Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires

Thibault Roy

UNIVERSITÉ de CAEN/BASSE-NORMANDIE

U.F.R. : Sciences

ÉCOLE DOCTORALE : SIMEM

THÈSE

présentée par

Thibault Roy

et soutenue

le 17 octobre 2007

en vue de l'obtention du

DOCTORAT de l'UNIVERSITÉ de CAEN

spécialité : Informatique

(Arrêté du 7 août 2006)

Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires

MEMBRES du JURY

M. Benoît Habert	Professeur	ENS Lettres et Sciences Humaines de Lyon	(rapporteur)
M. Pierre Zweigenbaum	Directeur de recherche	LIMSI-CNRS	(rapporteur)
Mme. Adeline Nazarenko	Professeur	Université de Paris XIII	
M. Jacques Vergne	Professeur	Université de Caen	(directeur de thèse)
M. Pierre Beust	Maître de conférences	Université de Caen	(co-directeur de thèse)

Mis en page avec la classe thloria.

Remerciements

Tout au long de mes trois années de thèse, j'ai pris beaucoup de plaisir à travailler sur un sujet passionnant. Durant ce travail, j'ai eu la chance de rencontrer et de collaborer avec un grand nombre de personnes, venant de différentes disciplines. Ces rencontres m'ont beaucoup apporté, aussi bien sur un plan scientifique que personnel, et c'est avec beaucoup d'émotion que j'essaie d'exprimer quelques remerciements dans les lignes suivantes.

Mes premiers remerciements vont naturellement à Pierre Beust. Ses conseils, ses encouragements, sa compréhension devant les difficultés ont constitué un encadrement idéal, que je souhaite à chaque doctorant, aussi bien pour sa pertinence scientifique, que pour ses qualités humaines. Je tiens à remercier Jacques Vergne, tout d'abord pour ses enseignements qui m'ont fait découvrir le traitement automatique des langues, il y a un petit moment déjà, mais surtout pour son encadrement de mon travail de thèse, me faisant confiance et me laissant beaucoup de liberté.

Je remercie vivement Benoit Habert et Pierre Zweigenbaum de m'avoir fait l'honneur de rapporter mon travail de thèse. Leurs différents conseils et jugements ont permis d'apporter une grande valeur ajoutée à cette thèse. Un grand merci à Adeline Nazarenko pour l'intérêt qu'elle a porté à ce travail en acceptant de faire partie de mon jury de thèse.

Mes remerciements vont ensuite à mes collègues avec qui j'ai eu la chance de travailler durant cette thèse, notamment aux cours des différentes expérimentations. Je remercie ainsi Stéphane Ferrari pour sa sympathie et surtout pour son regard très pertinent sur mes travaux, regard qui m'a grandement aidé tout au long de ma thèse. Un grand merci à Aurélie Névéol pour l'intérêt qu'elle a porté à mon travail et pour sa disponibilité dans le travail commun que nous avons pu réaliser, malgré les kilomètres qui nous séparent. Merci à Nicole Clouet, Georges Ferone, Eric Bruillard et Marie-Laure Compant-Lafontaine pour m'avoir permis de collaborer avec eux et ainsi de découvrir avec beaucoup de plaisir les Sciences de l'Éducation. Je remercie également Henri Roussel et Stéphane Breux pour les échanges passés et, je l'espère, futurs, sur la problématique du suivi du regard.

Enfin, un grand merci à mes collègues de l'équipe ISLanD, aux membres du groupe de travail *Enaction*, à mes collègues et amis du GREYC et du département d'informatique, ils ont constitué un environnement de travail et d'échange très agréable.

Je remercie bien évidemment mes amis et ma famille pour leurs encouragements durant cette thèse.

Pour finir, j'envoie mille pensées à Vanessa qui m'a constamment soutenu pendant ce travail de thèse.

À Vanessa.

Table des matières

Introduction

Contexte général 1
Plan de la thèse 2

Chapitre 1
Accéder à l'information textuelle dans des ensembles documentaires

Introduction : l'utilisateur seul devant un univers de textes 6
1.1 La problématique de l'accès à l'information dans des ensembles documentaires . . 7
1.1.1 Notions utilisées et définitions adoptées 7
1.1.2 Définition(s) de l'*accès à l'information* 10
1.1.3 Méthodes existantes pour l'accès au contenu de documents et d'ensembles documentaires 11
1.1.4 Un même accès à l'information et au contenu d'ensembles documentaires pour tous ? 17
1.2 La place de l'utilisateur dans des tâches d'accès à l'information documentaire . . 18
1.2.1 La place traditionnellement donnée à l'utilisateur dans des tâches d'accès au contenu 18
1.2.2 Les inconvénients liés à une faible place laissée à l'utilisateur dans de tels systèmes 19
1.2.3 Modèles et traitements pour des analyses personnalisées d'ensembles documentaires 21
1.3 Visualisation interactive d'ensembles documentaires 25
1.3.1 Visualiser des informations complexes 25
1.3.2 Visualiser des textes et des ensembles de textes *via* des systèmes informatiques 28
1.3.3 La cartographie d'ensembles documentaires 34
Conclusion : vers une meilleure prise en considération de l'utilisateur et de ses interactions 40

Chapitre 2
Un modèle d'analyse interprétative centrée utilisateur d'ensembles documentaires

Introduction 42
2.1 La Sémantique Interprétative comme modèle d'analyse subjective d'ensembles documentaires 42
2.1.1 La question du sens et de la sémantique en TAL 42
2.1.2 La Sémantique Interprétative : principes et positionnement 43
2.1.3 Notions utilisées et différents paliers textuels de la SI 46
2.1.4 Travaux en TAL exploitant la Sémantique Interprétative 49
2.2 *AIdED* : Analyse Interprétative d'Ensembles Documentaires 51
2.2.1 Vue globale sur le modèle *AIdED* et les propositions qu'il intègre 52
2.2.2 Propositions de représentation de l'intertexte et du contexte extralinguistique 53
2.2.3 La détection des isotopies intra et inter-textuelles comme élément de base aux traitements 57
2.2.4 Des visualisations cartographiques interactives comme pour l'accès personnalisé au contenu d'ensembles documentaires 63
2.2.5 Utilisations du modèle *AIdED* 65
Conclusion : une nouvelle façon de percevoir l'accès au contenu d'ensembles documentaires 67

Chapitre 3
Instrumentation logicielle du modèle

Introduction 70
3.1 Généricité visée de la mise en outils et en instruments 71
3.1.1 Indépendance par rapport aux langues 71
3.1.2 Indépendance par rapport aux codages des caractères 72
3.1.3 Indépendance par rapport aux formats des ensembles documentaires . . . 73
3.1.4 Propriétés des ressources textuelles et des transformations produites . . . 73
3.2 Aides logicielles à la construction de RTO personnelles 76
3.2.1 Une approche centrée-utilisateur de la construction et la description de ressources lexicales 76
3.2.2 Généralités sur les outils proposés pour l'extraction et la description de lexies 77
3.2.3 *Memlabor* et *ThemeEditor* : outils existants pour l'extraction et la structuration de lexies 79
3.2.4 *VisualLuciaBuilder* : construction interactive de dispositifs *LUCIA* 83
3.2.5 *FlexiSemContext* : outil pour la mise en contexte de lexies et de sèmes . . 89

3.3 *ProxiDocs* : projections de RTO personnelles dans des ensembles documentaires . 93
3.3.1 Présentation et objectifs . . . 93
3.3.2 Traitements mis en œuvre . . . 95
3.3.3 Méthodes de représentation d'ensembles documentaires à partir de RTO personnelles . . . 95
3.3.4 Projection et classification de l'ensemble documentaire . . . 97
3.3.5 Construction des supports de visualisation interactive de l'ensemble documentaire . . . 99
3.3.6 Mise en œuvre logicielle au sein de la plate-forme *ProxiDocs* . . . 108
3.3.7 Intégration et utilisation de *ProxiDocs* au sein de projets d'enseignement et de recherche . . . 112
Conclusion : mise en instruments du modèle d'analyse à travers des logiciels interactifs 112

Chapitre 4
Expérimentations et évaluations du modèle *AIdED*

Introduction . . . 116
4.1 La problématique de l'évaluation en TAL . . . 116
4.1.1 Méthodes d'évaluation traditionnelles . . . 117
4.1.2 Vers d'autres méthodes d'évaluation pour des systèmes interactifs et/ou centrés-utilisateur . . . 118
4.1.3 Évaluer des systèmes centrés sur leurs utilisateurs . . . 120
4.2 Recherche documentaire personnalisée . . . 122
4.2.1 Assistance dans une recherche d'information généraliste sur Internet . . . 122
4.2.2 Assistance dans une recherche d'information médicale . . . 128
4.3 Analyse d'expressions métaphoriques . . . 136
4.3.1 Étude de métaphores conceptuelles . . . 136
4.3.2 Ressources et corpus . . . 138
4.3.3 Analyses réalisées . . . 139
4.3.4 Vers une nouvelle phase du projet IsoMeta . . . 152
4.4 Étude de forums de discussion pédagogiques . . . 154
4.4.1 Observation de l'acquisition de l'identité professionnelle . . . 155
4.4.2 Observation des usages d'une terminologie professionnelle . . . 160
4.4.3 Apports de vues globales et interactives dans l'accès au contenu de forums pédagogiques . . . 166
4.5 Mesurer les premiers regards portés sur des cartes d'ensembles documentaires . . 167
4.5.1 Motivations et contexte . . . 167
4.5.2 Cadre expérimental . . . 169

4.5.3 Analyse des résultats et retour sur les cartes d'ensembles documentaires . 172
Conclusion : valeur ajoutée et flexibilité du modèle *AIdED* 177

Conclusion

Annexe A Extraits de fichiers XML décrivant les RTO utilisées 185
A.1 Extrait d'un fichier de RTO *LUCIA* simple 185
A.2 Extraits de fichiers décrivant un dispositif *LUCIA* 186

Annexe B Retour sur les méthodes de projection et de classification utilisées 187
B.1 Attribution d'un espace numérique à l'ensemble documentaire 188
B.2 Méthodes de projection de l'espace obtenu à l'issue du comptage 188
B.2.1 Méthodes de projection simples développées durant cette thèse 188
B.2.2 La méthode de l'analyse en composantes principales 190
B.2.3 La méthode de projection de Sammon 194
B.2.4 L'analyse factorielle des correspondances 195
B.3 Méthodes de classification de l'espace visualisé 197
B.3.1 La classification hiérarchique ascendante 197
B.3.2 La méthode des K-Means 198
B.3.3 Le choix du nombre de groupes d'une classification 199
B.4 Comparaisons des méthodes de projection et de classification 201
B.4.1 Cartes des textes 201
B.4.2 Cartes des groupes de textes 204

Annexe C Dispositifs LUCIA utilisés au cours des différentes expérimentations du modèle *AIdED* 207
C.1 Dispositifs utilisés pendant l'expérimentation de recherche documentaire sur Internet 208
C.2 Dispositifs utilisés dans l'étude des trois métaphores conceptuelle durant le projet *IsoMeta* 214
C.3 Ressources lexicales utilisées pendant les expérimentations sur les forums de discussions 217
C.3.1 Expérimentation sur la détection de l'identité professionnelle 217
C.3.2 Expérimentation sur l'usage d'une terminologie professionnelle 218

Annexe D Suivi du regard : les différentes diapositives proposées aux sujets 219

Bibliographie 227

Table des figures **245**

Liste des tableaux **251**

Table des algorithmes **253**

Introduction

Contexte général

Les œuvres littéraires, les livres, les nouvelles, les romans ont toujours été des « endroits » privilégiés où leurs auteurs expriment leurs connaissances, leurs idées, leurs opinions, leur imagination. Avec l'essor du numérique, l'expression textuelle a pris à la fois une toute autre forme et un nouveau support, permettant ainsi à chacun et à n'importe quel moment d'exprimer et de diffuser ce qu'il pense, ce qu'il ressent, ce qu'il imagine.

Une telle masse de données textuelles, prenant place directement dans un format électronique, vient alors rejoindre et « graviter » autour de la masse des ouvrages et des textes plus classiques et moins récents prenant traditionnellement place sur des supports papier. De tels ouvrages et œuvres sont en train de faire le chemin inverse : passer du papier au numérique. Ainsi, le projet de la Bibliothèque Numérique Européenne a pour objectif de rassembler 6 millions d'ouvrages dans un format numérique d'ici 5 ans[1]. La Bibliothèque Numérique de France *Gallica* propose dès maintenant 90 000 ouvrages numérisés et consultables en ligne[2]. Le projet *Google Books*[3] prend également place dans ce domaine. Devant cette masse de données textuelles, dont la taille croît chaque jour, il devient de plus en plus difficile pour le lecteur de choisir les ouvrages et textes pouvant lui apporter satisfaction et information, d'appréhender le contenu d'ensembles de textes toujours plus grands, etc.

Accéder au contenu de données textuelles, qu'elles soient numériques ou non, est une problématique suscitant un grand intérêt depuis longtemps. Les évolutions de l'écriture de sa naissance dans l'Antiquité à nos jours [Martin, 1988], la création des index au Moyen-Âge [Rouse et Rouse, 1982], des tables des matières à la Renaissance [Fayet-Scribe, 1997], des sciences bibliographiques [Otlet, 1903], etc., en sont des illustrations. Ces différents éléments ont tous pour objectif d'aider le lecteur à accéder à des éléments de textes préalablement identifiés. Ainsi, une tâche d'indexation, par exemple, est définie de la façon suivante [Witty, 1973] :

> *Étymologiquement,* indexer *signifie montrer du doigt quelque chose qu'on veut identifier à telle ou telle fin. À l'époque moderne, on désigne par ce mot l'action d'identifier tel ou tel aspect significatif du document quelle qu'en soit la nature, de façon que cet aspect ou ces aspects servent de clés quand on aura besoin, plus tard, de le rechercher au sein d'une mémoire. Pendant plusieurs siècles, cela s'est appliqué aux livres, l'auteur ou l'éditeur faisaient souvent suivre leur texte d'un index, et les bibliothécaires fournissaient des clés sous forme de listes ou de catalogues indiquant ce que contenaient leurs collections.* (traduction de l'anglais issue de [Foskett et Maniez, 1995])

[1] `http://www.theeuropeanlibrary.org/portal` (page consultée le 29 mars 2007).
[2] `http://gallica.bnf.fr` (page consultée le 29 mars 2007).
[3] `http://books.google.fr` (page consultée le 29 mars 2007).

Une telle tâche d'indexation, même si son visage a évolué depuis sa création [Otlet, 1903], revient alors à isoler certaines entités lexicales jugées pertinentes par un expert et à indiquer leurs contextes d'apparition (par exemple, les pages) dans l'ouvrage ou dans un ensemble d'ouvrages. De telles tâches, donnant aux lecteurs des points d'accès au texte, peuvent alors se révéler pertinentes pour faciliter ponctuellement un accès à un ouvrage ou à un ensemble d'ouvrages. Cela n'en constitue pas pour autant une aide universelle pour le lecteur dans l'accès au texte et plus précisément pour *son* accès au texte. Le personnage principal Guillaume de Baskerville du roman *Le Nom de la Rose* de Umberto Eco [Eco, 1980] se serait-il contenter de parcourir l'index ou la table des matières de l'exemplaire unique d'un texte d'Aristote sur l'humour et le rire (livre II de *la Poétique*), objet principal de sa quête ? La réponse peut sembler évidente même si le contexte évoqué est un contexte de fiction.

L'index et la table des matières sont propres à un unique texte. Ils ne permettent pas de relier le texte à d'éventuels textes « voisins ». Pourtant, devant le nombre de textes accessibles à chacun à l'heure actuelle, il devient de plus en plus fréquent de se retrouver face à un ensemble de textes plutôt qu'à un seul texte, par exemple lors d'une recherche documentaire. Pour lier différents textes entre eux selon leur contenu, un index et une table des matières ne seront certainement pas les moyens les plus pertinents.

De tels accès traditionnels aux textes paraissent donc assez limités et il semble nécessaire d'aller plus loin dans les assistances à l'accès aux textes et, surtout, dans la prise en considération de l'expérience textuelle de chaque individu, de son point de vue et de ses domaines de prédilection, tous ces éléments lui étant propres et influençant grandement son interprétation de textes. Guillaume de Baskerville, le héros du *Nom de la Rose*, aurait certainement réalisé une lecture de l'œuvre d'Aristote très différente de celle de son disciple, narrateur et protagoniste du roman d'Umberto Eco. Sa lecture aurait été guidée par ses précédentes et nombreuses lectures et par ses connaissances acquises, alors que son disciple ayant une « expérience textuelle » plus restreinte aurait certainement perçu moins finement le contenu et les références de l'œuvre.

Plan de la thèse

Ce travail de thèse se situe donc dans le cadre général de l'accès au contenu d'ensembles de textes, et particulièrement dans celui d'un accès tenant compte des particularités de l'individu et de son point de vue.

Le **chapitre 1** de cette thèse *Accéder à l'information textuelle dans des ensembles documentaires* présente certains de ces accès au contenu d'ensembles de textes. Ce chapitre met en évidence le fait qu'il ne peut pas y avoir accès à l'« information textuelle » s'il n'y a pas de prise en considération du point de vue propre de l'individu, seul ce point de vue permettant de rendre pertinent un accès à l'information. Les notions d'*ensemble documentaire*, de *personnalisation*, de *visualisation* et d'*interaction* sont respectivement abordées dans le chapitre et sont considérées comme des briques de base à un réel accès au contenu d'ensembles de textes.

Le **chapitre 2** intitulé *Un modèle d'analyse interprétative centrée utilisateur d'ensembles documentaires* mène encore plus loin cette réflexion autour des briques de base pour un réel accès à des contenus textuels. Un modèle d'analyse, le modèle *AIdED* (Analyse Interprétative d'Ensembles Documentaires), faisant des emprunts à la Sémantique Interprétative de François Rastier, est proposé en mettant en interaction ces briques élémentaires et primordiales de notre point de vue. Notre objectif, à travers ce modèle, est de permettre à l'utilisateur de visualiser et d'interagir avec son ensemble documentaire *via* la machine, cette dernière intégrant une représentation des domaines d'intérêt de l'utilisateur. C'est avec de telles aides visuelles, interac-

tives et personnalisées que nous proposons à l'utilisateur d'accéder au contenu de son ensemble documentaire.

Le **chapitre 3** *Instrumentation logicielle du modèle* détaille la conception et la réalisation informatique du modèle *AIdED* à travers des outils et des instruments logiciels centrés sur leurs utilisateurs. De tels éléments logiciels proposent à la fois une assistance dans la construction des domaines d'intérêt de l'utilisateur et la production de supports de visualisation et d'interaction sur un ensemble documentaire à partir de tels domaines.

Ces supports sont ensuite exploités dans le **chapitre 4** *Expérimentations et évaluations du modèle AIdED* par différents utilisateurs dans le cadre de différentes tâches d'accès aux contenus d'ensembles documentaires. La valeur ajoutée de nos propositions est ainsi interrogée à travers différentes expérimentations prenant place dans des contextes expérimentaux très variés.

Enfin, en conclusion de cette thèse, nous dressons un bilan du chemin parcouru durant un tel travail. De nombreuses perspectives de recherche sont également soulevées, aussi bien sur des aspects théoriques qu'appliqués. Ces perspectives visent toutes à positionner l'utilisateur toujours plus au centre de traitements dont l'objectif est de fournir des accès au contenu d'ensembles documentaires.

Chapitre 1

Accéder à l'information textuelle dans des ensembles documentaires

Sommaire

Introduction : l'utilisateur seul devant un univers de textes 6

1.1 La problématique de l'accès à l'information dans des ensembles documentaires 7

1.1.1 Notions utilisées et définitions adoptées 7

1.1.2 Définition(s) de l'*accès à l'information* 10

1.1.3 Méthodes existantes pour l'accès au contenu de documents et d'ensembles documentaires 11

1.1.4 Un même accès à l'information et au contenu d'ensembles documentaires pour tous ? 17

1.2 La place de l'utilisateur dans des tâches d'accès à l'information documentaire 18

1.2.1 La place traditionnellement donnée à l'utilisateur dans des tâches d'accès au contenu 18

1.2.2 Les inconvénients liés à une faible place laissée à l'utilisateur dans de tels systèmes 19

1.2.3 Modèles et traitements pour des analyses personnalisées d'ensembles documentaires 21

1.3 Visualisation interactive d'ensembles documentaires 25

1.3.1 Visualiser des informations complexes 25

1.3.2 Visualiser des textes et des ensembles de textes *via* des systèmes informatiques 28

1.3.3 La cartographie d'ensembles documentaires 34

Conclusion : vers une meilleure prise en considération de l'utilisateur et de ses interactions 40

Introduction : l'utilisateur seul devant un univers de textes

Le monde numérique ne cesse de gagner en ampleur et la masse de données textuelles électroniques produites, échangées et évoluant chaque jour entre des « utilisateurs-lecteurs » ne cesse de croître. En plus des bibliothèques numériques fleurissant à l'heure actuelle, Internet déborde de données textuelles. En juillet 2000, l'étude *Sizing the Internet* de la société Cyveillance[4], estimait le Web visible[5] à 2,1 milliards de pages, augmentant ainsi à un rythme soutenu de 7,3 millions de pages par jour. Selon ce taux de croissance, la taille du Web visible a été estimé à 3 milliards de pages en octobre 2000 et à 4 milliards en février 2001 soit un doublement de volume par rapport à juillet 2000 (ce qui correspond à une période de 8 mois). Si l'on estime que le Web visible double de volume tous les ans, il représenterait fin 2007 au moins 256 milliards de pages.

Cette estimation est bien évidemment très minorée puisque elle est basée sur un taux de croissance stable et donc très peu probable au vu des grandes possibilités de publication offertes à chacun au travers des *blogs*[6] et autres sites d'écriture collaborative comme les *wikis*[7]. Ainsi, l'encyclopédie en ligne Wikipédia[8], basée sur un principe d'écriture collaborative ouverte à chacun, totalise depuis sa création en janvier 2001 plus de 5 millions d'articles toutes langues confondues et plus de 9 000 articles sont créés chaque jour[9].

Les utilisateurs se retrouvent alors très souvent confrontés à une trop grande quantité de données textuelles où seule une petite partie peut se révéler « pertinente » à leurs yeux. Pour faire face à ces ensembles vertigineux de données textuelles, différents logiciels sont proposés aux utilisateurs afin de les guider dans cet espace selon le type d'éléments recherchés. De tels logiciels, ainsi que les différents types d'accès qu'ils proposent, sont détaillés dans une première partie de ce chapitre.

Les données textuelles contenues dans ces ensembles sont, de plus, étroitement liées à leurs « entourages ». Par exemple, la lecture d'un billet d'un *blog* peut nécessiter la lecture d'autres billets du même *blog* ou d'autres *blogs*, la lecture d'un texte dans une encyclopédie peut nécessiter la lecture d'autres textes de la même encyclopédie, etc. Plus généralement, quand un lecteur parcourt un texte en tant que tel, isolé, cela ne suffit pas toujours pour appréhender son contenu, il a souvent besoin de prendre en considération un niveau plus global, c'est-à-dire l'ensemble documentaire dans lequel le texte se positionne. De tels liens entre textes et entourages sont fortement guidés par l'individu et son point de vue. C'est ce dernier qui éprouvera le besoin de positionner un texte dans un ensemble documentaire, mais ce sera également lui qui construira cet ensemble toujours selon son point de vue, ses besoins et son passé textuel. La deuxième partie de ce chapitre met alors en évidence la place laissée ou devant être laissée à l'utilisateur dans des logiciels proposant des accès au contenu d'ensembles documentaires.

Une telle prise en considération de l'utilisateur et des ensembles documentaires nécessite d'utiliser des techniques permettant à l'utilisateur d'appréhender globalement et d'interagir avec

[4] `http://www.cyveillance.com/web/corporate/white_papers.htm` (page consultée le 5 avril 2007).

[5] Par le terme « Web visible », nous entendons les sites Internet référençables, accessibles à tous, à l'opposé du « Web invisible », contenant des sites accessibles uniquement par des utilisateurs autorisés. À noter également que le « Web dynamique », caractérisant des sites dont l'intégralité ou une partie est générée de façon automatique par différents programmes, pose de grandes difficultés de référencement.

[6] Un *blog* est un site Internet constitué d'un ensemble de billets triés par ordre chronologique où chaque billet contient du texte ajouté par l'auteur et où chaque lecteur peut laisser des commentaires.

