré. Dans tous les cas, ce qui est commun et qui n'est pas remis en cause est de rechercher en amont une formalisation (c'est-à-dire une théorie) de notre compréhension du fonctionnement de la langue permettant la reproduction, par un système artificiel, de la compétence linguistique de l'humain. C'est en cela qu'on peut qualifier ces approches de théoriques. Leur début est historiquement daté avec notamment les travaux sur la grammaire générative de Noam Chomsky (Chomsky 1971) ou encore la modularité de l'esprit de Jerry Fodor

28 Automatic Language Processing Advisory Committee

29 cf. la loi de Moore (Schuhl 2004) indiquant que la capacité des processeurs double tous les 18 mois.

30 En 1968, Stanley Kubrick met en scène dans le film « 2001, l'odyssée de l'espace » un ordinateur appelé HAL qui parle en langue naturelle. Pour se renseigner sur les avancées de l'intelligence artificielle en matière de langage, il contacte le MIT où Marvin Minsky, chercheur en IA, lui dit qu'à l'heure actuelle une interaction naturelle avec un ordinateur n'est pas réalisable mais que compte tenu des progrès fulgurant de l'informatique, cela devrait être le cas à un horizon d'une trentaine d'année … c'est-à-dire dans les années 2000 !

(Fodor 1975). Au même moment, McCulloch et Pitts (dans un article[31] fondateur de ce qui allait être la cybernétique) défendaient qu'il était possible de considérer le cerveau comme une réalisation matérielle d'une machine de Turing (Bachimont 2000).

Se donner une théorie sur un objet d'étude, c'est chercher une objectivité dans la façon de considérer cet objet. C'est faire l'hypothèse d'un consensus. La grande variabilité des langues ne plaide pas forcément pour l'existence de ce consensus. Ceci explique peut-etre qu'il est plus facile de chercher ce consensus dans la faculté de langage (c'est par exemple l'approche de Chomsky) que dans les langues. Lorsque l'objet d'étude concerne la sémantique des langues, force est de constater qu'on est très loin d'un consensus et d'une objectivité. Gérard Sabah, dans un article publié par l'ATALA (Sabah 1996), précise ce que peut être le sens du point du résultat visé par telle ou telle sémantique :

- Préciser des conditions de vérité (sémantique véric onditionnelle)
- Décrire une expression comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants (sémantique intensionnelle)
- Décrire une expression comme l'ensemble des objets ou des situations du monde de référence que cette expression peut désigner (sémantique dénotationnelle ou référentielle)
- Chercher à décomposer le contenu des mots en éléments de sens plus primitifs pour étudier les possibilités de combinaison de ces éléments (sémantique componentielle)
- Décrire une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné (sémantique procédurale)
- Mettre en évidence les marqueurs et les constructions utilisées pour qu'un énoncé puisse servir comme un argument en faveur d'un autre énoncé (sémantique argumentative)

Aucune de ces sémantiques n'a, à elle seule, complètement tort ou complètement raison. Par exemple, il existe bon nombre d'énoncés en langue naturelle dont le sens a un certain rapport avec la vérité ou encore la référence, mais cela ne veut pas dire que tout énoncé est interprétable en terme de vérité ou de référence.

Lorqu'elle s'applique au champ de la sémantique, l'approche théorique implique deux conséquences qui traduisent des hypothèses épistémologiques très fortes : la recherche du sens est un calcul et le sens est assimilable à une représentation[32] (qui justement est issue du calcul). Ce sont deux hypothèses que nous rejetons dans le cadre d'une approche centrée-utilisateur.

La vision calculatoire du sens est certainement à rechercher dans les bases du cognitivisme qui voit la pensée et la communication comme des processus de traitement de l'information. Les travaux de J. Fodor en sont tout à fait prototypiques.

Cette position a certainement été renforcée par le développement de l'informatique et notamment les théories de la compilation dans les langages formels. La tentative de rapprochement entre langues et langages formels s'installe avec les travaux de Montague et son article « *English as a formal language* » (Montague 1970) qui plaide pour une approche du sens des langues par la logique formelle. Pourtant ce rapprochement n'a rien d'éclairant car il mélange des registres très différents,

31 (McCulloch & Pitts 1943)

32 (Venant 2006)

notamment le fait que les langages formels sont des artefacts à la différence des langues (qu'on appelle justement langues naturelles[33]). Comme le montre (Nicolle 2005) ce sont les langues qui créent les langages, pas le contraire. La formalisation logique enferme le sens et la signification dans un rapport à la preuve et aux valeurs de vérité. Le modèle de la compréhension est celui du calcul de la vérité et c'est pour pour nous une simplification très réductrice de l'objet, qui le dénature. De plus, la valeur ajoutée d'un formalisme en terme d'explicitation d'un objet est tout à fait constestable. Formaliser, ce n'est pas, par principe, expliquer. Le calcul, parce qu'il est automatique et désincarné, se veut une explication. C'est sans doute une explication de l'axiomatique formelle mais il n'est pas du tout évident que ce soit une explication de l'objet qui a été formalisé. Les questions de sens et de significations trouvent des explications bien plus éclairantes en linguistique là où le rapport au calcul n'est pas forcément imposé.

Le calcul logique et l'axiomatique formelle imposent une démarche compositionnelle. Le calcul s'organise en différents sous-calcul fournissants des résultats qui s'agrègent dans le résultat global qui est le sens, à la manière dont un mur est construit à partir de briques. Là encore, c'est un modèle opératoire de la compilation mais pas de l'interprétation en langue comme le montrent les incidences contextuelles sur tous les niveaux de description linguistique. Le calcul compositionnel appliqué au TAL à donné lieu à une tradition de modèles de traitement que François Rastier appelle la tradition logico-grammaticale. Le traitement se décline en couches allant du « bas » niveau phonétique, phonologique ou lexical en fonction de l'entrée jusqu'au « haut » niveau sémantique ou pragmatique, chaque niveau fournissant la donnée du traitement de niveau supérieur. Cette façon de voir confère à tous le niveaux, y compris au dernier et donc au sens, un préjugé d'ordre représentationnel. A tous les niveaux, entrées et sorties sont des représentations, des concepts plus ou moins agrégés. Ces modèles, où l'interprétant n'est finalement qu'en bout de chaîne, ne sont pas sans poser des problèmes d'explosion combinatoire car, pris isolément, chaque niveau génère un bon nombre d'ambiguïtés (et donc de faux résultats) que les niveaux suivants ne peuvent pas toujours dissocier des analyses correctes. Le contexte, et donc les conditions d'interprétation, y est aussi relégué en bout de chaine, dans la phase d'analyse pragmatique. L'importance est d'abord donnée aux analyses co-textuelles (morphologie, syntaxe, sémantique) et la contextualisation consiste de façon réductrice à une ultime levée d'ambiguïtés.

L'approche logique est très liée à une approche ontologique. C'est un héritage en TAL de l'intelligence artificielle comme en témoigne le *Knowledge Level* de A. Newel (Newel 1982) où l'accent est mis sur les connaissances représentées comme des structures symboliques rationnelles interprétables de façon propositionnelle. Quand il est question de mettre au point un système expert, il faut recueillir la connaissance d'un domaine et la formaliser. Les notions fondamentales de concept et de relation que l'on trouve dans la majeure partie des ontologies sont très pertinentes. Quand on fait du TAL, le domaine c'est la langue. Modéliser les signes linguistiques par des concepts et les règles de fonctionnement de la langue par des relations entre ces concepts est une confusion (au même titre que la confusion interprétation = compilation). Termes et concepts ne sont pas de même nature, de même que la langue n'est pas une base de connaissances, mais une activité sociale et culturelle. La « vie » d'une ontologie est un parcours cumulatif où l'objectif est de connaître toujours plus de concepts pour en déduire plus de relations entre concepts. La diachronie d'une langue n'est pas du tout un parcours cumulatif. Les signes linguistiques ont une vie liée à leur usage. On en trouve notamment un bon exemple avec la description en terme de sèmes de la lexie *Outreau* relatée dans (Reutenaeur 2010). Les signes apparaissent quand une communauté parlante les instituent implicitement. Ils varient dans

33 En anglais (à la différence du français), il n'y a qu'un mot pour langue et langage (*language*). C'est aussi une possible explication du rapprochement des deux notions). Du coup, on parle de *langues naturelles* (ce qui en français est un pléonasme) en rapport à l'anglais qui lexicalise l'expression *natural language*.

leur signification par les rapprochements syntagmatiques qui les impliquent. Ils disparaissent quand les communautés parlantes les oublient[34]. Dans une ontologie un concept nouveau ne disparaît jamais car cela mettrait en jeu la cohérence des autres concepts avec lesquels il entretient des relations dans l'ontologie.

Si l'assimilation de la signification et du sens à la connaissance nous semble être une erreur, il n'en demeure pas moins que c'est un point de vue explicitement assumé dans bon nombre de travaux qui relèvent de la tradition logico-grammaticale. C'est le cas du système NELL[35] de Tom Mitchell à l'Université Canegie Mellon de Pittsburg aux Etats-Unis. NELL est l'acronyme de *Never Ending Language Learning*. Le principe est d'analyser un maximum de textes sur le Web et de manière continue pour en extraire une base de connaissance constituée de faits logiques tels que *San Francisco est une ville*. La gestion de la cohérence logique de cette base n'est d'ailleurs pas sans poser problème[36] (par exemple, le système a inféré le 21/3/11 le fait *important_elements is a protein*).

Ce courant des approches théoriques qui voient le sens comme une représentation issue d'un calcul est encore dominant dans le domaine du TAL et l'essor de l'Internet dans les dernières années lui a même donné un second souffle. En témoigne le projet du Web Sémantique[37]. En matière de recherche d'information et d'ingénierie documentaire sur Internet, la tendance actuelle est de faire du Web une vaste base de connaissances. C'est la démarche considérée dans le projet du Web Sémantique (et prolongée dans le projet *Linked Data*). L'objectif annoncé par Tim Berners-Lee[38], initiateur du projet et créateur du langage HTML (par la suite fondateur du W3C), est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents sur le Web avec des informations sur leur propre sémantique de manière que ces informations soient directement utilisables par des agents logiciels sans la supervision d'une interprétation humaine :

> *« The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in coopération » (Berners-Lee, mai 2001, définition du projet de Web Sémantique).*

L'objectif visé est de permettre notamment à des outils de veille et des moteurs de recherche d'avoir un « accés » au sens (d'une certaine façon à l'égal de l'humain pour les défenseurs du Web Sémantique) pour permettre une meilleure recherche d'information relativement à une requête. Il en découle souvent des propos très largement phantasmés sur les attendus du Web Sémantique, comme en témoignent les propos suivants rapportés par Jean-Paul Pinte sur le Web 3.0 vu comme une autre appellation du Web Sémantique :

> *« Selon Véronique Mesguich et Armelle Thomas[39], quelle que soit votre recherche, les*

34 D'après Jean-Marie Hombert (Hombert 2011), on estime que 20% du vocabulaire de base d'une langue varie tous les 1000 ans.

35 http://rtw.ml.cmu.edu/rtw/ consultée le 17/11/12.

36 http://www.lemonde.fr/technologies/article/2010/10/15/quand-la-machine-apprend-le-langage_1426969_651865.html consultée le 17/11/12.

37 http://www.w3.org/standards/semanticweb/ consultée le 17/11/12

38 http://fr.wikipedia.org/wiki/Tim_Berners-Lee consultée le 17/11/12.

39 V. Mesguish & A. Thomas, *Net Recherche 2009 : le guide pratique pour mieux trouver l'information utile et surveiller le Web*, ADBS Eds. 2009.

> *outils du Web 3.0 permettront non seulement aux humains d'interagir, mais offriront également une meilleure organisation de l'accès aux énormes gisements d'informations numérisées disponibles en modélisant le cheminement du cerveau humain. » (Pinte 2011, p. 103)*

Un des points de départ du Web Sémantique est le concept de métadonnées, comme données décrivant des données. Ici les contenus des pages web (textes, images, etc) sont vus comme des données qu'il faut annoter par des métadonnées pour en représenter le sens de manière computationnellement « manipulable ».

Le Web sémantique propose essentiellement des méthodes, des langages et des techniques pour d'une part enrichir le contenu des pages, d'autre part permettre l'exploitation de ces enrichissements sémantiques à l'aide de différentes ressources ontologiques. La représentation ontologique est au centre du projet et elle donne lieu à la création de plusieurs langages issus des technologies XML, par exemple RDF (Ressource Description Framework) , OWL (Web Ontology Language) couplé avec RDF, ou encore DAML (DARPA Agent Markup Language). Ces langages sont issus des travaux en logique formelle développés dès les débuts de l'IA. Notamment OWL est présenté comme basé sur les logiques de description, elles mêmes issues des réseaux sémantiques de Quillian (Quillian 1968) et des Frames de Minsky (Minsky 1975).

Dans la continuité des formalismes logiques, l'interprétation pour le Web Sémantique (permettant une annotation d'un document par son sens) deviendrait un mécanisme d'inférence basé sur l'ontologie. Ce mécanisme inférentiel part des contenus de pages web pour au final produire des valeurs de vérité. C'est une position complètement assumée comme en témoigne le fameux *Web Semantic Layer Cake*[40] (cf illustration 37).

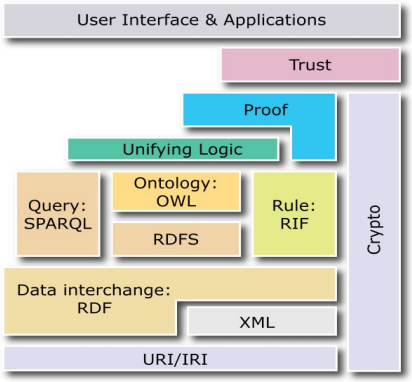


Illustration 37: Le « Web Semantic Layer Cake »
http://www-igm.univ-mlv.fr/~dr/XPOSE2009/Le%20Web%203.0/technologies.html consultée le 7/01/13
A la base, nous trouvons toujours les URI qui permettent d'identifier les ressources. RDF est le modèle de base pour l'échange des données et peut éventuellement s'appuyer sur XML, dans le cas de l'utilisation de la syntaxe RDF/XML. Au-dessus, on trouve SPARQL pour effectuer des requêtes, RDFS et OWL pour définir des vocabulaires RDF, assimilables à des ontologies et RIF, un langage de définition de règles. Au sommet, on trouve la logique, la preuve et les valeurs de vérité.

40 Le graphique date de 2007 et a évolué. Dans sa version de 2001, la base était Unicode, ce qui laissait encore une place aux textes (cf. Rastier 2011).

Les récentes évolutions du Web Sémantique consistent à mettre en relation les annotations sémantiques des documents les unes avec les autres. L'idée est de mettre en œuvre, comme caché derrière la dimension langagière des documents (dimension propre aux humains), un Web des données interconnectées. C'est le projet *Linked Data*[41].

Scientifiquement, le Web Sémantique n'est pas sans poser problème (qui plus est relativement à une démarche centrée-utilisateur comme la nôtre) et apparaît comme une ambition technologique sans beaucoup de recul épistémologique (en dehors de l'ancrage dans l'ontologie) :

- Le rapport de l'Action Spécifique 32 CNRS/STIC de 2003 (Charlet & al. 2003) dresse un constat sur l'avancement du projet du Web Sémantique où des investissements financiers et humains sont colossaux. La conclusion du document parle d'obstacles « particulièrement cruciaux pour les débuts même du Web sémantique » (ibid. p.123). Ces obstacles concernent entre autres l'absence de consensus sur les langages et ontologies à utiliser mais surtout, c'est la possibilité même de l'utilisation généralisée de méta-données qui est mise en doute : « la détermination et l'ajout, même de simples méta-données, n'est pas une activité naturelle pour la plupart des personnes. (…) Les expériences dans la construction d'ontologies sont (…) instructives et pourraient contribuer à lever quelques illusions.» (ibid. p.124).
- Les principes d'organisation et de représentation ontologique sur lesquels repose le Web Sémantique sont applicables à l'échelle d'un système d'information fermé comme une entreprise ou une bibliothèque. C'est un cadre limité dans lequel la description ontologique est appréhendable dans sa globalité et donc possible. C'est ce que Houssem Assadi (Assadi 1998) appelle les ontologies régionales. Par contre il est beaucoup plus douteux que ces principes puissent être applicables au monde dans son ensemble (c'est ce qu'a montré l'échec du projet CYC de Lenat[42]). On peut donc s'attendre à ce qu'elles ne fonctionnent pas ou mal à l'échelle du Web. Les raisons ne sont pas techniques mais bien plus sémiotiques et sociales, ce que les défenseurs du Web Sémantique ne veulent visiblement pas voir.
- D'un point de vue très pratique, force est de constater que des ressources très généralistes, valables pour tout type de traitement envisagé ainsi qu'à destination de tout rédacteur potentiel, ne sont pas facilement disponibles (sous forme électronique pour des traitements automatiques) et encore moins gratuites. Pour utiliser efficacement une ontologie, il faut bien la connaître à la manière dont un documentaliste connaît son fond documentaire et son catalogue. Ce n'est pas à la portée de tout créateur de contenu sur le Web.
- Le Web est loin d'être assimilable à une vaste base de connaissances. C'est encore plus vrai depuis l'essor récent de la blogopshère et des réseaux sociaux. Il est un reflet de la société, ni plus ni moins.
- Penser que le sens d'un document puisse être décrit de manière objective et consensuelle est un point de vue naïf, à moins qu'il ne s'agisse d'un projet d'uniformisation culturelle. Qui pourrait penser par exemple que les concepts de bien ou de mal soient définis de la même façon d'une culture à l'autre (même à une même époque donnée).
- Le sens n'est évidemment pas indépendant de son contexte d'interprétation. Sinon, cela revient à nier la dimension pragmatique des langues. Dans ces conditions il paraît difficile de décider

41 Consulter pour une présentation de Linked Data l'exposé de T. Berners-Lee aux conférences TED le 13 mars 2009 : http://www.youtube.com/watch?v=OM6XIICm_qo consultée le 17/11/12.

42 En 1986, 2 ans après le lancement du projet, Doug Lenat (Lenat 1986) estimait encore qu'il fallait encore enrichir CYC d'environ 250 000 règles et que cela demanderait 350 années-homme de travail !

quel est le sens d'un document avant d'en connaître ses conditions de réception.

- Le sens d'un document n'est pas uniquement le fait de son auteur. Ce que l'interprétant apporte avec lui est tout aussi important pour établir un sens. C'est ce qu'indique notamment la notion d'anagnose chez (Thlivitis 1998) ou encore la fonction muette du langage de (Coursil 2000).

Tous ces arguments sont bien sûr à opposer aux approches théoriques du sens dans leur ensemble et plus largement que le Web Sémantique. D'un point de vue plus philosophique, il nous semble que ces approches traduisent une vision dénaturée du sens. D'une part, le sens est une donnée, un résultat qui pourrait se passer du texte (c'est même explicitement le but de *Linked Data*). Il n'y a pas de moyens de dire le sens d'un texte en dehors d'un autre texte. Dit autrement, il n'y a pas de réalisations extralinguistiques du sens. D'autre part, nous rejetons une vision purement utilitariste du sens. Le sens ne nous sert pas à mener à bien telle ou telle tâche (communication, recherche d'information …) comme il pourrait servir à un agent logiciel pour, lui aussi, répondre à une tâche. Le sens est un milieu dans lequel on vit (comme l'eau est un milieu pour les poissons), un environnement sémiotique nécessaire à la vie humaine.

4.1.2. Les approches empiriques

Un autre courant que celui des approches théoriques est apparu en TAL dans les années 1990. Ce courant est inspiré par les méthodes de la linguistique qui consistent à observer des données attestées. Dans les approches théoriques, le chercheur est dans une position ambivalente : il cherche à formaliser le « fonctionnement » du langage et, en tant que sujet parlant, il se donne (dans une démarche quasi introspective) des exemples pour corroborer (ou pas) son formalisme. Les approches empiriques en mettant l'accent sur des productions réelles issues de « vraies » productions langagières ont choisi d'adopter une démarche scientifique plus satisfaisante. De ce fait, en analysant des productions langagières situées (i.e. dont on peut rendre compte du contexte) on peut notamment plus facilement observer (en toute objectivité) des variabilités contextuelles dans la signification.

La linguistique a depuis longtemps adopté cette démarche scientifique. Il aura fallu attendre les années 1970-1980 que les capacités de calcul des ordinateurs soient facilement mobilisables (notamment avec l'usage des tableurs destinés au plus grand nombre) et que des méthodes statistiques génériques[43] soient implémentées et disponibles pour que le TAL s'intéresse plus souvent au recueil et à l'analyse de vastes données linguistiques. Cette « démocratisation » du calcul n'a d'ailleurs de fait pas impacté que le TAL mais plus largement aussi la linguistique qui en a tiré des protocoles d'observation de données qui jusqu'alors n'étaient pas forcément à la portée du linguiste. Ceci engendre deux courants scientifiques :

- Celui de la linguistique informatique[44] qui cherche à mobiliser des outils informatiques (essentiellement statistiques) pour observer des données linguistiques recueillies en fonction d'hypothèses.
- Celui de l'informatique linguistique qui cherche à inférer à partir de régularités observées sur de vastes données linguistiques attestées des processus de traitement et d'analyse des langues. C'est ce courant du TAL que nous dénommons par « approches empiriques ».

43 Telles que les analyses factorielles des correspondances ou encore les analyses par composantes principales par exemple.

44 Les dénominations *linguistique informatique* et *informatique linguistique* sont empruntées à François Rastier (Rastier 1994).

Les approches empiriques en TAL sont théoriquement fondées sur la théorie du distributionnalisme de Harris (Harris 1951). Elles sont aussi très inspirées par la lexicométrie, dans le prolongement notamment les travaux de Zipf (Zipf 1949) qui expriment une décroissante linéaire sur une échelle logarithmique des occurrences des mots d'un texte classés du plus fréquent au moins fréquent. De cette loi de Zipf plusieurs interprétations sont possibles telles que :

- la fréquence du mot de rang n est à peu près 1/n de l'occurrence du mot de rang 1
- plus un mot est fréquent, plus il est court
- les mots qui indexent le mieux le texte sont ceux qui sont à la fois longs et occurrents

Les méthodes statistiques de l'analyse des données sont appliquées dans le domaine des données textuelles. Les méthodes d'analyse en composantes principales, de classification hiérarchique de données, d'analyse factorielle des correspondances sont notamment mises à profit pour mettre en évidence des régularités globales qui ne sont pas immédiatement perceptibles. Nous en avons mis certaines à profit dans le cadre de ProxiDoc. Des outils ont été développés pour mettre facilement en œuvre des observations statistiques de corpus. C'est le cas par exemple du logiciel Lexico3 de l'équipe d'André Salem[45].

Les approches sur corpus ont marqué une véritable évolution du TAL. Elles se sont développées grâce aux évolutions de la statistique textuelle mais également grâce à une plus grande facilité d'accès à des données textuelles nombreuses du fait de l'essor de l'internet. Le recueil de données textuelles constituant les corpus est beaucoup plus simple aujourd'hui grâce au Web mais il ne faut pas pour autant considérer le Web comme un grand corpus. La constitution d'un corpus répond à un projet de collection de textes relevant tous d'un même genre (articles de journaux, dépêches d'agence de presse, romans relevant d'une époque donnée, dialogue …).

La fouille de données et la recherche d'information sont des domaines qui ont beaucoup apporté aux approches empiriques, notamment en terme de méthodologie. Par exemple, les notions de corpus séparés pour l'apprentissage et les tests en témoignent. Les protocoles d'évaluation de la recherche d'information le montrent aussi, notamment avec les mesures de rappel, précision et f-mesure (Van Riesbergen 1979). Ce n'est d'ailleurs pas sans poser quelques problèmes de « standardisation scientifique » dans la mesure où il est maintenant difficile de faire accepter des publications n'affichant pas des taux de rappel et de précision en guise d'évaluation.

Entre le recueil facilité de corpus et les mesures d'évaluation « prêtes à l'emploi », on constate paradoxalement que certains travaux tendent à instaurer une séparation entre le TAL et la linguistique au motif que l'informaticien n'a plus besoin des connaissances du linguiste car en interrogeant directement des corpus il serait à même de corroborer telle ou telle hypothèse. C'est une tendance dont Cécile Fabre (Fabre 2010, p.139) estime les débuts dès les années 1990. Elle s'amplifie aujourd'hui avec le contexte de ce qu'on appelle *big data*[46]. Même en défendant un certain pragmatisme d'approche, faire ainsi l'impasse sur les connaissances linguistiques explicites est, bien entendu, extrêmement préjudiciable à une maturité pluridisciplinaire de l'objet d'étude. Il est à craindre que ce courant de *la linguistique computationnelle* ne reproduise des erreurs de « jeunesse » du TAL qui ne voyait dans les questions linguistiques que des problèmes de calcul. Tout comme Cécile Fabre, nous

45 cf. http://www.tal.univ-paris3.fr/lexico/lexico3.htm consultée le 17/11/12.

46 cf. notamment à ce sujet http://blog.veronis.fr/2012/10/conf-big-data-et-technologie-du-langage.html, consultée le 17/11/12.

défendons vivement ici que les corpus offrent au contraire un terrain de recherche commun, et enrichissent les échanges entre la linguistique et le TAL.