[7] Un *wiki* est un système de gestion de contenu de site Internet qui permet librement et facilement la création et la modification de pages de sites Internet par tous les visiteurs autorisés selon le système.

[8] `http://fr.wikipedia.org/wiki/Accueil`, page d'accueil de la version française de l'encyclopédie, page consultée le 6 avril 2007.

[9] Chiffres disponibles à l'adresse suivante : `http://stats.wikimedia.org/EN/ChartsWikipediaZZ.htm` (page consultée le 5 avril 2007).

les textes de ces ensembles. Pour apporter un début de réponse, nous présentons en quatrième partie de ce chapitre différentes techniques de visualisation interactive et en particulier de cartographie d'ensembles documentaires. Enfin, nous conclurons ce chapitre en mettant en évidence le travail ainsi visé par cette thèse : la prise en compte du point de vue de l'utilisateur dans l'aide interactive à l'interprétation d'ensembles documentaires.

1.1 La problématique de l'accès à l'information dans des ensembles documentaires

L'accès automatique à l'information dans des ensembles documentaires est une problématique qui suscite beaucoup d'intérêt dans différentes communautés, comme celles du traitement automatique des langues, de la documentation, de l'intelligence économique, etc. Cet « accès à l'information », souvent à l'intérieur d'ensembles documentaires, constitue la problématique générale de la thèse. Tout d'abord, nous présentons dans cette section l'objet d'étude visé : l'ensemble documentaire. Nous expliquons ensuite ce qui est communément désigné par les termes *accès à l'information* dans de tels ensembles documentaires, puis nous précisons notre point de vue par rapport à cet existant, en insistant sur la place primordiale qui doit être rendue à l'utilisateur afin de tenter de lui fournir une assistance pertinente pour l'accès au contenu d'ensembles documentaires.

1.1.1 Notions utilisées et définitions adoptées

La problématique de l'accès à l'information interroge les moyens de faciliter le parcours, la lecture, l'appropriation de données textuelles pour les utilisateurs. Avant de présenter une telle problématique, nous nous attardons sur les objets d'étude considérés.

Texte

L'une des notions élémentaires lorque l'on étudie des données textuelles est la notion de *texte*. La définition du texte que nous adoptons est issue de [Rastier, 2001a]. L'auteur donne d'abord des définitions négatives schématisant les erreurs souvent commises lors de la définition du concept de texte. Tout d'abord, un texte ne doit pas être considéré comme une chaîne de caractères (seulement une infime partie des chaînes de caractères sont des textes). Un texte ne doit pas être non plus considéré comme une suite d'instructions (à la manière d'un programme informatique), cette considération ramenant la compréhension d'un texte à l'exécution d'un programme par un ordinateur. Enfin, un texte n'est pas non plus une suite de schémas cognitifs, la lecture d'un texte suscite généralement la création de schémas mentaux, mais elle ne se limite pas à cela. F. Rastier donne ensuite une définition positive du texte :

> *Un texte est une suite linguistique empirique attestée, produite dans une pratique sociale déterminée, et fixée sur un support quelconque.*

Un texte n'est donc pas une création artificielle (comme pourrait l'être un exemple linguistique construit pour illustrer un fait de langue), il est créé dans le cadre d'une pratique sociale et il prend place sur un support (feuille de papier, fichier informatique, etc.). François Rastier dans [Rastier, 2005] propose alors de considérer le texte comme une unité minimale d'une linguistique « évoluée ».

Document

Du texte au document, le passage semble assez naturel. Pourtant, la notion (ou plutôt les notions) de *document*, et surtout de *document numérique*, ont provoqué de profonds changements dans l'utilisation et la diffusion de telles entités.

Dans [Buckland, 1997], l'auteur affirme qu'un document représente toutes bases de connaissance fixées sur un support matériel qui est susceptible d'être utilisé pour la consultation, l'étude ou la preuve. Par exemple, un manuscrit, un imprimé, une représentation graphique ou figurée, un objet tiré d'une collection, sont des documents. Avec l'arrivée du document numérique, Michael Buckland avoue dans [Buckland, 1998] un certain changement des usages du document et même de sa définition :

> *Quand on se réfère au document papier ou au papyrus ou au microfilm, la signification est claire. Cependant l'idée d'un document électronique est plus difficile à définir.*

Gwendal Auffret propose dans [Auffret, 2000] une définition plus complète et généraliste de la notion de document, électronique ou non. Le document est considéré comme une entité discrète provenant d'une activité éditoriale donnée. Selon l'auteur, le document est composé des cinq éléments suivants :

- le *support d'enregistrement* sur lequel les données composant le document sont stockées (le papier pour un texte analogique, le disque d'un ordinateur pour un texte numérique, etc.) ;
- la *forme d'enregistrement* selon laquelle les informations documentaires sont enregistrées sur le support d'enregistrement (le papier pour un texte analogique, un codage de caractères pour un texte numérique, un codage vidéo pour un film, etc.) ;
- le *support de restitution* grâce auquel un usager accède au contenu du document (toujours le papier pour un texte analogique, l'écran d'ordinateur pour un texte numérique, l'écran de télévision, de cinéma ou d'ordinateur et des hauts-parleurs pour des films audiovisuels, etc.) ;
- la *forme physique de restitution* selon laquelle le document va être présenté à son lecteur (l'encre sur le papier pour un texte analogique, le signal vidéo au audio analogique pour des documents audiovisuels, etc.) ;
- et enfin la *forme sémiotique de restitution* selon laquelle le document va être appréhendé par le lecteur (l'écriture alphabétique pour un texte, des images animées et des sons pour des documents audiovisuels, etc.).

Une telle définition, très complète, correspond à la vision que nous adoptons du document : est document tout ce qui a un support et une forme d'enregistrement, un support, une forme physique et une forme sémiotique de restitution. Dans cette thèse, nous faisons cependant la plupart du temps référence à des documents électroniques textuels, comme des pages de sites Internet, des articles de presse, des articles scientifiques, des messages de forums[10].

Corpus

Selon François Rastier [Rastier, 2005], l'unité du texte (et celle du document électronique textuel que nous étudions) prend tout son sens dans un contexte plus global : le *corpus*. L'auteur définit alors un corpus de la façon suivante :

[10] À noter que, depuis 2003, le Réseau Thématique Pluridisciplinaire du département STIC du CNRS « Document et contenu : création, indexation, navigation » a lancé une véritable réflexion autour du document électronique, allant de sa définition, sa création jusqu'à ses multiples utilisations. Description du RTP : `http://rtp-doc.enssib.fr` (page consultée le 14 mai 2007).

> *Un corpus est un regroupement structuré de textes intégraux, documentés, éventuellement enrichis par des étiquetages, et rassemblés : (i) de manière théorique réflexive en tenant compte des discours et des genres, (ii) et de manière pratique en vue d'une gamme d'applications.* [Rastier, 2005, page 32]

Dans [Pincemin, 1999, pages 415-427], l'auteur va plus loin dans la définition du corpus en expliquant qu'un regroupement de textes est un corpus s'il respecte un certain nombre de conditions :

- des *conditions de signifiance* : le corpus doit être constitué en vue d'une étude déterminée, qui porte sur un objet particulier et sur une réalité telle qu'elle est perçue sous un certain angle de vue ;
- des *conditions d'acceptabilité* : le corpus doit apporter une représentation fidèle, ne pas comporter d'éléments faisant office d'exceptions, et doit avoir une ampleur et un niveau de détail adaptés au degré de précision attendu en résultat de son analyse ;
- et des *conditions d'exploitabilité* : les éléments qui forment le corpus doivent être commensurables, le corpus doit comporter un nombre suffisant d'éléments pour pouvoir repérer des comportements significatifs.

Les deux définitions complémentaires du corpus que nous venons de citer sont celles que nous adoptons dans cette thèse. Un corpus contient donc des textes intégraux, regroupés selon une application visée et différents critères de pertinence, de volume, de représentativité, etc. Une telle acception de la définition du corpus n'est pas partagée par tous, certains travaux ne considérant pas le texte comme unité élémentaire mais plutôt la phrase, le syntagme, une fenêtre de n caractères, etc[11].

Une discipline de recherche à part entière est dédiée à l'étude linguistique de corpus de textes. Une telle discipline, appelée *linguistique de corpus*, a pour objectif général d'étudier la langue en contexte, le contexte étant approché par le corpus d'étude regroupant des textes. La linguistique de corpus s'est développée à partir des années 80 et a connu un essor tout particulièrement lié à l'arrivée de l'informatique.

Dans [Williams, 2005], Geoffrey Williams précise que la linguistique de corpus *est une discipline qui relève de la linguistique appliquée et qui cherche à comprendre les mécanismes de la communication et apporter des solutions à des questions pratiques.* Il ajoute que *la linguistique de corpus est un domaine qui s'intéresse aux textes, aux textes réels, c'est-à-dire produits pour des raisons de communication entre êtres humains et non par l'introspection des linguistes.* La linguistique de corpus constitue l'une des disciplines dans laquelle ce travail de thèse s'inscrit, comme nous aurons l'occasion de le voir par la suite.

Collection

Pour être considéré comme un corpus, un regroupement de textes doit donc répondre à un certain nombre de critères assez restrictifs. Avec la multiplication des documents numériques textuels et des moyens de partage et de parcours de tels documents, certains d'entre eux se retrouvent alors regroupés de façons plus ou moins fortuites et plus ou moins « naturelles ». C'est par exemple le cas des pages retournées par des moteurs de recherche sur Internet par rapport à la requête de l'utilisateur. C'est également le cas des courriers électroniques échangés par différentes personnes sur une même liste de discussion, comme c'est également le cas des messages échangés sur un même forum de discussion.

[11] Le terme de « corpus » est également utilisé pour désigner autre chose que des regroupements de textes, tels des regroupements de vidéos, de dialogues, d'images, etc.

Pour désigner de tels ensembles, nous utilisons le terme de *collection*. À l'instar d'un corpus, une certaine cohérence est présente au sein d'une collection, elle ne contient pas des éléments choisis au hasard[12] et regroupe des textes liés les uns avec les autres, du moins, selon un certain point de vue et pour un certain usage. À la différence d'un corpus, une collection n'est pas forcément une représentation fidèle d'un objet d'étude ou encore exempte d'éléments faisant exception, du fait de sa constitution moins rigoureuse, plus pragmatique.

Ensemble documentaire

Le travail présenté dans cette thèse a donc une position duale, en proposant d'étudier à la fois des corpus « bien formés » au sens linguistique et des collections obtenues de façon plus naturelle. Une telle étude nous amènera bien évidemment à tenir compte des différences importantes entre ces notions, mais surtout à tirer partie des points communs de ces regroupements de documents électroniques textuels. Pour désigner aussi bien des corpus que des collections, nous employons le terme d'*ensemble documentaire* (ou encore *ensemble de documents*). Ainsi, toute analyse basée sur des ensembles documentaires, dans le sens ainsi proposé, exploite des textes dont le regroupement, plus ou moins contraint, est pertinent selon un certain point de vue et pour un certain usage.

L'accès à l'information contenue dans des ensembles documentaires fait l'objet de nombreuses recherches depuis plusieurs années. La partie suivante de ce chapitre propose de définir cette notion d'accès à l'information.

1.1.2 Définition(s) de l'*accès à l'information*

La « transparence » des activités de nos sociétés et de leurs responsables devient de plus en plus un argument à la mode et ceci dans un grand nombre de domaines : l'éducation, la gestion et le financement des états, les activités éducatives et scientifiques, le monde politique, etc. Cette transparence est étroitement liée à l'accès à l'information prôné par de nombreuses personnes. Des décrets et des lois pour l'accès à l'information ont été votés par des responsables politiques de certains pays, des commissions dédiées ont été créées pour le respect de ces lois, etc[13].

La définition généralement adoptée pour l'« accès à l'information » est la mise à disposition de chacun sur des espaces publics (physiques, comme des bibliothèques, ou virtuels, comme des sites Internet) de l'« information » sur des domaines concernés [Bernhard, 1998]. Très généralement, cette « information » prend la forme de documents textuels. L'une des étapes importantes pour accéder à l'information consiste donc à lire attentivement les textes mis à disposition par des personnes « responsables » et à isoler dans ces textes, les informations semblant pertinentes aux yeux des lecteurs selon les domaines concernés et les objectifs visés, bien évidemment, si de telles informations sont présentes dans les textes aux yeux du lecteur (ce qui n'est pas toujours le cas, par exemple, en parcourant des textes juridiques assez inaccessibles pour les non-spécialistes).

[12] Par exemple, François Rastier utilise ainsi les termes d'*aire de stockage*, de *décharge publique* pour désigner Internet. Par ces termes, il nous semble que l'auteur veut désigner la très grande variété des documents textuels présents sur Internet, variété aussi bien en genres, en thèmes, en pertinence, en niveau d'expression, etc. Nous suivons également ce constat, c'est pourquoi un ensemble de documents extraits aléatoirement d'Internet ne pourra être considéré comme une collection, le regroupement n'étant pas guidé par des utilisateurs ou un outil de recherche.

[13] Le Canada est un pays particulièrement attentif à l'accès à l'information, cet intérêt est manifesté par une loi (`http://lois.justice.gc.ca/fr/A-1/166479.html`, page consultée le 16 avril 2007) visant à rendre accessible à chacun des documents décrivant les activités de l'état. Une commission dédiée à la promotion et au respect de l'accès à l'information a également été créée (`http://www.cai.gouv.qc.ca`, page consultée le 16 avril 2007).

Accéder à l'information est souvent assimilé par certains à l'accès à Internet. Un tel accès à l'information est alors encore plus fortement personnel car cette fois-ci, la masse de documents disponibles est bien plus grande que, par exemple, celle sélectionnée et mise à disposition par un responsable pour communiquer sur les activités d'un état ou d'une société. Ce sont alors les utilisateurs qui choisissent les ensembles documentaires où ils chercheront de l'information. La sélection de cet ensemble documentaire, et également son analyse, sont donc fortement guidées par l'utilisateur avec ses choix, ses centres d'intérêt et son expérience propre.

La partie suivante de ce chapitre présente les méthodes existantes proposant des accès à l'information dans des documents et des ensembles documentaires.

1.1.3 Méthodes existantes pour l'accès au contenu de documents et d'ensembles documentaires

Pour proposer un « accès à l'information » dans des données textuelles, de nombreuses méthodes exploitant différents niveaux et échelles d'analyse de documents et d'ensembles documentaires sont proposées. Certaines de ces méthodes sont présentées dans les parties suivantes, les différents traitements qu'elles proposent sont également discutés.

Types de méthodes proposées

Adeline Nazarenko établit dans [Nazarenko, 2005] quatre grandes familles de méthodes automatiques d'accès au contenu des documents :

- l'*extraction d'information*,
- les *méthodes de question/réponse*,
- le *résumé automatique*
- et l'*aide à la navigation*.

Par *extraction d'information*, il est question des méthodes consistant à retrouver dans des ensembles de textes très homogènes (par exemple, un corpus de dépêches d'actualité ou encore un corpus d'articles scientifiques) des informations dont on sait qu'elles s'y trouvent. Par exemple, lors de la conférence M.U.C. 7 (Message Understanding System) [M.U.C., 1998], il était demandé d'extraire d'un corpus d'actualités financières, les transactions de rachats, de fusions de sociétés, etc., et de remplir avec ces différentes informations des formulaires électroniques indiquant notamment qui a acheté qui, à quel prix, quand, etc. À noter également qu'un travail similaire a été mené dans le domaine des constats d'accidents de voitures [Enjalbert et Victorri, 1994] afin de mettre automatiquement en évidence les différents protagonistes et les circonstances de l'accident. Les utilisateurs visent alors, avec de tels outils, à alimenter de manière entièrement automatique des bases de données préexistantes à partir d'ensembles de textes soigneusement sélectionnés.

Les méthodes dites de *question/réponse* ont un objectif différent. Elles consistent le plus souvent à chercher un fragment de texte extrait d'un corpus volontairement assez généraliste dans lequel un utilisateur 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 » dans une chaîne de caractères, éventuellement en essayant tout de même de ne pas couper des mots en leur milieu. Lors des conférences d'évaluation TREC (Text REtrieval Conference)[14], les systèmes

[14]Le site `http://trec.nist.gov/` (consulté le 20 avril 2007) est dédié aux campagnes d'évaluation TREC et présente en détail les objectifs des différentes campagnes d'évaluation.

de questions/réponses avaient par exemple pour consigne de rendre des réponses de moins de 250 caractères à partir de 980 000 documents et de 700 questions. Plus récemment, la campagne EQueR (Evaluation Question/Réponse) menée en France en 2005 avait l'objectif double de mettre au point des systèmes de question/réponse interrogeant aussi bien des corpus très hétérogènes (des articles de presse de thématiques variées) que des corpus très homogènes prenant place dans le domaine médical (se reporter à [Ayache *et al.*, 2005] pour une vue d'ensemble de cette campagne). À la différence des méthodes d'extraction d'information, l'interprétation du sujet humain est particulièrement importante afin d'évaluer la pertinence de la réponse retournée.

Les méthodes de *résumé automatique* laissent également une large part à l'interprétation de l'utilisateur auquel le résumé est destiné. Assez 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 alors dans ce cas de rechercher des phrases que l'on pense significatives (par exemple, des phrases qui commenceraient par *en somme...*, *on constate que...* auraient de bonnes chances de synthétiser ce qui est dit précédemment) et de les juxtaposer dans un « résumé » où l'on espère que le lecteur pourra rétablir une certaine cohérence textuelle. D'autres techniques sont mises en œuvre afin de construire un nouveau texte résumant le premier, telle, par exemple, l'extraction de patrons morpho-syntaxiques, l'identification de la structure thématique du texte, etc. Nous renvoyons à [Minel, 2004] pour un panorama des différentes techniques de résumé automatique de textes.

Enfin la dernière « catégorie » de méthodes d'accès à l'information textuelle aborde l'*aide à la navigation*. Ces méthodes s'adressent en quelque sorte 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), contrairement aux méthodes précédentes s'adressant plutôt à la dimension rhématique (en cherchant à savoir ce qui est dit, où, quand, par qui, comment, etc.). Les applications les plus courantes de ces méthodes sont l'aide à la lecture de textes, la recherche documentaire, la gestion électronique de documents, etc.

Outils logiciels pour l'accès au contenu de documents et d'ensembles documentaires

Les différents types de méthodes d'accès au contenu d'ensembles documentaires présentés précédemment sont opérationnalisés à travers des logiciels aux objectifs assez différents. Nous présentons, dans cette section, certains de ces logiciels afin d'illustrer la catégorisation énoncée précédemment pour les types de méthodes d'accès au contenu.

Extraction d'information

Par ce type de méthodes d'accès au contenu d'ensembles documentaires, les systèmes informatiques proposés se contentent généralement de remplir automatiquement des formulaires définis par les usagers.

Dans le domaine bio-médical, les auteurs de [Alphonse *et al.*, 2004] proposent, dans le cadre du projet *Caderige*[15], des outils d'extraction d'information dans les bases de données bibliographiques médicales. Leur approche intègre des méthodes aussi bien issues de la fouille de données que du traitement automatique des langues. Les traitements sont réalisés sur des corpus de documents médicaux et sur des ontologies. Ils permettent d'isoler des interactions entre des gènes, des relations de synonymie entre noms de gènes, etc.

[15] Catégorisation Automatique de Documents pour l'Extraction de Réseaux d'Interactions GEniques : `http://caderige.imag.fr` (page consultée le 20 avril 2007).

L'extraction d'information s'intéresse à de nombreux domaines d'étude. Allant du domaine bio-médical au domaine de l'assurance, par exemple dans le cadre du projet *TACIT* [Victorri, 1998], où des outils proposant d'extraire des informations à partir de constats d'accidents automobiles ont été réalisés. Ces outils avaient pour objectif d'extraire automatiquement de ces constats, les informations sur les différents protagonistes et la situation, informations ensuite retournées sous forme de formulaires.

Le coté « tout-automatique » de tels outils d'extraction d'information semble être délaissé au profit d'outils d'extraction « semi-automatiques », laissant aux utilisateurs la possibilité de préciser et de réviser interactivement les objectifs de leur tâche. C'est par exemple ce que propose l'outil *Amilcare*[16] [Ciravegna, 2003]. Cet outil permet l'annotation de textes par des utilisateurs avec des étiquettes sémantiques et l'apprentissage automatique à partir de ces annotations. Une fois ces deux étapes réalisées, l'outil propose alors d'étendre l'annotation sur de nouveaux éléments et de permettre la formulation des requêtes portant sur les étiquettes afin d'extraire des informations.

Question/réponse

Les méthodes d'accès au contenu d'ensembles documentaires de type « question/réponse » sont également de plus en plus nombreuses. À partir de questions pouvant être de nature variée, les systèmes de question/réponse cherchent :

1. soit à trouver dans une base documentaire, une chaîne de caractères contenant une réponse à la question,
2. soit à construire une nouvelle phrase pouvant répondre à la question de départ.

En 2005, la campagne EQueR [Ayache *et al.*, 2005] a proposé une évaluation comparative de différents systèmes de la première catégorie. Cette évaluation a mis en concurrence différents systèmes sur des questions de différentes natures : questions simples (réponses : oui/non, date, etc.), questions plus complexes (réponses : définitions, listes, etc), etc. Les différents systèmes participant à cette évaluation (tels les systèmes *SQuAr* [Blaudez *et al.*, 2005], *Oedipe* [Balvet *et al.*, 2005] ou encore *FRASQUES* [Grau *et al.*, 2005]) se basent sur un ensemble documentaire constitué d'un très grand nombre d'articles de presse en français. À partir d'une question posée par l'utilisateur, ces systèmes vont mettre en œuvre différents traitements, différents algorithmes, afin d'extraire de leur base documentaire une fenêtre de caractères d'une certaine longueur (au cours de l'évaluation EQueR, la taille maximale autorisée des fenêtres était de 250 caractères).

La seconde catégorie de systèmes de question/réponse, évoquée précédemment, regroupe des outils proposant des réponses construites aux questions des utilisateurs et non des réponses sélectionnées dans des bases documentaires. Parmi ces systèmes, l'agent logiciel animé *Nestor* développé par France Télécom [Panaget, 2004] propose de répondre directement aux questions des utilisateurs portant sur une thématique réduite, celle de la localisation de restaurants dans Paris. La difficulté supplémentaire de tels systèmes est liée à la génération de phrases[17], à la fois correctement construites, mais aussi cohérentes par rapport aux questions des utilisateurs.

Sur un plan différent, des systèmes de question/réponse laissent place à une forte participation humaine. Ainsi, le système *Yahoo Answers*[18] montre que la majorité des questions posées par

[16] Outil disponible à l'adresse suivante : `http://nlp.shef.ac.uk/amilcare` (page consultée le 20 avril 2007).

[17] En décembre 2004, une journée ATALA (Association pour le Traitement Automatique des Langues) a été dédiée à la thématique de la génération d'énoncés en langue naturelle : `http://www.atala.org` (page consultée le 23 avril 2007), voir également le numéro de la revue T.A.L. [Grau et Magnini, 2005] pour plus de détails.

[18] `http://fr.answers.yahoo.com` (page consultée le 23 avril 2007).

des utilisateurs peuvent être résolues par d'autres. Le système se « contente » alors de mettre en relation les utilisateurs posant des questions et ceux y répondant, les rôles s'inversant à tout moment.

Résumé automatique

Tout comme les systèmes de question/réponse, les systèmes de résumé automatique se séparent en deux catégories distinctes : les systèmes qui sélectionnent des phrases jugées caractéristiques dans les textes à résumer, et les systèmes cherchant à reconstruire un nouveau texte.

Dans la première catégorie de systèmes, les traitements de textes les plus usuels, tels *Microsoft Word* ou *Open-Office Writer*, intègrent des modules de « résumé automatique » sélectionnant des phrases caractéristiques dans le texte à résumer. Par exemple, ces modules, assez basiques, exploitent entre autres des connecteurs linguistiques tels « en bref », « pour résumé », etc. Le système *REDUIT* décrit dans [Châar *et al.*, 2004] est un système plus avancé de filtrage de phrases caractéristiques d'un texte. Ce système prend à la fois en considération la structure thématique du texte, mais tient également compte d'ensembles de termes jugés pertinents pour les utilisateurs dans la tâche de résumé visée.