Les approches empiriques marquent néanmoins un renouveau du TAL qui en tirant profit des méthodes de la linguistique réaffirme la pertinence pluridisciplinaire du domaine. Nos travaux sont bien sûr très inspirés par les approches sur corpus (comme en témoignent les projets ThèmeEditor, ProxiDocs ou encore Canopée). Cependant, si les approches empiriques invitent à une observation de données réelles, l'accent est uniquement ciblé sur le matériau linguistique. C'est-à-dire que l'interprétant (l'utilisateur) n'est pas encore explicitement le centre du protocole d'observation. C'est à notre sens cette démarche centrée-utilisateur qui manque aux approches sur corpus. Mais, à la différences des approches théoriques, la démarche centrée-utilisateur est tout à fait conciliable avec l'approche empirique.

4.1.3. TAL et interaction

L'objectif d'origine du TAL est de faire faire automatiquement par des machines ce que l'humain (en tant qu'individu) « fait » avec le langage. Cette façon de poser la problématique du TAL a d'emblée repoussé au second plan la question des interactions puisque l'objectif premier n'est pas forcement de faire interagir par le langage les humains et les machines. Dés lors la démarche centrée utilisateur s'en trouve, elle aussi, reléguée à l'arrière-plan.

L'interaction personne/machine, proposée par les systèmes de TAL qui se disent interactifs (typiquement les systèmes de correction orthographique par exemple) est en effet le plus souvent préemptive, ce qui signifie que la machine requiert une information de l'humain pour pouvoir continuer son analyse. Les travaux en TAL où l'interaction n'a pu, par nature, être complètement éludée sont ceux sur le Dialogue Homme-Machine (DHM). Cependant là encore, on constate le plus souvent que l'interaction langagière telle qu'elle est mise en place dans les modèles de DHM est souvent dénaturée car également préemptive et finalement perd l'essentiel de son intérêt. Par exemple, la plupart des applications de DHM depuis ELIZA (Weizenbaum 1966) ont souvent consisté à mettre en place des échanges d'énoncés écrits (puisque tapés au clavier et affichés à l'écran) ce qui remet forcement en cause la production non préméditée des énoncés et la forme qu'ils prennent dans l'interaction (incluant des reprises ou encore des hésitations qui sont autant d'éléments signifiants pour l'intercompréhension). De même, l'enchaînement strict des tours de paroles (l'un bien après l'autre) est une forte contrainte qui fait que l'on est très éloigné de ce qu'est une conversation naturelle où les interlocuteurs parlent souvent au même moment ce qui consiste généralement en des procédures de régulation mutuelle et de co-construction d'un terrain commun. Enfin dans la plupart des systèmes de DHM, à l'exception notamment de COALA de J. Lehuen (Lehuen 1997), la machine n'est pas capable d'apprendre dans et par le dialogue lui-même et donc pas capable non plus de réagir aux erreurs et événement inattendus. La seule façon d'éviter ces désagréments consiste à contraindre l'interaction pour se prémunir des inattendus. On constate alors souvent que l'interaction n'est plus qu'un pâle reflet du modèle qu'on s'est fait de la tâche en cours dans le dialogue (c'est le cas des architectures de DHM les plus courantes basés sur la planification du dialogue par la tâche). C'est par exemple particulièrement flagrant dans le prototype de DHM pour la réservation de billets de train de la SNCF (système ARISE en collaboration avec le LIMSI/CNRS[47]) ou même dans le système SIRI de l'iOS[48]. Le relatif échec du DHM par rapport à ses objectifs initiaux doit certainement beaucoup à cette dénaturation de l'interaction.

47 (Lamel & al. 2000)

48 http://www.apple.com/fr/ios/siri/ consultée le 17/11/12.

Force est donc de constater une relative pauvreté des travaux de TAL tournés vers une problématique des interactions. Comparativement, d'autres courants de l'informatique ont plus travaillé la question des interactions. C'est par exemple le cas des modélisations multi-agents relativement aux interactions entre agents logiciels ou encore celui des Interfaces Homme-Machine (IHM) relativement aux interactions entre un système et son utilisateur. Certaines tentatives de « croisement » avec le TAL de ces courants existent mais restent relativement isolées (Maurel 2004, par exemple). On constate notamment que les communautés de chercheurs en IHM et en TAL ont une intersection très faible avec, par exemple, des revues et des conférences bien distinctes. Cela traduit de fait un positionnement épistémologique peu confortable en informatique pour l'étude des interactions langagières personnes-machines.

Profitant de l'énorme évolution technologique des ordinateurs et de leurs capacités de calcul de plus en plus importantes, le courant des IHM s'est majoritairement orienté vers une problématique technique de réalisation d'interfaces. Ce faisant elle s'est beaucoup plus rapproché scientifiquement des travaux en ergonomie que des travaux en TAL. Cependant, suivant le point de vue de Michel Beaudouin-Lafon (Beaudouin-Lafon 2004), cette façon de concevoir les IHM doit être aujourd'hui dépassée. La question principale n'est pas de proposer des interfaces toujours plus développées (mais finalement assez similaires aux premières interfaces graphiques au début des années 1980 car toujours basées sur les mêmes fonctionnalités : icônes, boutons, *drag & drop* ou encore copier/coller) mais de mettre en place des interactions avec l'utilisateur. Pour Beaudouin-Lafon, les interfaces doivent être aux IHM ce que les télescopes sont à l'astronomie, c'est-à-dire un moyen et non un but. Le but est l'interaction. L'interface idéale serait alors celle qui serait « invisible » laissant ainsi place à une interaction naturelle donnant notamment aux machines des capacités de résilience c'est-à-dire des capacités de gestion des imprévus (ce qui comme on l'a vu fait crucialement défaut aux modèles de DHM).

Dans la suite des travaux de Anne Nicolle et Jacques Coursil nous avons proposé une approche interactionniste du sens en TAL qui fait un constat assez similaire à celui de Michel Beaudoin-Lafon sur le recours nécessaire à l'interaction. Dans cette approche le sens est considéré comme une activité sémiotique au centre de l'interaction homme-machine. L'interface idéale est celle qui engendre un couplage personne-système où le sens est émergent. Instaurer ce couplage ne peut pas faire l'impasse d'une approche centrée-utilisateur où les corpus, les ressources, le contexte sont ceux de l'utilisateur. Il convient ainsi de subjectiver les systèmes interactifs.

Les approches centrées-utilisateurs dépassent largement la problèmatique du TAL. On en trouve plus généralement dans le domaine de la conception des logiciels et des systèmes d'information. C'est même un domaine d'expertise qui donne lieu à des activités commerciales, comme en témoigne notamment la cible visée (*User Experience Design*) de la société de développement informatique Axance[49]. Concilier les évolutions du TAL et l'approche centrée-utilisateur est un de nos objectifs. Pour convaincre il faut amener la preuve que notre type d'approche est vecteur de valeur ajoutée en terme de couplage personne-système et d'usage. Une évolution récente de l'Internet à récemment démontré implicitement que le passage à l'échelle des approches centrées sur les utilisateurs est possible et même vertueux. C'est le cas de ce que l'on nomme le Web 2.0.

4.2. Le cas du Web 2.0

Un des principes du Web 2.0 est l'idée du Web collaboratif, c'est-à-dire l'idée que l'usage qu'ont

49 http://www.axance.fr/ consultée le 7/01/13.

certains utilisateurs du Web peut être cumulé de façon altruiste pour alimenter l'usage d'autres utilisateurs. C'est une démarche dite d'UGC : *User Generated Content* ou encore de *crowdsourcing*. La force d'une telle approche réside dans le nombre d'internautes concernés car bien qu'on estime à seulement 1% la part des utilisateurs altruistes qui sont générateurs de contenus (les autres étant pour environ 10% des utilisateurs qui amendent les contenus créés par les premiers et les 89% restants étant visiteurs du contenu généré), l'aspect cumulatif et la masse des utilisateurs, très volontaires dans le création de contenus, fait qu'on arrive assez vite à des contenus très larges. Ainsi en s'appuyant sur le partage, la créativité, l'intelligence et le savoir-faire de tous on peut mettre en œuvre notamment un service de recherche d'information aussi utilisable que ce que pourrait proposer un moteur de recherche classique. C'est particulièrement ce que montrent des sites Web comme Delicious[50] ou encore Flickr[51]. Flickr est un site de partage et de recherche d'images. Delicious est un site de partage de favoris offrant un service de recherche d'information. Le point commun des deux sites réside dans le partage de *tags*, c'est-à-dire des étiquettes librement entrées (voire spontanément entrées) par les utilisateurs, qui tiennent lieu d'entrées d'index pour les images ou les pages web. Ainsi plutôt que d'indexer de l'information relativement aux concepts d'une taxonomie (une classification hiérarchisée à la façon d'une ontologie) on indexe les contenus avec des termes choisis par les utilisateurs. C'est ce qu'on appelle en opposition à la taxonomie, la *folksonomie*.

Pour bien mettre en évidence cette opposition, comparons par exemple le site de partage de favoris Delicious et l'annuaire de recherche d'information Yahoo[52] :

- Dans Yahoo l'indexation est faite par les catégories et sous catégories de l'annuaire, c'est-à-dire relativement à l'ontologie du Web que les auteurs de l'annuaire ont mis en place. Dans Delicious, l'indexation provient des tags choisis sans aucune contrainte ni justification par les utilisateurs quels qu'ils soient. Il en découle que les index considérés comme étant les meilleurs sont ceux qui reviennent statistiquement dans le plus grand nombre de tags (c'est-à-dire les consensus dans la communauté des utilisateurs). Yahoo est donc onto-centré là ou Delicious est statistiquement socio-centré.
- D'un point de vue de l'évolution de l'indexation, Yahoo et Delicious sont encore diamétralement opposés : pour Yahoo il faut maintenir une ontologie censée convenir à tous au risque de la voir devenir de moins en moins contrôlable (avec les problèmes de cohérence hiérarchique et de recouvrement de catégories) alors que dans le cas de Delicious les index une fois agrégés donnent une valeur ajoutée qui indique l'évolution d'un courant dominant consensuel où il s'opère une modération naturelle des avis non consensuels par simple fréquence d'occurrence.
- La nature des index est très différente d'un site à l'autre. Ils sont conceptuels chez Yahoo, comme le montre d'ailleurs l'usage exclusif de groupes nominaux pour étiqueter les concepts (par exemple, actualité, société, arts, sciences …). Ce sont des termes chez Delicious qui peuvent être contextuels et/ou temporaires. C'est le cas par exemple du tag « à_lire » (dont on peut constater au passage que ce n'est pas un groupe nominal) qui n'est pas assimilable à un concept et qui pourtant indexe 1244 pages (expérience faites le 21/11/12). C'est aussi le cas de tags s'apparentant à des expressions qui ne représentent pas non plus un concept (exemples : « bof », « mouais », « yeah »,« wouah » ...).
- Les tags sont importants essentiellement pour ce qu'ils excluent et pour leurs utilisations

50 http://www.delicious.com/ consultée le 21/11/12.

51 http://www.flickr.com/ consultée le 21/11/12.

52 http://fr.yahoo.com/ consultée le 21/11/12.

conjointes statistiquement significatives (les *related tags*) alors que les concepts sont justifiés par ce à quoi ils réfèrent. Les tags sont à ce titre assimilables à des valeurs (au sens de Saussure) qui se définissent par la place qu'ils occupent dans le système qu'ils forment avec les autres.

On constate aujourd'hui que le système de partage de favoris généraliste Delicious semble avoir réussi le pari de fédérer des millions d'utilisateurs à travers le monde. Plus généralement, c'est vrai de tous les sites Web 2.0 basés la génération collaborative de contenus. Delicious peut ainsi concurrencer un véritable moteur de recherche. D'ailleurs, on constate aujourd'hui que les annuaires classiques (de type Yahoo) sont de moins en moins nombreux et évoluent vers des systèmes à base de tags. Les tags sont une idée des plus simples. Ils s'approprient de manière naturelle par les utilisateurs qui implicitement « réinventent » par leur usage des tags la notion de sème de la linguistique.

Le pari gagné par les sites Web 2.0 en architecture UGC est l'énorme valeur organisationnelle de l'information que la simplicité de l'usage des tags engendre. En comparaison, la grande complexité que représente l'utilisation et la maintenance de vastes ontologies standardisées explique peut-être en partie que le Web Sémantique n'ait pas la même croissance que le Web 2.0. Le Web Sémantique que l'on présente parfois comme le Web 3.0 (cf. illustration 38 qui est un point de vue datant de 2007) serait de ce point de vue en train de perdre son pari. Hélas, pas complètement. L'effet pervers est que si les contenus échappent majoritairement à une indexation ontologique du fait des pratiques du Web 2.0, il ne reste plus au Web Sémantique qu'à indexer autre chose et l'on peut craindre que ce soit les profils d'utilisateurs qui en fassent l'objet. D'autant plus que, dans ce domaine, il y a de là une vrai demande commerciale de la part des systèmes de recommandation d'achats qui se sont multipliés avec les architectures Web 2.0.

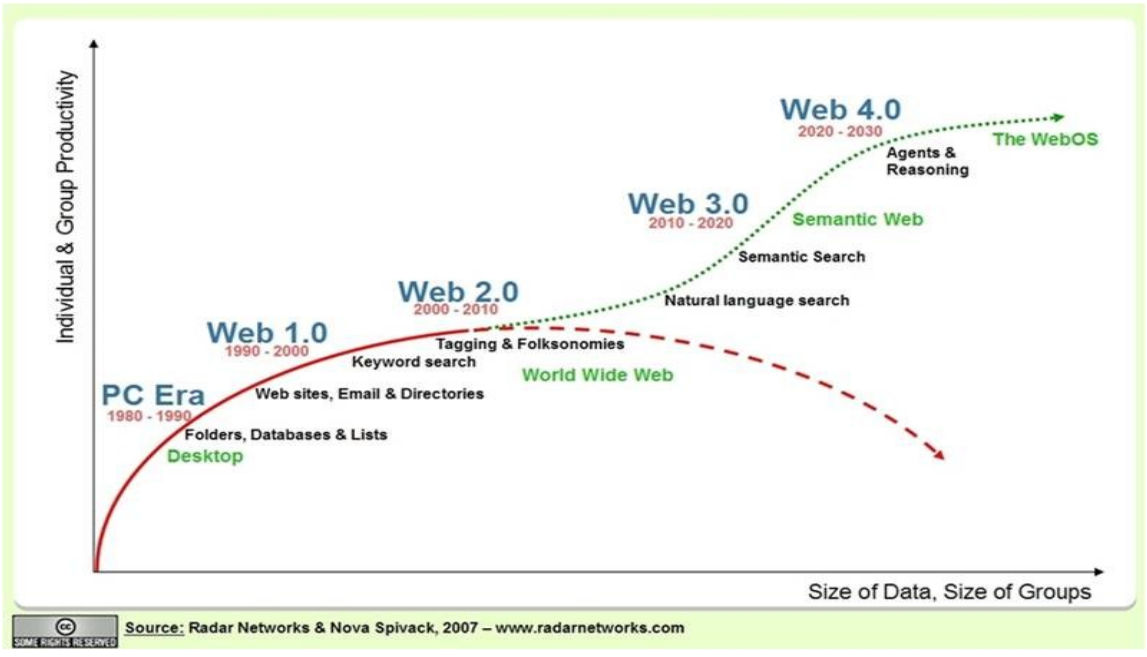


Illustration 38: Un point de vue sur les évolutions de l'Internet

La simplicité est donc vertueuse. C'est un des enseignements du Web 2.0. Ce n'est pas parce qu'une tâche est complexe, ce qui est indiscutablement le cas de la veille sur Internet, que les outils qui instrumentalisent cette tâche doivent aussi être complexes. Les outils les plus simples du point de

vue des concepts et de l'ergonomie sont bien souvent ceux qui rencontrent le plus grand succès auprès des utilisateurs novices ou même experts. Par exemple on peut remarquer que la popularité de Google et sa supériorité par rapport aux autres moteurs de recherche s'est accrue notamment parce que son interface utilisateur très simpliste tranchait par rapport à ses concurrents. Cette vertu de la simplicité se confirme aujourd'hui dans les nouveaux produits de Google. En témoigne encore le navigateur Chrome sorti par Google le 2 sept. 2008 et qui était présenté de la manière suivante :

> *« Google Chrome vous permet de naviguer sur Internet plus vite, plus facilement et en toute sécurité, à l'aide d'une interface très simple qui sait rester discrète ».*

Par ailleurs, les logiciels les meilleurs en terme de résultats techniques ne sont pas toujours les plus utilisés si leur appropriation par les utilisateurs n'est pas facilitée. Un exemple est celui des traitements de textes : tout le monde s'accorde pour dire que Latex est un excellent traitement de texte qui est notamment bien meilleur techniquement que Microsoft Word et pourtant ce dernier est de loin le plus utilisé par ce que le couplage avec l'utilisateur est plus immédiat.

Le Web 2.0 montre une approche très différente de celle du Web Sémantique. Elle est centrée utilisateur là où le Web Sémantique est techno-centré. Plutôt que de chercher à tirer profit des ontologies, le Web 2.0 avec la folksonomie cherche à tirer profit des utilisateurs qui, dans leurs activités numériques, sont très demandeurs d'une mise à disposition pour les autres de leur expérience propre. Il en découle que l'accès à des contenus (textes, images, vidéo) exploite une sémiotique de l'interprétation car les tags sont des traces *a posteriori* des interprétations des utilisateurs. C'est une différence de taille par rapport à l'indexation ontologique où le sens est vu avant tout du côté de l'auteur qui annote son document par des métadonnées extraites de l'ontologie.

Il nous semble que dans le domaine du TAL le même positionnement que celui du Web 2.0 relativement aux utilisateurs doit pouvoir être institué. Ce n'est pas dans les approches théoriques et empiriques qu'il faut chercher ce positionnement. Il faut l'importer d'une théorie de l'interprétation vers une problématique du TAL. Cette théorie de l'interprétation, nous la trouvons en linguistique. C'est la Sémantique Interprétative.

4.3. Sémantique Interprétative pour l'approche centrée-utilisateur

Selon Bruno Bachimont (Bachimont 2000), les expressions manipulées par un traitement informatique n'ont pas de sens en soi. Leur sens leur est conféré par le sémiotique et la linguistique, c'est-à-dire par l'humain.

> *Ainsi, l'informatique est autothétique. Cela signifie que les expressions qu'elles manipulent sont dépourvues de sens, comme le sont les nombres entiers. Pourtant, l'usage quotidien de l'ordinateur montre bien que les expressions symboliques manipulées possèdent un sens, et c'est ce sens associé qui justifie, la plupart du temps, la manipulation informatique. Selon nos considérations, cela est incompréhensible, voire absurde. C'est que le signe informatique se voit sur-déterminé par le sémiotique et linguistique de manière à ce que les constructions symboliques issues de l'effectivité informatique possèdent néanmoins un sens pour le sémiotique. (Bachimont 2000)*

Nous situons justement notre approche centrée-utilisateur en TAL dans le cadre d'une théorie linguistique et sémiotique qui puisse expliquer comment l'interprétation confère du sens aux signes et aux textes. C'est la Sémantique Interprétative (SI) de François Rastier (Rastier 1987) et plus généralement la sémantique componentielle dont elle hérite. La SI étant elle-même en filiation avec la sémantique structurale dont l'origine remonte aux travaux de Saussure.

Les apports de la SI au TAL sont aujourd'hui reconnus et d'autres travaux que les nôtres en témoignent également. Notamment, I. Kanellos et C. Mauceri développent une plateforme d'analyse interprétative des données en utilisant les principes de la SI. Leur projet, auquel nous adhérons, est « *de construire un outil en vue de donner corps, par son maniement, au cercle herméneutique de l'interprétation des données* » (Kanellos & Mauceri 2008, p. 42). Ils combinent à cette fin des outils interprétatifs tels que l'isotopie et des méthodes statistiques éprouvées comme l'analyse sémantique latente (Dumais 1998).

De la SI on retient principalement deux aspects. Premièrement, on y trouve un « appareillage » théorique fin et complet pour la description d'effets de sens. C'est la description sémique avec les différents types de sèmes, d'isotopies et plus largement d'opérations interprétatives. Deuxièmement (et c'est tout aussi important), on s'approprie de la SI un positionnement épistémologique relativement à la question du sens. Celui-ci consiste à préférer la tradition rhétorique et herméneutique à la tradition logico-grammaticale. La SI nous donne un cadre de pensée qui prend le contre-pied de tout ce que nous reprochons aux approches théoriques du TAL et, en même temps, elle nous permet de reprendre ce qui est positif dans les approches empiriques du TAL, en y plaçant au premier plan la question de l'utilisateur qui fait défaut aux approches empiriques.

4.3.1. Description sémique

Nous ne présenterons pas ici de manière détaillée les différents concepts opératoires de la SI. Nous renvoyons pour cela aux ouvrages de François Rastier (Rastier 1987, Rastier & al. 1994 notamment[53]). Simplement nous rappellerons en quoi ces concepts nous intéressent dans le cadre d'une approche centrée-utilisateur en TAL.

La notion de sème est la notion centrale de la sémantique componentielle où s'inscrit la SI. Le sème est défini comme toute paraphrase métalinguistique indiquant un aspect de la signification d'une lexie. Le sème présente plusieurs propriétés qui retiennent notre attention :

- Le sème est une notion générale qui se spécialise en deux notions très utiles d'un point de vue opératoire : les sèmes génériques et les sèmes spécifiques. Les sèmes spécifiques permettent de différencier les lexies au sein d'un même taxème (classe de lexies proches). Ainsi on pourrait dire par exemple que les lexies *vélo* et *mobylette* diffèrent uniquement par la présence ou pas du sème spécifique /moteur/. Les sèmes spécifiques apportent une grande finesse de description d'une part de la dimension paradigmatique, et d'autre part de la dimension syntagmatique (comme le montrent notamment l'interprétation des énumérations[54]). Les sèmes génériques indexent des lexies proches dans des mêmes champs lexicaux. Nos deux exemples *vélo* et

53 Nous renvoyons également à un glossaire en ligne des notions de la Sémantique interprétative consultable à l'adresse http://www.revue-texto.net/1996-2007/Reperes/Reperes.html consultée le 7/1/13.

54 Dans l'expression « *Vins, bières et boissons non comprises* » la lexie *boissons* se distingue de *bières* et *Vins* par un sème spécifique /sans alcool/ (qui ici est contextuellement afférent du fait de l'énumération)

mobylette portent notamment un même sème générique /moyen de locomotion/. Les sèmes génériques repérés dans l'enchainement syntagmatique permettent de dégager les grandes orientations thématiques (c'est notamment ce que montrent les coloriages réalisés avec ThèmeEditor).

- Le sème est une notion simple à appréhender pour un utilisateur qui veut décrire la signification de ses propres termes sans pour autant être linguiste. Il n'en serait pas forcement de même de quelqu'un qui voudrait définir un concept dans une ontologie. Cette simplicité du sème, loin d'en restreindre la portée opératoire, est un argument important dans une démarche centrée-utilisateur où le couplage personne/système passe également par une facilité à maitriser les notions opératoires du système.
- Le sème est une entité linguistique à part entière. On entend par là que la description de la signification est intralinguistique dans la mesure où on décrit le signifié avec du signifiant. C'est un avantage par rapport aux logiques formelles ou aux sémantiques référentielles. Insister sur la nature intralinguistique du signe consiste aussi à réaffirmer l'idée que la signification et le sens n'ont pas de réalisation non langagière.

La description de la signification en sème résulte pour la SI d'un processus de décontextualisation. C'est un principe qui convient très bien à une approche centrée-utilisateur car la description de ressources terminologiques par un utilisateur ne se fait pas sans une justification et un rapport fort à la tâche que doit mener l'utilisateur, c'est à dire à un contexte et un corpus préexistant. La conséquence de ce principe est de considérer le signe comme un passage de texte pris isolément, affirmant ainsi le statut premier de la textualité et de la parole. Par rapport aux approches logico-grammaticales, la logique est renversée car c'est la textualité qui construit les signes et pas le contraire de manière compositionnelle.