Proche de la première famille de systèmes de résumé automatique, nous pouvons également citer l'outil *Théma* développé au laboratoire GREYC de l'Université de Caen / Basse-Normandie[19]. Cet outil, basé sur une structuration thématique, procède à un découpage hiérarchique du texte en thème (de quoi il est question dans le texte) / rhème (ce qui en est dit)[20]. Le découpage d'un texte ainsi retourné à l'utilisateur prend place sur plusieurs niveaux. Au premier niveau, on trouve seulement deux « blocs » de textes, le thème et le rhème, à un deuxième niveau, le rhème est alors lui-même segmenté en deux blocs thème et rhème, et ainsi de suite (exemple de segmentation en figure 1.1).

FIG. 1.1 – Un exemple de segmentation hiérarchique réalisée avec le logiciel *Thema* sur un texte chinois.

Un tel découpage peut alors permettre aux utilisateurs d'avoir une sorte de vue condensée du texte, en ne sélectionnant par exemple que les parties thématiques de premiers niveaux. Dans

[19] Se reporter à `http://users.info.unicaen.fr/~nadine` pour plus de détails (page consultée le 30 avril 2007).
[20] Nous précisons dans le chapitre suivant la notion de thème.

[Sidir *et al.*, 2006], l'outil *Théma* est utilisé pour l'analyse de forums de discussion afin d'avoir des vues réduites de différents forums.

La seconde catégorie de systèmes de résumé automatique est celle regroupant les outils proposant la construction d'un nouveau texte résumant le texte d'origine. Ces systèmes, moins nombreux, possèdent des fonctions de filtrage et de sélection d'unités lexicales pertinentes dans le texte d'origine. Après cette étape, l'étape de génération d'un résumé entièrement nouveau fait intervenir des grammaires, des « patrons de textes », des formulaires, etc. Ainsi, les auteurs de [Farzindar *et al.*, 2004] proposent le système *LetSum* construisant des résumés automatiques à partir de textes juridiques. Ce système se base principalement sur une segmentation thématique des textes (les textes respectant tous une même structure) et une extraction des unités lexicales saillantes. Un patron de texte résumé est donné afin d'être complété à l'issue de l'analyse.

Aide à la navigation

Comme nous l'avons déjà énoncé précédemment, les systèmes d'aide à la navigation dans des documents ou des ensembles de documents interrogent plutôt une dimension thématique globale. Parmi ces systèmes d'aide à la navigation, deux catégories se distinguent : les systèmes de navigation intra-documentaire et ceux proposant une navigation inter-documentaire.

Les systèmes de la première catégorie proposent aux utilisateurs de naviguer au sein d'un seul document, souvent assez volumineux. Ainsi, le système *3D-XV* [Jacquemin et Jardino, 2002] propose de naviguer dans la structure et le contenu dans un grand document structuré via une interface présentée en figure 1.2.

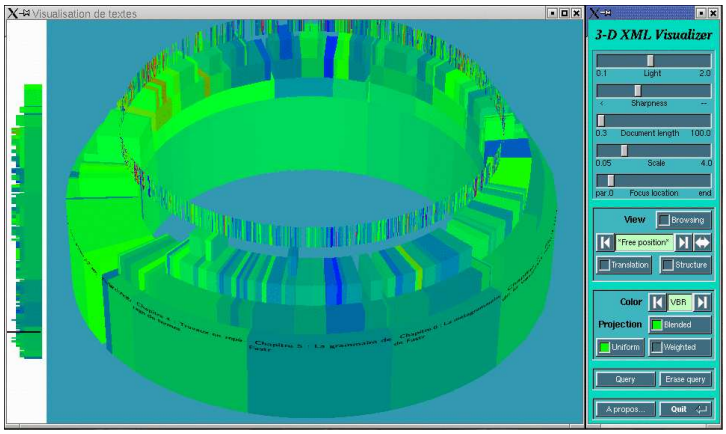


FIG. 1.2 – Un écran du système de navigation intra-documentaire *3D-XV*.

Dans le cadre de l'aide à la navigation dans des documents géographiques, Frédérik Bilhaut [Bilhaut, 2006] propose de produire des vues segmentées et coloriées d'un atlas géographique, ces vues tenant compte aussi bien de la structure thématique de l'atlas que des principales entités abordées et mises en relation.

La seconde catégorie regroupe des systèmes, plus nombreux, proposant de naviguer dans des ensembles de documents. Comme systèmes « pionniers » mais toujours pertinents à l'heure actuelle, nous pouvons citer les travaux décrits dans [Salton, 1989], [Hearst, 1995] ou encore [Hearst, 1999]. Ces travaux sont liés au domaine de la recherche d'information et décrivent des systèmes proposant d'avoir des vues plus globales sur les ensembles de documents et où il est possible d'observer (toujours d'une façon assez globale) des similarités et des différences entre les documents de l'ensemble.

Les moteurs de recherche traditionnels sur Internet (tels *Google*[21] et *Yahoo*[22]) peuvent également être considérés comme des systèmes d'aide à la navigation dans un ensemble de données textuelles gigantesques (les données de leurs index). Ces moteurs emploient des programmes (également appelés « robots ») parcourant un grand nombre de sites Internet. Pour une page donnée, ces programmes réalisent deux opérations :

- l'indexation du texte contenu dans la page ; différentes techniques peuvent être utilisées, l'indexation par mots ou suite de mots reste à l'heure actuelle la plus employée ;
- la récupération des liens vers d'autres pages présents dans la page courante afin de poursuivre leur parcours de la Toile.

À partir de mots-clés choisis par les utilisateurs pour décrire l'objet de sa recherche, les moteurs parcourent ensuite leur index afin de retourner des listes de pages jugées pertinentes par le système, listes que les utilisateurs pourront parcourir afin de tenter de satisfaire leur recherche d'information. Nous pouvons par exemple renvoyer à [Foenix-Riou, 2005] pour une présentation plus détaillée des moteurs de recherche traditionnels[23].

Depuis plusieurs années la communauté du traitement automatique des langues s'implique dans la recherche d'information sur Internet. Nous pouvons par exemple citer des travaux portant sur l'extension des requêtes des utilisateurs, l'indexation de termes complexes ou encore l'intégration d'un analyseur flexionnel, travaux qui sont entre autres rassemblés dans [Jacquemin, 2000].

Niveaux et empan d'analyse des méthodes mises en application

Les quatre types d'accès au contenu d'ensembles documentaires se basent sur différentes propriétés des données textuelles. Les principaux niveaux d'analyse suivants peuvent alors être isolés.

- Les **analyses morphologiques et syntaxiques**. Ces analyses sont réalisées pour repérer les catégories morphologiques des mots ainsi que les différentes relations pouvant exister entre eux, comme des relations sujet/verbe ou verbe/objet.
- Le ***pattern matching***. Une telle analyse consiste à repérer dans les textes des sous-chaînes de caractères données. Le repérage de ces sous-chaînes peut alors permettre d'isoler des marqueurs thématiques, sémantiques, des mots introducteurs de marques de discours, etc.
- L'**analyse de la mise en forme matérielle** (MFM). La structure du texte, comme ses titres, sous-titres, paragraphes, sa typographie peut apporter des informations utiles sur le contenu du texte.

Ces niveaux d'analyses assez différents sont rarement combinés. Leur empan est également assez limité. Par exemple, dans [Poibeau, 2004], l'auteur rappelle que pour une tâche d'extraction d'information, et notamment comme celles des conférences MUC, moins de 10% des données textuelles du corpus d'étude sont exploitées.

La place de l'utilisateur reste également très secondaire dans les niveaux d'analyse présentés précédemment. Il ne lui est laissé que la possibilité de définir quelques règles d'analyse, quelques termes d'un lexique et d'interpréter les résultats finaux produits, afin de tenter d'en extraire une valeur ajoutée selon sa tâche. Pourtant, c'est l'utilisateur qui est censé accéder au contenu d'ensembles documentaires. Une telle combinaison de différents angles d'analyse, avec une prise en considération des points de vue des utilisateurs, pourrait alors permettre de réaliser des analyses,

[21] `http://www.google.fr`

[22] `http://www.yahoo.fr`

[23] Voir également `http://www.commentcamarche.net/www/moteur-recherche.php3` (page consultée le 28 septembre 2007) pour une présentation plus courte et moins formelle.

à la fois plus fines, et surtout plus pertinentes aux yeux des utilisateurs. Adeline Nazarenko dans [Nazarenko, 2005, page 225] souligne cette absence d'une telle architecture d'analyse.

> *Il n'existe pas encore de modèle permettant de donner une représentation unifiée de ces différents niveaux d'analyse, de comprendre* ***le rôle et la contribution de chacun dans le résultat d'analyse global*** *(...). En réalité, les méthodes d'accès au contenu textuel ont été conçues de manière très pragmatique. En combinant les ingrédients de base, on a testé différentes recettes jusqu'à ce que « ça marche ».*

Ce pragmatisme peut effectivement être dommageable pour la conception de telles méthodes et architectures s'il borne, ou limite, la conception à un « optimum local », celui d'un état de fonctionnement *a priori* convenable (l'état du « ça marche »). Par contre, ce pragmatisme nous semble nécessaire et pertinent pour des données prises en entrée représentant le point de vue de l'utilisateur, ainsi que pour les interactions qui lui sont laissées sur les sorties d'analyse. Dans [Peirce, 1879], l'auteur propose la maxime suivante, qui sera considérée plus tard comme étant la maxime du pragmatisme :

> *Considérer quels sont les effets pratiques que* ***nous*** *pensons pouvoir être produits par l'objet de notre conception. La conception de tous ces effets est la conception complète de l'objet.*

Prendre en considération tous les effets produits par un outil d'accès au contenu d'ensembles documentaires est à notre avis indispensable. Ceci entraîne forcément une prise en considération de l'utilisateur, de son point de vue sur la tâche qu'il désire mener, de sa façon de parler, de ses centres d'intérêt, des interactions qu'il souhaite avoir avec l'ensemble documentaire, etc.

Nous expliquons, dans la partie suivante de ce chapitre, pourquoi il est absolument nécessaire de notre point de vue de personnaliser des accès au contenu de textes.

1.1.4 Un même accès à l'information et au contenu d'ensembles documentaires pour tous ?

L'accès à l'information, tel qu'il est vu traditionnellement, laisse entendre que de l'information est forcément contenue dans les éléments présentés aux individus, et ceci, d'une manière globalement similaire pour chacun. Accéder au contenu d'ensembles documentaires laisse paraître les mêmes propriétés, même si cette fois-ci, le type d'objet censé contenir l'information est précisé. Un très grand nombre de paramètres influence pourtant l'interprétation d'un texte, d'une information, pouvant même aller jusqu'à la remise en cause de la présence de l'information.

Dans [Fillol, 1999, page 25], l'auteur étudie la notion de « sens commun » (considéré comme les savoirs ordinaires, partagés par tous). Elle précise que même des énoncés dits de « sens commun » peuvent être interprétés de façons différentes selon le point de vue de l'individu, son expérience, ses compétences, etc. L'interprétation d'un texte peut également être différente selon le contexte général de sa lecture et tout particulièrement de l'actualité. Simona Constantinovici dans [Constantinovici, 2006] souligne l'importance de cette dimension temporelle dans l'interprétation de poésies. L'auteur insiste sur le fait qu'un même individu peut avoir une lecture et une appréciation différentes d'une poésie selon l'instant. Elle ajoute, reprenant l'écrivain roumain Tudor Arghezi, que « la temporalité domine la substance textuelle ».

Dans [Perlerin, 2004, page 13], l'auteur fait référence aux différentes interprétations du terme *axis of evil* prononcé par G.W. Bush dans son discours sur l'état de l'Union de janvier 2002. Le Président des États-Unis d'Amérique, G.W. Bush, y distinguait deux types de pays dans le monde : ceux de « l'axe maléfique » (traduction française de *axis of evil*), ces pays étant la Corée du Nord, l'Iran et l'Irak et les autres pays du Monde. Ces propos ont entraîné un grand nombre

de commentaires et de critiques. L'expression a été un grand nombre de fois réutilisée et a ainsi entraîné des interprétations différentes selon les individus et les contextes d'apparition, allant de la provocation pour certains, à la déclaration d'hostilités pour d'autres. Pourtant, avec ce terme, G.W. Bush ne s'appuie pas sur une référence au monde partagée par tous et trouve une légitimité qui s'est inscrite au fil de son discours avec des références explicites vers les pays désignés.

Ces différences d'interprétations peuvent donc aller très loin et être extrêmement forcées par le contexte, surtout lorsque ce dernier est influencé par des enjeux politiques. Lors du référendum du 29 mai 2005 en France sur le Traité Constitutionnel Européen (TCE), les différents candidats proposaient et argumentaient sur des interprétations du traité leur étant très personnelles. Certains voyait dans le TCE une avancée positive dans la consolidation de la structure de l'Union Européenne, d'autres y percevaient une « porte ouverte » à l'intégration de nouveaux pays qu'ils jugeaient inaptes, d'autres encore y interprétaient une libre circulation des individus, etc. Bien évidemment, le monde politique est très particulier et ces différences d'interprétations, très souvent motivées par des raisons idéologiques, sont présentes aussi bien sur des textes, comme pour le TCE, que sur des résultats numériques et statistiques (bilans budgétaires ou sondages, par exemple).

Pour François Rastier [Rastier, 2001a], une interprétation est un ensemble de parcours interprétatifs, les parcours interprétatifs étant des suites d'opérations permettant d'assigner un ou plusieurs sens à un passage, un texte, une image[24]. De tels parcours interprétatifs sont propres à chaque lecteur : chacun fait son propre cheminement afin d'arriver à une interprétation d'un élément à un moment donné. Accéder à une information, au contenu d'un texte ou d'un ensemble de textes, est loin d'être une tâche que l'on peut facilement assister et encore moins automatiser avec un logiciel. L'interprétation du lecteur-utilisateur est un élément central qui doit donc être pris en considération, en intégrant, par exemple, une description de sa tâche, de ses centres d'intérêts, de l'actualité, etc., en quelque sorte des facteurs pouvant influencer sa lecture. Les différents logiciels proposant de tels accès se doivent donc, selon nous, de prendre en considération de tels paramètres afin de donner la meilleure assistance possible à l'utilisateur.

1.2 La place de l'utilisateur dans des tâches d'accès à l'information documentaire

La section précédente de ce chapitre a présenté et illustré les travaux existants dans notre champ de recherche : l'accès au contenu d'ensembles documentaires. Nous avons ensuite souligné l'importance de l'utilisateur et de son interprétation dans de tels accès au contenu. Dans cette section, la place laissée aux utilisateurs dans les systèmes d'accès au contenu est abordée. Nous voyons ensuite les limites rencontrées par ces systèmes ainsi que les décalages, les incohérences et l'insatisfaction des utilisateurs, causés par une faible prise en compte de leur point de vue. Enfin, des systèmes plus récents d'accès aux contenus d'ensembles documentaires, particulièrement tournés vers les utilisateurs, sont présentés.

1.2.1 La place traditionnellement donnée à l'utilisateur dans des tâches d'accès au contenu

Dans les quatre familles d'accès à l'information énoncées précédemment, une place est bien évidemment laissée à l'utilisateur. Dans les systèmes d'extraction d'information et de résumé automatique, l'utilisateur a souvent pour rôle d'évaluer les résultats des systèmes, par exemple,

[24]Le chapitre suivant reviendra plus en détail sur la notion de parcours interprétatif.

les termes extraits ou le résumé produit. Les systèmes de question/réponse et les systèmes d'accès au contenu prennent forcément plus en considération l'utilisateur : dans le premier cas, c'est lui qui pose les questions et dans le second, c'est lui qui navigue dans l'interface proposée. Même si les systèmes de ces catégories sont résolument plus tournés vers les utilisateurs, très rares sont ceux qui prennent en considération individuellement chaque utilisateur et qui ne considèrent pas chaque utilisation du système comme indépendante.

Au niveau des interactions proposées aux utilisateurs, les systèmes d'accès au contenu d'ensembles documentaires sont également souvent assez rudimentaires. Par exemple, dans les systèmes d'interrogation d'ensembles documentaires par mots-clés (tels les moteurs de recherche sur Internet), il faut remarquer que l'utilisateur et son objectif de recherche sont uniquement considérés sous la forme d'une liste de mots (dont la casse et l'accentuation sont d'ailleurs rarement prises en considération) considérée 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 telle consultation, il convient en fait d'interroger successivement plusieurs fois le (ou les) système(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 convient de développer des stratégies efficaces (souvent itératives et interactives) pour choisir des mots-clés adaptés à sa recherche et ainsi accéder aux éléments qu'il recherche.

Le rôle d'un utilisateur de systèmes d'accès au contenu d'ensembles documentaires se limite alors le plus souvent à donner des entrées aux systèmes et à vérifier les sorties des systèmes. Dans le meilleur des cas, il « aiguille » les traitements des systèmes par des règles, des configurations, puis recommence. Les systèmes n'y « voient que du feu » et considèrent chaque questionnement comme une nouvelle utilisation par une nouvelle personne. L'utilisateur est donc vu comme un sujet interprétant du résultat final produit par les systèmes alors qu'il serait peut-être plus pertinent de le considérer comme un acteur à part entière dans le traitement. La partie suivante de cette section met alors en évidence les intérêts à aller plus loin dans la prise en considération de l'utilisateur en soulignant les inconvénients entraînés par l'absence d'une telle prise en considération.

1.2.2 Les inconvénients liés à une faible place laissée à l'utilisateur dans de tels systèmes

Les quatre familles de méthodes d'accès au contenu d'ensembles de documents présentées dans la section suivante regroupent des recherches où beaucoup d'intelligence est mise en œuvre, notamment du point de vue des collaborations interdisciplinaires, par exemple entre la linguistique et l'informatique. Cependant il faut constater que la grande partie de ces recherches est toujours à l'état de prototypes logiciels, et reste, jusqu'à présent, assez peu mise en application et évaluée dans des outils accessibles au plus grand nombre, par exemple sur Internet.

Ne pas se confronter aux utilisateurs a bien évidemment des conséquences comme le montrent les auteurs de [Lavenus et Lapalme, 2002] à propos des méthodes de question/réponse. Ces derniers mettent en évidence la différence entre les corpus de référence utilisés dans les conférences TREC et des questions posées dans la pratique par des utilisateurs aux systèmes. Les auteurs notent que les questions du corpus de référence sont toutes des interrogatives courtes (par exemple *What does a defibrillator do ?*) alors que la majorité des demandes de « vrais » utilisateurs sont le plus souvent des affirmatives complexes de la forme *je voudrais savoir....*

Les méthodes d'indexation utilisées par les moteurs de recherche se soucient également assez peu des utilisateurs pour associer des mots clés potentiels à des pages de sites Internet. Une telle indexation est en texte intégral dans le sens où tous les mots figurant dans un document sont

gardés comme entrée d'index pour ce document. Dans de telles conditions, les mots grammaticaux indexent une multitude de documents. Une expérience faite le 11 mai 2007 avec le moteur de recherche *Google* donne en réponse à une recherche, certes peu pertinente, avec le mot-clé *de* 3 920 000 000 réponses[25].

Ce genre de traitements a alors de gros inconvénients, telle la taille gigantesque des index que le moteur de recherche doit archiver et être capable de consulter rapidement[26]. Une prise en compte, même simpliste, de l'utilisateur permettrait d'isoler ce genre de cas et ainsi éviter des traitements coûteux et une occupation mémoire énorme.

Dans le domaine plus large de l'ingénierie documentaire, la tendance actuelle est de faire d'Internet une vaste base de connaissances. Par exemple, le moteur de recherche *Google* essaye depuis quelques années de mettre à profit la masse de données textuelles de son index afin de proposer un système de définition de termes[27]. Jean Véronis[28] illustre cette fonctionnalité afin d'obtenir des définitions du terme « femme ». Les définitions retournées par le moteur sont alors pour le moins plus que contestables et particulièrement douteuses.

Cette démarche, faisant d'Internet un vaste champ de connaissances, est celle considérée dans le projet du Web Sémantique où l'un des objectifs annoncés par Tim Berners-Lee, initiateur du projet et directeur du W3C, est d'enrichir (notamment au moyen des technologies développées autour du langage XML) les documents (à l'aide d'ontologies normalisées, soit automatiquement, soit en assistant leurs auteurs) avec des informations sur leur propre sémantique qui soient directement interprétables par des agents logiciels sans la supervision d'une interprétation humaine [Berners-Lee, 1998]. Ceci fait l'hypothèse que la valeur sémantique d'un passage de document est le fait de son auteur alors que c'est finalement bien plus celui de son lecteur. L'expérience sur la définition de *femme* nous apprend bien qu'une définition considérée comme telle par quelqu'un n'a pas pour autant cette valeur pour d'autres et, qu'au final, c'est celle de l'utilisateur du système qu'il faudrait considérer.

Le rapport de l'Action Spécifique 32 du département STIC du CNRS [Charlet *et al.*, 2003], présentant un état récent du Web Sémantique (WS), va également dans un sens similaire. Ce rapport précise un obstacle au projet du WS : la détermination et l'ajout, même de simples métadonnées, n'est pas une activité naturelle pour la plupart des personnes. Une certaine tradition logico-grammaticale, et en sémantique formelle et computationnelle, cherche à représenter et à produire, automatiquement ou non, des formes le plus possible objectivées des significations et du sens. Or, les différents constats que nous venons de réaliser vont à l'opposé de formes objectivées des significations et du sens, et mettent surtout en évidence un besoin de formes les plus subjectivées possibles.

La « valeur sémantique » que l'utilisateur peut affecter aux sorties d'un système est cependant assez délicate à déterminer. C'est pourtant ce qu'a essayé d'approcher Jean Veronis dans le

[25] Le chiffre ainsi retourné est très certainement une forte approximation du nombre exact de résultats, les mots grammaticaux comme « de » faisant sûrement l'objet d'un traitement spécifique par rapport aux autres mots.

[26] L'intérêt principal de cette indexation brutale réside dans le fait de pouvoir garder facilement comme entrée d'index, tout ce qui, dans un texte, ne peut être retrouvé dans un dictionnaire. C'est le cas des noms propres, des expressions temporelles, ou encore, des noms de quantité dont il est difficile de dresser un catalogue fiable et durable. Une telle question du repérage et même de l'étiquetage des entités nommées est un enjeu important du traitement automatique des langues aujourd'hui. De nombreux travaux de recherche abordent cette question avec des résultats intéressants mais leurs avancées ont peu de retombées sur les méthodes d'indexation utilisées par les moteurs de recherche sur Internet (des travaux comme celui détaillé dans [Aubin *et al.*, 2006] vont cependant dans une telle direction).

[27] À l'aide de l'opérateur `define` suivi du terme à définir.

[28] Description et « analyse » de cette petite expérience disponible à l'adresse : `http://aixtal.blogspot.com/2005/04/web-la-femme-selon-google.html` (page consultée le 12 avril 2007).

cadre d'une expérience sur différents moteurs de recherche sur Internet avec plusieurs utilisateurs [Véronis, 2006]. Les utilisateurs avaient à formuler un certain nombre de requêtes prenant place dans différents thèmes imposés sur 6 moteurs « classiques », dont *Google*, *Yahoo*, *MSN Search*. Un degré de pertinence entre 0, pour un mauvais résultat, et 5, pour un excellent, est alors donné par les utilisateurs pour chaque recherche. Comme nous pouvons le voir sur la figure 1.3, les scores obtenus par les différents moteurs ne sont pas très bons. Les moteurs de recherche bien connus comme *Google* ou *Yahoo* ont les meilleurs scores mais ces derniers sont tout de même inférieurs au score moyen sur l'échelle considérée. Une telle expérience illustre que même pour des tâches quotidiennes de recherche d'information généraliste sur Internet, avec des outils particulièrement réputés et autour desquels il y a un véritable consensus, les utilisateurs ne sont pas vraiment satisfaits.