Il découle de la notion de sème un concept central de la SI, hérité des travaux de Greimas (Greimas 1966), celui d'isotopie. L'isotopie est également une notion qui s'énonce facilement : c'est la redondance d'un même sème dans une zone de texte. En fonction de la nature du sème qui est récurrent dans la zone de texte, les isotopies peuvent être les isotopies spécifiques, génériques, ou mixtes[55] :

- Isotopie spécifique : une isotopie est spécifique quand le sème récurrent est un sème spécifique. C'est le cas de l'isotopie du trait /inchoatif/ dans le vers d'Éluard *L'aube allume la source* inhérent aux sémèmes 'aube', 'allume' et 'source'.
- Isotopie générique : une isotopie est générique quand le sème récurrent est un sème générique. En fonction du degré de généricité du sème, une isotopie générique peut être microgénérique, mésogénérique ou macrogénérique. Elle est microgénérique si le sème en jeu dans l'isotopie est microgénérique, c'est-à-dire qu'il indexe des sémèmes appartenant au même taxème (c'est le cas de l'isotopie du sème /degré de cuisson/ dans les sémèmes 'bleue', 'saignante', 'à point' et 'bien cuite' de l'énoncé *Et l'entrecôte, bleue, saignante, à point ou bien cuite ?*). L'isotopie est mésogénérique quand le sème est mésogénérique, c'est-à-dire qu'il indexe des sémèmes appartenant au même domaine (c'est le cas du sème /navigation/ dans les sémèmes 'amiral', 'carguer', et 'voiles' de l'énoncé *L'amiral Nelson ordonna de carguer les voiles*). Enfin, l'isotopie est macrogénérique quand le sème récurrent est macrogénérique, c'est-à-dire qu'il indexe des sémèmes appartenant à la même dimension (c'est le cas de /animé/ dans les sémèmes 'hérisson' et 'porc-épic' de *Le hérisson insectivore n'est pas de la même famille que le porc-épic*). Les

55 Les exemples que nous allons donner pour éclairer les isotopies génériques et spécifiques sont extraits de Sémantique Interprétative ([Rastier 87, p. 111-113]).

isotopies génériques sont liées à des paradigmes codifiés en langue. Elle induisent les impressions référentielles. Le coloriage thématique est une mise en évidence des isotopies génériques.

- Isotopie mixte : une isotopie est mixte quand elle n'est ni purement générique, ni purement spécifique tout le long de la chaîne, c'est-à-dire quand le sème récurrent est générique dans certains sémèmes de la chaîne et spécifique dans d'autres. C'est le cas de l'isotopie du sème /qui vole/ dans l'énoncé *Le Boeing 747 est un bon avion*, qui est générique dans le sémème 'Boeing 747' et spécifique dans le sémème 'avion'.

Plus qu'un constat, l'isotopie est à la base du processus d'interprétation. C'est ce qui lui confère son intérêt dans une démarche de TAL. En effet, interpréter un énoncé consiste à mettre en évidence une ou plusieurs isotopies (un faisceau d'isotopies constituant un *fond sémantique*). Il y a donc, dans le processus d'interprétation d'un énoncé, un préalable qui consiste en une présomption d'isotopie :

> *En général, on considère l'isotopie comme une forme remarquable de combinatoire sémique, un effet de la combinaison des sèmes. Ici au contraire, où l'on procède paradoxalement à partir du texte pour aller vers ses éléments, l'isotopie apparaît comme un principe régulateur fondamental. Ce n'est pas la récurrence de sèmes déjà donnés qui constitue l'isotopie, mais à l'inverse la présomption d'isotopie qui permet d'actualiser des sèmes, voire les sèmes. (Rastier 1987, p. 11)*

La présomption d'isotopie indique une recherche de cohérence (dans le but d'établir des fonds sémantiques) au cœur du processus interprétatif. C'est particulièrement observable dans le cas des messages obscurs. L'interprétation de messages énigmatiques fait appel à l'isotopie car l'enjeu dans la résolution de l'énigme est de reconstruire une isotopie qui n'apparaît pas lors d'une première interprétation de la chaîne (Mauger 1999). La présomption d'isotopie est ce grâce à quoi l'interprétant entre dans le jeu de l'énigme, en cherchant une isotopie cachée. Les textes non obscurs sont interprétés de la même façon et « fonctionnent » à la façon d'énigmes non énigmatiques (de mêmes que les tropes mettent en évidence les principes de l'interprétation non spécifiques aux tropes).

Chercher à établir une isotopie en concrétisant la présomption d'isotopie, c'est chercher à actualiser des sèmes en contexte (et par conséquence en virtualiser d'autres). C'est notamment ce que nous avons pu étudier dans nos travaux relatifs à la métaphore. En plus des actualisations et virtualisations de sèmes inhérents, l'établissement d'une isotopie peut se faire par afférence d'un sème nouveau au niveau d'une ou plusieurs lexies. C'est l'afférence cotextuelle ou socialement normée :

- L'afférence cotextuelle exprime l'enrichissement d'une lexie par un sème déjà redondant dans la zone de cotexte. C'est par exemple le cas dans le titre du livre de G. Lakoff *Women, Fire and Dangerous Things : What Categories Reveal about the Mind (1987)* où il y a afférence du sème /dangereux/ (co-occurrent dans *Fire* et *Dangerous Things*) au lexème *Women*.
- L'afférence socialement normée est alors le fait d'une norme sociale partagée au sein d'une communauté linguistique, un topos selon (Raccah 1997). C'est, par exemple, le cas du sème /tristesse/ afférent au sémème 'noir' dans *il broie du noir* ou encore le cas du sème /bonheur/ dans 'rose' dans *la vie en rose*.

L'afférence (et à travers l'afférence, l'isotopie) permet d'expliquer des opérations interprétatives fines comme les assimilations (actualisation d'un sème par présomption d'isotopie) et les dissimilations (actualisation de sèmes afférents opposés dans deux occurrences du même sémème, ou dans deux sémèmes parasynonymes ; par exemple *il y a musique et musique*). Plus généralement les concepts de

diffusion et de sommation (Rastier 2006, pp. 99-114), rendent compte des échanges sémiques entre fonds et formes (la diffusion est définie comme une transformation de formes en fonds et la sommation comme une transformation de fonds en formes).

Ces opérations interprétatives n'ont pas encore fait l'objet de mise en œuvre calculatoire. C'est un objectif à poursuivre dans le cadre d'une approche centrée-utilisateur où la description lexicale et l'analyse de corpus s'alimentent mutuellement dans la tâche de l'utilisateur. Nous suivons en cela les travaux de (Valette 2004) indiquant qu'il y a ici des perceptives pour des descriptions dynamiques du sens.

4.3.2. Ancrage épistémologique

Si les notions définies par la SI pour la description des opérations interprétatives s'avèrent tout à fait compatibles avec une approche centrée-utilisateur en TAL, les positionnements épistémologiques de la SI le sont également. C'est même à nos yeux presque plus important encore tellement les approches théoriques du TAL, approches dites logico-grammaticales, sont à l'opposé de ces positionnements et que nous cherchons justement à nous en démarquer. Selon F. Rastier, le cognitivisme constitue l'aboutissement contemporain de la tradition logique et grammaticale. La SI s'y oppose radicalement en se rapportant à la tradition rhétorique et/ou herméneutique qui, quant-à-elle, prend pour objet d'étude les textes et les discours dans leur production et leur interprétation. Il en découle une visée non réductionniste de la sémantique des langues.

Voici les points fondamentaux de l'ancrage épistémologique que nous héritons volontiers de la SI :

- L'interprétation est une perception sémantique individuelle selon Rastier. En tant que perception du sujet elle fait preuve d'une grande variabilité d'un sujet à un autre, variabilité éventuellement individuelle : il nous est tous arrivé de relire différemment un texte et d'en tirer des interprétations bien différentes. C'est que qu'explique Théodore Thlivitis dans son concept d'anagnose (Thlivitis 1998), tentative d'objectivation des rapports intertextuels des différentes interprétations du passé d'interprétant d'un individu qui conditionnent les interprétations à venir. Il en découle que l'interprétation des textes, comme des images d'ailleurs, est située : elle est liée à l'individu, à ses pratiques et donc, plus généralement, fortement contextualisée et culturellement déterminée. Dès lors qu'on parle de « vrais » textes, de « vrais » corpus, et pas simplement de phrases d'exemples artificiellement construites en dehors d'un contexte linguistique et pragmatique, la dimension perceptive personnelle fait évidence. Le sens ne repose pas dans le texte, mais dans les conditions d'interprétation. Il en résulte, à notre avis, que le sens ne peut être modélisé à la façon d'un résultat calculatoire qui serait plus ou moins complété ou dégradé d'un interprétant à un autre. Le sens n'est pas de nature symbolique ; c'est un processus sémiotique au centre de l'activité de l'interprétant. Ce processus sémiotique n'est pas un calcul déterministe prédictible, c'est une boucle perception-action. Serge Mauger (Mauger 2007), plagiant une formule bien connue de Malraux à propos de la culture, en tire l'assertion suivante : *« Le sens est ce qui reste quand on a tout oublié des formes ou des supports sémiques».* Nous le suivons volontiers dans la conséquence d'une impossibilité à formaliser une représentation du sens.
- Dans une filiation de la sémiotique de Saussure et des travaux de Hjelmslev (Hjelmslev 1943), il n'y a pas pour la SI de primat ontologique dans la signification. Des critères sociaux, culturels et praxéologiques sont souvent plus pertinents que des critères ontologiques. F. Rastier (Rastier, 1987) rapporte ainsi le résultat d'une étude sur le contenu du mot *caviar* suite à une enquête au

sein d'une population de collégiens. Il apparaît que le trait /luxueux/ est le plus fréquemment cité tandis que d'autres traits ontologiques tels que /texture granuleuse/ ou encore /salé/ ne sont jamais évoqués. La SI établit un rapport étroit entre la sémiotique des langues et les sciences de la culture et c'est à notre sens un éclairage important (qui manque notamment de manière cruciale aux travaux du type du Web Sémantique).

- L'établissement du sens consiste en des parcours interprétatifs que les alternances entre des fonds et des formes déterminent. Cela induit qu'il n'y a pas de sens sans texte (on pourrait également dire que par application de la présomption d'isotopie, il n'y a pas de texte sans sens[56]). C'est-à-dire pas de forme extra-linguistique du sens, contrairement aux logiques formelles. Beaucoup de travaux en sémantique formelle (logique, DRT , SDRT, etc.) ont depuis des années déployé beaucoup d'efforts et de moyens pour obtenir un « calcul du sens » acceptable, cherchant ainsi à remplacer le texte par une forme arbitrairement vue comme plus épurée. Force est de constater qu'un tel résultat reste hors de portée. Il ne s'agit pas ici que d'un problème d'évaluation dont on n'aurait pas encore bien mis en place la méthodologie mais d'un problème beaucoup plus profond : les alternances entre les fonds sémantiques et les formes textuelles sont constitutives de l'espace interprétatif. C'est vrai du point de vue de l'interprétation, comme celui de la parole qui n'est pas préméditée (elle se détermine en cours de production en réaction à elle-même comme l'a montré Jacques Coursil). C'est également, à sa façon, le cas de la production écrite, comme en témoigne les phénomènes de réécriture et de raturage en cours de production écrite et plus généralement, en témoigne aussi les phénomènes de reconstruction et de recomposition de textes antérieurs (Mayaffre 2002) actualisés dans de nouveaux corpus. Contrairement aux sémantiques formelles cherchant à résumer la donnée textuelle en une expression logique épurée, interpréter un texte, ce n'est donc surtout pas se passer du texte[57]. Au contraire, c'est bien plus y un projeter son anagnose et ainsi démultiplier les possibilités de relations intertextuelles.
- La SI va à l'encontre des approches compositionnelles en affirmant un principe de détermination du global sur le local. L'interprétation et l'établissement de parcours interprétatifs, procèdent principalement par contextualisation (et non par représentation du contexte). Elle rapporte le passage considéré à son voisinage, selon des zones de localité (syntagme, période) de taille croissante allant jusqu'au corpus dans sa globalité. C'est le rapprochement d'un passage de texte (éventuellement réduite à un signe) dans les zones de localité du co-texte proche ou de l'intertexte qui est générateur d'opérations interprétatives analysées en termes d'échanges et de récurrences de sèmes. En étendant ce principe de contextualisation aux ensembles documentaires et aux environnements numériques de travail, on peut tout à fait imaginer, par

56 On trouve dans (Bassi Acuña 1995, p. 122) une interprétation de l'énoncé sans cohésion apparente *D'incolores idées vertes dorment furieusement* qui conduit à former la paraphrase *Un vague espoir s'agite dans le subconscient* (incolore est ici interprété comme mal défini ou vague, vert comme espoir suivant la maxime selon laquelle le vert est la couleur de l'espérance, dormir comme faisant référence au subconscient, et furieusement comme ayant rapport à une agitation). Cet énoncé est à l'origine une traduction française de l'énoncé *Colourless green ideas sleep furiously* construit par Chomsky comme un exemple d'énoncé logiquement indéterminable.

57 Dans le vers *La terre est bleue comme une orange* (Premier vers du 7ème poème du premier chapitre "Premièrement" composant le recueil "L'amour la poésie", Paul Eluard, 1929) une interprétation formelle sous forme d'analogie conclurait à un « non-sens logique » (quelque chose que nous pourrions paraphraser de la sorte : *la terre est bleue comme l'orange est bleue !*). De toute évidence l'interprétation est très loin du non-sens et peut notamment s'expliquer linguistiquement comme une alternance fond-forme induite par une actualisation dans l'interprétation des lexies terre et orange d'un sème commun /sphérique/ (cf. http://photoblog.ludopics.com/index.php?showimage=14 pour une interprétation plus visuelle, consultée le 7/1/13).

exemple, qu'un article de presse présenté dans une source documentaire très sérieuse et classique ne sera pas perçu par l'interprétant de la même manière que le même article présenté dans une source plutôt satirique. Le contexte global instauré par les autres articles aiguillant forcément les parcours interprétatifs. Les textes qui composent ensemble un corpus influent les uns sur les autres, comme c'est également vrai des textes qui sont consultés avant, après ou plus ou moins en même temps que le texte en question (d'où une prise en compte nécessaire de l'activité de l'interprétant). Ce rapport du texte à la collection n'est pas seulement une détermination de la collection à partir des textes qui la composent. La globalité de la collection agit sur les conditions d'interprétation d'un document. C'est une détermination du global sur le local qui est tout aussi importante que la détermination inverse. La contextualisation émerge d'une activité de rapprochements des textes les uns relativement aux autres sans avoir dû en passer par une représentation du contexte. C'est exactement ce genre de détermination du global sur le local que nous avons mis en œuvre dans le projet ProxiDocs et dans ceux qui ont découlé.

Relativement à ce positionnement épistémologique de la SI, la question de la mise en œuvre d'une assistance logicielle à l'interprétation dans une approche centrée-utilisateur doit consister en une interaction entre le sujet et les textes. Le sens prend forme dans l'interaction entre un sujet, un corpus, des ressources lexicales et plus largement une culture. Cette interaction complexe dessine un espace dans lequel le sujet interprétant fait émerger du sens et tire profit d'une instrumentation informatique. C'est ce que nous appelons l'espace interprétatif du sujet et qu'il convient de médiatiser.

4.3.3. L'espace interprétatif

La SI nous permet de caractériser les dimensions de cet espace interprétatif du sujet interprétant qui, pour nous, est l'utilisateur d'une instrumentation informatique. Cet espace est structuré par des niveaux de textualité où s'opère la détermination du global sur le local et allant du signe (en tant que passage de texte) à l'anagnose (en tant que corpus de l'histoire interprétative du sujet). L'espace est également structuré de manière complémentaire par des conditions d'interprétation directement en lien avec l'activité de l'interprétant. François Rastier les décrit par une praxéologie (Rastier 2002), i.e une théorie de l'action dans et par le langage. Cette praxéologie situe les phénomènes interprétatifs dans un continuum allant de l'individualité à la culture. Elle nous enseigne qu'une approche centrée-utilisateur ne peut absolument pas être une approche de l'individu en tant qu'être isolé et qu'elle a forcement à intégrer la société au travers du prisme des actions de l'utilisateur.

4.3.3.1. Les niveaux de textualité

Les niveaux de textualité sont des outils analytiques de la SI pour caractériser la dynamique sémique à l'œuvre dans la détermination du local par le global. François Rastier décrit trois principes relevant de trois échelles de textualité :

- le principe de contextualité : deux signes ou deux passages d'un même texte mis côte à côte sélectionnent réciproquement des éléments de signification (sèmes). Cette sélection réciproque s'analyse notamment en termes de récurrences de sèmes appelées isotopies. La portée dans le texte de l'isotopie détermine une zone de localité sémantique renforcée par l'étendue dans le discours des relations de prédication et d'anaphore. Cette zone, appelée Période ([Rastier & al. 94, p. 116]), définit le champ d'étude de la mésosémantique.
- le principe d'intertextualité : deux passages de textes différents, si brefs soient-ils, et fussent-ils réduits à la dimension d'un signe, sélectionnent réciproquement dès qu'ils sont mis côte à côte,

des éléments de signification (sèmes). Le principe d'intertextualité généralise la notion d'isotopie au-delà de la linéarité syntagmatique du texte;

- le principe d'architextualité : Tout texte placé dans un corpus en reçoit des déterminations sémantiques, et modifie potentiellement le sens de chacun des textes qui le composent.

Ces 3 échelles sont des principes bien plus que des paliers, c'est pourquoi il ne faut pas entendre qu'il faille absolument partir du « barreau » de dessous pour aller au « barreau de dessus ». Le principe de contextualité est bien à l'œuvre dans les deux autres principes mais il n'est pas pour autant une étape préalable. L'interprétation d'un document au sein de sa collection est le produit des trois échelles. Une métaphore en terme d'échelle plus appropriée serait en une sorte d'échelle circulaire à deux sens où le dernier barreau permettrait de monter au premier ainsi que de descendre du premier barreau au dernier (une sorte d'escalier à la Escher). Ce qu'illustre cette métaphore c'est que l'appréhension des rapports « global – local » dans l'interprétation procède à tous les niveaux de contextualité, d'intertertextualité et d'architextualité d'alternances entre les formes et les fonds. Ainsi une forme n'a pas de fond tant qu'il n'y a pas interprétation et un fond doit nécessairement s'actualiser sur une ou plusieurs formes. L'instrumentation logicielle que nous visons doit donc médiatiser cette alternance en considérant notre échelle circulaire comme un cercle vertueux où se déploient les richesses d'interprétation du lecteur.

4.3.3.2. Les niveaux de contexte d'interprétation

Le sens n'est pas de nature symbolique, mais dynamique. Un texte de même qu'un signe n'a pas de sens en tant que tel mais, fait sens pour un interprétant. On retrouve ici la triade sémiotique de Pierce, sur laquelle (Morand 2004) établit une méthode pour la conception en informatique. Pour Rastier, la dynamique du sens s'inscrit dans ce rapport à l'interprétant et dans le contexte des tâches qu'il mène, son contexte praxéologique. Cette praxéologie place l'interprétation dans un espace fait de trois zones autour du sujet :

- La zone identitaire : c'est la dimension personnelle et subjective de l'interprétation consciente ou inconsciente. C'est là que s'exprime l'individualité, les points de vue particuliers. C'est la zone marquée dans le langage de la première personne verbale (je, nous) dans l'ici et le maintenant.
- La zone proximale : elle représente l'environnement linguistique immédiatement proche du sujet. C'est là que se trouve la communication avec l'autre. Dans les approches interactionnistes en psychologie (Brassac & al. 1996), le contexte de cette zone se ramène à la notion de terrain commun mis en évidence par (Clark & al. 1986, Vivier 1992). Le terrain commun permet de borner le contexte, évitant ainsi le recours à une description du monde dans son ensemble. C'est la zone linguistiquement marquée par la deuxième personne verbale (tu, vous), par la fonction phatique[58], par le passé proche et futur immédiatement probable.
- La zone distale : elle représente les normes et ancrages culturels du sujet. Les deux premières zones sont celles des expériences propres du sujet. La zone distale est celle de ce qui est étranger mais connu. C'est là qu'entre en jeu dans l'interprétation toute la dimension sociale et culturelle du langage. C'est linguistiquement la zone de la troisième personne verbale, du récit, du non-présent et de l'irréel.

Ces trois zones ou, niveaux praxéologiques des sujets, éclairent les apports de la dimension

58 La fonction phatique (Jacobson 1981) du langage vise à maintenir le contact entre le locuteur et l'allocutaire (exemples : « Allô ? », « Vous m'entendez ? » …).

subjective aux phénomènes interprétatifs. Considérer que l'interprétation ne fait intervenir que des connaissances ontologiques c'est ne voir que la troisième zone (qui plus est, d'un point de vue particulier) et passer sous silence tout ce qui relève des réalités empiriques du sujet, à savoir les deux premières zones. Il est donc fondamental dans un environnement d'assistance à l'interprétation que les trois zones puissent être utilement mises en œuvre. Cela a pour conséquence que les ressources qui guident les processus calculatoires qui interviennent dans l'interaction (extractions statistiques, visualisations, etc.) doivent être très massivement personnalisables pour donner toute l'importance qu'il convient aux deux premières zones. Au sein de la tâche du sujet, c'est par un travail d'aller-retours entre des corpus et des ressources termino-ontologiques personnelles que l'environnement logiciel peut amener à caractériser des formes sémantiques sous l'angle de moments stabilisés de parcours interprétatifs.

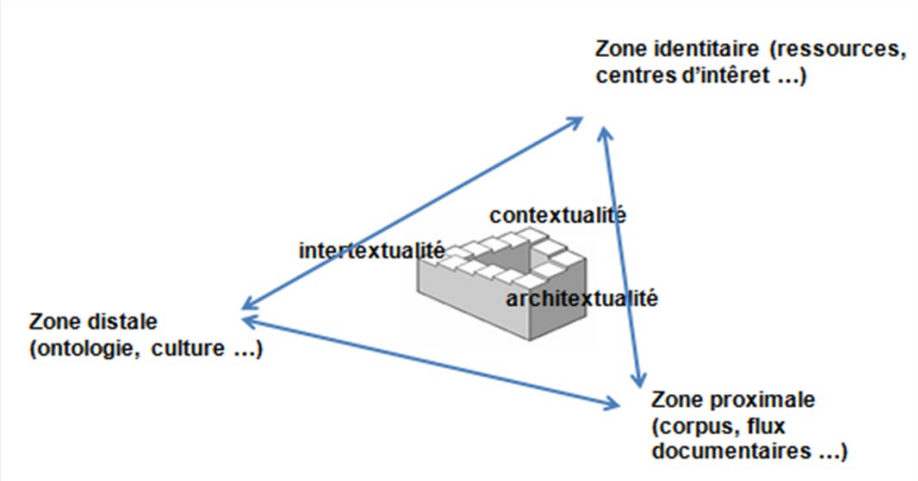


Illustration 39: L'espace interprétatif : une double triade

Au final l'espace interprétatif s'apparente à une double triade où zones de contextualité et échelles de textualité se répondent mutuellement dans la perception sémantique du sujet (cf. illustration 39).

Aucun des trois pôles de chacune des triades ne peut être éludé. Le sens émerge et se déploie dans cet espace au complet. Ceci a pour conséquence qu'on ne peut pas se passer des textes et qu'on ne peut pas non plus se passer de la subjectivité. En somme, on ne peut pas confier à autrui (humain ou machine) sa propre expérience interprétative. Comprendre un texte, interpréter un message c'est donc d'une certaine façon **s'y retrouver**, replacer sa subjectivité et sa culture dans l'espace interprétatif. C'est bien ce qu'indique également Serge Mauger dans la phrase suivante :

> *Dans l'interprétation dialogique et concertée, il y a un individu-sujet qui « s'y retrouve », en tant qu'il se restitue à lui-même par réinscription dans le social. (Mauger 2009, p. 48).*

Comprendre un texte, c'est aussi **le comprendre autrement** parce que l'on est jamais dans les mêmes conditions d'interprétation (dans le même espace) qu'un autre, ni même de soi-même à un autre moment de son histoire interprétative. L'interprétation n'est pas réplicable à l'identique.

C'est pour toutes ces raisons que nous nous tournons vers la SI pour trouver un cadre linguistique compatible avec une approche centrée-utilisateur en TAL. Implémenter la SI dans son ensemble en

TAL serait un enjeu complexe, certainement trop complexe car il n'est pas du tout garanti qu'une théorie de l'interprétation puisse être automatisée, sinon s'agirait-il encore d'une véritable interprétation ? De toute façon, là n'est pas forcément le but à poursuivre et la SI se veut avant tout descriptive, pas implémentable. La SI est un cadre de pensée dans lequel on peut chercher à se focaliser sur tel ou tel aspect d'un point de vue opératoire : simuler des opérations interprétatives, mettre en évidence des rapprochements de textes pour aborder la détermination globale-locale, développer des jeux de langages où textes et cultures se croisent (c'est notamment ce qu'à fait Serge Mauger sur les chaînes sémiques dans les charades[59]). Dans tous les cas, c'est l'espace interprétatif qui est médiatisé et la pluralité de ses possibilités de modélisation opératoire est plus une chance qu'un problème. La possibilité de médiatisation de l'espace interprétatif du sujet que nous avons entreprise jusqu'à présent est de permettre à l'utilisateur de tirer profit d'une interaction homme-machine où on lui suggère par des analyses statistiques sur ses corpus (textes, emails, etc.) et ses ressources des visualisations stimulant ses facultés d'oppositions, de rapprochements et de regroupements. En somme, le but est d'installer un couplage personne-système incitant à déployer des facultés interprétatives.