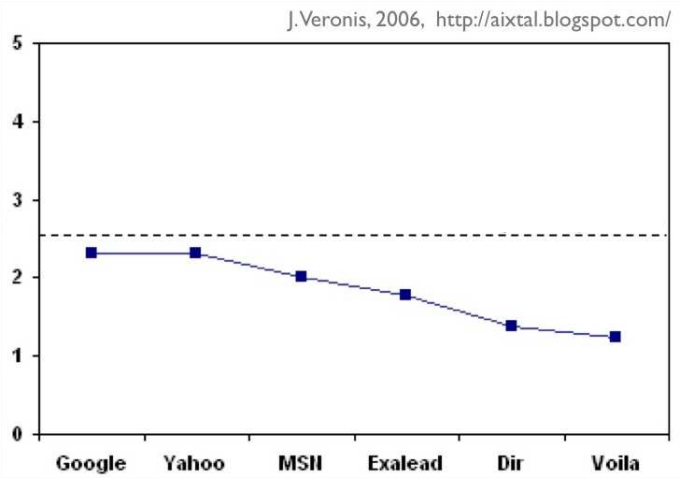


FIG. 1.3 – Résultats d'une évaluation sur six moteurs de recherche sur Internet.

1.2.3 Modèles et traitements pour des analyses personnalisées d'ensembles documentaires

Mieux prendre en considération l'utilisateur dans une tâche d'accès documentaire

Dans le rapport mentionné précédemment de l'AS 32 traitant du WS [Charlet *et al.*, 2003], un article rédigé par S. Galarti et Y. Prié traite de l'adaptation et de la personnalisation du WS aux utilisateurs [Garlatti et Prié, 2003]. Ils abordent le besoin de prendre en considération le point de vue de l'utilisateur afin de rendre le WS plus opérationnel :

> *(...) tous les utilisateurs ne sont pas intéressés par les mêmes informations et n'ont pas les mêmes attentes, connaissances, compétences, centres d'intérêts, etc. Ils ne sont capables de comprendre ou d'accepter que des services et des documents dont l'organisation, le contenu, les modes d'interaction et la présentation sont adaptés à leurs besoins.*

Pour mieux prendre en considération le point de vue des utilisateurs dans des tâches d'accès au contenu de documents et d'ensembles de documents, les auteurs font alors les trois recommandations suivantes, que nous pouvons résumer respectivement autour des trois termes-clés *cibler*, *guider* et *comprendre* :

1. **Mieux cibler les résultats des systèmes selon les besoins des utilisateurs afin d'effectuer un filtrage des ressources.** Les auteurs partent du constat que la plupart des systèmes d'accès au contenu d'ensembles documentaires, comme les moteurs de recherche, misent plus sur la quantité des résultats que sur leur qualité. L'utilisateur se retrouve alors face à un très grand nombre de résultats, et même si ces derniers sont pertinents, leur nombre les rend peu accessibles.
2. **Guider/orienter l'utilisateur dans l'espace d'information / adapter l'accès à l'information en fonction de ses besoins.** Dès qu'il est proposé à l'utilisateur de naviguer dans un document ou un ensemble documentaire de taille importante, une certaine perte de repères et une désorientation peuvent être ressenties et rendre difficile l'accès à l'information.
3. **Assister l'utilisateur dans la compréhension des documents selon son point de vue.** Afin d'augmenter la cohérence et de diminuer le surcoût cognitif[29] des accès au contenu de documents et d'ensembles de documents, les auteurs préconisent de donner à l'utilisateur différents repères l'aidant à identifier et à relier les éléments de contenu et de structure d'un document ou d'un ensemble documentaire selon le point de vue de l'utilisateur.

De telles recommandations nous semblent bien résumer les types d'aides devant être apportées aux utilisateurs pour accéder au contenu de documents et d'ensembles documentaires.

Afin de les mettre en œuvre, deux moyens complémentaires sont utilisés. Le premier consiste à proposer à l'utilisateur un certain nombre d'**interactions** sur un document ou un ensemble documentaire. Par ces interactions, prévues à l'avance par les concepteurs, l'utilisateur peut alors orienter et cibler son parcours des éléments qui lui sont présentés. Une mémorisation des parcours interactifs de l'utilisateur peut également être effectuée afin de déterminer certaines caractéristiques de l'utilisateur, comme nous pourrons l'illustrer dans la partie suivante.

Le second moyen, étroitement lié au précédent de notre point de vue, consiste à proposer des **modélisations de l'utilisateur** comme ses caractéristiques, son point de vue, ses préférences, etc. Une telle modélisation peut alors être vue comme une base de données sur ce dernier [McTear, 1993], une ontologie qui lui est propre [Razmerita, 2003], des ressources termino-ontologiques représentant son point de vue [Perlerin, 2004], etc. Les données et/ou les termes isolés et mis en relation peuvent représenter différentes caractéristiques de l'utilisateur. Par exemple, les connaissances qu'il possède sur un domaine ou encore son expérience et ses compétences, c'est-à-dire son niveau de savoir-faire et d'aisance avec le système qui lui est présenté. Ses préférences peuvent également être représentées, ainsi que ses objectifs et ses intentions sur la tâche envisagée. Nous renvoyons à [Razmerita, 2003, pages 25 à 28] pour une présentation des différentes modélisations de l'utilisateur envisagées jusqu'à présent.

Modèles et systèmes principalement basés sur les interactions utilisateurs / systèmes

Dès ses débuts, Internet s'est tourné vers les utilisateurs en leur permettant, via les hypermédias et les liens hypertextes, de réaliser leurs propres parcours de lecture. Bien évidemment, de tels parcours de liens sont bornés par les choix des concepteurs des documents. Pour essayer de fournir aux utilisateurs des documents plus adaptés, Peter Brusilovsky propose dans [Brusilovsky, 1996] la notion d'hypermédias adaptatifs. De tels hypermédias se détachent des hypermédias traditionnels du fait qu'ils ne présentent pas les mêmes pages et les mêmes hyperliens à tous les utilisateurs, en modifiant le contenu en fonction de leurs caractéristiques. Dans un même temps, Thomas Gruber [Gruber et Vemuri, 1996] définit ce qu'il appelle les documents

[29] Termes de *cohérence* et de *surcoût cognitif* définis dans [Thüring *et al.*, 1995].

virtuels personnalisables. Ce sont des documents hypermédias qui sont générés à la demande en fonction de plusieurs sources d'information et en réponse à une demande de l'utilisateur. Les technologies Web comme PHP, ASP, JSP, Flash ou encore AJAX ont largement contribué au développement de nouveaux hypermédias plus interactifs et adaptables aux utilisateurs.

L'interaction est donc souvent utilisée afin de permettre aux utilisateurs, par de simples actions sur le système, d'aiguiller leur parcours des documents, de centrer ce parcours sur certains éléments. Par exemple, le moteur de recherche sur Internet *Exalead*[30] propose à l'utilisateur différents termes jugés associés à sa recherche lors de l'affichage des résultats. De telles propositions permettent à l'utilisateur d'appréhender plus globalement le contexte de sa recherche et ainsi de s'orienter plus facilement et utilement dans l'ensemble des résultats retournés. L'appréhension du contexte global de la tâche visée par l'utilisateur est une problématique particulièrement abordée par des systèmes de visualisation de l'information, comme nous le verrons dans la partie suivante.

Les interactions entre les utilisateurs et les systèmes sont très informatives sur les utilisateurs comme le montre, par exemple, [Delépine, 2003]. Certains systèmes essaient alors de mémoriser et d'exploiter de telles actions pour représenter d'une certaine manière l'activité de l'utilisateur et pour l'exploiter dans des utilisations futures. Ainsi, la méthodologie *Hera* [Frassincar et Houben, 2002] a été utilisée pour la conception de systèmes d'information intelligents et adaptatifs sur Internet, l'adaptation se faisant, entre autre, par rapport à l'historique de navigation des utilisateurs. Dans [Nicolle *et al.*, 2002], les auteurs proposent un modèle de mémoire basé sur des interactions homme-machine. Des travaux, comme ceux détaillés dans [Fortier et Kassel, 2004], s'intéressent à la création de mémoires d'entreprise en capitalisant les connaissances et les interactions de chaque utilisateur. Nicolas Durand propose dans [Durand, 2004] d'exploiter l'historique des consultations de documents sur un réseau d'entreprise afin de faire émerger des communautés d'usagers et d'exploiter ces communautés dans un système de recommandation de documents. Dans ce cas, des catégories d'utilisateurs sont construites et décrites, l'un des objectifs est de positionner le nouvel utilisateur dans l'une des catégories existantes, la personnalisation est alors moindre que dans les modélisations considérant individuellement chaque utilisateur.

Même pour des usagers « classiques », de tels systèmes alliant interaction et mémorisation peuvent se révéler très utiles. Pour des usages plus quotidiens, nous pouvons également citer le cas des correcteurs orthographiques des traitements de textes, comme *Microsoft Word* ou *Open Office Writer* par exemple, où chaque utilisateur a la possibilité d'ajouter interactivement des termes de son choix dans son propre dictionnaire. Le moteur de recherche sur Internet *Ujiko*[31] permet également de capitaliser les différentes utilisations de l'internaute, en lui permettant d'évaluer les sites retournés (bons ou mauvais) et de visualiser des regroupements de sites autour de termes. Une capitalisation de l'expérience de l'utilisateur est également présente afin de lui proposer de nouvelles fonctionnalités au fur et à mesure des usages du moteur de recherche. Des systèmes à base de *tags* (étiquettes) sont également souvent proposés aux utilisateurs afin que ces derniers « marquent » leurs photos (avec, par exemple, les logiciels *Flickr*[32] et *iPhoto*[33]) ou leurs morceaux de musique (à l'aide des logiciels *iTunes*[34] ou encore *Amarok*[35]) à l'aide de mots-clés de leur choix. Une fois ce marquage réalisé, les systèmes proposent différents tris et sélections basés sur ces informations.

[30] `http://www.exalead.com` (page consultée de 18 mai 2007).
[31] Moteur de recherche disponible à l'adresse suivante : `http://www.ujiko.com` (page consultée le 18 mai 2007).
[32] `http://www.flickr.com` (page consultée le 18 mai 2007).
[33] `http://www.apple.com/fr/ilife/iphoto` (page consultée le 18 mai 2007).
[34] `http://www.apple.com/fr/itunes/download` (page consultée le 18 mai 2007).
[35] `http://amarok.kde.org` (page consultée le 18 mai 2007).

Modèles et systèmes basés sur des utilisateurs

Malgré toute l'intelligence mise en œuvre par certains systèmes pour s'adapter dynamiquement à leurs utilisateurs, certaines tâches semblent tout de même nécessiter un besoin d'une représentation de l'utilisateur, de son point de vue, de ses préférences. Dans [Poslad et Zuo, 2006], les auteurs présentent un modèle formel pour représenter le point de vue de l'utilisateur et utilisent ce modèle dans un système de question-réponse. Les auteurs de [Vallet *et al.*, 2006] proposent une méthode, de plus en plus utilisée à l'heure actuelle, consistant à partir d'une ontologie existante et à « positionner » l'utilisateur dans cette ontologie selon les concepts et les liens entre concepts lui semblant les plus pertinents. Un tel positionnement est alors utilisé afin de faire des communautés d'utilisateurs. Dans [Popescu *et al.*, 2006], les auteurs exploitent une technique similaire en utilisant la base de données lexicales *WordNet*[36] afin d'enrichir et d'adapter les requêtes de l'utilisateur pour une tâche de recherche d'images sur Internet.

Les approches précédentes sont plutôt basées sur des modèles formels et conceptuels du point de vue de l'utilisateur. Les approches basées plutôt sur des ressources terminologiques nous paraissent mieux correspondre à une certaine réalité linguistique dans laquelle chaque individu évolue.

C'est par exemple le cas pour le système *The Brain*[37]. Ce système est fondé sur les trois notions de *contenu*, de *pensée*, de *relation*, ainsi que sur des associations entre ces notions. Des notes et des documents peuvent alors être associés aux différentes entités, permettant ainsi de mettre en place une organisation en réseau des notes et des documents dans un objectif de travail collaboratif.

Le système *Porphyry*[38] permet un enrichissement itératif d'ensembles documentaires par différentes structures hypermédias. Ces structures, prenant la forme de graphiques, de notes, d'indications de parcours de lecture, etc., sont construites par les utilisateurs en fonction de leur tâche et de leur point de vue. L'objectif de ce système est d'aider les utilisateurs dans la lecture d'articles scientifiques.

Le logiciel *Pastel*[39] de Ludovic Tanguy [Tanguy, 1997] s'inspire de la Sémantique Interprétative (SI) de François Rastier [Rastier, 1987][40]. *Pastel* a pour objectif d'assister son utilisateur dans l'interprétation d'un texte, en lui permettant à la fois de décrire des entités lexicales du texte à l'aide des principes issus de la SI et d'exploiter ces descriptions pour l'analyse du texte.

Ces trois logiciels laissent à l'utilisateur la possibilité de décrire différents éléments d'un domaine ou d'un texte de son choix avec ses propres mots. Ils nous semblent mieux adaptés à la réalité linguistique dans laquelle nous sommes plongés. C'est dans cette voie que nous positionnons nos travaux et faisons nos propositions dans le chapitre suivant de cette thèse.

Dans cette partie, nous avons souligné l'importance de prendre en considération l'utilisateur, son point de vue, ses centres d'intérêt, pour des tâches d'analyses de documents et d'ensembles documentaires. Les interactions possibles entre l'utilisateur et le matériau étudié *via* le système informatique sont également très importantes car elles permettent à l'utilisateur de réaliser son propre parcours, sa propre appropriation du matériau. La partie suivante de ce chapitre développe l'importance des interactions entre l'utilisateur et son ensemble documentaire, mais souligne également la grande nécessité d'avoir un regard global et multi-échelle sur le matériau étudié à l'aide de représentations graphiques.

[36] `http://wordnet.princeton.edu` (page consultée le 18 mai 2007).

[37] `http://www.thebrain.com` (page consultée le 18 mai 2007).

[38] `http://www.porphyry.org/prototypes/expert/index_html/view` (page consultée le 18 mai 2007).

[39] « Programme d'Aide à l'Analyse Sémantique de TExtes, même Littéraires », `http://www.revue-texto.net/Inedits/Tanguy/Tanguy_these.html` (page consultée le 18 mai 2007).

[40] Nous revenons en détail sur la SI dans le chapitre 2 de cette thèse.

1.3 Visualisation interactive d'ensembles documentaires

Depuis très longtemps, des techniques de visualisation sont utilisées pour présenter aux lecteurs, aux experts, des vues sur des objets. Dans cette section, nous présentons la problématique de la visualisation de données complexes et plus particulièrement de données textuelles. Nous conclurons en expliquant l'intérêt que nous portons tout particulièrement aux techniques de cartographie d'ensembles documentaires.

1.3.1 Visualiser des informations complexes

Comme le remarque le vieil adage populaire « une image vaut mieux qu'un long discours », les images, les représentations graphiques de données complexes sont souvent pertinentes pour en proposer un ou plusieurs regards à leurs lecteurs. Cette pertinence est liée principalement à la spatialisation des informations qui facilite la mise en évidence de relations, aidant ainsi à une meilleure mémorisation [Card *et al.*, 1999]. Un grand nombre de techniques de visualisation ont été élaborées, telles des techniques de cartographie pour représenter des données spatiales, des techniques issues des mathématiques et des statistiques pour synthétiser des données numériques, des techniques issues de l'art pour représenter des données abstraites, etc.

Traditionnellement, deux types de données sont distinguées : les données quantitatives et les données qualitatives. Une étude des données de la première famille peut permettre de décrire l'objet, le phénomène qu'elles représentent, alors qu'une étude de données de la seconde famille cherche plutôt à apporter une explication de l'entité représentée par les données. Edward Tufte propose dans [Tufte, 2004] de passer en revue les principales techniques de visualisation de données quantitatives[41]. L'auteur s'intéresse tout particulièrement à la visualisation de données quantitatives car, selon lui, les visualisations proposées doivent tendre à décrire un objet, un phénomène plutôt qu'à chercher à l'expliquer[42].

L'auteur dégage ainsi trois familles de visualisation pour de telles données quantitatives :

1. les cartes de données (*data maps*) représentant des informations sur des éléments situés spatialement, comme par exemple des données relatives aux habitants des différentes régions d'un pays ;
2. les séries temporelles (*time series*) représentant l'évolution des données au fil du temps, telle, par exemple, l'évolution dans le temps du prix d'un produit ;
3. les graphiques narratifs de l'espace et du temps (*narrative graphics of space and time*) couplant à la fois des informations spatiales et temporelles afin de représenter un objet ou un phénomène évoluant temporellement et spatialement.

La carte présentée en figure 1.4 fait partie de cette dernière catégorie. Cette carte, réalisée par Charles Joseph Minard en 1869, symbolise les pertes des troupes françaises pendant leur déplacement lors de la campagne de Russie. L'évolution de la température est également mise en relation avec les pertes de troupes lors du trajet de retour. Une telle carte, ayant pourtant presque un siècle et demi, illustre parfaitement l'intérêt de représentations graphiques pour avoir un regard global sur des informations représentées par des données.

[41] Page personnelle d'Edward Tufte présentant ses travaux sur la visualisation : `http://www.edwardtufte.com/tufte` (page consultée le 1er juin 2007).

[42] Ce dernier souligne d'autant plus la faible confiance qu'ont les lecteurs sur des visualisations de données. Si ces dernières cherchent en plus à proposer une explication quelconque de l'entité représentée, elles seront d'autant plus critiquées [Tufte, 2004, page 53].

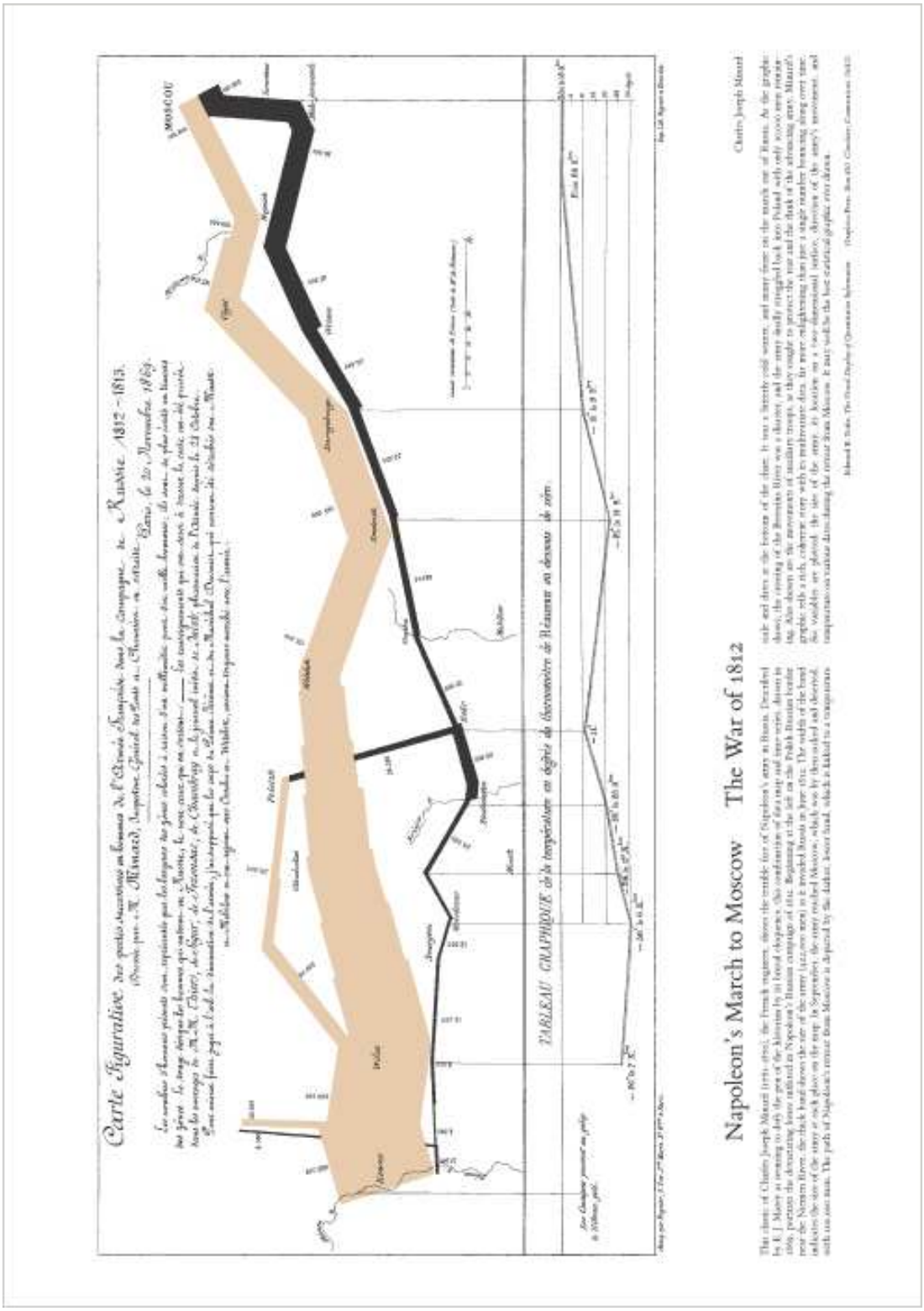


FIG. 1.4 – Carte des pertes humaines de l'Armée Française lors de la campagne de Russie.

Une « bonne » représentation graphique se doit de respecter certains principes[43] :

> *Graphical excellence is the well-designed presentation of interesting data - a matter of* substance*, of* statistics*, and of* design*. Graphical excellence consists of complex ideas communicated with clarity, precision, and efficiency. Graphical excellence is that which gives to the viewer the greatest number of ideas in the shortest time with the least ink in the smallest space. Graphical excellence is nearly always multivariate. And graphical excellence requires telling the truth about the data.* [Tufte, 2004, page 51]

De tels principes, certes généraux, résument les différentes contraintes régissant la création de vues sur des données de tous types.

Avec le développement des systèmes informatiques, des techniques de visualisation ont été particulièrement utilisées afin de mettre en évidence les différentes entités représentées. Les importantes nouveautés offertes par cette informatisation sont les grandes possibilités d'interactions offertes aux utilisateurs sur les représentations graphiques. De telles représentations deviennent alors de véritables interfaces permettant à l'utilisateur, non seulement de visualiser les objets représentés, mais aussi de les manipuler. La problématique de la construction de représentations graphiques diverses, interactives ou non, a alors suscité un grand intérêt, pour un très grand nombre de tâches. Différents travaux ont alors émergé, notamment pour la création des interfaces de systèmes d'exploitation[44] et la création d'œuvres numériques interactives ou non[45].

Parmi cette très grande variété de travaux, ceux qui nous intéressent le plus sont liés à la visualisation de textes ou d'ensembles de textes au format électronique. Le développement des systèmes informatiques conjoint à la multiplication des documents textuels électroniques a entraîné le développement de différentes interfaces de lecture, de navigation dans des textes et des ensembles de textes. Un grand nombre de travaux, dépassant grandement le cadre du traitement automatique des langues, ont alors tenté d'apporter aux utilisateurs différentes façons d'appréhender visuellement le contenu de textes ou d'ensembles de textes. Ces travaux, largement motivés par les limites atteintes par les interfaces en ligne de commandes, proposent différentes fonctionnalités, pour répondre à différents besoins. Ben Shneiderman présente dans [Shneiderman, 1996] différentes fonctionnalités devant être accessibles aux utilisateurs dans des systèmes de visualisation de textes ou d'ensembles de textes. L'auteur définit alors les sept fonctionnalités suivantes :

- avoir un aperçu global de l'ensemble de documents ;
- pouvoir zoomer sur une collection d'éléments ;
- pouvoir supprimer certains éléments sur différents critères ;
- pouvoir obtenir des détails sur demande sur un groupe d'éléments ;
- voir les relations entre les éléments ;
- pouvoir extraire des sous-collections d'éléments sur certains critères ;
- consulter un historique des actions réalisées.

De telles fonctionnalités nous semblent alors particulièrement bien résumer les manipulations devant être proposées aux utilisateurs par les différents systèmes de visualisation d'ensembles documentaires. Nous présentons certains de ces systèmes dans la partie suivante de ce chapitre.

[43] En plus de la prise en considération de contraintes perceptives, physiologiques et psychologiques expliquées, par exemple, dans [Bertin, 1983] et [Cleveland, 1993].

[44] Par exemple, selon le paradigme WIMP (Window, Icon, Menu, Pointer) encore très présent dans les systèmes actuels - `http://en.wikipedia.org/wiki/WIMP_(computing)` (page consultée le 1^{er} juin 2007).

[45] Par exemple, l'organisation *Levitated* propose de telles œuvres : `http://www.levitated.net/gravityIndex.html` (page consultée le 1^{er} juin 2007).