~

Notre approche centrée-utilisateur en TAL trouve avec la SI un appui théorique et épistémologique fort. Le rapport de l'utilisateur aux autres et à la société est un point important qui place bien l'approche centrée-utilisateur dans un environnement culturel et linguistique. En somme, il en découle que le caractère centré-utilisateur n'est surtout pas synonyme d'une approche de l'utilisateur isolé. L'utilisateur n'est jamais seul et c'est même bien souvent un petit groupe d'utilisateurs partageant et développant ensemble des ressources dont il est en fait question. En multipliant les utilisateurs à plus grande échelle on glisse du « centré-utilisateur » au « socio-centré » mais sans remettre en cause l'ancrage aux sciences de la culture et au langage. C'est bien l'enseignement qu'on tire du mouvement du Web 2.0. En donnant aux utilisateurs une liberté de production de contenus, d'annotation et de création de ressources, on produit des services où la multiplicité des subjectivités est une richesse. La folksonomie nous montre que dans cette liberté les utilisateurs ont réinventé avec la notion de tag celle de sème. C'est aussi un argument pour revisiter la SI (et plus généralement la sémantique componentielle) sous l'angle d'un modèle explicatif des phénomènes linguistiques portés par les environnements numériques.

La SI nous invite à voir le sens comme une perception (d'ordre sémantique). En tant que perception, elle est individuelle et située. Elle est aussi, de manière indissociable, liée à l'action du sujet sur son environnement. C'est la problématique de la boucle perception-action qui est bien connue des sciences cognitives. En complément de la SI un « détour » vers les sciences de la cognition s'impose pour assoir le projet scientifique d'une approche centrée-utilisateur. C'est ce que nous allons faire dans le cadre du chapitre suivant en nous rapportant au courant le l'*énaction*.

59 (Mauger 1999), (Mauger & Luquet 2005)

5. Vers des ENT énactifs

Nous avons argumenté dans les chapitres précédents pour qu'une approche centrée-utilisateur en TAL procède continuellement par des allers-retours ; des allers-retours corpus ↔ ressources par exemple ou encore, relativement au sens, des allers-retours fonds ↔ formes. C'est un point de vue interactionniste sur le TAL où l'on estime que les traitements alimentent des boucles interactives qui ne sont pas nécessairement finalisées dans le temps et qui demeurent tant que l'utilisateur les fait se prolonger. Au sein d'une même interaction le rapport entre plusieurs traitements pose la question de leur interopérabilité. Classiquement en TAL, l'interopérabilité est vue sous le prisme des chaînes de traitement (c'est le cas par exemple des plateformes de TAL du type LinguaStream[60]) : les traitements ont des entrées et des sorties qui doivent être compatibles pour viser le calcul d'un résultat global. Les allers-retours de l'approche centrée-utilisateur donnent conceptuellement une autre idée de l'interopérabilité. A tout moment, tout traitement doit pouvoir être mobilisé pour apporter une information supplémentaire dans l'environnement. C'est-à-dire que c'est l'utilisateur à travers son Environnement Numérique de Travail (ENT) qui demande une interopérabilité et ce n'est pas un calcul finalisé qui organise l'interopérabilité. L'objectif est donc bien de s'intéresser aux applications du TAL qui peuvent être prévues pour s'intégrer à la demande dans des boucles interactives et faire que ce soit l'ENT qui détermine l'organisation des traitements, et pas le contraire. En qui concerne le TAL, c'est un constat que partage notamment Adeline Nazarenko en conclusion de son HDR :

> *J'ai souligné les limites du tout automatique à plusieurs reprises. Il faut donc à la fois cerner au mieux le champ de l'automatisable et inventer des processus coopératifs, sans perdre l'utilisateur de vue. (Nazarenko 2004, p. 88)*

Ainsi l'ENT permet à l'utilisateur une participation active en lui donnant la possibilité de se

60 cf. http://www.linguastream.org/ consultée le 7/01/13.

construire une perception de ses corpus, de ses ressources, de sa tâche et au final de lui-même. En retour, cette perception est celle qui organise l'interaction avec l'environnement. Nous sommes ici dans le cas d'un environnement qui se construit dynamiquement dans un rapport étroit entre la perception et l'action. En sciences cognitives, une théorie se pose justement la question de la boucle action-perception au sein de l'environnement du sujet (environnement qui pour le coup n'est pas forcément numérique). C'est l'énaction.

Le but de ce chapitre est de décrire en quoi les problématiques des ENT et de l'énaction se rejoignent dans le projet d'une approche centrée-utilisateur du TAL. Nous expliquerons comment nous abordons ce rapprochement avec les collègues du groupe NU (Nouveaux Usages). Nous verrons également les conséquences que cela a sur des questions très actuelles comme celles de l'évaluation.

5.1. Environnements Numériques de Travail (ENT)

L'acronyme ENT désigne en fait deux expressions différentes *Environnements Numériques de Travail* ou *Espaces Numériques de Travail*. Les deux sont utilisées le plus souvent comme des synonymes, mais en ce qui nous concerne la notion d'environnement indique, beaucoup mieux que celle d'espace, là où s'opère la boucle perception action. Nous utiliserons donc ENT dans son acceptation *Environnements Numériques de Travail.*

La problématique des ENT provient essentiellement du monde de l'éducation. Elle a été initiée par des initiatives dans le secondaire pilotées par le ministère de l'éducation nationale comme notamment le cahier de texte en ligne[61] (premières expérimentations dans l'académie de Aix-Marseille à la rentrée scolaire 2009). Depuis elle est massivement reprise dans les milieux de l'enseignement supérieur pour offrir aux communautés d'utilisateurs (avec différents profils d'utilisateurs) des services numériques personnalisés. L'ENT doit, en outre, pouvoir favoriser la mutualisation des services et ressources au niveau inter universitaire (UNR, UNT, ...), avec des partenaires publics et privés, en France, en Europe ou au niveau international.

L' ENT est un dispositif informatique global permettant à un usager d'un établissement (étudiant, enseignant, personnel technique et administratif) d'accéder à l'ensemble des ressources et des services numériques en fonction de son profil et en rapport avec son activité. Par exemple, c'est le cas pour tous les personnels des messageries électroniques et c'est aussi le cas de manière spécifique aux étudiants aux emplois du temps et convocations aux examens. L'ENT, accessible depuis n'importe quelle connexion internet, offre à l'usager un point d'entrée unifié et sécurisé, avec une authentification unique, qui structure et présente de façon cohérente les services du système d'information[62] (SI) de l'établissement. En masquant la complexité technique, l'ENT permet aux utilisateurs de se concentrer sur les apports des TIC pour la pédagogie et l'organisation de l'établissement d'enseignement.

L'ENT s'inscrit dans un cadre de cohérence global défini par le ministère de la Jeunesse, de l'Éducation nationale et de la Recherche dans le cadre du SDET (Schéma Directeur des Espaces Numériques de Travail). Le SDET[63] vise à guider les démarches des collectivités, des industriels, des

61 http://www.educnet.education.fr/veille-education-numerique/septembre-2010/cahier-de-textes-numerique consultée le 7/01/13.

62 Ensemble des éléments participant à la gestion, au stockage, au traitement, au transport et à la diffusion de l'information au sein d'une organisation.

63 Ce schéma fournit donc un ensemble de préconisations fonctionnelles, organisationnelles et technologiques

académies et des universités pour fournir un cadre de cohérence aux nouveaux services numériques (intranet, mail, forums, publication de site Web, cours en ligne, etc.), qui viennent s'ajouter aux services « traditionnels » rendus par l'informatique de l'établissement. À la différence du SI qui structure les informations stratégiques de l'établissement indépendamment des usages, l'ENT permet notamment leur consultation de manière sécurisée à travers un environnement que l'utilisateur peut personnaliser.

D'un point de vue technique, l'ENT est composé d'un socle (par exemple Esup-Portail[64]) et de services numériques. Le socle de l'ENT est chargé d'orchestrer les services numériques, de les présenter de manière structurée et cohérente, et fournit à ces derniers un certain nombre de fonctionnalités communes de bas niveau (annuaire, identification et authentification des usagers, personnalisation des services offerts, etc.). Les services offerts classiquement sont :

- les services de communication (courrier électronique, forum de discussion, listes de diffusion, annuaire, chat, visioconférence)
- les services du bureau numérique (carnet d'adresses, espace de travail et de stockage, agenda, publication Web, gestion de signets, outils de bureautique)
- les services de scolarité, d'orientation
- les services documentaires
- les services « vie universitaire »

D'un point de vue organisationnel, l'ENT permet de passer d'une « culture métier » (verticalité des applicatifs) à une « culture de l'usager » (transversalité et complémentarité) : l'usager est placé au cœur d'un système personnalisé lui présentant toutes les ressources auxquelles il doit pouvoir accéder.

L'accès à des cours et activités pédagogiques en ligne sont souvent une des principales catégories de service à l'usager, principalement aux étudiants et enseignants. Ces services sont mis en œuvre dans le cadre de plateformes de gestion de contenus spécifiques aux activités pédagogiques (des LMS, *Learning Management System* ou encore VLE, *Virtual Learning Environment*). Ces plateformes prévues pour être intégrées dans les ENT sont aussi en soi des ENT dédiés dans le sens où l'accès aux ressources pédagogiques et aux moyens de communication (forum, chat, webconférence …) et des fonctionnalités de travail collaboratif sont aussi fournis de manière personnalisée sur authentification des utilisateurs. Ces ENT médiatisent l'espace pédagogique et instituant les rôles et identités de chacun des utilisateurs (enseignants, tuteurs, étudiants) et en organisant les actes d'enseignement et les actes d'apprentissage (Cf. la conférence en vidéo de Jean-Claude Bertin[65] aux Universités Vivaldi tenues à Caen en mai 2011).

---

aux acteurs impliqués dans le déploiement de ces plates-formes et services en ligne : établissements, rectorats, fournisseurs de plates-formes, de services, de contenus, collectivités territoriales. Document évolutif, le SDET est destiné à devenir un instrument de dialogue pour l'Education nationale et ses partenaires autour de la formalisation du concept ENT (cf. http://www.educnet.education.fr/services/ent/sdet consultée le 7/01/13).

64 http://fr.wikipedia.org/wiki/Esup-Portail consultée le 7/01/13.

65 http://www.canal-u.tv/video/centre_d_enseignement_multimedia_universitaire_c_e_m_u/dispositifs_acteurs_accompagnement_une_approche_par_la_modelisation.8746 consultée le 15/12/12

Une tendance actuelle est de développer conjointement aux plateformes LMS des environnements PLE (*Personnal Learning Environnement*). Ces PLE, aussi appelés « e-portfolios », sont des systèmes qui permettent à un utilisateur de gérer un espace personnel en ligne de dépôt des traces de son apprentissage, d'exposition de ses compétences et de partage de ses productions. L'objectif visé est multiple : inciter encore plus au travail collaboratif, rendre les utilisateurs réellement détenteurs de leurs contenus (ce qui n'est pas le cas dans l'usage de réseaux sociaux type Facebook par exemple) et donner des moyens numériques pour élaborer une stratégie d'insertion professionnelle.

Aujourd'hui une des plus utilisées (par environ 75% des universités françaises) des plateformes LMS est celle du projet Moodle[66] (*Modular Object-Oriented Dynamic Learning Environment*). En juin 2009, les chiffres officiels[67] communiqués par le consortium Moodle étaient :

- Sites enregistrés : 56 715 plateformes en utilisation
- Nombre de pays : 210
- Cours : 3 047 685 d'espaces de cours en ligne
- Utilisateurs[68] enregistrés sur les plateformes : 32 815 756

A l'échelle de l'université de Caen Basse-Normandie, le CEMU (que je dirige) a mis en place en juin 2009 pour toute l'université une plateforme Moodle devant devenir la brique « pédagogie en ligne » du futur ENT de l'établissement. En décembre 2010, la plateforme comptait 15 805 utilisateurs (dont 848 enseignants et 14 957 étudiants sur les environ 24 000 de l'établissement), elle regroupait environ 1 300 cours et comptait des journées avec des pics à plus de 4 500 connexions par jour. Depuis l'année universitaire 2011-2012, la plateforme Moodle de l'établissement est couplée à une plateforme de e-portfolio (plateforme libre de type Mahara[69]).

Le constat d'un tel développement de la plateforme en à peine plus d'un an, montre le succès rencontré par le déploiement des ENT. Cela indique bien plusieurs choses :

- proposer un environnement intégré où l'utilisateur trouve l'ensemble des tâches qu'il a à gérer est exactement l'attente des utilisateurs,
- proposer des environnements personnalisables à l'extrême est aussi une attente forte des utilisateurs,
- proposer un environnement de travail, c'est proposer à l'utilisateur un outil de collaboration avec les autres (partageant la même institution).

Dans ces trois cas, l'attente des utilisateurs est de se retrouver, individuellement et socialement, dans les outils qu'on leur propose. Quand c'est le cas, le couplage personne – système se crée. Une autre raison de la mise en place du couplage est que les ENT sont des systèmes actifs dans le sens où l'utilisateur interagit dans et par le système. Il y construit des ressources (par exemple des supports de

66 http://moodle.org/ consultée le 7/01/13.

67 http://www.elearningcyclops.com/2009/06/how-big-is-moodle.html consultée le 7/01/13.

68 Le nombre d'utilisateurs est toujours un indicateur intéressant pour le choix d'une plateforme, même s'il est loin d'être le seul. Un comparatif de 4 LMS open source (Moodle, Sakai, Ganesha, Claroline) propose une série de critères pertinents. cf. http://www.projet-plume.org/files/Choix_plateforme_a2l.pdf consultée le 7/01/13.

69 https://mahara.org/ consultée le 21/11/12.

cours) qui permettent d'alimenter d'autres ressources (par exemples des activités d'évaluation en ligne). L'utilisateur n'est pas uniquement dans une situation de réception d'information mais il est actif par son action dans l'environnement. Là aussi, il s'agit d'un couplage, c'est le couplage action perception que décrit la théorie de l'énaction.

5.2. L'énaction

La théorie de l'énaction (ou cognition incarnée) a été proposée dans les années 1980 par le biologiste, neurologue et philosophe chilien Francesco Varela (Varela 1996). Varela propose le concept d'énaction pour appréhender l'action adaptative de tout organisme vivant à son environnement. En réaction au cognitivisme, l'énaction fait prédominer l'action et la boucle action-perception à la réception passive des formes perçues, des connaissances et des normes ou valeurs sociales. Ainsi la cognition, du point de vue de l'énaction et dans une perspective constructiviste, n'est pas la représentation d'un monde pré-donné mais l'avènement conjoint d'un monde et d'un être vivant à travers l'histoire des diverses actions et interactions accomplies dans le monde par cet être. C'est la négation même de l'idée d'un matériau ontologique du monde préexistant à l'action.

L'énaction est au départ une théorie du vivant construite à partir de la propriété d'*autopoièse* (Maturana & Varela 1973) décrivant la cellule biologique. L'autopoièse est la propriété d'un système à se produire lui-même, à se maintenir dans son milieu, à se définir lui-même.

Non cantonnée au domaine du vivant, l'énaction est une théorie qui peut être réinvestie dans d'autres champs de recherche. Par exemple en systémique, il découle de l'énaction qu'un système autonome n'est jamais un système isolé car il est nécessairement couplé à un milieu ambiant et que dans ce milieu il maintient et fait évoluer continuellement son organisation en dépit des perturbations occasionnées par le milieu. Au sein des sciences cognitives l'énaction constitue un nouveau paradigme scientifique par rapport au cognitivisme. Là où le cognitivisme prône un modèle de la cognition inspiré par la métaphore de l'ordinateur, l'énaction ramène la cognition dans un modèle inféré des organismes vivants, ce qui paraît indéniablement plus cohérent.

Dans une inspiration phénoménologique au sens de Husserl (Husserl 1913) et Merleau-Ponty (Merleau-Ponty 1945), l'énaction met l'accent sur l'expérience en tant qu'intuition sensible des phénomènes pour en décrire l'essence. Ainsi, la connaissance serait le résultat d'une interprétation permanente qui émerge des capacités de compréhension, elles-mêmes enracinées dans l'histoire des relations du sujet à son environnement. Ces capacités s'avèrent alors inséparables du corps, du langage et de l'histoire culturelle. Elles permettent de donner un sens au monde.

L'énaction s'impose peu à peu comme un paradigme unificateur des sciences cognitives. En témoigne la constitution d'une communauté scientifique avec une réelle cohérence intellectuelle. L'école thématique du CNRS organisée par l'ARCo (Association pour la Recherche Cognitive) du 29 mai au 3 juin 2006 à Ile d'Oléron (France) et intitulée « Constructivisme et énaction - Un nouveau paradigme pour les sciences cognitives » en est une preuve.

Au delà d'un positionnement théorique dans les sciences cognitives, l'énaction est une source d'inspiration pour beaucoup d'applications dans le champ des interfaces et des interactions homme machine :

- Un exemple parmi les plus emblématiques est le cas du TVSS (Tactile Vision Substitution System) de Paul Bach-y-Rita (Bach-y-Rita 1972). Le système de TVSS est un exemple de système dit de suppléance perceptive (Gapenne & Gaussier 2005) pour personnes non voyantes, c'est-à-dire un système permettant de percevoir des propriétés du monde inaccessibles par la vue. Le système permet de transformer une image captée par une caméra vidéo en un stimulus tactile sur le thorax ou sur le front[70] produit par une matrice de 400 tacteurs (20 lignes, 20 colonnes de picots de un millimètre de diamètre). Les sujets équipés du système ont assez rapidement été capables de distinguer des formes (lignes horizontales ou verticales) et cibles très basiques en mouvement et de les pointer. Pour une reconnaissance de formes géométriques et d'objets usuels les temps d'apprentissage deviennent plus longs et peuvent largement atteindre une dizaine d'heures de pratique. Ce qui est le plus important dans les expériences de Bach-y-Rita, c'est que si le sujet ne manipule pas la caméra il n'est pas en mesure de bien faire la part des choses parmi les stimulus engendrés par la forme elle-même et par son entour dans l'image. Pour y arriver, et donc pour enclencher le processus d'apprentissage, le sujet doit pouvoir orienter à souhait le dispositif de captation en même temps qu'il en perçoit les effets sur la matrice tactile. Ainsi, la perception et la cognition ne peuvent être conçues comme des traitements de l'information simplement reçue de la part de l'environnement. Il n'y a pas de perception dans l'environnement sans action couplée dans l'environnement. Ainsi, Bach-y-Rita apporte une preuve empirique aux principes de l'énaction.

- Un autre exemple de l'implication de la démarche énactive dans la conception de systèmes d'interactions homme-machine réside dans les projets et développements en réalité virtuelle. Mettre un utilisateur dans un environnement où il se retrouve acteur en reproduisant la boucle action-perception est bien un moyen de mettre à profit d'une tâche applicative la démarche énactive naturelle des utilisateurs. C'est capitaliser l'expérience vécue dans et par un environnement numérique fondé par l'énaction (De Loor & Tisseau 2011).

Prendre en compte l'énaction dans une perspective de conception de systèmes informatiques amène à dédoubler les conditions de couplage. D'une part, le système informatique est un environnement pour son utilisateur (c'est encore plus vrai pour les ENT) et en tant que tel le **couplage action/perception** s'y produit. D'autre part, si l'environnement donne à l'utilisateur un accès actif à ses corpus, ses ressources, ses tâches alors il agit comme un prolongement cognitif de l'utilisateur lui permettant de faire émerger du sens. Il en découle que l'environnement en tant qu'intégré à la perception de l'utilisateur s'efface pour permettre de percevoir autre chose à la façon dont une paire de lunettes permet de voir sans qu'on la voie en tant que telle. La complexité de la vision appareillée est « dissoute » dans la perception (Holzem & Labiche 2010). Ce prolongement cognitif de l'utilisateur dans le système est une seconde forme de couplage : le **couplage personne/système**.

Concernant le couplage personne/système, l'énaction conduit à examiner particulièrement la question des usages et des contournements d'usage dans l'interaction. Le contournement des usages par les utilisateurs n'est par principe jamais anticipé car seul ce pour quoi le logiciel est prévu est considéré dans la conception classique des logiciels sur le modèle «spécifications – cahier des charges – développements – tests». Pourtant l'informatique est depuis toujours un « lieu » de contournement d'usages par excellence car après tout les machines initialement conçues pour calculer servent aujourd'hui bien plus souvent à communiquer. Par ailleurs les logiciels qui par leurs fonctionnalités permettent de faire évoluer les usages pour lesquels ils sont prévus rencontrent en général un grand succès. C'est par exemple le cas des traitements de texte. Aujourd'hui, on n'écrit plus de la même

70 Des versions ultérieures ont également utilisé des matrices tactiles posées sur la langue.

façon avec les traitements de texte qu'avec un crayon et une feuille de papier car on tape souvent partiellement des bouts de texte que l'on complète, que l'on déplace, que l'on met en forme ou que l'on reformule en plusieurs étapes. Cette nouvelle façon de rédiger n'est pas explicitement prévue par le logiciel, elle est induite de ses fonctionnalités par les utilisateurs. C'est un cas d'affordance (Gibson 1977) au sens de capacité d'un objet à suggérer sa propre utilisation. L'environnement logiciel présente des saillances pour notre système perceptif et sensorimoteur. Ces saillances suggèrent des actions au détriment d'autres et prescrivent ainsi des actions possibles. Les affordances s'inscrivent dans des interactions et dans un couplage personne/système. Un autre exemple est celui des moteurs de recherche sur Internet que l'on peut tout à fait contourner pour vérifier si une expression se dit ou pas[71] ou encore si une traduction est bonne ou pas[72]. Cette idée que l'interaction personne/système amène à re-concevoir les logiciels est celle recherchée dans ce qu'on appelle les interfaces énactives. Les interfaces haptiques basées sur le toucher (et donc le couplage action-perception) et largement développées par les smartphones sont tout à fait dans ce courant. Cependant la communauté des chercheurs travaillant sur des interfaces énactives est un courant des Interactions Homme-Machine qui n'a pas beaucoup d'intersections avec le TAL.

Le champ des applications numériques évolue avec les attentes actuelles des utilisateurs. La prise en compte de ces attentes et la façon dont les utilisateurs s'approprient leurs outils au sein de sphères d'activités devenues numériques est une problématique informatique actuelle très forte qu'illustre bien le Web 2.0. En fait de développer des logiciels et des machines comme des dispositifs les plus autonomes possibles, il convient de proposer des environnements d'interaction avec les utilisateurs et entre les utilisateurs eux-mêmes. Dans cet objectif, une simplicité conceptuelle et ergonomique des outils logiciels proposés aux utilisateurs est un atout pour un meilleur couplage et finalement de bien meilleurs résultats. En permettant de théoriser le couplage personne/système on pourrait considérer l'énaction comme une théorie qui sous-tend ou explique une approche centrée-utilisateur comme la nôtre.

5.3. Expérimentations en cours

De la même manière que Paul Bach-y-Rita a réalisé une mise en oeuvre opérationnelle de l'énaction dans le domaine de la suppléance perceptive, nous voudrions pouvoir concevoir un système fonctionnel d'accès au contenu centré-utilisateur pour montrer le bien fondé de l'énaction dans une démarche de TAL. C'est le but d'une expérimentation en cours que nous présentons ici à travers un projet de recherche visant à améliorer les interactions d'utilisateurs de différentes catégories professionnelles avec un système d'information dédié au droit du transport et de la logistique qui repose sur un corpus de textes réglementaires et de compte-rendus de jurisprudence. L'objectif vise à mettre au point un ENT destiné à un public professionnel (entreprises de la filière logistique, juristes, *risk managers*, assureurs, avocats, …) et non professionnel (usagers ou salariés des transports). Ce projet intitulé AIDé (Aide à l'Interprétation de Documents énactive) est réalisé par un collectif de chercheurs normands (dont je fais partie) : le groupe NU (Nouveaux Usages)

71 Expérience faite sur Google le 21/11/12 : si on cherche l'expression "tirer avec des boulets rouges" on a 6 réponses alors que "tirer à boulet rouge" ramène plus d'un million de réponses. On sait alors quelle est la bonne expression même si l'estimation du nombre des résultats n'est pas toujours très fiable.

72 Expérience faite sur Google le 21/11/12 : si on veut retrouver que la meilleur traduction en anglais de traitement automatique des langues est « Natural Language Processing », il suffit de comparer différentes recherches : "natural language processing" (5 170 000 réponses) et "natural language computation" (74 100 réponses).

5.3.1. Le groupe Nouveaux Usages

Le groupe NU a été constitué en 2006 dans le cadre du PUN (Pôle Universitaire Normand). C'est un groupe de recherche pluridisciplinaire[73] entre Basse et Haute Normandie. Il vise à proposer de nouveaux modes d'accès au contenu dans les collections de documents numériques en se basant principalement sur une approche énactive des interactions homme-machine. Il est à l'origine de plusieurs propositions de projets dans le cadre d'appel à financements (hélas infructueuses à l'heure actuelle) auprès notamment du CNRS, de l'ANR ou encore du FUI (Fonds Unique Interministériel).