1.3.2 Visualiser des textes et des ensembles de textes *via* des systèmes informatiques

Dans cette section, nous présentons différents systèmes proposant des visualisations de données textuelles de tous types pour différentes tâches allant de la gestion électronique de documents à la recherche et la veille documentaire, ou encore à l'analyse lexicale, statistique ou non. Différents types de visualisations sont présentés dans cette section à l'exception des visualisations cartographiques que nous détaillons tout particulièrement dans la section suivante.

Proposer des visualisations de textes et d'ensembles de textes est une tâche visée dans de nombreuses recherches touchant à différents domaines comme la gestion électronique de documents, l'analyse statistique de textes et la recherche et la veille documentaire. Les auteurs du projet *Visual...Catalog* décrivent dans [Papy et Chauvin, 2005], l'intérêt de techniques de visualisation et d'ergonomie appliquées aux bibliothèques numériques. Dans le cadre de la navigation dans ces bibliothèques, [Fox et Kipp, 1997] présente le compte-rendu d'expériences menées pour la navigation en 3 dimensions dans un grand corpus national de thèses numérisées.

Certains outils nécessitent des appareillages dépassant le simple ordinateur personnel. C'est par exemple le cas du *Stereoscopic Field Analyser* (figure 1.5) de [Ebert *et al.*, 1996], qui, en s'inspirant de techniques issues de la réalité virtuelle, propose à l'utilisateur de s'équiper de lunettes et de capteurs électroniques pour permettre à ce dernier d'évoluer et de manipuler un ensemble documentaire.

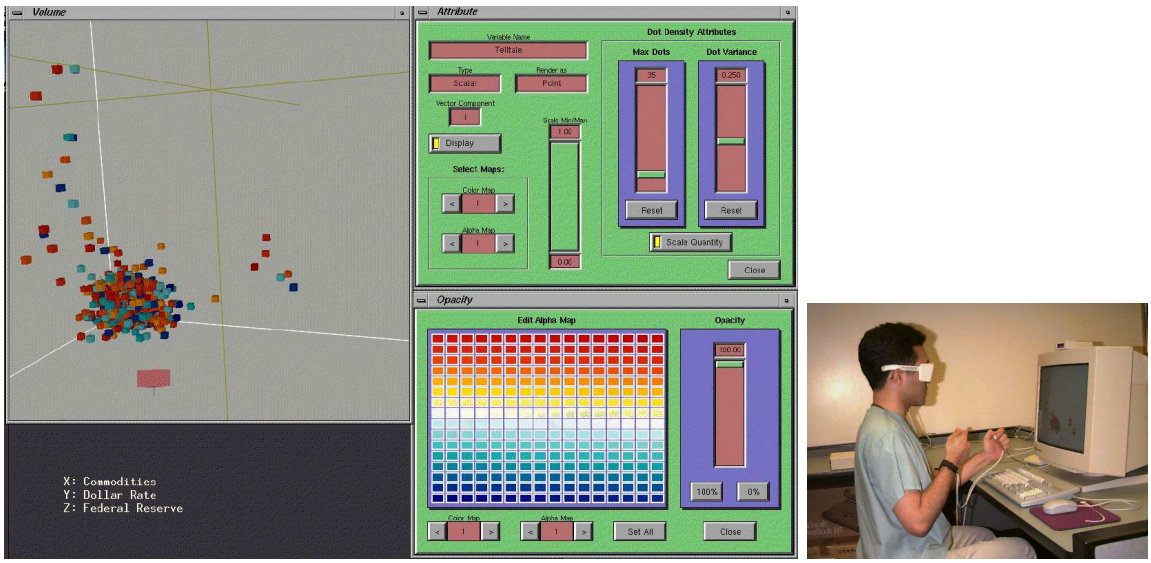


FIG. 1.5 – *Stereoscopic Field Analyser* de [Ebert *et al.*, 1996].

L'interface *3DXV* [Jacquemin et Jardino, 2002], citée précédemment, s'inscrit dans ce domaine en permettant la visualisation de grands documents XML. Dans *LibViewer* (figure 1.6, partie gauche), détaillé dans [Rauber et Bina, 2000], le résultat de recherches dans des bibliothèques numériques est affiché sous la forme d'étagères de livres en 3 dimensions.

Dans [Cubaud et Bénel, 2006], les auteurs proposent un atelier logiciel de lecture permettant à l'utilisateur d'avoir une vue sur les ouvrages d'une bibliothèque et d'en parcourir certains de façon interactive. L'originalité du système est de proposer à l'utilisateur un espace où ce dernier peut stocker des ouvrages repérés, placer des marque-pages, le tout *via* une interface en 3 dimensions (figure 1.6, partie droite).

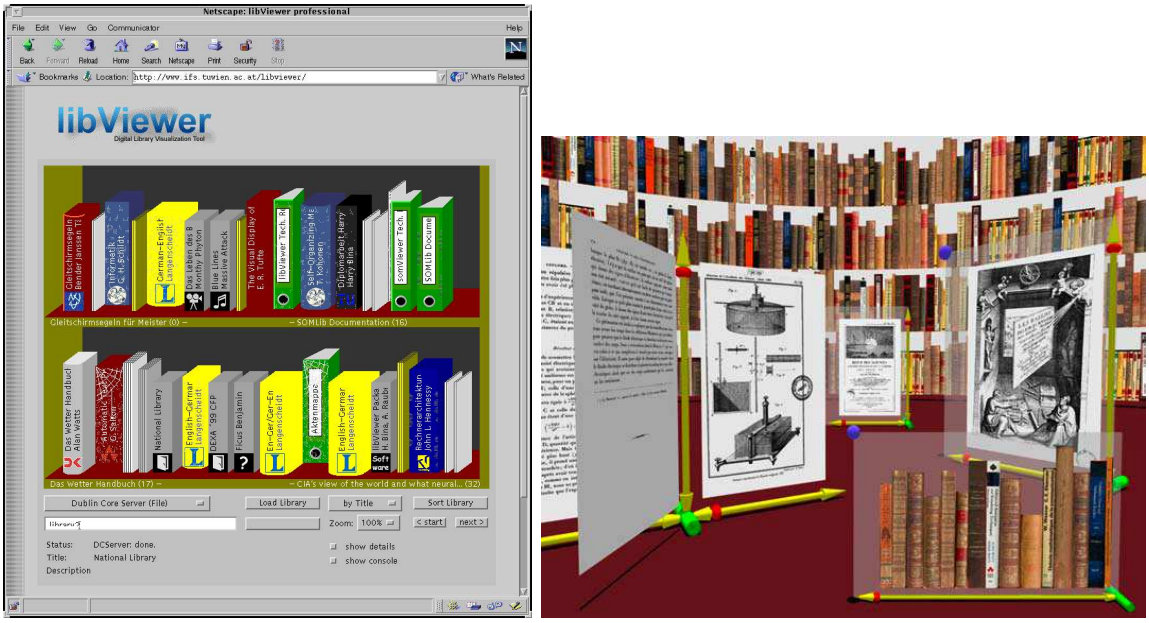


FIG. 1.6 – À gauche proposé par *LibViewer* [Rauber et Bina, 2000], à droite, l'atelier de lecture de [Cubaud et Bénel, 2006].

Les systèmes cités précédemment étudient tout particulièrement des ensembles de textes. Certains systèmes portent leur intérêt au niveau du texte. Le logiciel *Document Lens* propose une visualisation d'un texte mis en perspective avec des textes jugés voisins par le système [Mackinlay et Robertson, 1993]. Dans [Perlerin, 2004, pages 199 à 205], l'auteur propose différentes représentations du texte portant aussi bien sur sa structure que sur son contenu, en schématisant la structure par différentes boîtes de couleurs et en mettant en œuvre des techniques de visualisation en 2 et en 3 dimensions du texte au contenu colorié selon des ressources utilisateurs (cf. figure 1.7).

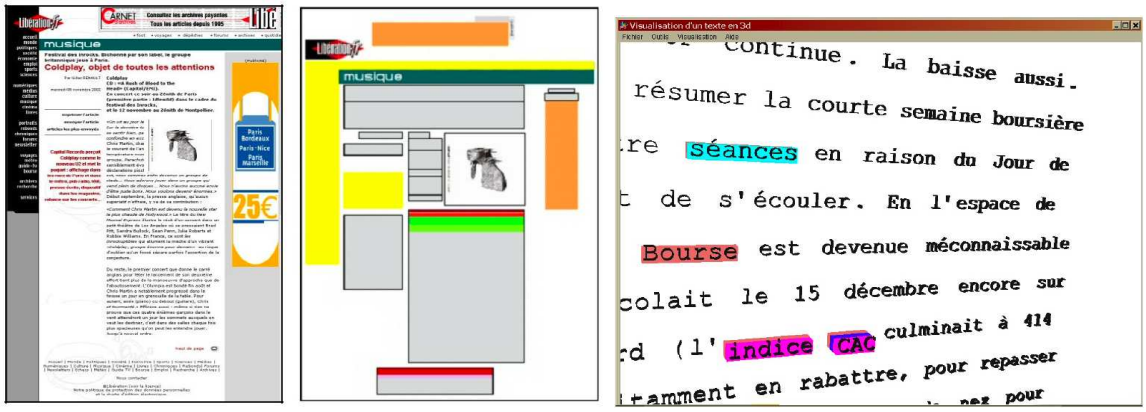


FIG. 1.7 – Visualisations au niveau du texte proposées dans [Perlerin, 2004, pages 199 à 205]. À gauche, schématisation d'un document, à droite, vue en 3 dimensions d'un texte colorié.

La visualisation de textes et d'ensembles de textes a fait appel à des techniques de visualisation de données hiérarchiques, de graphes et de réseaux. Des techniques comme celles détaillées dans [Westerman *et al.*, 2005] ont été utilisées pour présenter des ensembles de documents[46]. De telles visualisations sont, par exemple, les *Cone Trees* de [Robertson *et al.*, 1991] ou encore les *Hyperbolic Trees* de [Lamping, 1995], permettant une visualisation de hiérarchies respectivement en 3 et 2 dimensions (voir figure 1.8).

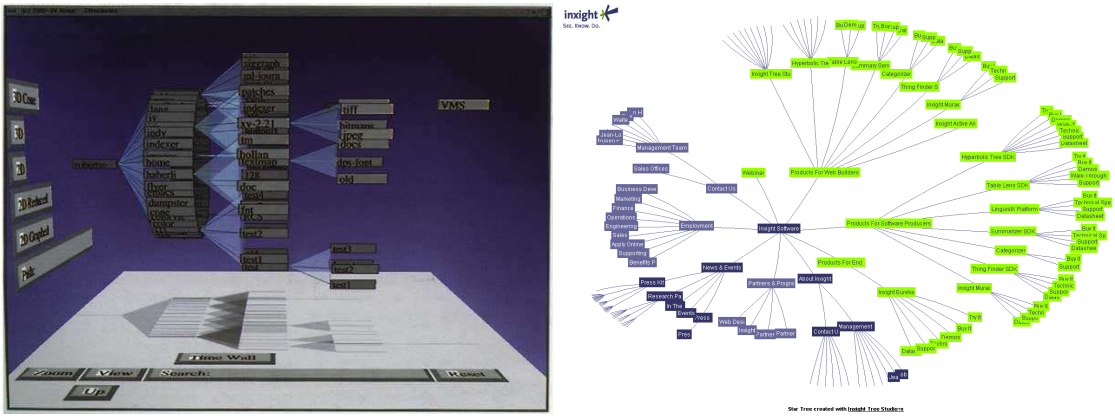


FIG. 1.8 – À gauche, le *Cone Trees* selon [Robertson *et al.*, 1991], à droite, les *Hyperbolic Trees* selon [Lamping, 1995].

Le dictionnaire des synonymes[47] du laboratoire de Sciences du Langage CRISCO de l'Université de Caen propose une visualisation des espaces sémantiques sous forme d'hypergraphes [Ploux et Victorri, 1998] (figure 1.9, partie gauche). Jean Véronis propose dans [Véronis, 2003] un algorithme permettant de déterminer automatiquement les différents usages d'un mot dans une base textuelle sans utilisation d'un dictionnaire. Cet algorithme est associé à une technique de représentation graphique permettant à l'utilisateur de naviguer de façon visuelle à travers le champ lexical associé au mot (figure 1.9, partie droite).

Le domaine des statistiques lexicales est également très actif dans l'utilisation de techniques de visualisation de données textuelles. De nombreux logiciels dédiés à l'analyse de données textuelles sont proposés, tels par exemple les logiciels *Hyperbase*[48] d'Etienne Brunet, *Sphinx Lexica*[49] de la société *Le Sphinx*, *Lexico3*[50] de l'équipe CLA^{2}T de l'université Paris III, *Tropes*[51] de la société *Acetic* ou encore *Alceste*[52] de la société *Image* (quelques visualisations d'*Alceste* sont proposées en figure 1.10). Chacun de ces systèmes propose des visualisations de résultats d'analyses de données textuelles qui dépassent les simples listes de mots ou de textes. Ces programmes informatiques calculent des projections en Analyse de Composantes Principales (ACP), des courbes, des graphiques en secteurs, etc. Beaucoup de ces systèmes permettent une interaction avec l'utilisateur pour par exemple changer des paramètres de configuration comme les couleurs, ou un déplacement de focus sur les données visualisées (navigation, zoom, etc.).

[46] Dans [Holten, 2006], Danny Holten présente un état de l'art sur les techniques de visualisation de données hiérarchiques et propose de nouvelles techniques de visualisation de grands graphes dont l'adaptation à notre problématique est tout à fait envisageable.

[47] `http://elsap1.unicaen.fr/cgi-bin/cherches.cgi` (page consultée le 2 juin 2007).

[48] `http://ancilla.unice.fr/~brunet/pub/hyperbase.html` (page consultée le 11 juin 2007).

[49] `http://www.lesphinx-developpement.fr/page.php?article=6&langue=fr` (page consultée le 11 juin 2007).

[50] `http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW` (page consultée le 11 juin 2007).

[51] `http://www.acetic.fr/tropes.htm` (page consultée le 11 juin 2007).

[52] `http://www.image.cict.fr/index_alceste.htm` (page consultée le 11 juin 2007).

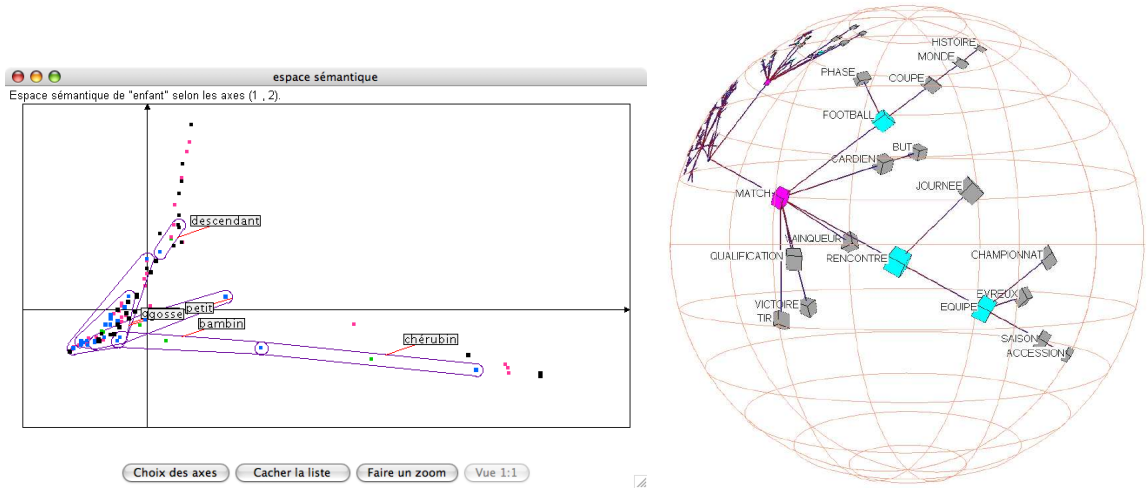


FIG. 1.9 – À gauche, l'interface proposée par le dictionnaire des synonymes pour la visualisation des espaces sémantiques, à droite, la visualisation des champs lexicaux proposées par [Véronis, 2003].

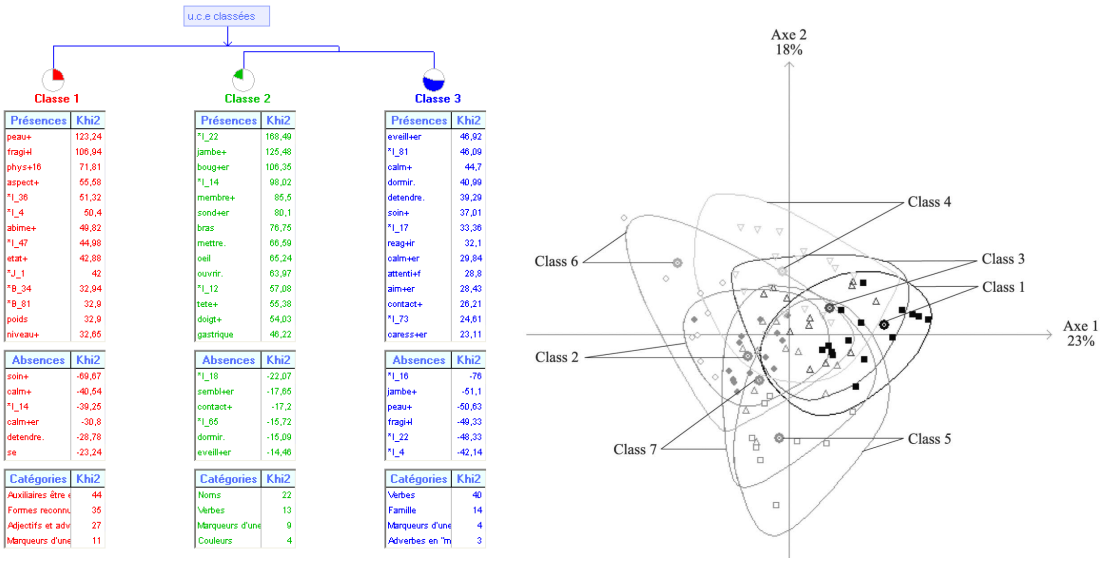


FIG. 1.10 – Différentes vues sur des données textuelles proposées par le logiciel *Alceste* de la société *Image*.

Dans le domaine de la recherche documentaire, Gerard Salton propose dans [Salton, 1989] de « projeter » sur le périmètre d'un cercle les pages obtenues en réponse à une requête. La visualisation ainsi obtenue donne une information sur les proximités entre pages ainsi retournées à l'utilisateur (figure 1.11, partie gauche). Le même auteur propose dans [Salton *et al.*, 1995] de projeter sur le périmètre d'un cercle les différents passages d'un texte et de relier ces passages selon leur similiarité (figure 1.11, partie droite).

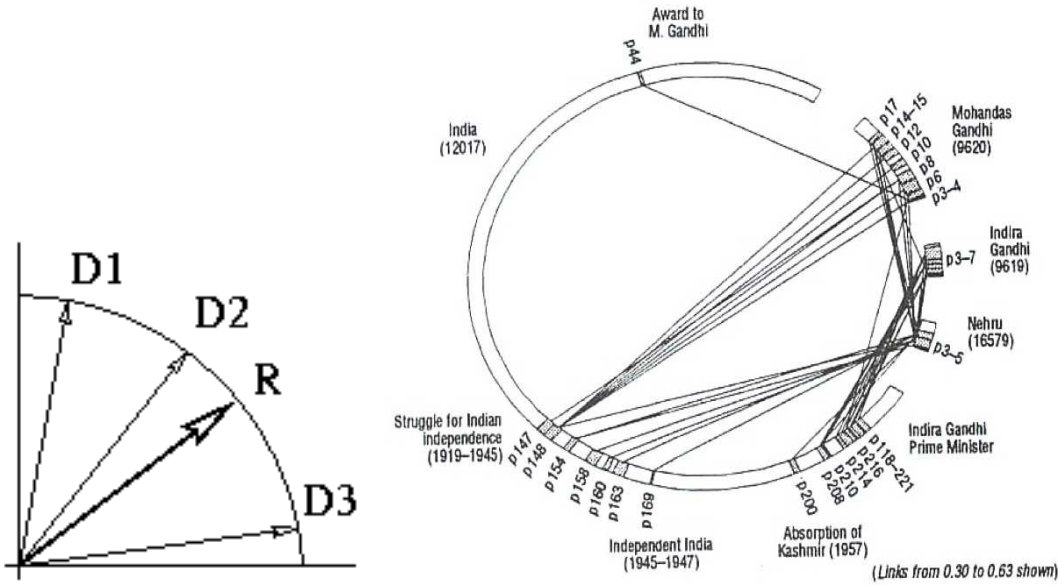


FIG. 1.11 – À gauche, projection d'après [Salton, 1989] sur un cercle des documents D répondant à une recherche d'information R, à droite, projection et lien sur un cercle entre différents passages d'un texte d'après [Salton *et al.*, 1995].

Toujours dans ce domaine, Hearst propose dans [Hearst, 1995] en réponse à une recherche un ensemble de rectangles correspondant chacun à un document jugé « pertinent » par le système. Dans ces rectangles quadrillés, chaque ligne correspond à un mot-clef de la requête et chaque colonne est grisée en fonction de la fréquence du mot-clef au sein du segment de document qui lui est associé.

Plus récemment, *The Big Picture*[53] propose une visualisation interactive sous forme de graphes d'un ensemble de dépêches d'agences de presse (figure 1.12, partie gauche). Le site Internet *TagCloud*[54] propose une vue en forme de « nuages » des termes fréquemment utilisés dans des *blogs*. Sur les nuages, les termes sont positionnés dans l'ordre alphabétique et leur taille est proportionnelle à leur nombre d'occurrences dans le texte des *blogs* considérés (figure 1.12, partie droite). Cette visualisation en nuages de mots a été reprise par Jean Véronis afin de mettre en évidence les termes abordés dans des *blogs* ou dans des articles de presse[55].

[53] `http://news.com.com/The+Big+Picture/2030-12_3-5843390.html` (page consultée le 11 juin 2007).

[54] `http://www.tagcloud.com` (page consultée le 11 juin 2007).

[55] Voir par exemple : `http://aixtal.blogspot.com/2005/11/blogs-un-nuage-sur-les-banlieues.html` (page consultée le 11 juin 2007).

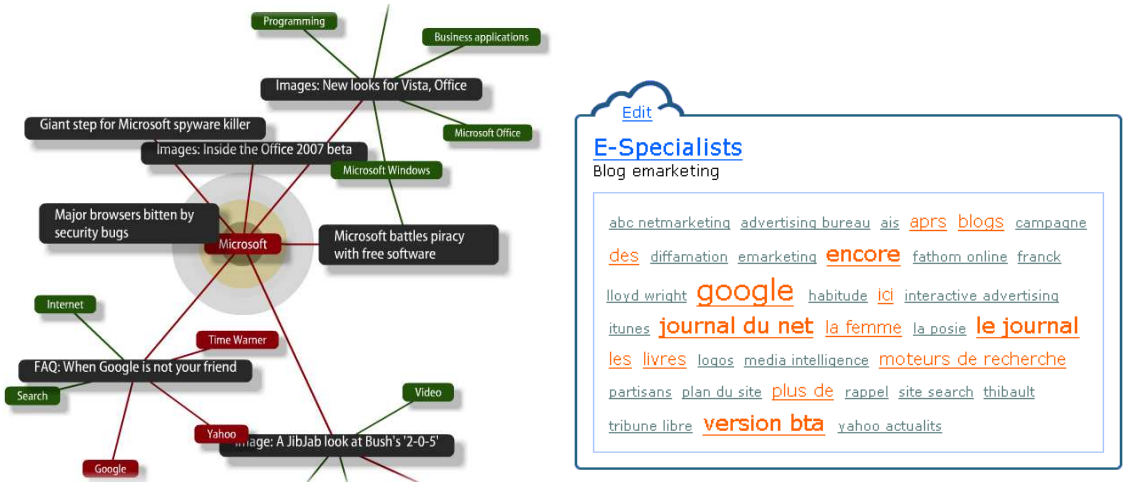


FIG. 1.12 – À gauche, visualisation des actualités proposée par *The Big Picture*, à droite, visualisation en nuages de mots du contenu de *blogs* proposée par *TagCloud*.

Dans sa thèse [Viégas, 2005], Fernanda Viégas a développé le logiciel *Mountain*[56] permettant de visualiser des archives de courriers électroniques sous forme de montagnes dont les différentes « couches géologiques » correspondent aux contacts et l'épaisseur des couches correspond à l'ancienneté du contact (cf. figure 1.13, partie gauche). Sur son *blog*, Jean Véronis propose plusieurs outils de visualisation d'ensembles documentaires. L'un de ces outils, appelé le *Chronologue*[57], permet de visualiser en fonction du temps le niveau de citation de différents termes dans des articles de grands quotidiens français (cf. figure 1.13, partie droite).

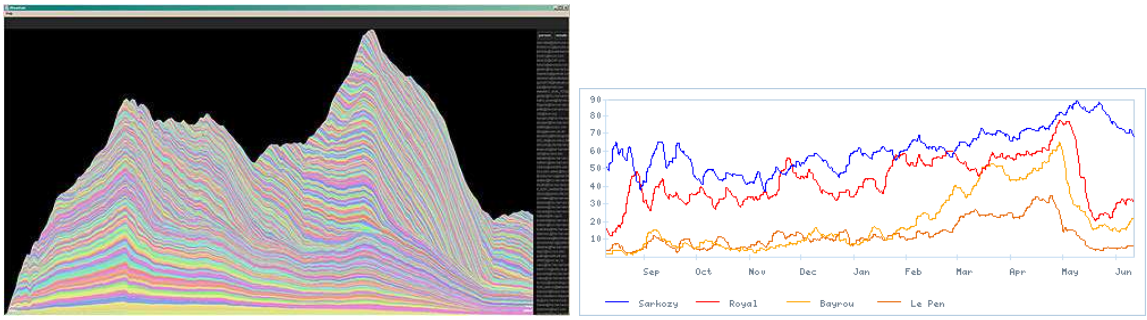


FIG. 1.13 – À gauche, *Mountain* de [Viégas, 2005], à droite, le *Chronologue* de Jean Véronis.

Les visualisations présentées dans cette section mettent en évidence la très grande variété de techniques et de cas d'applications. Une autre « famille » de visualisations, volontairement non présentée dans cette section, s'inspire de techniques issues de la cartographie géographique. Ces techniques, appliquées à la visualisation d'ensembles documentaires, ont entraîné le développement d'un grand nombre de systèmes de visualisation. Quelques uns de ces systèmes sont présentés en partie suivante.

[56] `http://alumni.media.mit.edu/~fviegas/projects/mountain/index.htm` (page consultée le 11 juin 2007).
[57] `http://www.up.univ-mrs.fr/veronis/Presse2007` (page consultée le 11 juin 2007).

1.3.3 La cartographie d'ensembles documentaires

Une technique de visualisation particulière, appelée « cartographie », est également utilisée dans un grand nombre de travaux pour des objectifs très variés. Le Comité Français de Cartographie (CFC)[58] propose de définir la cartographie de la façon suivante :

> *La cartographie est l'ensemble des études et des opérations scientifiques, artistiques et techniques, intervenant à partir des résultats d'opérations directes ou de l'exploitation d'une documentation, en vue de l'élaboration et de l'établissement de cartes, plans et autres modes d'expression, ainsi que dans leur utilisation.*

De cette définition, le CFC déduit la définition suivante d'une carte :

> *La carte est une représentation géométrique conventionnelle, généralement plane, en positions relatives, de phénomènes concrets ou abstraits, localisables dans l'espace ; c'est aussi un document portant cette représentation ou une partie de cette représentation sous forme d'une figure manuscrite, imprimée ou réalisée par tout autre moyen.*

Dans sa thèse [Tricot, 2006], Christophe Tricot[59] présente un état de l'art très complet des différentes techniques de cartographie, nous renvoyons donc à ce document pour plus d'information sur ces techniques et leur histoire. L'auteur présente également un panorama des différentes façons de passer de connaissances ontologiques à une représentation cartographique de ces connaissances. Le site *Places & Spaces*[60] réalisé par l'Université de l'Indiana (États-Unis) se propose également de recenser différents travaux exploitant la cartographie de données (géographiques ou non, textuelles ou non). Les auteurs du site proposent alors trois catégories de cartes présentées en figure 1.14 :

1. des cartes *géographiques* mettant en évidence des territoires ;
2. des cartes *conceptuelles* représentant des entités immatérielles ;
3. des cartes de *domaines* représentant des entités physiques de tout type.

Un grand nombre de cartes est proposé pour chaque catégorie, certaines cartes étant construites de façon manuelle, d'autres de façon automatique. Les trois cartes de la figure 1.14 illustrent les différentes catégories abordées ci-dessus.

Dans le domaine plus particulier de la visualisation d'ensembles documentaires, Daniel Bihanic aborde dans [Bihanic, 2003] la notion d'*hypermédias cartographiques*. Il définit de tels hypermédias comme des supports proposant des visualisations cartographiques interactives de données textuelles, documents ou ensembles documentaires. Selon lui, ce type d'hypermédia *démontre l'intérêt de visualiser un ensemble très large d'items pour accéder à l'un ou l'autre le plus rapidement possible et également de proposer, dans certains cas, des représentations adaptables aux besoins des différents utilisateurs dans l'objectif d'améliorer la compréhension du fonctionnement des processus spatiaux*. Deux points nous semblent tout particulièrement ressortir de la citation précédente : le besoin de vue globale sur les données textuelles et le besoin d'adaptabilité aux utilisateurs ; ces points mettent, selon nous, en évidence les principales fonctionnalités devant être proposées par un support de visualisation cartographique.

[58] `http://www.lecfc.fr/index.html` (page consultée le 11 juin 2007).
[59] Site Internet de l'auteur : `http://ontology.univ-savoie.fr/tricot` (page consultée le 11 juin 2007).
[60] `http://www.scimaps.org/` (page consultée la 11 juin 2007).

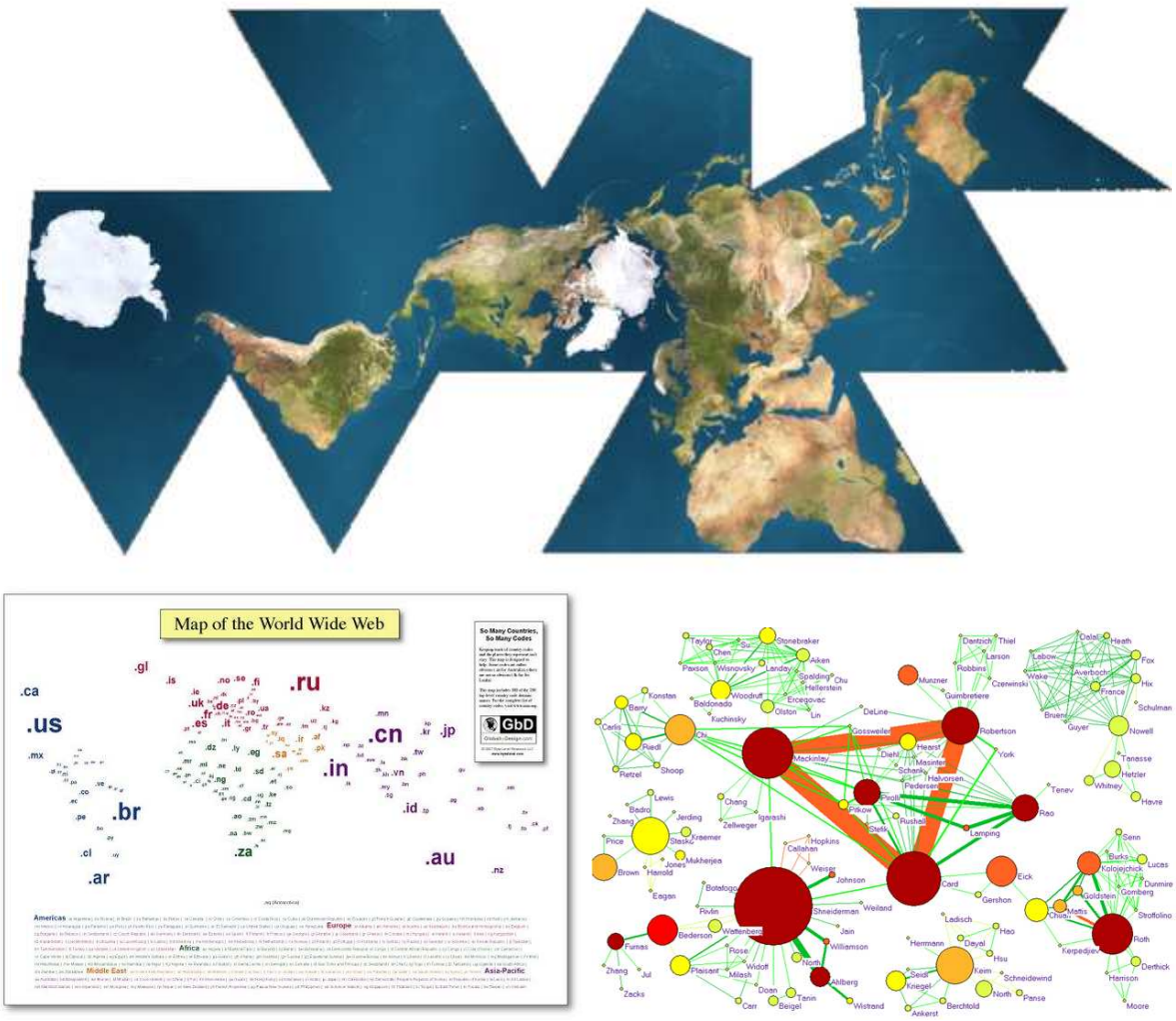


FIG. 1.14 – En partie supérieure, un exemple de carte géographique mettant en évidence les différents continents ; en partie inférieure gauche, une carte conceptuelle représentant la répartition des noms de domaines de sites internet ; en partie inférieure droite, une carte de domaines mettant en évidence les distances et les collaborations entre différents auteurs.

Selon nous, une carte d'un ensemble documentaire met en évidence des proximités et des liens entre entités textuelles (comme, par exemple, des mots, des textes, etc.) au sein de cet ensemble, un peu à la manière d'une carte routière mettant en évidence des proximités et des liens entre villes. La frontière entre certaines visualisations présentées précédemment et des visualisations cartographiques est cependant assez fine. La notion de distance entre éléments de l'ensemble analysé ainsi que les notions de vue globale et d'interactivité sont ce qui différencie une carte d'un autre type de visualisation.

De nombreux outils, tout particulièrement dédiés à la cartographie d'ensembles documentaires, sont également disponibles. En 2007, Laurent Baleydier, Président Directeur Général de la société *KartOO* S.A., a dressé une cartographie très complète des outils de visualisation cartographique accessibles aux utilisateurs[61]. Ce panorama, présenté en figure 1.15, permet d'observer à la fois le nombre d'outils existants, mais également leur variété.

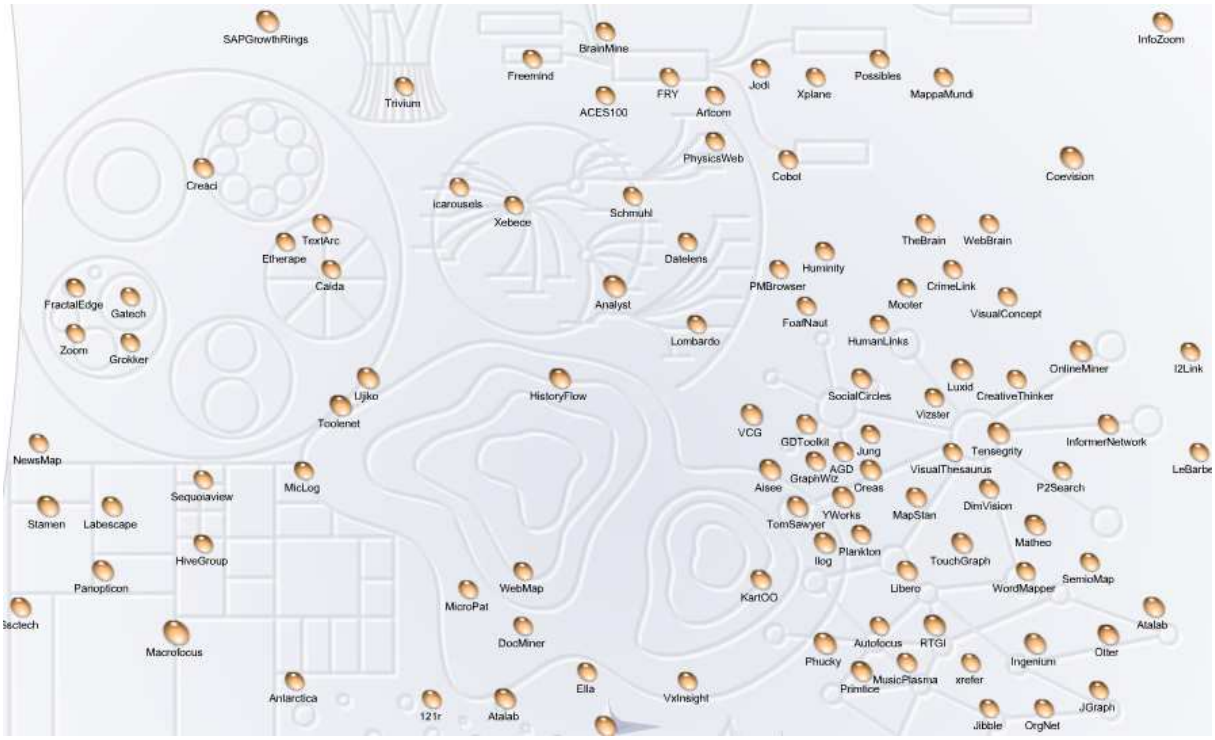


FIG. 1.15 – Cartographie des outils de cartographie proposée par Laurent Baleydier.

Le plus connu de ces outils est sans doute le métamoteur de recherche sur Internet *KartOO*[62]. Lancé en avril 2001 par la société française de Clermont-Ferrand du même nom, *KartOO* est un métamoteur de recherche sur l'Internet mondial et français. Il présente les résultats sous forme d'une carte reliant entre eux les « concepts » voisins du sujet de la recherche explicité par quelques mots-clés. *KartOO* effectue simultanément une recherche sur des moteurs de recherche « majeurs » dont *Google*, *Yahoo !*, *AlltheWeb*, etc. Fin 2003, *KartOO* s'est orienté vers la personnalisation et la veille avec des fonctions de personnalisation des résultats, de mémorisation et d'alerte.

[61] Cartographie des outils de cartographie accessible à l'adresse suivante : `http://www.mapdream.com/carte.htm` (page consultée le 11 juin 2007).

[62] `http://www.kartoo.com` (page consultée le 11 juin 2007).

Comme dans tout métamoteur de recherche, un utilisateur de *KartOO* saisit sa requête en respectant une certaine syntaxe, cette requête est ensuite transmise à différents moteurs de recherche qui fournissent chacun une liste de liens en réponse à la requête. Le fonctionnement de *KartOO* suit le schéma suivant :

- Les listes de liens retournées par les différents moteurs de recherche sont traitées et un classement de ces liens est alors réalisé ;
- Des « termes-clés » sont alors extraits des pages désignées par les liens en tête de classement ;
- Ces pages sont ensuite organisées sous forme d'un graphe : deux pages seront liées dans le graphe si elles possèdent plusieurs « termes-clés » en communs ;
- La carte est ensuite construite, les « termes-clés » les plus occurrents peuvent servir à indexer certains liens ou groupes de liens et deviennent des « concepts » selon le vocabulaire employé par les auteurs de *KartOO*.

Les points abordés précédemment présentent de façon très schématique le fonctionnement du métamoteur de recherche, nous renvoyons à [Chung *et al.*, 2002] pour une présentation plus détaillée du fonctionnement de ce métamoteur. La figure 1.16 illustre la carte obtenue en réponse à une recherche effectuée sur le mot-clé *cartographie*. On peut ainsi remarquer que des concepts se distinguent sur la carte, comme par exemple les concepts de « cartes », d'« analyse », de « logiciels », etc. Ces différents concepts donnent une information utile à l'utilisateur qui, en cliquant sur les liens contenus dans un concept particulier, aura une première idée du contenu des pages désignées par ces liens.

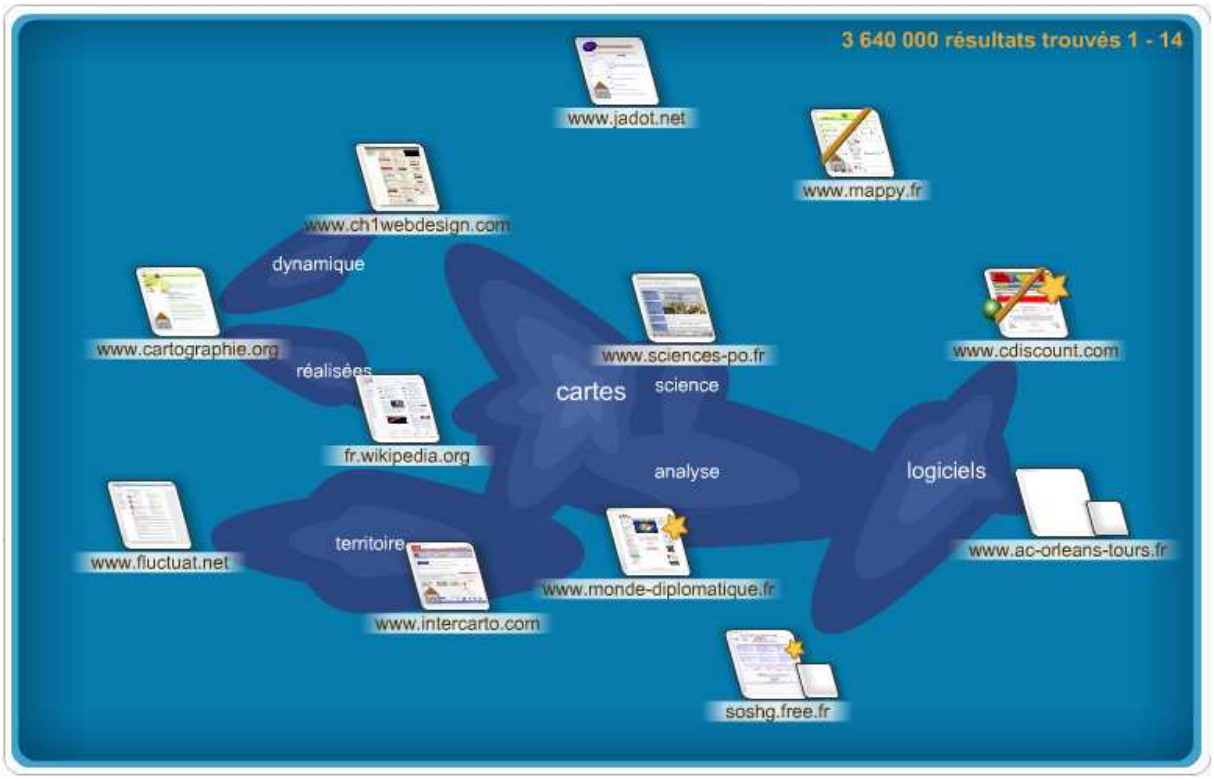


FIG. 1.16 – Résultat d'une recherche sur *KartOO* avec le mot-clé *cartographie*.

D'autres outils de cartographie sont accessibles sur Internet, tel le métamoteur de recherche *MapStan*[63] lancé en décembre 2001 par la société du même nom. *MapStan* est un outil de recherche cartographique qui propose une « représentation » d'Internet sous forme de plan de ville (quartiers, bâtiments, routes, ...) et une fonctionnalité de capitalisation des connaissances. La représentation visuelle choisie par les concepteurs, qui cartographie les résultats de *Google*, propose une ville avec des places (cercles de couleur) plus ou moins grandes selon le trafic des sites. Les places regroupent un ou plusieurs sites proches. Après l'entrée d'un mot-clé, apparaîtra une carte d'une « région » d'Internet comprenant différentes couleurs selon la proximité ou non des sites avec votre requête. La technologie utilisée par *MapStan* a été appelée le WPS (*Web Positionning System*) et fait appel à des algorithmes de filtrage collaboratif, d'analyse de corrélation et de cartographie d'informations. La technologie utilisée par *Mapstan* étant brevetée, peu de détails sur le principe de fonctionnement de l'outil sont disponibles. Cependant, selon l'un des auteurs de l'application dans [Spinat, 2002], le principe de fonctionnement de *MapStan* serait assez proche du fonctionnement de *KartOO*, avec, toujours selon l'auteur, une « capitalisation de connaissances ». Par ces termes, l'auteur désigne tout simplement le fait qu'un archivage systématique des requêtes passées sur le métamoteur est effectué à des fins de réutilisation et de complément de recherche future. La figure 1.17 illustre la carte obtenue en réponse à une recherche effectuée sur le mot-clé *intelligence économique*.

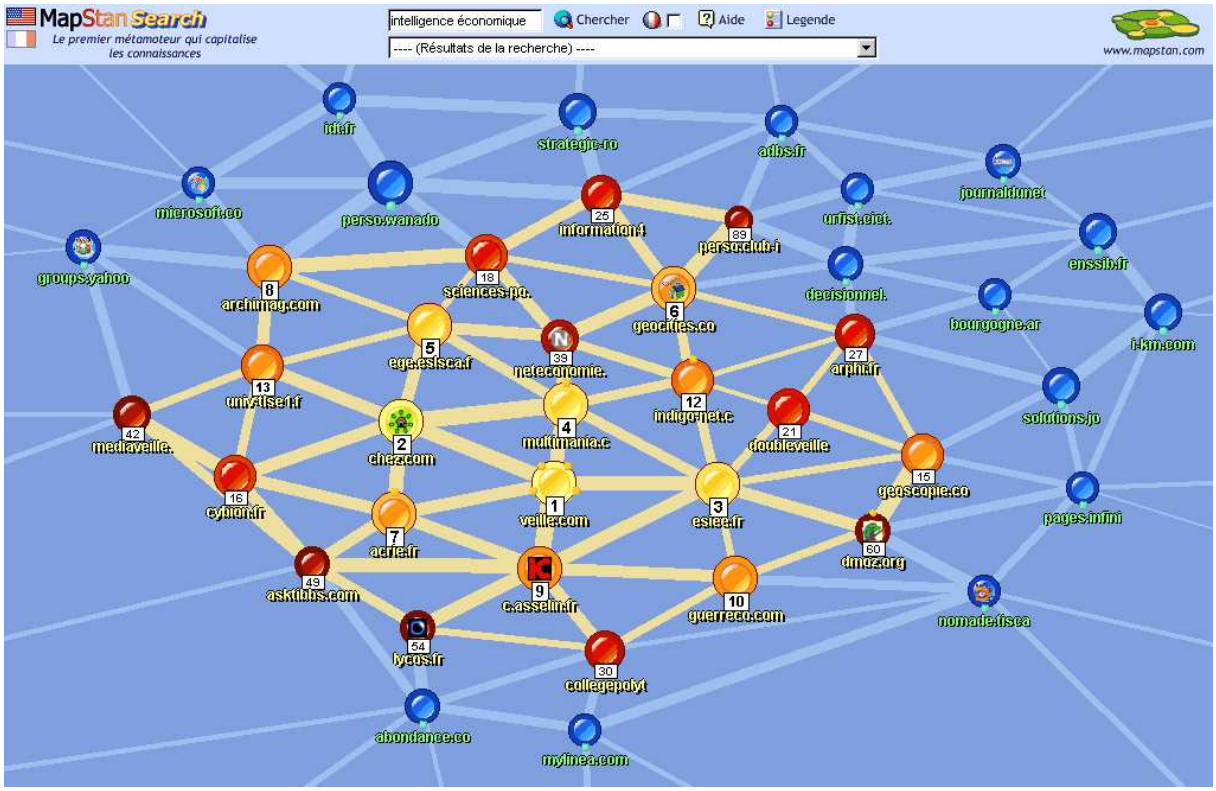


FIG. 1.17 – Résultat d'une recherche sur *MapStan* avec le mot-clé *intelligence économique*.

[63] `http://search.social-computing.com` (page consultée le 11 juin 2007).

Les deux outils précédents sont proposés pour répondre à une tâche de recherche documentaire sur Internet. D'autres tâches sont visées, un parcours de la carte des outils de cartographie présentée en figure 1.15 permet d'ailleurs de les appréhender très rapidement. Dans une tâche de veille documentaire, Abdenour Mokrane propose dans [Mokrane *et al.*, 2004] d'utiliser une technique de cartographie afin de visualiser les liens entre les principaux termes présents dans un ensemble de dépêches d'agences de presse (cf. figure 1.18, partie gauche). Le site *NewsMap*[64] propose d'avoir un regard global sur les actualités issues de grands quotidiens (cf. figure 1.18, partie droite).