Le groupe NU regroupe des chercheurs en informatique, en linguistique et en psychologie (N. Baudouin, P. Beust, N. Chaignaud, D. Dionisi, S. Ferrari, M. Holzem, D. Jacquet, J-P. Kotowicz, J. Labiche, S. Mauger, F. Maurel, E. Trupin, Y. Saidali) relevant des institutions et laboratoires suivants :

- Laboratoire GREYC CNRS - UMR 6072 (Groupe de recherche en informatique, image, automatique et instrumentation de Caen) / Université de Caen Basse-Normandie
- Laboratoire PALM JE 2528 (Psychologie des Actions Langagières et Motrices) / Université de Caen Basse-Normandie
- Pôle Modesco (Modélisations en Sciences Cognitives) de la MRSH Caen / Université de Caen Basse-Normandie
- Laboratoire LITIS EA 4108 (Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes) / Université de Rouen, Université du Havre & INSA rouen
- Laboratoire LiDiFra EA 4305 (Linguistique, Didactique, Francophonie) / Université de Rouen
- Rouen Business School
- Esigelec, Rouen

5.3.2. Le projet AIDé

Le projet AIDé est un programme de recherche et de développement mettant en collaboration le groupe NU avec des partenaires extra-universitaires :

- l'Institut du Droit International du Transport (IDIT)
- le pôle de compétitivité Logistique Seine Normandie (devenu Nov@log)
- le groupe SAVE (prestataire de services issus d'un regroupement de compagnies d'assurances et dédié à l'analyse des risques inhérents aux flux de marchandises et d'informations)

Avec l'apparition des techniques numériques et de l'internet, la distance entre les utilisateurs et l'information économico-juridique tend apparemment à se réduire. De nombreux sites proposent aujourd'hui une large gamme d'informations générales ou spécialisées. Sur le plan juridique, par exemple, il est aujourd'hui possible d'accéder à un large pan de la réglementation et de la jurisprudence, qu'elles soient françaises ou étrangères. Mais ce rapprochement technique n'est pas pour autant signe d'une meilleure maîtrise et appropriation de l'information, bien au contraire.

73 auto-qualifié d'*indisciplinaire*

L'herméneutique juridique demande des connaissances métiers importantes. La tâche d'interprétation d'une loi consiste à concrétiser cette loi dans chaque cas particulier. Cette concrétisation est d'ailleurs le thème central de la jurisprudence. Une loi ne demande pas en effet à être comprise historiquement mais doit se concrétiser dans sa valeur juridique à travers des cas particuliers d'interprétation. Nous constatons que les systèmes actuels (interface de dialogue, bases de données, …) conduisent à une interaction système/utilisateur forcément appauvrie, parce qu'ancrée dans un environnement prédéfini sous forme de thésaurus qui réorientent la question de l'utilisateur. Le projet AIDé jette les bases de la conception d'un ENT capable de s'enrichir d'apports successifs dus aux interactions de plus en plus denses et complexes au sein de sphères d'activités devenues numériques. Cela nous conduit à sortir de la problématique du mot-clé, ou du figement lexical (représentation de connaissances), pour celle de la thématique des textes et de l'interprétation située.

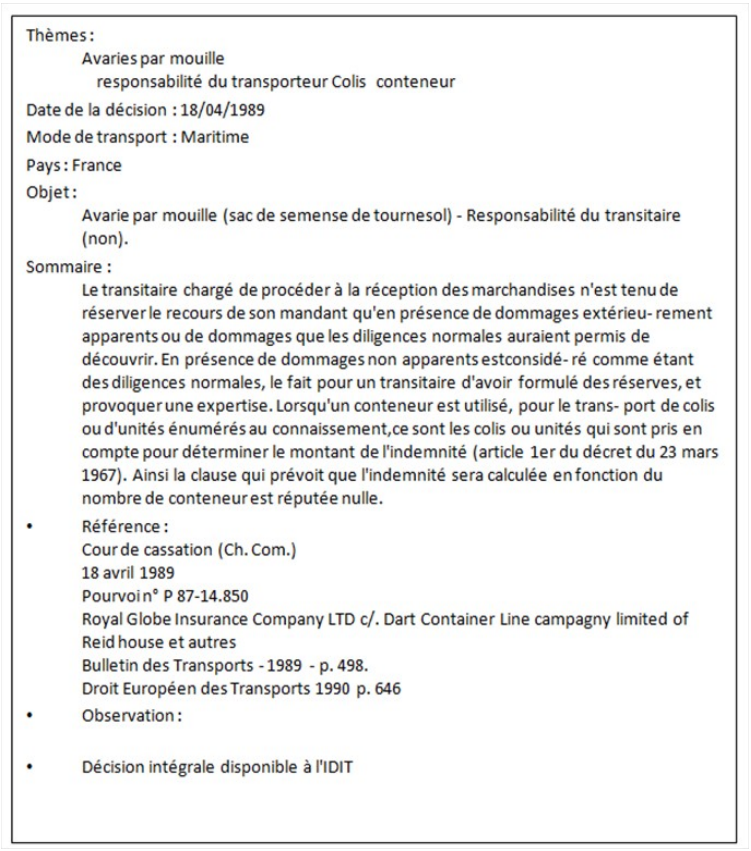
Le système d'information de l'IDIT est conçu pour diffuser des informations aux adhérents (services payants) afin qu'ils puissent gérer dans les meilleures conditions leur entreprise et sécuriser leurs activités. Ces acteurs du transport et de la logistique se doivent de rechercher et d'analyser des informations de plus en plus nombreuses. Ainsi le suivi et l'anticipation des cadres juridiques communautaires et nationaux sont des éléments de gestion incontournables. Or, la fragmentation de l'information relative au droit des transports et de la logistique qui couvre des domaines aussi variés que le droit commercial, le droit des sociétés, le droit de l'environnement, le droit administratif, le droit pénal, le droit social, rend difficile l'accès à l'information (information éparse, réglementation pléthorique, accès difficile et coûteux…), d'où la nécessité d'une mise en relief (signalement pour interprétation) de celle-ci assurée par des veilleurs juridiques à l'IDIT.

Les veilleurs de l'IDIT utilisent quotidiennement et alimentent une vaste base documentaire hétérogène :

- 19571 fiches de comptes-rendus de jurisprudence, i.e. des documents secondaires interprétant des documents primaires de cour d'appel, de cour de cassation depuis 1971 (cf. illustration 40),
- 17846 Articles spécialisés,
- 2449 textes officiels,
- 944 documents de fond documentaire,
- des guides des transports maritime et routier « destinés à tous ceux qui sont confrontés aux problèmes juridiques liés à une opération de transport de marchandises »,
- des fiches Lexisnexis du transport maritime et routier.

Le but applicatif est de concevoir un ENT à destination des veilleurs de l'IDIT leur permettant plusieurs fonctionnalités dans le cadre d'une boucle interactive non nécessairement finalisée :

- permettre d'interagir avec la base documentaire,
- permettre de prototyper et projeter sur corpus des ressources personnelles,
- permettre de tracer l'activité de l'utilisateur,
- permettre d'intégrer des outils en interopérabilité,
- permettre de faire émerger des usages non prescrits.



Thèmes :
Avaries par mouille
responsabilité du transporteur Colis conteneur
Date de la décision : 18/04/1989
Mode de transport : Maritime
Pays : France
Objet :
Avarie par mouille (sac de semense de tournesol) - Responsabilité du transitaire (non).
Sommaire :
Le transitaire chargé de procéder à la réception des marchandises n'est tenu de réserver le recours de son mandant qu'en présence de dommages extérieu- rement apparents ou de dommages que les diligences normales auraient permis de découvrir. En présence de dommages non apparents estconsidé- ré comme étant des diligences normales, le fait pour un transitaire d'avoir formulé des réserves, et provoquer une expertise. Lorsqu'un conteneur est utilisé, pour le trans- port de colis ou d'unités énumérés au connaissement,ce sont les colis ou unités qui sont pris en compte pour déterminer le montant de l'indemnité (article 1er du décret du 23 mars 1967). Ainsi la clause qui prévoit que l'indemnité sera calculée en fonction du nombre de conteneur est réputée nulle.
- Référence :
Cour de cassation (Ch. Com.)
18 avril 1989
Pourvoi n° P 87-14.850
Royal Globe Insurance Company LTD c/. Dart Container Line campagny limited of Reid house et autres
Bulletin des Transports - 1989 - p. 498.
Droit Européen des Transports 1990 p. 646
- Observation :
- Décision intégrale disponible à l'IDIT

Illustration 40: Exemple de fiche de l'IDIT

L'approche centrée-utilisateur et personnalisable, notamment en terme de ressources lexicales, est une demande très forte de la part des veilleurs de l'IDIT. Pour que le couplage personne système s'installe énactivement il faut que les veilleurs puissent projeter leurs points de vue spécifiques et leurs connaissances métiers dans l'environnement. Le droit à beau être le même pour tout le monde, l'interaction et la « navigation » ne se fait pas de la même façon pour quelqu'un qui travaille dans le domaine depuis des années ou pour quelqu'un qui effectue une recherche ponctuelle en marge de son activité principale. De plus, même en considérant qu'une connaissance juridique puisse être capitalisée (notamment à partir d'une ontologie), il est fort peu probable qu'on puisse trouver un lexique adapté à des milieux professionnels aussi précis que ceux dont il est question ici, à savoir la jurisprudence du droit des transports essentiellement dans le domaine maritime. Par exemple, la fiche IDIT reproduite dans l'illustration 40 montre le résumé d'un cas jugé pour un problème d'*avarie par mouille*, c'est-à-dire un problème de recherche de responsabilité entre un affréteur et un transporteur dans le cas d'une cargaison contenue dans un container ayant pris l'eau durant le transport au point d'altérer la cargaison. Une ontologie ou base terminologique donnant une définition de la lexie *avarie par mouille* est très peu probable, alors que la lexie est ici essentielle pour le veilleur. L'approche centrée-utilisateur est bien ici une nécessité.

Les outils dont nous disposons pour partie actuellement (par exemple, ThèmeEditor et ProxiDocs) et d'autres que nous envisageons de développer seront intégrés et mis à disposition incrémentalement dans le socle de l'ENT afin que l'utilisateur puisse formuler des requêtes et être aidé dans

l'interprétation des résultats qui lui sont proposés pour affiner et/ou reformuler sa recherche d'information. Finalement, la réalisation de cet ENT repose bien sur une nouvelle approche dans la conception d'interfaces interactives et cognitives où l'engagement cognitif de l'utilisateur sera concentré sur son activité et non plus déplacé sur la machine.

L'intégration dans un même environnement d'outils en interopérabilité est un travail technique conséquent pour lequel le projet AIDé cherche à trouver des financements. Pour l'instant un portail Web orienté Web2.0 avec intégration de Web-services a été mis au point (cf. illustration 41) et demande encore à être enrichi. Nous avons récemment (depuis septembre 2012) obtenu au GREYC le recrutement d'un collègue sur contrat post-doctoral (Alexandre Labadié) financé par la région Basse-Normandie et le FEDER pour conduire des développements logiciels qui pourront notamment enrichir le projet AIDé.

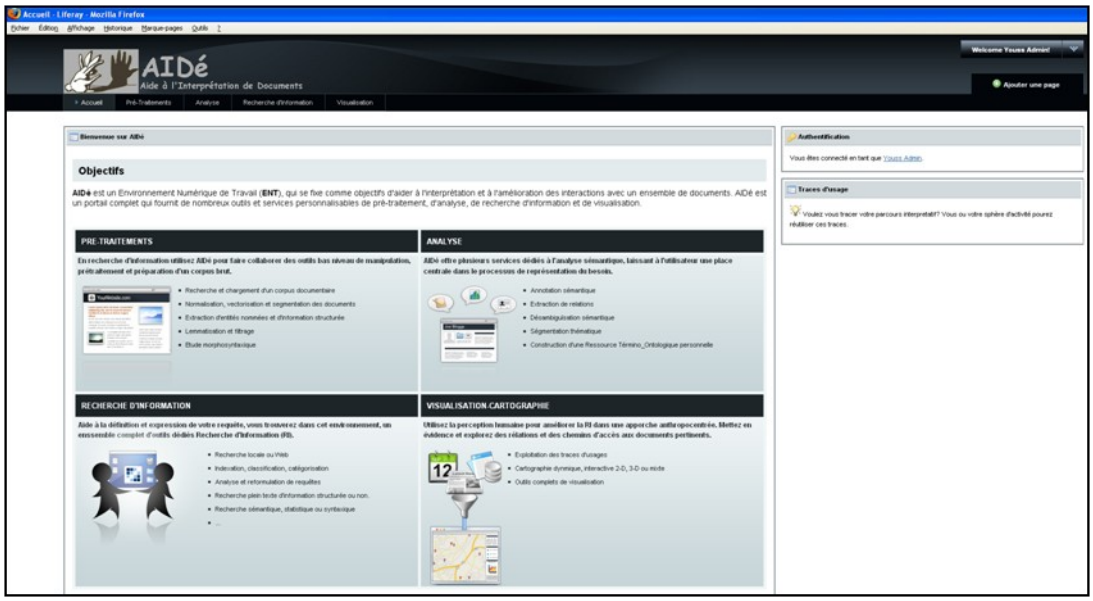


Illustration 41: Le portail d'intégration AIDé

Une fois l'ENT fonctionnel (au moins dans une première version) nous aurons avec l'IDIT un contexte d'expérimentation riche avec des « vrais » utilisateurs que l'on pourra observer dans leurs interactions avec l'environnement, autant individuellement que collectivement. De cette expérimentation nous chercherons à définir des protocoles d'évaluation pour des systèmes centrés utilisateurs énactifs.

5.4. L'évaluation

Incontestablement, l'évaluation est une notion à la mode dans le fonctionnement actuel de nos structures sociales. Comme d'autres préoccupations actuelles telles que le principe de précaution par exemple, on retrouve l'évaluation impliquée dans un nombre important de domaines de la vie publique : les sciences et technologies, les institutions et loi, les placements et transactions financiers, le commerce, la publicité[74] …

74 Dans des méthodes publicitaires l'évaluation est avancée comme un argument commercial. Ainsi on tente de vendre des services en les faisant évaluer et accréditer par des organismes de normalisation reconnus (c'est le

Dans la plupart des domaines de l'informatique, l'évaluation des programmes a toujours été un souci premier car les conditions de leurs exécutions en découlent. Ainsi on a tout intérêt à savoir concevoir des programmes efficaces plutôt que des programmes qui en fonction des données qu'on leur soumet peuvent engendrer une explosion combinatoire.

Dans le cadre particulier du TAL, les méthodes classiques d'évaluation (issues de la problématique de la Recherche d'Information) visent principalement la quantification de la qualité des résultats produits par un système. Sur cette question, les approches théoriques et empiriques en TAL se rejoignent. Des taux sont proposés pour rendre compte de la réussite ou de l'échec d'un programme produisant un résultat à partir du traitement d'une donnée linguistique : le Rappel[75], la Précision[76], la F-mesure[77] (Van Riesbergen 1979).

Ces mesures sont principalement mises en œuvre dans le cadre d'évaluations comparatives menées lors de campagnes où sont mis en rivalité différents systèmes. Ainsi on a très fortement recours au rappel et à la précision dans les conférences sur la compréhension automatique de message (MUC), la désambiguïsation de mots (SENSEVAL) ou l'extraction d'information (TREC) alors même que les objectifs opératoires spécifiques de ces conférences peuvent être très différents. Par exemple, il peut s'agir d'être capable de rechercher un passage d'un document où se situe la réponse à une question posée (tâche de Question/Réponse dans TREC) ou encore de remplir une base de données à partir d'une analyse sémantique de courts documents (par exemple lors de certaines conférences MUC, on a cherché à extraire des dates, des noms de lieux par exemple à partir de comptes-rendus d'attentats terroristes). Dans quelques cas particuliers, des variantes du rappel et de la précision sont mises en place. C'était le cas lors de l'action GRACE (Adda & al. 1999) où il était question de comparer différents analyseurs morpho-syntaxiques du français contemporain avec les mesures de précision et de décision (en remplacement du rappel). La décision proposée dans GRACE mesure la proportion d'étiquetages stricts parmi l'ensemble de tous les étiquetages (stricts et ambigus). Ainsi un taux de décision de 100% correspond à une analyse où toutes les formes sont étiquetées de façon non ambiguë, c'est-à-dire avec une seule catégorie.

L'omniprésente utilisation des métriques classiques (rappel, précision, F-mesure et variantes telles que le bruit, le silence) laisse croire que leur généricité pour une application à toute forme d'évaluation n'est pas contestable. Ce n'est pas vrai. Il faut garder à l'esprit que ces métriques ne sont utilisables que lorsque qu'un corpus de test a pu être constitué et c'est loin d'être toujours le cas (sinon le rappel n'est pas calculable précisément, au mieux « approximable » par extrapolation sur un échantillon connu de corpus[78]). De plus, dans beaucoup de cas, quand un corpus de tests est réalisable, il ne faut pas sous-estimer le coût qu'il représente en temps et en investissement. A titre d'exemple, lors de

cas de certaines normes ISO) où encore on cherche à vendre un yaourt plus qu'un autre parce qu'on argumente que des chercheurs en ont évalué (sans bien sûr dire comment) les retombées bénéfiques pour la santé. La mise en place d'une évaluation a ainsi pour but de donner l'impression d'une démarche sérieuse parce qu'on prend un recul sur ce qui est fait, parce qu'on met en place une démarche réflexive.

75 Proportion de bons résultats fournis par le programme par rapport aux bons résultats qu'il aurait dû idéalement fournir.

76 Proportion de résultats corrects parmi l'ensemble des réponses fournies par le programme.

77 Mesure synthétisant le rappel et la précision en fonction d'une variable de préférence β :
F-mesure = $(\beta^2+1)PR/ (\beta^2P+R)$

78 Par exemple, il n'est pas possible de calculer de façon incontestable le rappel d'un moteur de recherche sur Internet car personne ne peut établir en dehors de l'utilisation d'un moteur de recherche sur un échantillon de documents combien il existe précisément de documents qui répondent réellement à une requête.

l'action GRACE, une des étapes les plus importantes en temps a été de se mettre d'accord sur un jeu d'étiquettes morpho-syntaxiques commun à tous les systèmes et au corpus de test alors qu'intuitivement trouver la catégorie morpho-syntaxique d'un mot dans une phrase semble ne pas poser de problèmes. On imagine dans ces conditions quelle serait la difficulté pour constituer des corpus de références pour d'autres types d'évaluation beaucoup moins consensuelles (par exemple, il semblerait difficile de produire un corpus d'énoncés de dialogue avec étiquetage des actes de langage tellement la dimension interprétative subjective paraît ici première). Il faut aussi veiller à ce que le corpus de référence reste en accord avec les conditions réelles d'utilisation des systèmes pour lesquels il doit servir à l'évaluation. Comme le montre Spark Jones (Spark Jones 2001) une évaluation en laboratoire est fondée sur des présupposés concernant le contexte de la tâche, ce qui a un impact important sur le bien-fondé de l'évaluation et sur les résultats de performance.

Les protocoles d'évaluation classiques sont restrictifs car ils ne s'appliquent qu'à des systèmes considérés comme finalisés dont il convient de quantifier les résultats, mais il ne serait pas prévu d'évaluer la capacité d'apprentissage d'un système et la montée en qualité de ses résultats au fur et à mesure de cet apprentissage. Les campagnes d'évaluation ne s'appliquent donc qu'à des logiciels fournis « clés en main » et laissent de coté les systèmes dits « en amorce » et les logiciels d'étude conçus en laboratoires. Plutôt que de considérer l'évaluation comme étant forcément une tâche finale sans retour sur le logiciel (dont du coup l'intérêt peut être contestable et n'a finalement qu'une valeur de validation), il serait intéressant de chercher à l'intégrer à la conception même du logiciel en imaginant par exemple des méthodes d'évaluation intégrées au cycle de conception et d'usage du logiciel où l'utilisateur visé, du coup, aurait sa place. C'est le cas par exemple des techniques de modélisation du dialogue en magicien d'Oz dans les projets de dialogue Homme-Machine.

Les systèmes centrés sur l'utilisateur s'opposent aux systèmes technocentrés où l'utilisateur n'a principalement qu'un rôle de réception des résultats de la machine sans qu'il puisse déterminer la façon de les produire. Dans un environnement centré-utilisateur les traitements sont déterminés par l'expression d'un point de vue particulier, celui de son utilisateur, sur une tâche particulière. II ne s'agit pas simplement de permettre à l'utilisateur de personnaliser son application (ce qui reviendrait à prévoir d'avance une liste exhaustive de profils d'utilisateurs et de sélectionner l'un de ceux-là en fonction des choix faits). Les ressources utilisées sont produites de façon endogène dans la boucle d'interaction système-utilisateur à partir des corpus observés et analysés. Elles n'ont pas forcément l'objectif d'être adéquates à un plus grand nombre ni d'être cumulatives car c'est leur spécificité en tant qu'expression de l'utilisateur qui prime. Le problème de l'évaluation se trouve donc ici déplacé. Il ne s'agit pas tant d'évaluer une application que la faisabilité et l'efficacité, du point de vue d'une tâche, d'une interaction entre un utilisateur humain et un dispositif 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, comme les taux de compétence et d'efficacité (Luzzati 1996).

Prenant en compte que c'est le point de vue particulier de l'utilisateur qui est primordial, l'idée même de constituer un corpus de référence sur le résultat idéal qu'il conviendrait d'obtenir n'est plus tenable. En somme, il n'y aurait personne d'autre que l'utilisateur lui-même qui serait bien placé pour dire ce qui émerge de son utilisation du logiciel. Le problème n'est donc pas une question de comparaison à un quelconque résultat mais une question d'auto-évaluation, voire d'introspection ou d'un regard réflexif de l'utilisateur sur l'expression de son point de vue relativement à sa tâche et de ce qui en découle dans le système. Dans ces conditions les mesures de rappel, de précision et autres variantes ne sont pas applicables et il convient plutôt de se tourner vers les méthodes d'expérimentation pratiquées dans les sciences humaines : expérimentation avec plusieurs sujets,

recueil des résultats et entretiens avec les sujets et enfin analyses.

Une interprétation est une perception personnelle qui n'est pas observable (on peut tout au plus observer les conditions et les effets d'une interprétation). De plus, en tant que liée à une expérience vécue, elle n'est pas reconductible à l'identique (car l'expérience vécue n'est plus la même). Nous sommes donc dans une situation où un observateur ne peut rien dire d'une interprétation d'autrui. Il faut donc nous tourner vers des méthodes d'expérimentation et d'entretiens des sujets interprétants eux mêmes en les rendant acteurs de l'évaluation. Dans la logique d'une approche centrée-utilisateur, si l'environnement est centré-utilisateur alors il doit en être de même de l'évaluation.

Cette réflexion rejoint la méthodologie d'entretien en première personne pour l'exploration de l'expérience vécue proposée par la psychologue Claire Petitmengin (Petitmengin 2007). Dans un premier temps, on conduit le sujet à stabiliser son attention pour l'amener à décrire une expérience singulière vécue (sans avoir recours par catégorisation à une classe d'expériences prévues d'avance par autrui), pour enfin dans une troisième phase conduire le sujet à retourner son attention du *quoi* vers le *comment*. L'auteur a pu expérimenter la pertinence de cette approche auprès de malades souffrant d'épilepsie afin de les conduire à prévenir les crises. L'expérimentation en première personne peut permettre une exploration rigoureuse de la réflexivité en « revenant sur un soi agissant » non comme retour vers un déjà vécu mais comme un moyen de décentrer la compréhension que l'utilisateur a de lui-même, un moyen de refocalisation sur un autre soi-même.

Dans une tradition phénoménologique tout à fait en cohérence avec l'énaction, des travaux comme ceux de C. Petitmengin nous incitent à passer de la culture de l'évaluation à celle de l'expérimentation. C'est un choix épistémologique que nous assumons volontiers mais, plus encore, c'est un renversement de culture[79] incontournable en conséquence directe de l'approche centrée-utilisateur.

~

Ce chapitre montre bien qu'adopter une démarche centrée utilisateur en TAL est un projet scientifique de décloisonnement disciplinaire. La pluridisciplinarité en TAL est souvent réduite aux inter-relations entre informatique et linguistique. Nous avons bien fait un pont disciplinaire entre le TAL et la linguistique dans le recours qui est le nôtre à la Sémantique Interprétative. Mais, plus loin encore, l'approche centrée utilisateur nous plonge maintenant dans les sciences cognitives avec la théorie de l'énaction et insidieusement aussi dans la philosophie avec le courant de la phénoménologie. Il en ressort que la collaboration interdisciplinaire ne se réduit pas à un échange de modèle. Les pratiques sont également transposables de manière intéressante, comme c'est le cas par exemple de l'expérimentation. Plus globalement, c'est un échange de culture qui s'opère.

Dans ce brassage de culture scientifique, le projet de conception et d'expérimentation d'ENT énactifs nous apparaît comme un programme unificateur tout à fait en lien avec les attentes que créent les évolutions des sphères d'activités numériques. Nous poursuivons notamment dans ce projet l'idée de (Dionisi & Labiche 2006) qui consiste à caractériser de manière indissociable des processus logiciels impliqués dans des processus expérientiels, eux-mêmes révélant des processus cognitifs.

79 On peut s'en convaincre aisément à en juger par les difficultés à publier des travaux centrés utilisateurs dans des conférences de TAL ou encore à obtenir des financements institutionnels types ANR.

6. Conclusions

« L'homme ne comprend que dans la mesure où il est créateur »
(E. Cassirer, Logique des sciences de la culture , p. 84)

Il nous semble que dans la plus grande partie des développements en TAL la place donnée à l'utilisateur le cantonne en tant qu'observateur à l'extérieur du système. Pourtant au final c'est bien lui que l'on cherche à aider et c'est bien lui à qui l'on cherche à donner des facultés de compréhension plus vastes. Si l'on suit l'avis de Cassirer dans la citation ci-dessus, l'utilisateur ne peut rester observateur passif et doit être impliqué dans le système. C'est ce à quoi nous cherchons à contribuer.

Etudier la place de l'utilisateur en tant que créateur au sein de l'interaction homme-machine nous amène à nous intéresser à des dispositifs logiciels plus évolués qu'un programme qui rendrait un résultat à partir des données qu'on lui fournirait. De ce point de vue le mot *Traitement* dans l'acronyme TAL est très réducteur et se veut annonciateur d'une très faible prise en compte de l'utilisateur. C'est en nous invitant à chercher d'autres formes d'instrumentations logicielles que notre approche centrée sur l'utilisateur nous amène aux Environnements Numériques de Travail. Dans son ENT l'utilisateur est par définition acteur et cela ouvre des portes pour des applications sémiotiques.

A quoi bon chercher à calculer du sens de manière computationnelle et automatique quand un utilisateur est là et qu'il ne demande qu'à être créateur dans l'interaction homme-machine notamment en déployant naturellement ses facultés interprétatives ? La question de l'effectivité sémiotique de la machine se trouve déplacée dans les ENT : il ne s'agit pas de calculer du sens à « servir » à

l'utilisateur, mais il est question de constituer des signes qui, dans l'interaction homme-machine, seront interprétés par l'utilisateur qui y projettera du sens. C'est par exemple exactement le cas des cartographies ProxiDocs ou Canopée qui proposent des résultats de calcul sous forme de visualisations **qui font sens** pour l'utilisateur. Ici l'utilisateur est doublement acteur : il est acteur parce que c'est lui qui interprète les signes et projette du sens mais il est acteur également plus en amont en réunissant les corpus et en constituant interactivement des ressources terminologiques. Cette double action dans l'environnement participe d'un couplage sémiotique fructueux car c'est dans et par son action que l'utilisateur peut **se retrouver** dans les prolongements de l'interaction.

L'approche centrée sur l'utilisateur en TAL et son implication dans les ENT nous invite à considérer que les traitement sémantiques ont vocation à enrichir l'entour sémiotique de l'utilisateur lui permettant ainsi notamment d'établir des rapprochements inter-textuels, d'évaluer un rapport de similitude et de différence entre telle ou telle information potentielle ou encore d'effectuer une recherche dans un espace documentaire. L'ENT devient l'outil qui permet l'accès à l'information et non plus simplement une finalité à l'interaction. C'est ce que nous avons appelé un principe de **sémantique augmentée**, à la façon où les outils dits de réalité augmentée enrichissent la réalité pour finalement mieux l'appréhender. Dans cette sémantique augmentée les phénomènes linguistiques et cognitifs sont au cœur de la tâche menée. L'ENT doit évidemment les prendre en compte dans les processus mis en œuvre (c'est par exemple le cas de la question de la détermination du global sur le local). Mais plus encore, l'ENT doit favoriser, au délà des usages prévus, le fait que l'on puisse observer des phénomènes émergents interprétables d'un point de vue linguistique et cognitif et même plus largement culturel. C'est l'idée que les sciences de la culture sont non seulement embarquées dans les ENT mais plus profondément **impliquées** dans et par les ENT. Le courant du Web Sémantique porte l'idée que c'est l'ontologie (i.e. le matériau ontologique du monde, le réel …) qui est commun dans l'intercompréhension entre les producteurs d'informations et ceux qui l'interprètent (quand bien même cette distinction serait valide). C'est faux. Ce sont les pratiques sociales et culturelles qui sont au centre des mécanismes interprétatifs. Si le Web 2.0 a pris de l'ampleur, comparativement plus que le Web Sémantique, ce n'est pas tant parce qu'il a choisi une autre voie que celle de l'ontologie, mais bien parce qu'il est en appui sur des réseaux sociaux et plus généralement des cultures.

Les ENT induisent des changements dans les pratiques informationnelles des utilisateurs (comme les SMS ont également introduit des changements dans les façons d'écrire). C'est le propre de la dimension linguistique d'être synchro-diachronique et d'être ce par quoi elle change. Dans les pratiques des utilisateurs à travers le numérique la notion de lecture est en train de se transformer. On constate que l'on passe de moins en moins de temps à des lectures extensives pour multiplier les lectures discontinues et fragmentaires. Ce ne sont pas pour autant des lectures baclées car elles sont presque encore plus que les lectures classiques dans une interprétation nécessairement inter-textuelle. Il faut accompagner ces nouveaux usages plutôt que de les ignorer ou de les dénigrer. De ce point de vue on pourrait tout à fait considérer que les travaux de coloriage et de cartographie thématiques que nous avons réalisés s'apparentent à des interfaces favorisant la lecture rapide. Notre démarche s'inscrit dans des processus de développement en aller-retours entre des interfaces de lecture rapide d'ensembles documentaires, des corpus d'étude et des utilisateurs (ou des groupes d'utilisateurs), les uns étant conditionnés par les autres. Les corpus utilisés sont à la fois à l'origine des ressources lexicales construites et constituent en même temps le matériau d'expérimentation de ces ressources. Ce que l'utilisateur peut gagner en temps de lecture, il le « paye » par un temps d'interaction et de couplage avec l'ENT, mais sans le regretter.

Relativement au domaine du TAL et plus précisément de la sémantique, il faut dépasser l'héritage scientifique du courant dominant logico-grammatical qui cherche une voie médiane entre terminologie

des concepts et conceptualisation des termes. Nous argumentons dans le sens qu'il faille redonner son importance première à l'utilisateur car c'est lui en fonction de sa tâche, et plus particulièrement de son étendue (entre la recherche d'actualité sur le web et le classement d'un domaine), qui décidera de manipuler des termes ou d'utiliser des concepts, sans certainement a avoir à s'en créer un problème. D'autre part, il n'est pas prouvé que les utilisateurs soient en attente de machines qui sachent extraire et « manipuler » à leur place le sens des textes. Les usages sur le Web montrent autre chose. Les utilisateurs ont besoin d'outils qui participent à leur compétence interprétative propre interactivement via un couplage personne/système. Ce sont les utilisateurs qui « projettent » du sens sur les textes relativement à leur pratiques sociales et à leur culture. Une machine peut tout à fait aider un utilisateur dans ce but sans nécessairement devoir inclure une formalisation du sens et une représentation du monde, vaste, incontrôlable et finalement inaccessible. Pour permettre à l'utilisateur cette projection de sens sur les textes par jeux d'alternances entre fonds et formes, il faut que l'environnement interactif donne une place particulière à la textualité et en faisant par exemple un lieu d'interaction. C'est l'idée que nous avons poursuivie notamment avec les techniques de coloriage thématique. Dans bon nombre de travaux en TAL, les analyses sur les données textuelles consistent à se passer du texte (c'est le cas du Web Semantique) ou encore à « dé-textualiser » le matériau linguistique. C'est le cas par exemple de « *l'observatorium*[80] », système qui mesure des similarités entre textes sur le Web de J. Louçã qui commence par retirer de tous textes les mots de moins de 3 lettres et les ponctuations (Louçã & Rodrigues 2011). Outre le fait que les ponctuations peuvent être tout à fait significatives statistiquement (Valette 2004) et qu'il n'y a pas d'autre forme de sens que les signes et les textes, la textualité est « précieuse » et le numérique ne doit pas l'altérer. C'est dans et par le texte que l'utilisateur construit son interprétation

Dans les applications du TAL, l'engagement cognitif et « l'appétit » interprétatif des utilisateurs sont une chance qu'il faut pouvoir mobiliser. La simplicité conceptuelle, fonctionnelle et ergonomique des outils de TAL intégrés aux ENT va, espérons-nous, dans cette direction. Plus un outil est simple, plus le couplage avec l'utilisateur est au rendez-vous et plus les contournements fructueux sont envisageables. Les domaines de la sémiotique, des interactions homme-machine et des sciences cognitives sont évidemment complexes, mais ce n'est parce que les domaines sont complexes que les outils doivent l'être également. La complexité peut se trouver dissoute dans le couplage qu'un outil simple induit. L'expérience par l'utilisateur de cette dissolution dans le couplage est un champ de recherche pluridisciplinaire motivant et finalement, grâce à cette expérience, on pourrait faire l'hypothèse que la façon d'accéder à un contenu définit déjà le contenu. Il en découle qu'une instrumentation centrée sur l'utilisateur en sémantique n'est pas un modèle opératoire prédictif. Il est certainement bien plus intéressant de mettre en place des ENT produisant des interactions sémiotiquement suggestives.

Dans le projet d'une contribution de l'approche centrée-utilisateur au TAL, les barrières entre disciplines académiques tombent. La problématique est indiscutablement au croisement de plusieurs champs disciplinaires, comme nous l'avons vu avec les apports de la linguistique (notamment la Sémantique Interprétative) et des sciences cognitives (notamment l'énaction). Il nous semble intéressant de proposer l'approche centrée-utilisateur dans les ENT comme un projet fédérateur de l'informatique (et notamment le TAL et les IHM), la linguistique, les sciences cognitives et plus généralement les sciences de la culture. C'est dans cette direction que s'ouvrent les perspectives de recherche que nous précisons par la suite.

80 http://theobservatorium.eu consultée le 21/11/12.

7. Projets et perspectives

L'habilitation à diriger des recherches est une étape dans le parcours professionnel, pas une fin heureusement. Elle dresse un bilan dont le but est d'identifier dans quelles directions prolonger le parcours et quelles pistes de recherche explorer. Aujourd'hui, il me semble que quatre directions de travail principales sont identifiables en étant, bien entendu, non exhaustives. Il s'agirait de :

- Prolonger les projets engagés,
- Se rapprocher d'un domaine de recherche connexe de nos problématiques, les EIAH en nous focalisant particulièrement sur la question des traces d'usage,
- Caractériser et problématiser les contournements d'usage,
- Etudier les rapports entre une approche centrée-utilisateur et l'idée d'une intelligence collective dans la composante numérique des sociétés actuelles.

7.1. Projets engagés

Comme cela apparaît à la lecture du présent mémoire, des projets de recherche et de développement sont bien engagés et il convient de les poursuivre car ils sont prometteurs.

C'est le cas du projet AIDé (avec le groupe NU) visant une mise en place et une expérimentation d'un environnement centré utilisateur énactif pour l'aide à l'interprétation au sein de collections de documents numériques textuels. Le développement logiciel du projet devrait avancer sensiblement dans les mois à venir avec l'arrivée au GREYC d'un chercheur post-doctoral travaillant dans le cadre du groupe SemComp.

SemComp est un projet soutenu par le FEDER et la région Basse-Normandie réunissant le laboratoire GREYC (S. Ferrari, P. Beust, G. Dias, F. Maurel, S. Mauger, A. Labadié), le laboratoire en psychologie développementale PALM (D. Jacquet) et la jeune entreprise Noopsis (T. Roy). La problématique générale concerne l'accès à l'information dans des collections numériques variées. Nous proposons la mise en place d'un ensemble d'outils logiciels, dont notamment un module d'analyse sémantique automatique et des « interfaces interactives » visant à rendre possible la visualisation des ressources (et leur modification), la visualisation des résultats d'analyses sémantiques, tant à l'échelle du texte que de la collection de textes. Les attentes sont multiples :

- étudier la pertinence de l'utilisation de ressources sémantiques componentielles pour fournir une aide à l'interprétation dédiée à des utilisateurs qui ne sont pas des linguistes maîtrisant les modèle componentiels;

- analyser dans ce cadre l'automatisation des processus d'actualisation et de virtualisation de sèmes, d'afférence, en regard avec la notion de parcours interprétatif, à l'échelle du texte (infra-documentaire) et de la collection de textes (supra-documentaire) ;

- proposer un environnement interactif permettant à un utilisateur de s'approprier les ressources, d'observer leurs effets sur les textes d'une collection donnée, et de les modifier en conséquence.

Deux cadres applicatifs distincts sont envisagés pour permettre la validation de l'approche : le tourisme culturel nomade et l'apprentissage sur des plate-formes de e-learning. L'approche componentielle est tout à fait dans la lignée des projets LUCIA et ISOMETA que nous avons présentés au chapitre 3.

Dans le cadre des 2 projets AIDé et SemComp qui se rejoignent, il s'agira de travailler ici à des modélisations opératoires d'opérations interprétatives (dissimilations et assimilations par exemple), de la sommation et de la diffusion dans le cadre de faisceaux d'isotopies. La spécificité de cette modélisation est qu'elle devra être pensée comme donnant lieu à des visualisations à différents niveaux de localité textuelle qui interviennent dans une dynamique interactionniste avec l'utilisateur. L'objectif, particulièrement stimulant ici, est de penser une dualité fonctionnelle entre un signe produit dans une interaction qui en est à la fois un résultat en même temps qu'un élément interactif, vecteur potentiel de prolongation de l'interaction. Le concept d'isotopie avec sa simplicité et sa portée opérationnelle est un bon exemple de cette dualité en étant, dans le cadre d'un coloriage textuel, un résultat et un support d'interaction d'une interprétation inter-textuelle.

Un autre aspect stimulant du projet d'ENT énactif est son caractère (volontairement assumé) ouvert en terme de développements informatiques. L'ENT doit s'enrichir de différents modules apportant chacun des signes potentiels à l'interaction et à l'interprétation. Il y a là des questions techniques à régler autour de l'intégration des web services et des formats d'échange XML. Il y a aussi à se poser des questions d'apports qualitatifs de tel ou tel module de traitement, d'extraction ou de visualisation. Si l'approche énactive et centrée-utilisateur est un apport au TAL, il est évident que la réciproque est également vraie. On pourrait par exemple chercher à alimenter les traitements de l'ENT par une extraction d'entités nommées, problématique qui revient régulièrement en recherche d'information. C'est un problème particulièrement difficile en TAL quand il est posé de manière « 100% automatique ». Comment anticiper une variabilité extrême des entités nommées quand celle-ci dépend d'une réalité diachronique socio-linguistique plus que de propriétés du signifiant. Il est fort probable

que sur ce type de problématique pour laquelle aucune base exhaustive n'existe et ne peut exister, l'approche centrée-utilisateur et la conception de ressources en cours d'usage soit une bonne approche. En somme, nous défendons ici la vertu du « 100% couplage personne-système » plus que du « 100% automatique ».

Parmi les projets en cours, le développement industriel avec la société eXo maKina est également extrêmement prometteur. Les deux projets ne sont d'ailleurs pas absolument indépendants dans la mesure où il serait envisageable (par exemple dans le cadre d'un projet type ANR liant nos laboratoires de recherche et la société eXo maKina) que des fonctionnalités de Canopée intègrent l'ENT du projet AIDé.

Les questions actuelles que nous nous posons avec le partenaire industriel eXo maKina sont variées et intéressantes tant d'un de point de vue de recherche que d'un point de vue plus pragmatique de finalisation technique d'une version 1 qui déjà ouvre des possibilités pour les versions à venir. Elles concernent par exemple des méthodes exploratoires de visualisation 3D d'espaces documentaires (projection sur des espaces sphériques ou encore navigation dans un espace documentaire à la manière d'une interface de type Street View Google). Elles concernent également les notions de taille de corpus pertinents. A titre d'exemple, nous avons traité à des fins de démonstration les 3 premiers mois de l'année 2011 des dépêches AFP en français et en anglais. Ce corpus totalise quelques 152 000 documents. Ceci représente un flux quotidien de presque 1700 dépêches en moyenne. Le volume des documents est à ce point croissant qu'on pourrait imaginer que des problèmes techniques d'optimisation vont très vite se trouver incontournables. Cependant, en se rapportant (comme toujours) à l'utilisateur et à ses conditions d'usage avant tout, il est peu probable que l'utilisateur demande tous les jours à cartographier les dernières année de l'AFP (et qui en plus de le demander souhaiterait avoir le résultat dans la seconde). Dans le cadre d'un environnement de travail impliquant des actions répétées régulières, il est plutôt enclin à anticiper des demandes de cartographie des documents du jour, de la semaine passée ou d'une période donnée (par exemple une période de congés). En somme, une quantité d'information raisonnable ne donnant pas lieu a une explosion combinatoire. On retrouve bien là dans l'approche centrée-utilisateur la vertu d'une sémantique légère (au sens de Perlerin 2004), légère au niveau des ressources comme des traitements.

Une autre question concerne enfin le rapport entre la sémiotique textuelle et la sémiotique de l'image. Dans le cadre du projet Tungstène déjà commercialisée par eXo maKina à propos de la détection des images contrefaites on s'aperçoit que l'interprétation d'une altération/contrefaçon d'une image est localement guidée par des signes que le logiciel produit mais qu'elle est aussi et principalement déterminée globalement par un discours et des rapports à d'autres images. La similitude entre l'interprétation des images et l'interprétation des textes et flagrante et, en même temps, elle est normale dans la mesure où c'est la même sémiosis qui est en œuvre (cf. Rastier 2011b[81]). Il est d'ailleurs fort possible que l'interprétation d'une vidéo sur le Web relève également de cette sémiosis de la détermination du local par le global. La société eXo maKina a d'ailleurs en projet de produire un outil dédié à l'assistance à l'interprétation des vidéos potentiellement falsifiées sur le même modèle que Tugnstène à propos de l'image. Un objectif en terme de développement informatique pourrait être de rendre les outils actuels Tungstène et Canopée, ainsi que ceux à venir, plus interfaçables au sein d'un environnement d'interprétation augmentée personnalisée multimédia.

C'est un objectif que nous allons pouvoir poursuivre dans le cadre d'un projet soutenu par l'ANR

81 *« La problématique interprétative dépasse les textes et peut s'étendre à d'autres objets culturels, comme les images (susceptibles des mêmes méthodologies : recueil de corpus, détermination des genres, indexation par des traits de l'expression). »* (Rastier 2011b, p. 25)

qui commence actuellement. C'est le projet DocScope. DocScope a pour objectif de travailler à l'identification de documents falsifiés en développant notamment des méthodes de recherche d'indices visuels directement apparents ou dissimulés. Il réunit des entreprises (Hologram Industries, eXo maKina) et des laboratoires de recherche (ENST Brest, GREYC). Un tel projet et l'exploitation des résultats (mise en évidence d'incohérences, d'intrusions, de modifications frauduleuses) se conçoit aussi en tant qu'il intègre une phase interprétative qui permette de préciser le (les) signification(s) des manipulations éventuellement constatées. Les travaux issus de la sémiotique générale et plus particulièrement, ceux qui procèdent de l'analyse d'image, ont ainsi leur rôle à jouer et peuvent apporter un éclairage analytique particulier. C'est l'apport des collègues du GREYC impliqués dans le projet (S. Mauger, P. Beust, G. Dias, N. Lucas).

7.2. Vers les EIAH et les traces d'usage

Dans nos projets de recherche et développement il y a un dédoublement de problématique : l'environnement numérique à développer est l'objet de toutes nos forces de travail et de développement, mais en fait il n'est en même temps qu'un moyen pour appréhender un autre champ d'étude : l'utilisateur et ses capacités interprétatives (elles mêmes ayant des répercussions sur l'environnement numérique). A ce stade de nos recherches il nous semble que deux pistes de recherche doivent être entreprises pour rendre fructueux ce dédoublement :

- celle dites des Environnements Informatisés pour l'Apprentissage Humain (EIAH)
- celle du traçage des activités des utilisateurs dans les environnements.

Le domaine des EIAH est certainement à l'heure actuelle le champ de recherche le plus actif parmi les travaux qui s'intéressent aux ENT dans la mesure où les projets d'ENT sont presque tous des environnements pour la pédagogie (ce qui est souvent dénommé de manière plus ou moins heureuse par l'anglicisme *e-learning*). Les EIAH peuvent nous apporter des exemples d'études comparatives pour le recensement des cas d'usage dans tel ou tel type d'ENT (par exemple, les différents scénarii pédagogiques mis en œuvre dans un même ENT de type Moodle comme celui dont nous disposons, via le CEMU, à l'université de Caen Basse-Normandie). En retour, nous pouvons enrichir le courant des EIAH avec l'apport, par la dimension énactive, d'une théorisation cognitive de la question du couplage personne-système. C'est un aspect qui ne semble pas vraiment défendu comme une position théorique forte dans la plupart des travaux en EIAH (cf. pour s'en convaincre des programmes des conférences reconnues dans le domaine[82]). Pourtant les travaux en EIAH sur les *serious games* (Michael & Chen 2005) par exemple, portent par nature la question de l'engagement ludique comme moyen de couplage avec l'utilisateur.

L'ancrage épistémologique commun aux EIAH est celui du socio-constructivisme issu des travaux de Vigotski (Vigotski 1978) selon lesquels l'apprentissage est le fruit d'une interaction sociale. Le socio-constructivisme[83] (souvent opposé au *behaviorisme*) souligne l'impact de la collaboration, du contexte social et des négociations sur la pensée et sur l'apprentissage. C'est un point de vue résolument socio-centré qui ne nous paraît pas pour autant incompatible avec une théorisation individu-centrée des usages. En outre, le détour que nous proposons en direction des EIAH se propose d'unifier la problématique des ENT car il nous semble qu'il n'y a pas vraiment d'ENT dans lesquels les utilisateurs n'apprennent rien. Sinon, il n'y a pas de couplage et au final pas d'ENT. Les EIAH n'ont

82 Par exemple http://eiah2009.univ-lemans.fr/ consultée le 7/01/13.

83 Courant défendu notamment de manière totalement affirmée et revendiquée par Martin Dougiamas, fondateur du projet Moodle.

certainement pas à y gagner à se démarquer d'une part des autres champs de l'informatique comme le TAL par exemple, comme d'autre part des sciences cognitives. La motivation de l'utilisateur apprenant est un facteur déterminant dans ce couplage personne-système qui de toute évidence conditionne la réussite de l'étudiant. La nature plus ou moins autotélique de l'étudiant et le sentiment de *flow* (Heutte 2011) sont donc des facteurs tout aussi importants à étudier que les conditions du contexte social.

Dans la problématique des EIAH la progression pédagogique de l'apprenant est une question qui requiert naturellement beaucoup d'attention. Cette progression pédagogique a l'intérêt d'être observable, à la différence de l'interprétation qu'un utilisateur peut se faire d'un document. On peut l'observer par des progressions de taux de réussite sur des exercices et des cas pratiques. On peut aussi en rendre compte en suivant et par anglicisme en traçant dans l'environnement l'activité et le parcours des utilisateurs. Dans les plateformes pédagogiques de type Moodle les utilisateurs sont déjà abondamment tracés. Les traces sont enregistrées dans la base de données de l'application et décrivent les activités effectuées, le temps passé sur la consultation d'une ressource ou encore sur la fréquence de connexion[84]. Ce type de trace est ce qu'on appelle des traces passives, c'est-à-dire des traces captées « à l'insu » de l'utilisateur ou au moins sans en rendre compte à l'utilisateur.

Alain Mille et Yannick Prié (Mille & Prié 2006) défendent un autre type de traçage des utilisateurs dans le domaine des EIAH. C'est la trace active. La trace active est volontairement assumée par les utilisateurs. Elle leur est présentée dans l'environnement même où l'utilisateur est tracé en temps réel. Un tableau de bord personnel de l'achèvement des activités de l'apprenant dans un environnement de e-learning est déjà une forme relativement simple de trace active (Temperman 2012). En plus d'intéresser éventuellement un observateur analyste extérieur ou un tuteur, un enseignant, elle vise surtout à enrichir les éléments qui alimentent la perception de l'utilisateur et donnent une dimension réflexive à l'environnement. En cela, elles peuvent prendre des formes bien plus complexes qu'un tableau de bord. En outre la trace active est également un moyen pour un groupe d'utilisateurs impliqués dans une tâche collaborative en ligne d'évaluer en temps réel l'apport de chacun au groupe. Selon Mille et Prié, la trace active apporte une valeur ajoutée indéniable dans les EIAH dans le sens où l'on réifie un objet réflexif support de la construction de connaissances individuelles et collectives.

Nous sommes convaincus que des méthodes de traçage assumées et actives dans les ENT, qui plus est ceux qui nous intéressent en premier lieu où un rapport à l'activité interprétative est central, sont d'un intérêt indiscutable et potentiellement vecteur d'une grande valeur ajoutée. Une perspective importante, tant du point de vue fonctionnel que du point de vue expérimental, dans la suite de nos travaux est de mettre en œuvre ce type de méthodes. Comme nous l'avons déjà signalé, interpréter et faire émerger du sens à partir de ce que renvoie l'environnement c'est, du point de vue de l'utilisateur, *s'y retrouver*. La trace active en tant qu'entité réflexive participe sûrement à cet objectif. Les expérimentations que nous aurons à mener auront notamment à répondre à cette question : à quel point une trace active permet-elle de stimuler positivement les phénomènes interprétatifs en temps réel ? Premièrement, la notion de trace active va tout à fait dans le sens de la sémantique augmentée que l'on cherche à mettre en place dans les ENT. Deuxièmement, la trace active apporte un avantage de taille pour l'évaluation de l'ENT. Nous avons argumenté pour une évaluation des ENT en terme d'entretiens menés en première personne selon le modèle proposé par C. Petitmengin. La possibilité de reproduire dynamiquement un ensemble de traces que l'utilisateur ne découvre pas mais revisite a postériori est certainement une bonne façon d'engager l'évaluation réflexive. Cette trace conscientisée en devenir n'est constituée en tant que trace que lorsqu'elle est interprétée. Elle a pour fonction de faciliter l'interprétation d'un utilisateur en « lui permettant de se revoir agir ». C'est ce que permettent certains

84 Dans le contexte de la formation continue ces traces ont un aspect stratégique car elles permettent de justifier l'activité de l'étudiant vis à vis des partenaires financeurs de la formation.