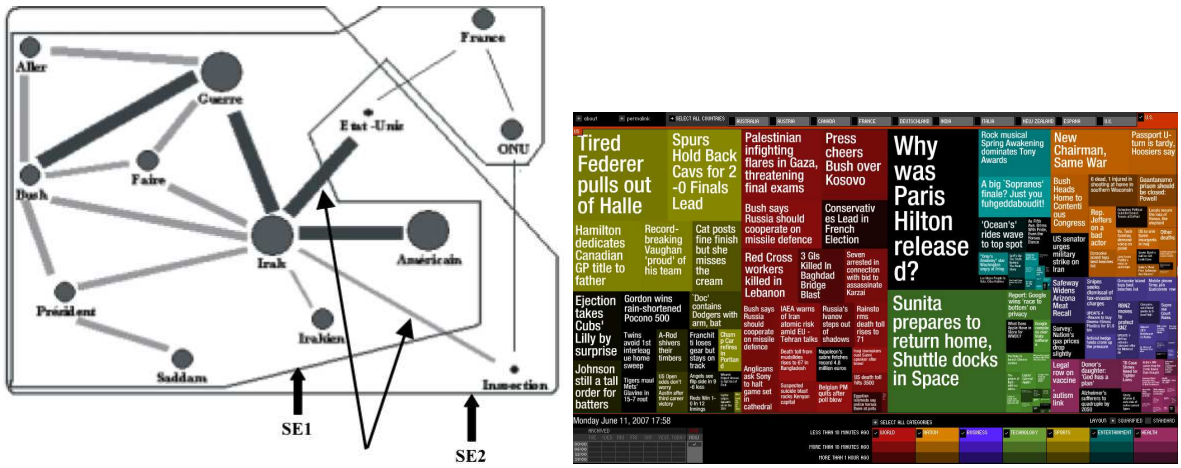


FIG. 1.18 – À gauche, une cartographie de dépêches de presse proposée dans [Mokrane *et al.*, 2004], à droite, l'interface de *NewsMap*.

Dans [Lelu et Aubin, 2001], les auteurs présentent la plate-forme *Neuronav*, de la société *Diatopie*[65], proposant un environnement cartographique d'analyses statistiques de données textuelles. La plate-forme met en œuvre différents traitements statistiques afin de produire des cartes des thèmes dans les textes à analyser. Une carte obtenue est présentée en figure 1.19.

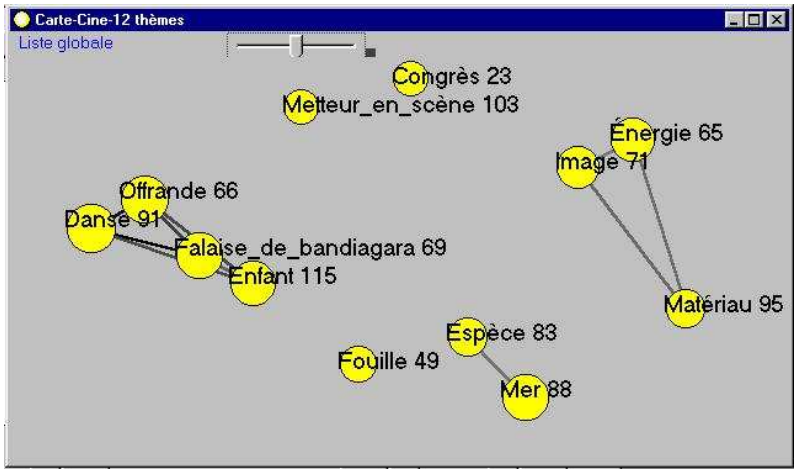


FIG. 1.19 – Cartographie thématique produite par *Neuronav* [Lelu et Aubin, 2001].

[64] `http://www.marumushi.com/apps/newsmap/newsmap.cfm` (page consultée le 11 juin 2007).
[65] `http://www.diatopie.com/ProdNeuroNav.htm` (page consultée le 11 juin 2007).

Nous avons présenté dans cette section différentes visualisations cartographiques de données de tout type et plus particulièrement d'ensembles documentaires. Pour l'accès au contenu d'ensembles documentaires, des cartographies de ces ensembles nous paraissent particulièrement pertinentes, puisqu'elles permettent, par définition, d'avoir une vue globale et métrique sur l'ensemble documentaire, et également d'interagir avec ce dernier. L'appropriation de l'ensemble documentaire par l'utilisateur s'en voit alors grandement facilitée. La cartographie nous semble alors être un bon médium entre l'utilisateur et l'ensemble documentaire qu'il souhaite analyser[66].

Conclusion : vers une meilleure prise en considération de l'utilisateur et de ses interactions

Dans ce chapitre, nous avons positionné notre travail sur l'accès au contenu d'ensembles documentaires. Nous avons insisté sur la place devant être laissée aux utilisateurs dans de telles tâches : c'est l'utilisateur et lui seul qui accède au contenu d'un ensemble documentaire. Afin de proposer une réelle personnalisation de l'accès à des ensembles documentaires, nous avons présenté différentes techniques de visualisation, en insistant finalement sur des visualisations cartographiques d'ensembles documentaires, positionnant mieux l'utilisateur au cœur de l'ensemble analysé et lui permettant d'interagir avec ce dernier.

Les notions d'ensembles documentaires, de personnalisation, de visualisations cartographiques interactives, que nous avons présentées dans ce chapitre, constituent les briques de base de notre modèle, présenté dans le chapitre suivant de cette thèse.

[66] Nous nous approchons ici du point de vue de [Ware, 2000], qui considère les cartes comme étant *un médium idéal entre un grand nombre d'informations et l'esprit.*

Chapitre 2

Un modèle d'analyse interprétative centrée utilisateur d'ensembles documentaires

Sommaire

Introduction **42**

2.1 La Sémantique Interprétative comme modèle d'analyse subjective d'ensembles documentaires **42**

2.1.1 La question du sens et de la sémantique en TAL 42

2.1.2 La Sémantique Interprétative : principes et positionnement 43

2.1.3 Notions utilisées et différents paliers textuels de la SI 46

2.1.4 Travaux en TAL exploitant la Sémantique Interprétative 49

2.2 *AIdED* : Analyse Interprétative d'Ensembles Documentaires **51**

2.2.1 Vue globale sur le modèle *AIdED* et les propositions qu'il intègre 52

2.2.2 Propositions de représentation de l'intertexte et du contexte extralinguistique 53

2.2.3 La détection des isotopies intra et inter-textuelles comme élément de base aux traitements 57

2.2.4 Des visualisations cartographiques interactives comme pour l'accès personnalisé au contenu d'ensembles documentaires 63

2.2.5 Utilisations du modèle *AIdED* 65

Conclusion : une nouvelle façon de percevoir l'accès au contenu d'ensembles documentaires **67**

Introduction

Dans le chapitre précédent, nous avons mis en évidence notre problématique de travail, celle de l'accès au contenu d'ensembles documentaires. Après avoir présenté un certain nombre de travaux prenant place dans cette problématique, nous avons souligné les manques de prise en considération de l'utilisateur et de vues globales et interactives sur les ensembles documentaires. Dans ce chapitre, nous proposons de donner plus de poids à l'utilisateur et à ses interactions dans une telle tâche d'accès au contenu. Ces propositions prennent place dans un modèle d'analyse que nous détaillons.

Dans une première partie de ce chapitre, nous présentons les différentes notions exploitées par notre modèle. Ces notions se positionnent principalement au sein de la Sémantique Interprétative (SI) de François Rastier [Rastier, 1987]. Différents travaux en TAL guidés par cette théorie sont également présentés. Certains d'entre eux, particulièrement proches de notre problématique, sont mis en avant afin d'en souligner certains éléments intégrés à notre modèle d'analyse.

La seconde partie de ce chapitre présente le modèle proposé. Nos différentes propositions sont alors détaillées. Ces propositions abordent aussi bien les liens sémantiques existant entre les textes d'un même ensemble documentaire, que la prise en considération du contexte instauré par l'utilisateur et ses domaines d'intérêt. Les différents moyens employés pour faire ressortir des informations sur le contenu de l'ensemble documentaire sont ensuite présentés, de même que les différents types de vues que nous proposons de retourner à l'utilisateur pour lui permettre une appropriation interactive de l'ensemble documentaire.

2.1 La Sémantique Interprétative comme modèle d'analyse subjective d'ensembles documentaires

Dans cette section, nous abordons la notion de « sens » et les théories sémantiques cherchant à construire, à détecter du sens dans des textes. Nous mettons ensuite l'accent sur la SI de François Rastier qui nous semble être une théorie pertinente dans notre tâche d'accès personnalisé au contenu d'ensembles documentaires. Après avoir justifié notre choix pour la SI et les notions qu'elle fait intervenir, nous revenons sur les différentes utilisations de cette théorie en informatique, et plus particulièrement en TAL, pour mieux singulariser notre approche.

2.1.1 La question du sens et de la sémantique en TAL

La construction, ou encore la détection, de façon automatique, d'éléments de sens ou de signification dans des textes, est une question très fréquemment abordée en TAL. La notion de « sens » en elle-même reste cependant assez floue. Dans [Ogden et Richard, 1923], les auteurs tentent de recenser les différentes définitions pouvant être attribuées à ce mot. Ces définitions proposent entres autres de considérer le sens d'un mot comme : une propriété intrinsèque, les autres mots liés au mot, les connotations du mot, ce à quoi l'utilisateur d'un symbole réfère, ce à quoi l'interprète d'un symbole réfère, etc. Le flou régnant autour de cette notion de sens a entraîné différentes théories, originaires de différentes disciplines, telles la linguistique ou encore la logique. Gérard Sabah dresse un état de l'art autour de cette notion dans [Sabah, 1997] et il précise ce que peut être le sens du point de vue de différentes théories appelées « sémantiques » :

- dans le cadre d'une *sémantique vériconditionnelle*, le sens d'une expression est vu comme ses conditions de vérité ;
- en *sémantique intensionnelle*, le sens d'une expression est considéré comme l'ensemble des propriétés théoriques que possèdent les concepts correspondants ;

- en *sémantique extensionnelle* (appelée également *sémantique dénotationnelle* ou encore *sémantique référentielle*), il s'agit de décrire une expression comme l'ensemble des objets, ou des situations, du monde de référence que cette expression peut désigner ;
- dans le cadre d'une *sémantique componentielle*, pour trouver le sens d'une expression, on cherche à décomposer le contenu des mots la composant en éléments de sens plus primitifs, pour étudier leurs possibilités de combinaison ;
- en *sémantique procédurale*, il convient de décrire une expression comme l'ensemble des actions à effectuer pour trouver l'objet désigné ;
- dans le cadre d'une *sémantique argumentative*, il convient de mettre en évidence les marqueurs et les constructions utilisés dans une expression pour qu'elle puisse servir comme un argument en faveur d'une autre expression ;
- etc.

Une telle catégorisation n'est pas exhaustive et confirme la grande variété de travaux portant sur la notion de sens.

Aucune de ces théories sémantiques ne permet de représenter de façon parfaite le sens de différents énoncés linguistiques. Par exemple, il existe un grand nombre d'énoncés en langue naturelle dont le sens a un certain rapport avec la vérité, mais cela ne veut pas dire que tout énoncé est interprétable en terme de vérité. Ainsi, dans [Beust, 1998, page 20], l'auteur donne l'exemple d'un énoncé simple pouvant pourtant provoquer une incohérence selon certaines théories sémantiques. Il considère tout d'abord le mot *apéritif* auquel il associe le concept de boisson alcoolisée que l'on consomme avant un repas. Dans l'énoncé suivant : *Je prendrai un apéritif sans alcool*, une incohérence peut subvenir, le mot *apéritif* étant associé au concept de boisson alcoolisée, alors que, dans l'énoncé, il est associé au qualificatif *sans alcool*. De telles « incohérences » de représentation par des théories sémantiques ne sont pas rares et nous rencontrons et produisons de tels énoncés quotidiennement. De façon plus caricaturale, nous pouvons également citer les deux phrases suivantes, la première prononcée en 2004 par Jean-Louis Debré à propos de la Corse, et la seconde attribuée à François Hollande en 2007[67] :

> *Je n'imagine pas un instant cette île séparée du continent.*
> *Jack Lang avait toutes les qualités pour briguer la Présidence de la République. C'est pour cela que je l'ai chaudement encouragé à se retirer.*

Là encore, des représentations des deux phrases, par exemple, vériconditionnelles, impliqueraient de fortes incohérences dans les interprétations associées.

Dans [Beust, 1998, pages 13 à 41], l'auteur présente une étude détaillée et exemplifiée des principales théories sémantiques énoncées précédemment. Nous y renvoyons pour plus de détails. Selon nous, il est particulièrement important de garder à l'esprit que l'une ou l'autre de ces théories sémantiques ne permet de rendre compte que d'une certaine partie du sens et ne le couvre pas dans son ensemble.

2.1.2 La Sémantique Interprétative : principes et positionnement

Principes généraux

Dans la catégorisation de Gérard Sabah présentée précédemment, nous situons notre approche dans le cadre d'une sémantique componentielle et plus précisément dans un héritage scientifique de la Sémantique Interprétative (SI) de François Rastier [Rastier, 1987], elle-même liée avec les travaux en sémantique structurale dont l'origine remonte à Algirdas-Julien Greimas

[67] Phrases sélectionnées dans le cadre du Prix *Press Club* respectivement en 2004 et en 2007.

[Greimas, 1966] et à Louis Hjelmslev [Hjelmslev, 1971]. Deux principaux aspects de la SI nous semblent la différencier des autres théories sémantiques :

- la SI propose un « appareillage » théorique fin permettant la description d'effets de sens avec, entre autres, les notions de sèmes et d'isotopies que nous définissons par la suite[68] ;
- la SI se positionne par rapport à la question du sens en préférant la tradition rhétorique et herméneutique à la tradition logico-grammaticale. De cette manière, il est défendu un principe selon lequel le global détermine le local (marquant une rupture avec le principe de compositionnalité). Ensuite, il est considéré que le sens ne peut pas être intégralement objectivé [Rastier, 1998][69].

La SI rejoint donc le point de vue de [Nicolle, 2005] selon lequel le sens n'est jamais capturé par des représentations :

> *Comme pour parler du sens d'une expression, on utilise bien sûr une autre notation, ou une paraphrase, ou une traduction, le point de vue de Pierce sur l'objet du signe comme signe lui-même, en ce qu'il est partageable dans une communauté linguistique est précieux pour comprendre que le sens n'est jamais capturé par ses représentations.*

Toute représentation du sens est alors incomplète et il n'y a donc pas de langage formel qui puisse reproduire fidèlement le sens d'un énoncé en langue naturelle alors que tout énoncé formel peut être reformulé en langue naturelle. Anne Nicolle en tire la conséquence que la langue est un langage terminal : aucun autre langage ne peut parfaitement représenter la langue. En poursuivant dans ce sens, Jacques Coursil [Coursil, 2000] définit les principes de non préméditation et de non consignation de la chaîne parlée. De tels principes expriment que les énoncés linguistiques prononcés par chacun ne sont ni entièrement construits avant leur prononciation, ni entièrement mémorisés après leur prononciation. Il est ainsi défendu un principe, que nous partageons, qu'il n'y a pas de forme non-linguistique du sens.

Notre positionnement sur la SI

Nous avons choisi de nous positionner au sein de la SI tout particulièrement en raison de la place laissée à la subjectivité de l'individu. Dans le cadre de la SI, l'interprétation est considérée comme une perception sémantique individuelle, dont toute tentative d'objectivation est une sommation incomplète de points de vue. Le sens d'un texte est une interprétation à un moment donné, dans une tâche donnée pour un individu donné. Un tel positionnement est orthogonal à un grand nombre de travaux en sémantique formelle, comme la DRT[70] [Kamp, 1981] ou la SDRT[71] [Asher, 1993]. Ces travaux déploient beaucoup d'intelligence depuis des années pour obtenir un « calcul du sens » acceptable, même s'il est possible de constater qu'un tel résultat n'est toujours pas atteint à l'heure actuelle. Il ne s'agit pourtant pas ici d'un problème d'évaluation dont on n'aurait pas encore bien mis en place la méthodologie, mais d'un problème beaucoup plus profond. Dès lors qu'on parle de « vrais » textes, et pas simplement de phrases exemples artificiellement construites en dehors d'un contexte linguistique et pragmatique, il convient de se rendre compte que la dimension interprétative personnelle fait qu'il n'y a pas systématiquement de consensus évident sur ce qu'est ou n'est pas le sens d'un texte. Il en résulte, à notre avis, que le

[68] Par exemple, dans l'énoncé précédent *je prendrai un apéritif sans alcool*, la SI proposerait de ne pas tenir compte du « contenu sémantique » d'*apéritif* lié à l'alcool (une telle opération est une *virtualisation* comme nous le verrons par la suite).

[69] Ce dernier point allant totalement à l'encontre des théories cherchant à représenter le sens de manière formelle, comme cela est souvent le cas dans la tradition logico-grammaticale.

[70] *Discourse Representation Theory.*

[71] *Segmented Discourse Representation Theory.*

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. Ce ne sont pas tant les caractéristiques propres des mots, des phrases ou des paragraphes qui priment dans le sens des textes mais c'est ce que les interprétants en attendent ou y projettent.

La SI suit une approche anti-compositionnelle du sens, se distinguant du principe de compositionnalité fondateur d'une certaine tradition logico-grammaticale. D'après [Partee *et al.*, 1990], le principe de compositionnalité du sens est énoncé de la façon suivante :

> *Le sens d'une expression composée ne dépend que du sens de ses composants et des règles syntaxiques par lesquelles ils sont combinés.* (traduction de l'anglais)

Un tel principe a été suivi par un très grand nombre de travaux basés sur des représentations formelles d'énoncés. Notamment, Noam Chomsky l'exploite pour représenter la structure syntaxique de phrases dans sa grammaire générative [Chomsky, 1969]. À contre-courant de ces travaux, la SI se positionne avec un principe de détermination du local par le global [Rastier, 2001a]. Par exemple, le sens d'une phrase n'est pas vu comme la composition du sens des lexies la composant. L'inverse est proposé : c'est le contexte global, par exemple celui de la phrase, qui permet d'attribuer du sens à des éléments locaux, comme, par exemple, les lexies. Une telle « attribution de sens » se fait principalement par des mises en contexte, des rapprochements et des différenciations entre éléments locaux dans le contexte global. Ce principe de détermination du local par le global, et l'importance donnée à la contextualisation, sont exprimés par François Rastier dans [Rastier, 2001a, page 92] :

> *L'activité interprétative procède principalement par contextualisation. Elle rapporte le passage considéré, si bref soit-il (ce peut être un mot) à son voisinage selon des zones de localité (syntagme, période) de taille croissante ; à d'autres passages du même texte, convoqués par des procédures d'assimilation et de contrastes ; enfin à d'autres passages d'autres textes choisis dans le corpus de référence, et qui entrent ainsi dans le corpus de travail.*

C'est donc le positionnement en « global » d'un élément local qui permet son interprétation[72]. Cette importance assumée du global et son influence sur le local sont des arguments que nous partageons. De notre point de vue, de tels arguments s'inscrivent pleinement dans notre problématique afin de rendre au contexte global de l'ensemble documentaire un rôle prépondérant dans des tâches d'accès à son contenu.

De cette détermination du local par le global, François Rastier propose les trois principes suivants [Rastier, 2001a, page 92] :

1. le principe de contextualité : *deux signes ou deux passages d'un même texte mis côte à côte sélectionnent réciproquement des éléments de signification (sèmes).* ;
2. 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).* ;
3. et 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.*.

Les deux premiers principes illustrent l'importance du rapprochement d'éléments locaux différents dans un même contexte global afin de faire ressortir des éléments de signification. Le dernier principe, le principe d'architextualité, nous intéresse tout particulièrement puisqu'il exprime une

[72] Ce principe de détermination du local par le global a été abordé par Ferdinand de Saussure précédemment aux travaux de François Rastier : *La valeur d'une forme est tout entière dans le texte dont on la puise.* [de Saussure, 1972, page 351] cité dans [Rastier, 2007].

détermination du local par le global entre les paliers du texte et du corpus. Un texte placé par un individu dans un corpus va ainsi hériter des propriétés de ce corpus. Une double détermination est cependant énoncée dans ce principe, outre le global qui détermine le local, le local va lui aussi influencer le global : un texte ajouté par un individu dans un corpus va modifier légèrement la cohérence de ce dernier du fait de ses particularités.

La tâche d'accès personnalisé à des ensembles documentaires que nous visons fait grandement intervenir un tel rapport entre le global (l'ensemble documentaire) et le local (le texte). En schématisant ce principe d'architextualité, et en l'étendant aux ensembles documentaires, on peut, par exemple, facilement imaginer qu'un article de presse présenté dans un journal très sérieux ne sera pas perçu par le lecteur de la même manière que le même article présenté dans les colonnes d'un journal satyrique, le contexte global instauré par les autres articles aiguillant forcément la lecture.

Cette reconnaissance de critères globaux externes aux textes et influençant leur interprétation dépasse largement le cadre de la SI. Par exemple, dans le domaine de la recherche d'information sur Internet, *Google* fait intervenir, dans le classement des pages retournées aux utilisateurs, l'algorithme *PageRank* [Page *et al.*, 2001]. Cet algorithme conditionne tout particulièrement la pertinence d'une page à l'aide du principe suivant : plus il existe de pages qui ont un lien vers la page donnée, plus cette page est pertinente (indépendamment de son contenu et des mots-clés de la requête). En intégrant un tel algorithme, le moteur de recherche est en quelque sorte « socio-centré » en donnant plus de poids à des pages particulièrement désignées par d'autres pages et par leurs auteurs. Ceci met alors en évidence que quelque chose ne suivant pas un principe entièrement compositionnel, tel que le *PageRank*, n'en est pas moins calculable.

2.1.3 Notions utilisées et différents paliers textuels de la SI

La SI propose de définir (et de redéfinir) un grand nombre de notions liées à l'analyse sémantique de textes. Nous présentons certaines de ces notions en rapport avec notre modèle dans cette section[73].

Lexies

Une lexie est une unité fonctionnelle (que l'on peut considérer également comme une entrée lexicale) regroupant plusieurs morphèmes et qui peut correspondre à plus d'un mot. Deux types de lexies peuvent être distingués : les lexies simples et les lexies complexes selon qu'elles consistent en un mot graphique (comme les lexies *eau* et *marcher*) ou en plusieurs (comme la lexie *pomme de terre*). Dans [Rastier *et al.*, 1994], la lexie est vue comme « une unité de signification » et les auteurs la définissent comme un élément de base à toutes analyses sémantiques de textes. Une lexie est représentée dans un texte par une ou plusieurs formes graphiques : ses flexions et éventuellement d'autres graphies. Ainsi, la lexie *cassette* peut apparaître en ensembles documentaires sous plusieurs graphies : *cassettes* (flexion plurielle de la lexie) ou encore *K7*, *K-7*, etc.

Sèmes et isotopies

Le sème est défini comme la plus petite unité de signification [Rastier *et al.*, 1994], il est le plus souvent exprimé sous la forme d'un trait (entre deux « / »). Par exemple, la lexie *chien* pourra porter les sèmes /mammifère/, /possède des crocs/, /animal domestique/, etc. Comme

[73] Pour une présentation plus détaillée des différentes notions et des notions liées, nous renvoyons au « Petit glossaire du sémanticien » de Louis Hébert présenté sur le site de la revue *Texto !* : `http://www.revue-texto.net/Reperes/Glossaires/Glossaire_fr.html` (page consultée le 20 juin 2007).

nous le présentons ci-dessous, différents sèmes sont distingués en SI : les sèmes inhérents et afférents, et les sèmes spécifiques et génériques, les deux premiers pouvant qualifier les deux seconds.