Serious Games où l'on a la possibilité de voir rejouer une partie, revivre le déroulement d'une séquence d'apprentissage (par exemple Dr Geo, un jeu sérieux pour l'apprentissage de la géométrie, cf. Carron 2011).

Mettre en place un traçage actif en lien avec des évaluations en première personne est une perspective stimulante qui place bien la problématique dans une dimension pluridisciplinaire à laquelle nous adhérons assurément.

7.3. Les contournements d'usage

La question du couplage personne-système amène évidemment à s'intéresser aux usages. Il y a les usages directement prévus par les outils et ceux qui ne sont pas explicitement prévus et que les utilisateurs « induisent » par exploitation des affordances de l'outil. Les usages prévus sont largement étudiés dans toutes les méthodes de conception de systèmes informatiques. Les usages induits non prévus sont quant à eux naturellement bien moins envisagés. Pourtant les technologies de l'information et de la communication sont des « lieux » de contournement d'usage par excellence. Par exemple, tout le monde à déjà fait l'expérience d'interroger un moteur de recherche pour voir à quel rang de classement dans les réponses une page ou un site connu figure. Dans ce cas, l'objectif n'est pas de rechercher un site étant donné qu'on le connaît déjà ; c'est plus essayer de mesurer une certaine accessibilité d'un site, voir une popularité. On pourrait donc penser qu'il y a là un usage détourné ou contourné car n'étant pas explicitement celui prévu. Pourtant, c'est certainement un usage très fréquent.

Pour tenter de préciser les choses, nous pourrions essayer de dresser une typologie des usages du numérique observés dans environnements de travail. Une première entrée en matière serait de chercher à mieux identifier ce qui relève précisément d'un détournement et d'un contournement. A titre d'hypothèse qu'il conviendrait de valider par observation d'usages, nous proposons une piste de distinction entre détournement et contournement :

- Quand on utilise un outil globalement prévu pour faire quelque chose et qu'on le contraint ou qu'on le limite consciemment à faire quelque chose de différent de ce pour quoi il est prévu, il y aurait alors détournement d'usage. Par exemple, s'envoyer un e-mail pour garder trace d'une idée ou réflexion à la façon d'un boc-note (la fonctionnalité principale d'envoi de message étant quasiment vu comme un résultat collatéral). Nous allons donner deux autres exemples de détournements ci-dessous dans le domaine du TAL.
- Quand on prend un outil pour faire ce qu'il permet de faire mais que c'est un moyen, peu ou pas conscient, d'accéder à autre chose, une autre forme de résultat, alors il y aurait contournement d'usage. Par exemple, quand on forme des étudiants à l'usage des traitements de textes (notamment dans le cadre du C2i) on s'aperçoit qu'avant d'être formé aux techniques de mise en forme et de styles quasiment tous utilisent la touche de retour à la ligne pour réaliser de l'espacement avant et après les paragraphes (on observe le même genre de contournement avec l'utilisation de la touche espace en lieu et place d'un retrait). Un autre exemple est assez visible chez les jeunes dans leurs usages de la téléphonie mobile. Une étude[85] récente de l'ARCEP (Autorité de régulation des communications électroniques et des postes) montre qu'en moyenne les adolescents envoient 2500 SMS par mois, ceci revenant à un peu plus de 80 SMS par jour. L'étude montre que les jeunes interrogés déclarent en majorité passer assez peu de coups de

85 cf. http://etudiant.lefigaro.fr/vie-etudiante/news/detail/article/les-ados-envoient-2500-sms-par-mois-196/ consultée le 09/12/2012.

téléphone (et qui en général restent assez courts). On peut donc voir ici une forme de contournement d'usage où le téléphone pensé initialement pour principalement téléphoner se retrouve aujourd'hui principalement utilisé par certains utilisateurs comme un moyen d'envoyer des messages, des *tweets*, etc … Bien sûr ce contournement est certainement dû à la fois aux usages et aux pratiques de tarification des opérateurs (formules « temps de conversation limité et SMS illimités »). La question serait ici de savoir si c'est les pratiques commerciales qui engendrent les contournements ou le contraire.

Des cas de détournements d'usages en TAL sont bien connus et finalement assez fréquents. Dans une réflexion sur l'outillage de la linguistique par des outils du TAL (Habert 2006), Benoit Habert aborde cette question en relatant deux exemples d'outils qui se sont au final retrouvés détournés :

- Le premier exemple est celui de l'étiquetage morpho-syntaxique. Cordial, développé par la société Synapse, était au départ un correcteur orthographique et grammatical. A la suite d'une interaction avec des équipes universitaires (notamment dans le cadre de l'action GRACE d'évaluation des étiqueteurs grammaticaux), la société Synapse a rendu disponible une version « université » permettant l'étiquetage de corpus. Cet étiqueteur s'est même retrouvé appliqué à de l'oral là où Cordial visait l'écrit. La transcription orthographique s'avère rapidement insuffisante dans le traitement de l'oral car les spécificités de l'oral (répétitions, absence de ponctuation) peuvent amener beaucoup d'erreurs si l'on applique « brutalement » à l'oral des outils développés pour l'écrit. A. Valli et J. Véronis (Valli & Véronis, 1999) ont « détourné » Cordial alors qu'il a été conçu pour l'écrit. Un pré-traitement modifie la transcription manuelle (enlèvement des pauses silencieuses ou remplies, des amorces de mots interrompus, des indications d'événements non linguistiques, etc.). Cordial étiquette alors cette version aménagée. Un post-traitement aligne enfin les formes conservées dans la version aménagée et leurs étiquettes avec la version de la transcription manuelle. On obtient alors un étiquetage partiel mais fiable pour les mots étiquetés.
- Un autre exemple est un détournement de l'outil Lexter (Bourigault 1993) qui isole des syntagmes nominaux, principalement de type « *N1 de N2* ». Lexter a été développé initialement pour l'extraction de terminologie dans des corpus très spécifiques au sein d'EDF, ces corpus demandant par leur nature une robustesse de traitement très forte. C'est cette robustesse qui a été l'objet du détournement car plusieurs chercheurs en TAL, notamment via le groupe de travail Terminologie et Intelligence Artificielle, (TIA[86]) ont utilisé Lexter comme un analyseur syntaxique limité (au groupe nominal) mais permettant une grande plage d'utilisation en genre. Lexter, ainsi détourné, à donné lieu à un autre outil, Syntex (Bourigault & Fabre, 2000), un analyseur syntaxique en dépendances.

La question des contournements des usages (notamment dans les ENT) est moins facilement observable que les détournements car ils sont plus fortuits. Comparativement, les cas de détournements sont des démarches souvent techniques donc réfléchies et anticipées, là où les détournements sont beaucoup moins conscientisés. Les usages grand public des outils numériques se prêtent certainement beaucoup plus à des cas de contournements qu'à des détournements (c'est une hypothèse à vérifier).

Les contournement d'usage sont, à notre connaissance, un domaine assez peu exploré. Il nous semble que la démarche centrée-utilisateur que nous défendons, construite sur les apports épistémologiques de la sémantique interprétative et de l'énaction, est un cadre intéressant pour étudier scientifiquement cette question des contournements des usages. Le contournement comme une preuve

86 http://tia.loria.fr/TIA/ (consultée le 14/06/11)

de couplage et donc d'appropriation par l'utilisateur est plus une réalité empirique qu'un paradoxe.

L'étude des contournements n'est certainement pas sans lien avec le courant de la sérendipité qui étudie les cas de découvertes scientifiques fortuites. Van Andel et D. Bourcier (Van Andel & Bourcier 2009) décrivent un nombre impressionnant de découvertes scientifiques où quelque chose d'inattendu intervient à un moment dans un processus qui au final donne lieu à une découverte (on peut citer par exemple les inventions de l'hélice de bateau, l'imprimante à jet d'encre ou le four à micro-ondes). Mais la sérendipité n'est pas uniquement le fruit du hasard. La découverte par sérendipité repose sur le fait qu'un hasard n'est rien tant qu'il n'est pas perçu comme tel par les capacités de curiosité, d'observation, d'ingéniosité et de questionnement du chercheur. A la manière dont une trace d'animal dans la terre est sémiotisée comme trace par un chasseur qui l'identifie, le hasard n'est reconnu comme hasard que parce qu'il est sémiotisé comme tel par le chercheur. Ainsi, au cœur de la découverte par sérendipité réside la capacité interprétative comme une forme d'intelligence sémiotique. Cette interprétation ne reconnaît pas de représentations dans l'absolu, elle les créé par « arrachement » à un continuum perceptif. De ce point de vue, il y aurait bien une forme de processus interprétatif et énactif à l'œuvre dans la sérendipité et probablement de la même façon dans les contournements d'usage.

Une perspective est de chercher des protocoles expérimentaux interrogeant les contournements d'usage dans les ENT. Il y a là une problématique d'une grande richesse scientifique qui par nature est pluridisciplinaire. Une hypothèse que nous pourrions chercher à valider serait de savoir si, finalement, les utilisateurs ne sont pas plus attachés (du fait du couplage) à leurs environnements de travail pour les contournements qu'ils offrent, plus que pour leurs fonctionnalités natives.

Cette perspective de recherche est une des dimensions d'un projet qui démarre au GREYC et qui fait l'objet d'un financement ANR. C'est le projet ART-ADN (Accès par Retour Tactilo-oral Aux Documents Numériques) coordonné par Fabrice Maurel. Ce projet vise la construction et l'expérimentation de dispositifs innovants d'accès non visuel aux textes, principalement pour des utilisateurs non-voyants. L'idée est de développer une interface haptique capable de générer automatiquement, à partir du Web, une transformation intelligente des contrastes lumineux émis par l'écran en vibrations tactiles. C'est-à-dire donner à un sujet à expérimenter une représentation tactile « vibrante » d'une page web. Les différentes étapes de ce processus sont :

- l'extraction de la structure visuelle des documents ;
- l'augmentation de la réalité du document numérique par un environnement contrasté construit à partir de la structure extraite dans l'étape précédente ;
- le contrôle de l'interaction sur tablette numérique par le survol digital du document augmenté, via des effecteurs vibrants placés sur une partie sensible de l'usager ;
- l'évaluation du couplage homme-machine, généré par cette nouvelle possibilité d'interaction, entre le non-voyant et le document à travers cette « prothèse sensorielle ».

Des premières expérimentations prometteuses d'un prototype ont été réalisées (Maurel & al. 2012).

La question des usages d'une interface haptique pour la consultation de pages Web est bien plus large que l'application visée pour le public non-voyant. En fonction des contextes d'usages tout

utilisateur même voyant peut se trouver avec ses capacités visuelles non mobilisables (exposition en plein soleil, conduite automobile … etc). Dès lors il est fort probable que des usages non attendus puissent émerger éventuellement même par contournement de l'interface haptique. C'est ce qu'il sera notamment intéressant d'évaluer dans une phase d'expérimentation en contexte écologique. Nous sommes ici dans une démarche de conception d'interactions homme-machine où l'on conçoit d'abord pour expérimenter ensuite avec des retours éventuels, ce qui tranche avec les modes de conception où l'on expérimente d'abord pour cadrer la conception (par exemple les méthodes de magicien d'Oz en dialogue homme-machine). En cela la démarche s'apparente aux approches nouvelles de conception informatique constatées dans le domaine des environnements d'apprentissage éducatifs dites *Design-Based-Research*[87].

7.4. Approche centrée utilisateur et intelligence collective

Comme nous avons déjà eu l'occasion de l'écrire dans ce rapport, l'approche centrée-utilisateur n'est absolument pas synonyme de l'approche de l'utilisateur seul. Chaque utilisateur est, d'une certaine façon, dans ses centres d'intérêts et ses spécificités le produit de ses inter-relations avec les autres. C'est notamment quelque chose qui est tout à fait flagrant dans l'usage des réseaux sociaux sur le Web. F. Martin-Juchat & J. Pierre (Martin-Juchat & Pierre 2011) relatent à ce titre une observation sur Facebook où, en moyenne, les utilisateurs ont 150 « amis » dans leurs relations[88].

Dans les milieux industriels notamment, les spécificités des utilisateurs et leurs connaissances métiers ont un aspect stratégique évident. Si une structure veut se constituer en affirmant une particularité et une culture d'entreprise qui la différencie des autres, il est évident que cela doit se construire via un partage de points communs entre les collaborateurs. De même, la transmission des connaissances en cas de changement des membres d'une équipe participe de cet objectif. L'approche centrée-utilisateur peut aider à favoriser le partage et la transmission des connaissances métiers et des points de vue. Quand un utilisateur doit consacrer une partie de son temps à réifier ses centres d'intérêts et ses connaissances pour alimenter son ENT, il en découle que cette réification devient explicitée et partageable. L'échange avec d'autres à propos, par exemple, d'une ressource terminologique personnelle peut amener l'utilisateur à se rendre compte de positions implicites et aider les autres à mieux cerner la façon de voir de l'utilisateur. Ainsi les environnements numériques qui sont construits par les utilisateurs peuvent aussi agir pour les utilisateurs collectivement en leur permettant d'argumenter leurs choix dans un partage avec les autres.

L'intelligence artificielle, le TAL et plus largement l'informatique ont souvent pris pour principe d'apporter une forme d'intelligence aux utilisateurs. Pourtant, il nous semble que dans bon nombre d'usages, l'intelligence que les utilisateurs attendent des outils numériques, ils l'ont en eux et ne peuvent même pas l'attendre d'autrui. Une perspective de recherche qui s'ouvre avec la problématique des ENT est d'apporter des solutions pour permettre à l'utilisateur de trouver en lui et dans ses relations avec les autres ce dont il a besoin. Ainsi, nous affirmons encore que l'informatique (de même que la linguistique d'ailleurs) doit se trouver impliquée dans les questions sociétales autour du numérique, non pas comme « fournisseur de moyens techniques » mais plutôt comme champ de recherche à forte dimension sémiotique. Un champ de recherche dans lequel nous proposons de creuser le sillon d'une stimulation par cas d'usages induits de « l'agir interprétatif » dans les environnements numériques,

87 Cf http://www.designbasedresearch.org/ par exemple (consultée le 15/12/12)

88 Un anthropologue, R. Dunbar, a évalué (sur la base de la taille du néocortex, ce qui n'est peut-être pas le critère le plus pertinent ici …) que l'homme peut difficilement entretenir directement (et réellement) plus de 148 relations sociales (Dunbar 1993).

pour soi et pour les autres.

~

Pour toutes ces perspectives de prolongements du travail de recherche, de développement logiciel et d'expérimentation, la quantité de travail à fournir reste bien entendu considérable (et c'est une chance). Il est dès lors fort probable que plusieurs sujets de thèses (en informatique bien sûr mais pas uniquement) pourront se dégager dans les années à venir. C'est en soi une raison suffisante pour avoir mené ce travail de préparation de l'habilitation à diriger des recherches.

8. Références

Adda G., Mariani J., Paroubek P., Rajman M., Lecomte J., 1999, *Métrique et premiers resultants de l'évaluation GRACE des étiqueteurs morpho-syntaxiques pour le français*, actes de TALN'99, p. 15-24.

Aristote. *Organon I. Les catégories*. Trad. J. Tricot, 1969, Coll. Bibliothèque des textes philosophiques. Ed. VRIN : Paris.

Assadi H., 1998, *Construction d'ontologies à partir de textes techniques – application aux systèmes documentaires,* Thèse de doctorat en Informatique - Université Paris 6 http://www.lip6.fr/lip6/reports/1998/lip6.1998.048.ps.tar.gz (consultée le 11/10/12).

Bach-y-Rita P., 1972, *Brain mechanisms in sensory substitution*, New York, Academic Press.

Bachimont B., 2000, *Connaissance et support d'inscription : entre raison graphique et raison computationnelle*, Septième École d'été de l'ARCo, Bonas, 10-21 juillet 2000.

Bassi Acuña A., 1995, *Un modèle dynamique de la compréhension de textes intégrant l'acquisition de connaissances*, Thèse de l'Université Paris XI, Orsay.

Beaudouin N., 2008, *Problèmes d'ergonomie linguistique et traitement d'images : une approche socioterminologique*. Thèse de doctorat en linguistique de l'université de Rouen.

Beaudouin-Lafon M., 2004, *Designing Interaction, not Interfaces*, Proceeding of the working conference on Advanced Visual Interfaces, May 25-28.

Beust P., 1998, *Contribution à un modèle interactionniste du sens. Amorce d'une compétence*

interprétative pour les machines, Thèse en Informatique de l'Université de Caen Basse-Normandie.

Beust, P., 2002, *Un outil de coloriage de corpus pour la représentation de thèmes*. Actes des 6èmes Journées internationales de l'Analyse statistique de Données Textuelles (JADT 2002), 1:161–172.

Beust P., Breux S., Roussel H., Roy T., 2008, *Une expérimentation pluridisciplinaire sur le suivi du regard dans une démarche de conception de logiciel*, Journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels) "Expérimentation et sytèmes complexes" Megève, France.

Bommier-Pincemin, B., 1999, *Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Paris IV Sorbonne.

Bourigault, D., 1993, *« Analyse syntaxique locale pour le repérage de termes complexes dans un texte »*, Revue TAL, 34(2).

Bourigault D. et Aussenac-Gilles N., 2003, *Construction d'ontologies à partir de textes*, Actes de Traitement Automatique des Langues Naturelles (TALN), Tome 2, pp. 27-47.

Bourigault, D. & Fabre, C., 2000, *« Approche linguistique pour l'analyse syntaxique de corpus »*. Cahiers de grammaire, (25), p.131-151.

Bourion, E., 2001, *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Nancy 2.

Bouroche, J.-M. et Saporta, G., 1980, *L'analyse des données*. Presses Universitaires de France, Paris.

Brassac C., Stewart J., Février 1996, *Le sens dans les processus interlocutoires, un observé ou un co-construit ?*, Cinquièmes Journées de Rochebrune "Du collectif au social", Megève.

Carron T., 2011, Observation dans les Environnements Informatiques pour l'Apprentissage Humain, Rapport d'Habilitation à Diriger de Recherches en Informatique de l'université de Chambéry, https://publications.lip6.fr/index.php/publications/show/8934 (consultée le 15/12/12).

Cassirer E., 1991, *Logique des sciences de la culture*, tr. J. Carro et J. Gaubert, Paris, Cerf.

Charlet J., Laublet P., Reynaud C,. 2003, *Web Sémantique*. Rapport de l'Action Spécifique 32 CNRS / STIC.

Chomsky N., 1971, *Aspects de la théorie syntaxique*, Seuil, ISBN 2020027402

Clark H. H., Wilkes-Gibbs D., 1986, *Referring as a Collaborative Process*, Cognition, n° 22, p. 1-39.

Claveau V., 2003, *Acquisition automatique de lexiques sémantiques pour la recherche d'information*, Thèse de doctorat en Informatique, Université de Rennes 1.

Condamines A. (sous la direction de), 2005, *Sémantique et corpus*, Hermès, Paris, ISBN : 2-7462-

1055-X.

Coursil J., 1992, *Grammaire analytique du français contemporain. Essai d'intelligence artificielle et de linguistique générale*, Thèse en Informatique de l'Université de Caen Basse-Normandie.

Coursil, J., 2000, *La fonction muette du langage*. Ibis Rouge Editions, Presses Universitaires Créoles, Petit-Bourg (Guadeloupe), ISBN 2-84450-090-0.

De Loor P., Tisseau J., 2011, Réalité virtuelle et énaction. Journal de l'Assocation Francaise de Réalité Virtuelle, 10:1–4, http://hal.archives-ouvertes.fr/docs/00/60/39/93/PDF/DeLoorTisseauJAFRV2011.pdf (consultée le 15/12/12)

Dionisi D., Labiche J., 2006, *Enaction et informatique : les enjeux de l'opérationnalisation technologique d'une théorie de la cognition*, in Actes du colloque ARCo 2006, 6 au 8 Décembre 2006 – Bordeaux.

Dumais S., 1988, *Using Latent Semantic Analysis to improve access to textual information*, in Proceedings of the Conference on Human Factors in Computing Systems (CHI'88), New York, ASSN for Computing Machinery, p. 281-285.

Dunbar R., 1993, *« Coevolution of Neocortical Size, Group Size and Language in Humans »*, Behavioral and Brain Sciences, n°16, p. 681-694.

Eco U., 1988, *Sémiotique et philosophie du langage*, Paris, Presses Universitaires de France.

Fabre C., 2010, *Affinités syntaxiques et sémantiques entre les mots - apports mutuels de la linguistique et du traitement automatique des langues*, Mémoire d'habilitation à diriger des recherches en linguistique, Université Toulouse 2 Le Mirail.

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 Université de Paris Sud.

Ferrari S., 2010, *Études pour le Traitement Automatique des Langues. De la rhétorique des figures à la rhétorique du discours*, mémoire d'habilitation à diriger des recherches, Université de Caen.

Ferrari S., 1997, *Méthode et outils informatiques pour le traitement des métaphores dans les documents écrits*, Thèse de l'Université Paris XI, Orsay.

Fodor J., 1975, *The Language of Thought*, Harvard University Press, ISBN 0-674-51030-5

Gapenne O., Gaussier P., 2005, *Suppléance perceptive et interface : une introduction*, Arob@se, vol. 1, pp. 1-7, revue en ligne : http://www.univ-rouen.fr/arobase/ consultée le 07/01/13.

Gibson J.-J., 1977, *The Theory of Affordances*. In Perceiving, Acting, and Knowing, Eds. Robert Shaw and John Bransford, ISBN 0-470-99014-7

Gosselin L., 1996, *Sémantique de la temporalité en français, un modèle calculatoire et cognitif*, Louvain-la-neuve, Duculot.

Greimas A. J., 1966, Sémantique structurale, Paris, Larousse (ré-édité aux PUF en 1986).

Habert B., 2005, *Portrait de linguiste(s) à l'instrument*, Revue Texto! en ligne, décembre 2005, vol. X, n°4. Disponible sur : http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html (Consultée le 07/1/13).

Harris Z., 1951, *Methods in Structural Linguistics*, University of Chicago Press, Chicago.

Hearst, M. A., 1995, *TileBars: Visualization of Term Distribution Information in Full Text Information Access*. In proceedings of CHI'95, the Conference on Human Factors in Computing Systems CHI'95. ACM. Pp59-66.

Heutte J., 2011, *La part du collectif dans la motivation et son impact sur le bien-être comme médiateur de la réussite des étudiants. Complémentarités et contributions entre l'autodétermination, l'auto-efficacité et l'autotélisme*. Thèse de doctorat en Sciences de l'éducation, Université Paris Ouest Nanterre La Défense.

Hjelmslev L., 1943, *Prolégomènes à une théorie du langage*, trad. française, 1968, Paris, Ed. de Minuit.

Holzem M. Labiche J., 2010, *« En marchant se construit le chemin » : manifeste pour une approche culturelle du couplage sujet-environnement numérique de travail*, L'homme sémiotique, Namur 19-22 avril 2010.

Husserl E., 1913, *Idées directrices pour une phénoménologie (Ideen I)*, traduction Paul Ricœur, Gallimard.

Indurkhya B., 1987, *Approximative semantic transference : a computational theory of metaphors and analogy*. Cognitive Science, 11(4):445-480.

Jacquet G. 2005, *Polysémie verbale et calcul du sens*, Thèse de doctorat en Sciences Cognitives, Paris, EHESS, http://www.sudoc.fr/151609055

Jakobson R., 1963, *Essai de linguistique générale*, trad. Ruwet, Éditions de Minuit, Paris

Jakobson R., 1981, *Eléments de linguistique générale (1 et 2)*, Éditions de Minuit, Collection Double, Paris, ISBN 2-70730-579-0.

Kanellos I., Mauceri C., 2008, *Une conscience interprétative face à un univers de textes. Arguments en faveur d'une analyse de données interprétative*. In Syntaxe & Sémantique, vol 9. « Textes, documents numériques, corpus. Pour une science des textes instrumentée », J. François et N. Le Querler eds, ISBN 978-2-84133-318-9, p. 37-52.

Lakoff G., Johnson M., 1985, *Les métaphores dans la vie quotidienne*, Paris, Éditions de Minuit.

Lamel L., Rosset S., Gauvain J. L., Bennacef S., Garnier-Rizet M., Prouts B., 2000, *The LIMSI Arise system*, IVTTA'98 Interactive Voice Technology for Telecommunications Applications. Workshop N°4 (29/09/1998), Torino, Italie, vol. 31, n°4, pp. 339-353.

Lancieri L., 2009, *De l'analyse de traces à l'exploitation des phénomènes d'intelligence collective*, chapitre du livre,Analyse de traces et Personnalisation des EIAH,Traité IC2 - Série Informatique ed Hermes Science. Coordonné par A. Mille et J.C. Marty, ISBN: 2-7462-2120-9; Pdf

Landauer T.K., Foltz P. W. et Laham D., *Introduction to Latent Semantic Analysis*, dans Discourse

Processes, vol. 25, 1998, p. 259-28

Lavenus K., Lapalme G., 2002, *Evaluation des systèmes de question réponse*, revue TAL, Vol. 43, n°3/2002, p. 181-208.

Lehuen J., 1997, *Un modèle de dialogue dynamique et générique intégrant l'acquisition de la compétence linguistique*, Thèse de doctorat en Informatique de l'université de Caen Basse-Normandie.

Lenat D., 1986, *Understanding Computers: Artificial Intelligence*. Amsterdam: Time-Life Books. p. 84. ISBN 0-7054-0915-5.

Loiseau, S., 2006, *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*, Thèse de doctorat, Paris X-Nanterre.

Louçã J., Rodrigues D., 2011, *The Observatorium. Observation et analyse de réseaux de communication a grande échelle*. Journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels) "Echelles et modèlisations multi-niveaux", Megève, France, à paraître aux éditions "Les chemins de traverse".

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

Martin-Juchat F. & Pierre J., 2011, *Facebook et les sites de socialisation : une surveillance librement consentie*, dans *L'Homme trace*, Perspectives anthropologique des traces contemporaines, sous la direction de Béatrice Galinon-Mélénec, CNRS ÉDITIONS, Paris, 2011, pp 105-125 , ISBN : 978-2-271-07104-0

Maturana H.R., Varela F.G., 1973, *De máquinas y seres vivos*, [en anglais "Autopoiesis: the organization of the living", in Autopoiesis and Cognition by Maturana H.R. and F.G. Varela, rédédition 1980].

Mauceri, Chr., 2007, *Indexation et isotopie : vers une analyse interprétative des données textuelles*, Thèse de doctorat, ENSTB.

Mauger S., 1999, L'interprétation des messages enigmatiques. Essai de semantique et de traitement automatique des langues, Thèse de doctorat en linguistique de l'université de Caen.

Mauger S. Luquet P.-S., 2005, Réflexivité dans un processus d'interaction homme-machine. Exemple de résolution de charades assistée par ordinateur, Journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels) "Réflexivité et auto-référence", Megève, France. Publication de l'ENST 2005 S 001.

Mauger S., 2007, *La catastrophe, Alpha et Omega du signe et du sens*, Journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels) "Catastrophes, discontinuités, ruptures, limites, frontières" Megève, France. Publication de l'ENST ISSN 1242-5125 ENST S (Paris).

Mauger S., 2009, *Concertation, dialogue et illusion interprétative*, Actes du colloque de l'ARCo « Interprétation et problématiques du sens », Rouen, 9-11 décembre 2009, p. 41-48.

Maurel F., 2004, *Transmodalité et multimodalité écrit/oral : modélisation, traitement automatique et évaluation de stratégies de présentation des structures « visuo-architecturales » des textes*,

Thèse de doctorat en Informatique de l'université de Toulouse 2, https://maurelf.users.greyc.fr/documents/envrac/these.pdf

Maurel F., Dias G., Routoure J.-M., Vautier M., Beust P., Molina M., Sann C., Haptic Perception of Document Structure for Visually Impaired People on Handled Devices, Special Track on Involving People in Web Accessibility Evaluating and Accessibility Solutions, DSAI 2012, Publication in Procedia Computer Science, ELSEVIER.

Mayaffre D., 2002, *Les corpus réflexifs : entre architextualité et hypertextualité*, Corpus, n°1, novembre 2002, revue en ligne : http://corpus.revues.org consultée le 07/01/13.

McCulloch W., Pitts W., 1943, *A Logical Calculus of Ideas Immanent in Nervous Activity*, Bulletin of Mathematical Biophysics 5:115-133.

Merleau-Ponty M., 1945, *La Phénoménologie de la perception*, Paris, Gallimard.

Michael D., Chen S., 2005, *Serious Games. Games that educate, train and inform*, Boston MA: Course Technology PTR

Mille A., Prié Y., 2006, *Une théorie de la trace informatique pour faciliter l'adaptation dans la confrontation logique d'utilisation/logique de conception*. in 13eme Journées de Rochebrune - Traces, Enigmes, Problèmes : Emergence et construction du sens - Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels, jan 2006, Rochebrune, 12 pp.

Minsky M., 1975, *A framework for representing knowledge*, In P. Winston ed., The psychology of Computer Vision, New-York, McGraw-Hill.

Montague R., 1970, *English as a formal language*, Formal philosophy, Yale university press, , p. 188.

Morand B., 2004, *Logique de la conception. Figures de sémiotique générale d'après Charles S. Peirce*, L'Harmattan, ISBN : 2-7475-6366-9.

Nazarenko A., 2004, *Donner accès au contenu des documents textuels Acquisition de connaissances et analyse de corpus spécialisés*, mémoire d'habilitation à diriger des recherches, Université Paris-Nord.

Névéol, A., 2005, *Automatisation des tâches documentaires dans un catalogue de santé en ligne*. Thèse de Doctorat en Informatique, INSA de Rouen, Rouen.

Newel A., 1982, *The Knowledge Level*, Artificial Intelligence, Vol. n°18, p. 87-127

Nicolle A., 1996, *L'expérimentation et l'intelligence artificielle*, Intellectica, n° 22, p. 9-19, Association pour la Recherche Cognitive (ARC).

Nicolle N., 2005, *Comparaison entre les comportements réflexifs du langage humain et la réflexivité des langages informatiques*, 12e journées de Rochebrune (Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels) "Réflexivité et auto-référence", Megève, 24-28 Janvier. S. Stinckwitch (Ed.) Paris, ENST. (ENST 2005 S 001).

Nicolle, A., Beust, P., et Perlerin, V., 2002, *Un analogue de la mémoire pour un agent logiciel*

interactif. In Cognito. 21. pp.37-66.

Peirce C. S., 1978, *Écrits sur le signe*, rassemblés, traduits et commentés par Gérard Deledalle, Paris, Seuil.

Perlerin, V. et Beust, P., 2003, *Pour une instrumentation informatique du sens*. In Variation, construction et instrumentation du sens, pages 197–229. Ed. M. Siksou, Hermes, Paris.

Perlerin, V., 2004, *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat en Informatique – Université de Caen Basse Normandie (http://www.revue-texto.net/1996-2007/Inedits/Perlerin/Perlerin.html consultée le 7/01/13).

Petitmengin C., 2007, *Découvrir la dynamique de l'expérience vécue*, Bulletin de psychologie, t. 60, p.114-118.

Pichon, R., Sébillot, P., 1999, *Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience*. In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'1999), pages 279–288.

Pinte J.-P., 2011, *Le Web invisible : l'antre du cybercrime*. Dossier n°70 « Police scientifique », revue « Pour la Science », Janvier-Mars 2011.

Poudat, C., 2006, *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, Thèse de doctorat, Orléans.

Quilian M., 1968, *Semantic memory*, in M. Minsky : Semantic Information Processing, Cambridge, MIT Press, p. 216-270.

Raccah P. Y., 1997, *L'argumentation sans la preuve : prendre son biais dans la langue*, Cognition et Interaction, volume 2, n° 1-2, Nancy.

Rastier, F., 2001, *Arts et sciences du texte*. Presses Universitaires de France, Paris.

Rastier F, 2002, *Anthropologie linguistique et sémiotique des cultures*, in Rastier F. et Bouquet S. (sous la dir.), Une introduction aux sciences de la culture, Paris, PUF.

Rastier F., 1987, *Sémantique interprétative,* Paris, Presses Universitaires de France.

Rastier F., 2011a, *Web Sémantique, surdocumentarisation et complexité — Situation et propositions*, à paraître.

Rastier F., 2011b, *La mesure et le grain. Sémantique de corpus*. Paris : Champion, Collection Lettres numériques, ISBN 978-2-7453-2230-2.

Rastier F., Cavazza M., Abeillé A., 1994, *Sémantique pour l'analyse*, Paris, Masson.

Rastier, F., 2006, *Formes sémantiques et textualité*, Langages, n°163, Paris.

Reutenaeur C., Jacquey E. Lecolle M, Valette M., 2010, *Sémème au macroscope : genèse et variation sémiques d'une unité lexicale*, JADT 2010 : 10es Journées internationales d'Analyse statistique des Données Textuelles, Université de la SAPIENZA – Rome (Italie), 9-11 Juin

2010.

Rossignol, M., 2005, *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de doctorat, Rennes 1, disponible sur http://www.revue-texto.net consultée le 7/01/13.

Roussel H., Beust, P., Roy T., 2010, *Le comportement oculomoteur de l'utilisateur du logiciel de cartographie de corpus textuels ProxiDocs : Analyse de traces de la réussite ou de l'échec.* 2ème Atelier ICT (Interactions, Contextes, Traces), dans le cadre de la conférence INFORSID 2010, Le 25 mai 2010, Marseille.

Roussel H., Beust, P., Roy T., 2011, *Analysis of the oculomotor behaviour of Proxidocs users*. Revue ERAP (European Review of Applied Psychology), à paraître.

Roy T. Ferrari S., 2008, *User Preferences for Access to Textual Information: Model, Tools and Experiments,* livre Advances in Semantic Media Adaptation and Personalization, Studies in Computational Intelligence, Springer Verlag, pp 285 à 305.

Roy T., 2007, *Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires*, Thèse de doctorat en Informatique – Université de Caen Basse Normandie (http://roythibault.free.fr/these/index.html consultée le 7/01/13)

Roy, T. et Beust, P., 2004, *Un outil de cartographie et de catégorisation thématique de corpus*. In Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data, volume 2, pages 978–987.

Roy, T. et Beust, P., 2005, *La cartographie thématique de corpus : une solution aux problèmes de veille documentaire ?* In Chapitre français de International Society for Knowledge Organization (ISKO-France 2005), pages Article tiré en dehors des actes, distribué au début de la conférence. Version électronique : http://roythibault.free.fr/rech/ROY_BEUST_ISKO_2005.pdf consultée le 7/01/13.

Roy, T. et Névéol, A.,2006, *Cartographie d'un corpus de domaine médical*. In Actes des XIIIèmes Rencontres de la Société Francophone de Classification (SFC'06), pages 185–189, Metz, France.

Roy, T., Beust, P. et Ferrari, S., 2007, *User-centered analysis of corpora using semantic features redundancy*. In Proceedings of the fourth Corpus Linguistics Conference (CL'07), to appear. Birmingham.

Sabah G., 1996, *Le sens dans les traitements automatiques des langues — le point après 40 ans de recherches*, conférence invitée, journée ATALA du 14/12/96 « un demi-siècle de traitement automatique des langues », http://www.limsi.fr/Individu/gs/textes/ATALA-14.12.96/LePointSurLeSens.html consultée le 7/1/13.

Salem A., 1993, *Méthodes de la statistique textuelle, Thèse pour le doctorat d'État ès lettres et sciences humaines*, Université de la Sorbonne nouvelle - Paris 3, 3 vol, 998 p.

Salton G., M.J. McGill, 1983, *Introduction to modern information retrieval*, McGraw-Hill, ISBN : 0070544840.

Salton G. et Yang C.S., 1975, *A vector space model for automatic indexing*, Communication of the

ACM, vol. 18 (11), nov., p. 613-620.

Saussure F. de, 1915, *Cours de linguistique générale*, Paris 1986, Ed. Mauro-Payot.

Schuhl A. 2004, *Les ordinateurs de demain*, Éditions Le Pommier, Paris, ISBN 274650183X.

Sparck Jones K., 2001, *Automatic language and information processing : rethinking evaluation*, in Natural Language Engineering, Cambridge, Cambridge University Press, n°7, p. 29-46.

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 l'Université de Rennes 1.

Temperman G., De Lièvre B., Depover C., De Stercke J., 2012, Effets des modalités d'intégration d'un outil d'auto-régulation dans un environnement d'apprentissage collaboratif à distance, Actes de TICE 2012, Lyon, 11-13 décembre 2012, http://tice2012.univ-lyon1.fr (consultée le 13/12/12).

Thlivitis T., 1998, *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*, Thèse de l'Université de Rennes 1.

Berneers-Lee T., Hendler J., Lassila O., 2001, *The Semantic Web*, Scientific American, May 2001

Valette, M., 2004, *Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet. Approches sémantiques du Document Electronique*, In actes de CIDE7, le 7e Colloque International sur le Document Numérique. Europia. La Rochelle. pp.215-230.

Valette, M., Slodzian, M., 2008, *Sémantique des textes et Recherche d'Information*, Extraction d'information : l'apport de la linguistique, A. Condamines & Th. Poibeau, éds., Revue Française de Linguistique Appliquée (volume XIII-1 / juin 2008), 119-133.

Valli, A. & Véronis, J., 1999, *« Etiquetage grammatical des corpus de parole : problèmes et perspectives »,* Revue française de linguistique appliquée, IV(2), p. 113-133.

Van Andel P. & Bourcier D., 2009, *De la sérendipité dans la science, la technique, l'art et le droit. Leçons de l'inattendu*, Edition L'Act Mem.

Van Rijsbergen C.J., 1979, *Information Retrieval*, 2nd edition. University of Glasgow.

Varela F., 1996, *Invitation aux sciences cognitives*, Editions du Seuil.

Venant F., 2006, *Représentation et calcul dynamique du sens*, Thèse de doctorat en Sciences Cognitives, Paris, EHESS, http://www.sudoc.fr/160862752

Vivier J., 1992, Faire et dire ce qu'il faut faire pour ... Physique naturelle et discours : Projet d'étude developpementale. Diplôme d'habilitation à diriger des recherches, Université de Caen.

Vygotsky, L. S., 1978, *Mind in society*. Cambridge, MA: Harvard University Press.

W3C, 2001, *Synchronized Multimedia Integration Language (SMIL 2.0) – W3C Recommendation.* http://www.w3.org/TR/2001/REC-smil20-20010807/ consultée le 25/01/07.

Weizenbaum J., 1966, *ELIZA - a computer program for the study of natural language communication between man and machine*, In : Communication of the ACM, n. 9, pp. 26-45.

Zipf, H. G. K., 1949, *Human Behavior or the Principle of Least Effort.* Hafner Publishing Co., New-York, USA.

9. Index des auteurs

10. Annexes

10.1. Curriculum Vitae

10.1.1. Situation professionnelle

Depuis septembre 1999 (titularisé le 1/9/2000) : Enseignant-chercheur

- Maître de conférences en Informatique (27e section du CNU) au département LEA (Langues Etrangères Appliquées) de l'UFR des LVE (Langues Vivantes Etrangères) de l'Université de Caen Basse Normandie
- Membre du laboratoire d'informatique GREYC CNRS UMR 6072 (Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen), équipe DLU (Documents, Langues, Usages)
- Membre du pôle ModeSCo (Modélisations en Sciences Cognitives) de la Maison de la Recherche en Sciences Humaines (MRSH) de l'Université de Caen Basse Normandie
- Directeur du CEMU (Centre d'Enseignement Multimédia Universitaire) depuis février 2009, service commun de l'UCBN en charge de la FOAD et des TICE.

10.1.2. Cursus universitaire

- 1999 : Qualifications CNU en section 27 (Informatique) et en section 7 (Sciences du langage)
- 1998 :Thèse de doctorat en Informatique de l'Université de Caen Basse Normandie (UCBN)

- Titre : *Contribution à un modèle interactionniste du sens. Amorce d'une compétence interprétative pour les machines*
- Thèse soutenue le 22/12/1998, obtenue avec la mention très honorable
  - Jury : A. Nicolle (dir. de thèse), L. Gosselin (co-directeur), J. Sallantin (rapporteur),F. Rastier (rapporteur), B. Levrat (rapporteur), J. Coursil, J.M. Pierrel, B. Victorri.
- 1993-95 : Service national en tant que volontaire à l'aide technique attaché en tant qu'enseignant à l'université des Antilles et de la Guyane (Schoelcher, Martinique).
- 1993 D.E.A. Intelligence Artificielle et Algorithmique (UCBN)
- 1992 Maîtrise d'Informatique (UCBN)
- 1991 Licence d'Informatique (UCBN)
- 1990 DEUG de Mathématiques (UCBN)
- 1988 Baccalauréat Série C

10.1.3. Activités d'enseignement

Durant les dernières années j'ai eu l'occasion d'effectuer les enseignements suivants dans le cadre de mon service statutaire d'enseignant-chercheur :

- Au **département LEA** (UFR des LVE) :
  - 1e année de Licence LEA : CM et TD « Initiation, documents électroniques, applications bureautiques »
  - 2e année de Licence LEA : CM et TD « Modèles de calcul, structure et contrôle des données et initiation à l'algorithmique »
  - 3e année de Licence LEA : CM et TD « Gestion de bases de données et analyse et conception des systèmes d'information »
  - 1e année de Master Pro. LEA : CM et TD « Traitement automatique des langues et conception de site Web, Web 2.0 »
  - 2e année de Master Pro. LEA : CM et TD « Recherche d'information et veille documentaire sur Internet »
- Au **département Informatique** (UFR Sciences) :
  - 1e année de Master Informatique : TD « Systèmes à bases de connaissances »
  - encadrement de projets annuels d'étudiants en traitement automatique des langues.
- Au **département Linguistique** (UFR Sciences de l'Homme) :
  - 1e année de Master Recherche « Sciences du langage » : CM et TD « initiation aux outils et techniques de traitement automatique des langues ».
- Dans la filière **Administration Economique et Sociale AES** (Faculté de Droit et Sciences Politiques) :
  - 3e année de Licence AES : CM et TD de préparation au C2I (Certification Informatique & Internet).
- Au **CEMU**:
  - Enseignement dans le cadre de la certification C2I Niveau 1 : culture numérique, logiciels bureautiques, réseaux.
  - Formation des tuteurs C2I

En dehors de mes activités d'enseignement à l'université de Caen Basse-Normandie, je dispense quelques heures de formation en 2e année de BTS “Management” sur les bases de données et leurs applications sous MS-Access pour le compte de l'IFAG/AFTEC Caen (13 rue A. Cavelier - 14000 CAEN).

10.1.4. Responsabilités administratives

Depuis le début de ma carrière d'enseignant-chercheur à l'UCBN, j'ai assuré et je continue encore à assurer différentes responsabilités administratives liées à la recherche et à l'enseignement.

10.1.5. Administration de la recherche

- Membre d'un Comité de Sélection (COS) pour le recrutement d'un maitre de conférences en Informatique (2009).
- Membre titulaire de la commission de spécialistes n°27 de l’université de Caen Basse Normandie (2004-2008).
- Membre extérieur suppléant de la commission de spécialistes n°27 de l’université du Maine au Mans (2004-2008).
- Membre du conseil du laboratoire GREYC CNRS UMR 6072 (1999-2004).
- Membre du bureau du pôle ModeSCo (Modélisations en Sciences Cognitives) de la Maison de la Recherche en Sciences Humaines (MRSH) de l’Université de Caen Basse Normandie.
- Responsable du thème Interactions de l'équipe ISLanD (2004-2010)
- Responsable de l'axe Interactions du pôle Modesco de la MRSH (2000-2010)

10.1.6. Administration de l’enseignement

- Depuis février 2009 : **Directeur du Centre d'Enseignement Multimédia Universitaire**[89] (CEMU), Service commun de l'Université de Caen Basse-Normandie en charge de la formation à distance[90] (FOAD) et des technologies de l'information et de la communication appliquées à l'enseignement[91] (TICE) : http://cemu.unicaen.fr
- Depuis janvier 2007 : **Chargé de mission pour la mise en place du C2I**[92] (Certificat Informatique et Internet) sur toute l’université de Caen Basse Normandie et **correspondant national C2I-1** auprès de la MINES (Mission Numérique pour l'Enseignement Supérieur - MESR) pour l'UCBN

89 Le CEMU est un service regroupant 24 personnes parmi lesquelles des enseignants-chercheurs, des ingénieurs, des techniciens et du personnel administratif.

90 Le CEMU organise l'offre diplômante à distance de l'UCBN. Il gère une vingtaine de diplômes à distance du DAEU au master 2 pour environ 500 étudiants (dont environ un tiers en formation continue)

91 Parmi les actions en terme de TICE mises en place par le CEMU, on peut citer le déploiement sur toute l'université de la plateforme d'activités en ligne Moodle qui compte aujourd'hui plus de 15000 utilisateurs réguliers dont près 800 enseignants.

92 Cette année on compte 1625 étudiants inscrits au C2I-1, issus de toutes les composantes de l'UCBN ou directement inscrits au titre de la formation continue.

- Depuis 2010 : Membre du comité de pilotage Environnement Numérique de Travail (ENT) puis Système d'Information (SI) de l'UCBN.
- Depuis 2010 : Chargé de mission pour l'UCBN dans les actions n°3 (indexation de ressources pédagogiques) et 10 (accompagnement des utilisateurs) de l'Université Numérique en Région (UNR) RUNN (Réseau Universitaire Numérique Normand)
- Depuis 2010 : Membre élu du conseil d'administration de la Fédération Inter-universitaire de l'Enseignement à Distance (FIED)
- 2000-2008 : Représentant au Bureau du département LEA.
- 2000-2004 : Représentant à la « commission des nouvelles technologies » de l'UFR des LVE
- 2004-2006 : Responsable du parcours SCDD « Stratégies de Communication du Développement Durable » du Master 2 Professionnel du département LEA (Master Pro MPM « Management de Projets Multiculturels).
- 2003-2006 : Président du jury de la 1e année de licence LEA.
- 2004-2005 : **Directeur du département LEA**[93].
- 2003-2004 : Directeur adjoint du département LEA.
- 1999-2009 : Responsable de l'UE TIL14B (et coordinateur d'une équipe d'intervenants comptant 9 personnes universitaires et extra-universitaires)
- Depuis 1999 : Responsable de l'UE TIL34A (et coordinateur d'une équipe d'intervenants comptant 3 personnes universitaires)
- 2008-2009 : Enseignant référent de L1

10.1.7. Publications scientifiques

cf. https://beust.users.greyc.fr/Papiers/publi.html (consultée le 7/01/13)

93 Nb : le département LEA regroupe des formations pluridisciplinaires liant les langues et des matières d'application (Droit, Eco/Gestion, Informatique, Communication). Il totalise une trentaine d'enseignants titulaires et environ 700 étudiants.

Résumé :

Notre problématique de recherche est ancrée en Traitement Automatique des Langues (TAL). Au sein du TAL, nous nous intéressons à la conception centrée-utilisateur d'environnements où les ressources et les processus mobilisés sont avant tout construits autour et en fonction des attentes et capacités interprétatives de l'utilisateur. La conception centrée-utilisateur n'est pas une posture théorique mais c'est déjà une réalité dans des applications utilisées quotidiennement. C'est le cas des architectures Web 2.0 comme c'est également le cas des Environnements Numériques de Travail (ENT). Notre recherche vise à analyser, concevoir et expérimenter des applications centrées-utilisateur dans les ENT où les capacités interprétatives s'enrichissent des éléments d'interaction dans l'environnement. Ce faisant nous cherchons à faire enrichir le TAL d'interconnexions avec les Interactions Homme-Machine et les EIAH (Environnements Informatiques pour l'Apprentissage Humain).

La problématique de l'interprétation est ici omniprésente et elle nous incite à tirer des ponts entre disciplines : entre l'informatique et la linguistique, plus précisément le courant de la sémantique interprétative et entre l'informatique et les sciences cognitives, plus précisément le courant de l'énaction. L'interprétation dans un environnement numérique n'est pas dissociable d'un couplage personne-système et de l'action de l'utilisateur dans cet environnement. Il en découle que nos objets d'étude sont principalement des usages et même des contournements d'usages vertueux par sérendipité. Les perspectives de recherche ouvertes s'orientent donc naturellement vers une mise en valeur de « l'agir interprétatif » dans les environnements numériques.

Mots clés : Traitement Automatique des Langues, Sémantique Interprétative, Énaction, Couplage Personne-Système, Environnements d'Apprentissage, Usages.

Abstract :

This work relates how semiotics improves the field of Natural Language Processing (NLP). Within this context, we are particularly interested in the user-centered design of environments, where the resources and the processes are built based on the users' interpretative abilities. User-centered design does not follow a theoretical approach, but rather faces the reality of everyday workspaces usage. Many digital workspaces such as Web 2.0 architectures already follow this approach. Interactions between users and their workspaces are essential to foster emerging interpretative abilities and help their blooming. Our research aims to conceive, analyze and experiment these semiotic interactions. As such, we search how studies in NLP, Human-Computer Interaction and Workspaces for e-learning can interconnect.

The main focus of our research is on interpretation, which led us to take advantage of interdisciplinary. Within this context, Computer Science, Linguistics (especially the field of Interpretative Semantics) and Cognitive Sciences (especially the field of Enaction) must share their domain of expertise. The interpretation process in a digital workspace mainly relies on a loop between users' actions and perceptions within a given workspace. Therefore, our domain of research includes the analysis of many forms of usage as well as understanding different ways to produce results that had not been foreseen in the application design (mainly by serendipity). As such, our perspectives of research aim to show how to produce digital workspaces that give rise to sense emerging through the action-interpretation couple.

Keywords : Natural Language Processing, Semiotics, e-learning, Enaction, User-System Interaction, Digital Workspaces, Online Education Environments, Usage.