Une isotopie est, par définition dans [Rastier, 1987], un *effet de la récurrence d'un même sème*. Cet effet de récurrence permet alors de caractériser et d'identifier l'importance de certains sèmes dans une phrase, dans un texte. Dans l'énoncé *le facteur m'a donné une lettre*, le sème /courrier/ est actualisé dans le contenu de *lettre* parce qu'il se répète dans le contenu de *facteur*, formant ainsi une isotopie. Cette actualisation permet de retenir la signification pertinente de *lettre* dans l'énoncé (on ne retient donc pas, par exemple, la signification de *lettre* en tant que caractère de l'alphabet). Ainsi, dans cet exemple, le sème /courrier/ est renforcé dans l'énoncé alors que ce n'est notamment pas le cas du sème /en papier/ appartenant également au contenu de *lettre*. À l'inverse, il y aura probablement actualisation de ce sème dans *il a chiffonné sa lettre* (et pas du sème /courrier/).

La virtualisation est l'opération inverse de l'actualisation. Elle décrit une neutralisation d'un sème en contexte. Par exemple, dans le syntagme *le chat immortel*, on dira que le sème /mortel/, appartenant au contenu de *chat*, est virtualisé car, non seulement, il n'est pas répété dans l'énoncé, mais de plus, il est invalidé par le contenu sémique de *immortel*. L'actualisation et la virtualisation jouent un rôle important sur la mise en évidence dans le texte de contenus sémiques définis en langue, que la SI appelle les sèmes *inhérents*.

La notion de sème inhérent est à opposer à celle de sème *afférent*. Dans des contextes particuliers, on peut actualiser des sèmes qui ne font pas partie des contenus des unités du texte. Ces sèmes sont dits afférents (l'opération qui consiste à les actualiser porte le nom d'afférence). L'afférence consiste en la production d'un sème qui vient créer ou renforcer une isotopie. Ainsi, dans l'énoncé *je vous ai mis une tarte aux fraises, un éclair au chocolat et un escargot*, le sème afférent /pâtisserie/ est attribué à *escargot*[74].

Thème générique et thème spécifique

Les sèmes inhérents et afférents interviennent dans la mise en place des isotopies. La SI propose de distinguer de tels sèmes en sèmes *génériques* et sèmes *spécifiques*. Un sème générique marque l'appartenance d'une lexie à une classe sémantique. Par exemple, la lexie *femme* possède le sème générique /être humain/. Un sème spécifique différencie une lexie au sein de la classe. Ainsi, le sème /sexe féminin/ permettra de singulariser *femme* dans la classe des êtres humains. La notion de thème est alors étroitement liée aux sèmes génériques et spécifiques et à leurs récurrences au sein d'isotopies génériques et spécifiques.

François Rastier distingue alors le thème générique et le thème spécifique. Il définit tout d'abord, dans [Rastier, 2001a, pages 38-39], le thème générique comme *un sème ou une structure de sèmes génériques récurrents*. Une telle récurrence définit ainsi une ou plusieurs isotopies

[74] Pour Thierry Mézaille dans [Mezaille, 2005], il y a afférence d'un sème dès lors qu'il y a une relation sémantique d'assimilation ou de dissimilation. L'assimilation est une afférence de sème(s) générique(s) dans un co-texte pour renforcer une répétition déjà établie, comme dans l'exemple précédent *je vous ai mis une tarte aux fraises, un éclair au chocolat et un escargot* où *escargot* s'est vu attribuer le sème afférent générique /pâtisserie/. La dissimilation est une afférence de sème(s) spécifique(s) pour différencier les contenus sémantiquement proches. Par exemple, dans l'énoncé *il y a musique et musique*, la dissimilation permet de distinguer deux significations de *musique*, la première avec le sème spécifique /agréable/ et la seconde avec le sème /désagréable/. De telles afférences sont dites contextuelles (le contenu sémique d'une lexie est enrichi à l'aide du contenu d'autres lexies du contexte). Des afférences socio-normées sont également proposées en SI, l'enrichissement est, cette fois-ci, externe au texte, il est lié à une norme sociale partagée au sein d'une communauté linguistique. C'est, par exemple, le cas du sème /tristesse/ afférent au contenu de *noir* dans *il broie du noir*, ou encore, le cas du sème /bonheur/ dans *rose* de *la vie en rose*.

génériques déterminant le ou les sujets abordés dans le texte. Les thèmes spécifiques sont vus comme *des groupements récurrents de sèmes spécifiques*. Ils ne sont pas nécessairement dépendants d'une lexicalisation particulière. Ainsi, dans [Rastier *et al.*, 1994, page 178], l'auteur prend pour exemple *L'Assommoir* d'Émile Zola où un thème spécifique est défini par un groupement sémique récurrent : /jaune/, /chaud/, /visqueux/ et /néfaste/, lexicalisé par *alcool*, *sauce*, *morve*, *huile*.

Contrairement aux différentes notions présentées jusqu'ici, la notion de thème est particulièrement exploitée en dehors de la SI. Une définition assez simple revient à considérer les thèmes d'un texte comme les différents sujets qui y sont abordés [Pichon et Sébillot, 1999] (on retrouve alors le thème générique de la SI). Cette considération se retrouve également dans l'opposition thème/rhème que l'on peut faire au sein d'une unité textuelle. Le thème représente alors de quoi on parle dans cette unité et le rhème représente ce que l'on en dit. Jean-Marie Schaeffer dans [Ducrot et Schaeffer, 1995, page 638] souligne la difficulté de définir la notion de thème et attire l'attention du lecteur sur l'absence de consensus autour de cette notion. Les auteurs citent deux « définitions » du thème. La première, tirée de [Richard, 1961], considère le thème comme *un principe concret d'organisation, un schème (...) autour duquel aurait tendance à se constituer et à se déployer un monde*. La seconde, tirée de [Collot, 1988], définit le thème comme *un signifié individuel, implicite et concret*. Ces deux définitions considèrent le thème plutôt comme un concept que comme une propriété linguistique possédée par une unité textuelle. François Rastier dans [Rastier *et al.*, 1995, page 224] considère les définitions précédentes comme étant plutôt d'ordre philosophique que d'ordre linguistique.

Sens, interprétation et parcours interprétatif

Le thème selon François Rastier repose essentiellement sur la notion d'isotopie. En désignant des récurrences de sèmes (génériques ou spécifiques) au sein de suites linguistiques, les isotopies mettent ainsi en évidence des informations sur le contenu, sur le sens des suites considérées. Le sens est vu en SI comme un ensemble de sèmes inhérents et afférents actualisés dans un énoncé, dans un texte. Il est admis également que le sens se détermine relativement à un contexte et à une situation, notamment au sein d'une pratique sociale. Le contexte d'une unité sémantique est défini comme l'ensemble des unités qui ont une incidence sur elle (contexte actif), et sur lesquelles elle a une incidence (contexte passif).

L'interprétation d'une suite linguistique est ainsi perçue comme l'assignation d'un sens à cette suite. Cette assignation est réalisée de façon dynamique, par exemple lors de la lecture linéaire d'un livre, en parcourant une encyclopédie, lors d'une navigation hypertextuelle entre des pages de sites Internet, etc. Elle se fait en plusieurs « aller-retours » avec le matériau textuel parcouru, le temps de révéler les isotopies appropriées. Cette activité est désignée en SI par la notion de parcours interprétatif. Un parcours interprétatif est une séquence d'opérations permettant d'assigner un ou plusieurs sens à une suite linguistique. Ces opérations sont, par exemple, des opérations d'actualisation, de virtualisation ou d'afférence que nous avons abordées précédemment.

Paliers textuels et contextualisation du sens

Les parcours interprétatifs sont réalisés à différents niveaux ou paliers textuels, comme la phrase, le texte, etc. Comme nous l'avons énoncé précédemment, notamment avec le principe d'architextualité, la SI considère des paliers supérieurs au texte.

Le principe de détermination du local par le global, propre à la SI, est un principe de mise en

contexte du sens. La question de la contextualisation du sens d'une unité linguistique (un mot, un syntagme ou un paragraphe, par exemple) est d'une grande importance dans la perspective de la SI. Elle constitue la base de l'établissement de parcours interprétatifs. Selon François Rastier, la contextualisation est à la base de l'interprétation d'une unité linguistique [Rastier, 2001b] :

> *L'interprétation procède principalement par contextualisation. Elle rapporte le passage considéré, si bref soit-il ? ce peut être un mot : (i) à son voisinage, selon des zones de localité (syntagme, période) de taille croissante ; (ii) à d'autres passages du même texte, convoqués soit pour des tâches d'assimilation, soit de contraste ; (iii) enfin à d'autres passages d'autres textes, choisis (délibérément ou non) dans le corpus de référence, et qui entrent, par ce choix, dans le corpus de travail. Aucune de ces trois contextualisations n'est déterministe, au sens de mise en Intelligence Artificielle, qui suppose un parcours linéaire mot à mot.*

Pour décrire une telle contextualisation du sens, trois notions sont donc à prendre en considération : le *co-texte*, le *contexte extralinguistique* et l'*intertexte* [Rastier, 2001b]. Ces trois notions peuvent être définies de la façon suivante :

- on entend par *co-texte* d'une unité linguistique son « entourage » dans le texte, c'est-à-dire un passage de texte : une zone de localité sémantique pertinente autour d'une unité. Cette zone est appelée *période* [Rastier *et al.*, 1994, page 116] et elle est délimitée par l'étendue des relations d'isotopies, de prédications et d'anaphores ;
- le *contexte extralinguistique* regroupe les conditions pragmatiques liées à l'interprétation du texte.
- et l'*intertexte* rassemble tous les textes que l'utilisateur estime liés à un texte du point de vue de son interprétation.

Contextualiser revient à établir au sein du co-texte des parcours interprétatifs qui tiennent compte du contexte extralinguistique et de l'intertexte. Ainsi, l'analyse de la détermination du local par le global consiste à identifier localement des sèmes pertinents issus du global. Une telle identification locale de sèmes pertinents issus du global est abordée dans la section de ce chapitre dédiée au modèle d'analyse que nous proposons. Avant cela, et afin de positionner un peu plus notre approche, nous présentons dans la section suivante un certain nombre de travaux en TAL exploitant la SI.

2.1.4 Travaux en TAL exploitant la Sémantique Interprétative

La SI a été exploitée dans un certain nombre de travaux en TAL, pour des tâches assez variées allant de l'aide à l'interprétation de textes à la recherche d'information en passant par la catégorisation de textes.

Ainsi, dans [Tanguy, 1997], l'auteur exploite la SI afin de proposer une assistance à l'interprétation de textes[75]. Dans sa thèse, Ludovic Tanguy propose à l'utilisateur de déclarer préalablement des traits (ou sèmes) qu'il juge pertinents et de les associer aux lexies de son choix. Une fois cette tâche réalisée, ces traits sont projetés sur le texte à analyser afin de mettre en évidence les isotopies pouvant y apparaître. La lecture « isotopique » du texte apporte à l'utilisateur une aide à l'interprétation en permettant, par exemple, l'éclaircissement de certains points ou en mettant en évidence certaines thématiques.

Théodore Thlivitis fonde également ses propositions sur la SI dans sa thèse [Thlivitis, 1998]. L'auteur propose un modèle informatique d'assistance à la compréhension de textes, modèle

[75] Une telle assistance étant mise en place *via* le logiciel *Pastel*, présenté dans le chapitre précédent de cette thèse.

appelé la Sémantique Interprétative Intertextuelle. L'auteur donne une place particulièrement importante à l'utilisateur en lui permettant, tout comme dans [Tanguy, 1997], de définir des traits et de les associer à des lexies, mais également, en lui permettant de positionner réellement le texte analysé dans son intertexte, prenant ainsi en considération la double influence texte/intertexte énoncée dans le principe d'architextualité. Différentes applications du modèle sont ensuite abordées tels l'analyse et l'enseignement littéraire, la conception publicitaire, la traduction, etc.

Dans sa thèse [Beust, 1998], Pierre Beust apporte une contribution à la modélisation informatique de l'interprétation d'énoncés linguistiques en langue naturelle. L'auteur se positionne dans la problématique du dialogue homme-machine. Le modèle proposé pour l'interprétation d'énoncés linguistiques a pour objectif d'en extraire les contraintes sémantiques conditionnant les enchaînements conversationnels. Ces contraintes sémantiques sont exprimées à l'aide de notions très proches de la SI, dont les notions de sème et d'isotopie. Ces contraintes sont alors isolées dans les énoncés par une mise en évidence de liens entre les contenus sémantiques des lexies y apparaissant. Ces contenus sémantiques, stockés en machine et déterminés par l'utilisateur en interaction avec cette dernière, sont des systèmes hiérarchiques de tables produites respectant un modèle différentiel proposé dans la thèse (le modèle *Anadia*). Le processus d'interprétation consiste alors à rechercher des isotopies dans les énoncés afin de réduire la polysémie lexicale et de garantir leur cohésion, ceci afin d'apporter à l'utilisateur une assistance dans l'interprétation des énoncés dans le domaine du dialogue homme-machine.

Dans la suite de la thèse précédente, Vincent Perlerin, dans [Perlerin, 2004], met en place les principes d'une « sémantique légère » pour le TAL qui suppose une limitation à la fois des ressources et des processus pour proposer aux utilisateurs des services personnalisés d'accès aux documents. Par le terme de « sémantique légère », l'auteur propose de déplacer vers l'utilisateur les tâches de description de domaines sémantiques, allant ainsi à l'opposé des approches ontologiques cherchant à positionner ces tâches sur la machine. Il est alors proposé un modèle de représentation de domaines et d'analyse de textes nommé *LUCIA* (pour *Located User-Centred Interpretative Analyser*), évolution du modèle *Anadia* de [Beust, 1998]. Le modèle *LUCIA* propose de représenter les domaines d'intérêt de l'utilisateur par des ensembles structurés de lexies caractérisées selon un principe sémique différentiel. De tels ensembles structurés, construits en interaction utilisateur-machine, permettent de représenter le point de vue d'utilisateurs sur des domaines de leur choix. Le modèle propose ensuite de projeter ces domaines en textes afin d'assister l'utilisateur dans l'accès au contenu de ces derniers. Ce modèle a été évalué dans deux champs d'application : la veille documentaire et l'analyse d'expressions métaphoriques. Dans le premier cas, une implantation du modèle dans un métamoteur de recherche permet un filtrage et un réordonnancement personnalisé de documents inconnus. Dans le second, les résultats obtenus ont permis d'envisager des aides à l'interprétation et à la détection d'expressions métaphoriques dans un corpus thématique.

Dans sa thèse, Bénédicte Pincemin s'intéresse à une tâche de diffusion ciblée d'informations sur l'Intranet d'une entreprise [Pincemin, 1999]. Cette tâche consiste à trouver les personnes les plus concernées par un texte donné ou les personnes expertes sur un sujet donné. Les propositions de l'auteur sont guidées par la SI. Elle exploite ainsi tout particulièrement la notion d'intertextualité. Pour caractériser les textes de son ensemble d'étude, elle cherche à isoler des unités dites « descriptives », plus pertinentes que de simples mots-clés, de telles unités prenant en considération la détermination du local par le global et la formation des isotopies. Ainsi, l'auteur propose deux nouvelles étapes à réaliser lors de l'analyse de textes : la *construction* et l'*élection* [Pincemin, 1999, pages 340 à 346]. La première étape consiste tout d'abord à partir de l'analyse locale de chaque texte, puis à attendre le retour de l'analyse globale sur l'ensemble

des textes, pour réaliser la *construction* des unités descriptives des textes du corpus. La seconde étape, l'*élection*, consiste à choisir, pour chaque texte, les unités descriptives le caractérisant le mieux de façon globale. Ces deux étapes permettent d'améliorer de façon significative le système étudié de diffusion ciblée d'informations.

Mathieu Valette, dans [Valette, 2004], base sur la SI ses travaux sur la détection de documents racistes et xénophobes dans le cadre du projet *PRINCIP*[76]. Il exploite l'opposition fond sémantique/forme sémantique[77]. L'auteur formule l'hypothèse principale que les textes racistes et anti-racistes partagent un même fond sémantique commun : du point de vue thématique, des propos racistes ou anti-racistes sont proches les uns des autres. Par contre, les textes racistes et anti-racistes se distinguent par des formes sémantiques différentes. L'auteur met ainsi en évidence un ensemble de critères de formes sémantiques permettant de différencier les sites racistes de sites anti-racistes : la ponctuation (on remarque de façon statistiquement significative que les sites racistes utilisent fortement le point d'exclamation et que les sites anti-racistes utilisent plutôt des points de suspension), le type de police de caractère (la police de caractère Arial semble caractéristique des sites racistes), les couleurs de fond et de police de caractères (le rouge et le noir semblent caractéristiques des sites racistes), les contenus des images entourant le texte (la thématique de l'animal dans les textes racistes est corrélée avec des dessins montrant des animaux, souvent connotés de façon péjorative, le rat par exemple), etc. En combinant de tels critères avec d'autres de différents types, comme la présence de certains morphèmes caractéristiques aux sites racistes, tels les morphèmes *ouill-* (comme dans les mots *magouille* et *fripouille*) ou encore *crass-* (comme dans *crasseux*), l'auteur atteint un taux de précision de 97% et de rappel de 74% pour son système de détection de sites racistes. Ce travail illustre notamment que des informations sur le contenu des textes sont portées par autre chose que les mots (traditionnellement considérés), et en particulier par la mise en forme matérielle du texte.

En donnant une place particulièrement importante à l'utilisateur, son point de vue et ses interactions, ainsi qu'au contexte global de l'ensemble documentaire, ces différents travaux, et en particulier ceux de Vincent Perlerin et de Théodore Thlivitis, nous montrent la pertinence de la SI pour des travaux liés au TAL. Des éléments théoriques liés à une meilleure prise en considération des domaines d'intérêt de l'utilisateur [Perlerin, 2004] et du contexte global de l'ensemble documentaire [Thlivitis, 1998] nous semblent tout particulièrement pertinents à exploiter dans nos propositions. Ces éléments sont alors repris dans la section suivante de ce chapitre, dédiée au modèle d'analyse que nous proposons.

2.2 *AIdED* : Analyse Interprétative d'Ensembles Documentaires

Le chapitre 1 de cette thèse a permis de dresser un panorama de différents modèles et outils informatiques pour l'accès au contenu d'ensembles documentaires, certains exploitant des techniques de visualisation interactive. Ces différentes propositions nous ont permis de mettre en évidence la faible place laissée à l'utilisateur aussi bien dans la prise en considération de son point de vue que de ses interactions. Dans la première partie de ce chapitre, nous avons mis l'accent sur la SI de François Rastier qui nous semble être un cadre théorique particulièrement adapté à

[76] *Plate-forme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet*, plus d'informations sur le projet : `http://www.crim.fr/princip.htm` (page consultée le 14 juillet 2007).

[77] Dans la SI, le fond sémantique est assimilé aux isotopies tandis que les formes sémantiques correspondent à une autre catégorie d'unités textuelles que sont des groupements stables de sèmes spécifiques (groupe non nécessairement lexicalisé)

des analyses portant sur le contenu de textes. Nous avons illustré son influence dans différents travaux de TAL donnant tous une place très importante à l'utilisateur et à son interprétation.

Le modèle que nous proposons, intitulé *AIdED* pour *Analyse Interprétative d'Ensembles Documentaires*, a pour objectif de réellement positionner l'utilisateur au cœur de la tâche d'accès au contenu d'un ensemble documentaire. Ce modèle prend place en continuité des travaux de l'équipe ISLanD du laboratoire GREYC-CNRS UMR 6072 de l'Université de Caen / Basse-Normandie sur l'analyse sémantique de textes. Là où les travaux de Vincent Perlerin portaient sur l'analyse personnalisée de contenu au niveau du texte, nous visons avec *AIdED* une analyse du palier textuel supérieur, celui de l'ensemble de textes. La prise en considération de ce palier textuel de niveau supérieur nous a amené à étudier des récurrences sémantiques dépassant le niveau du texte traditionnellement étudié. Ces récurrences, ainsi que les différents paliers textuels considérés, sont présentées dans la suite de cette section.

2.2.1 Vue globale sur le modèle *AIdED* et les propositions qu'il intègre

AIdED a pour objectif de permettre à l'utilisateur un accès personnalisé au contenu d'ensembles documentaires. Pour cela, nous mettons tout particulièrement l'accent dans nos propositions sur deux points : la personnalisation de l'accès au contenu et la possibilité d'avoir une vue interactive globale sur l'ensemble documentaire. La prise en considération de ces points dans notre modèle a pour objectif de proposer à l'utilisateur une véritable appropriation de son ensemble documentaire. La démarche interactive nous semble primordiale pour permettre une telle appropriation. C'est par la manipulation de vues sur l'ensemble documentaire, vues augmentées de connaissances propres à l'utilisateur, que nous envisageons la construction de parcours interprétatifs et donc un véritable accès au contenu de l'ensemble documentaire, comme l'illustre la figure 2.1 schématisant les principes généraux d'*AIdED*.

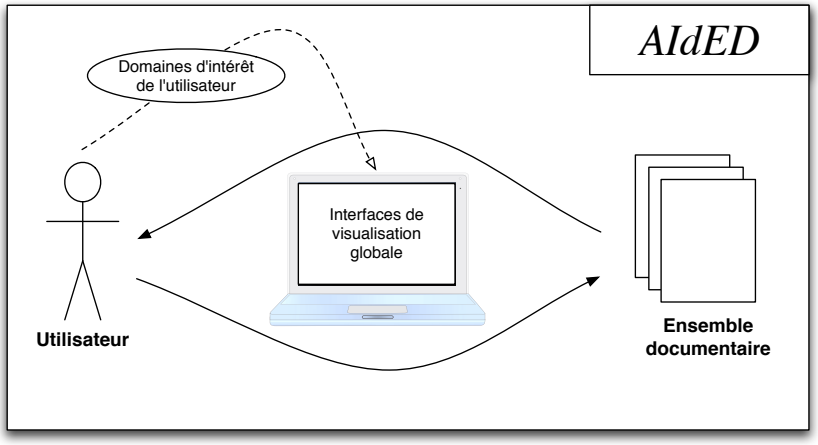


FIG. 2.1 – Principes généraux du modèle *AIdED* : interactions entre l'utilisateur et l'ensemble documentaire *via* la machine et des interfaces de visualisation interactive globale.

Les interactions utilisateur/ensemble documentaire se font à travers des interfaces de visualisation interactive globale codées en machine. L'élaboration de ces interfaces se fait en prenant en considération le point de vue de l'utilisateur sur des domaines de son choix. L'utilisateur

« construit » de tels domaines avec les lexies (et formes graphiques associées) de son choix et les descriptions qu'il souhaite y apposer. Nous utilisons la représentation des domaines proposée par le modèle *LUCIA* de Vincent Perlerin [Perlerin, 2004] permettant à l'utilisateur d'exprimer finement son point de vue sur ses domaines d'intérêt. Les traitements nécessaires à la construction des visualisations vont alors « projeter » ces représentations de domaines en ensembles documentaires. Les répétitions d'éléments de signification à l'intérieur des textes, mais également entre les textes, servent alors de base à la construction des visualisations de l'ensemble documentaire. Ces visualisations prennent principalement la forme de cartes interactives, qui, comme nous l'avons abordé dans le chapitre précédent, proposent des représentations métriques globales particulièrement adaptées à la tâche que nous visons.

Nous présentons plus en détail les différentes propositions constituant le modèle *AIdED* dans les sections suivantes de ce chapitre. Ces propositions correspondent à des éléments de modélisation de l'intertexte, du contexte extralinguistique et du co-texte. Une prise en considération de ces éléments, indispensables dans la contextualisation du sens et donc à l'interprétation, est ainsi menée, toujours dans l'objectif d'appropriation par l'utilisateur du contenu de son ensemble documentaire.

2.2.2 Propositions de représentation de l'intertexte et du contexte extralinguistique

L'ensemble documentaire comme approximation de l'intertexte

Par la définition que nous avons adoptée dans le chapitre 1, l'ensemble documentaire est, pour nous, plus que la simple somme des textes qui le composent. Que l'ensemble documentaire désigne un corpus ou une collection, les textes le composant possèdent une cohérence les reliant les uns aux autres. Nous rejoignons ici le principe général que le tout est différent de la somme de ses parties, principe fondamental de la théorie psychologique *Gestalttheorie*[78] [Ellis, 1938].

Cette cohérence, attribuée à l'ensemble documentaire par l'utilisateur, peut nous permettre de le considérer comme une réelle « société de textes » où chaque texte possède une influence sur les autres. Théodore Thlivitis illustre, dans [Thlivitis, 1998], ce terme de « société de textes ». L'auteur considère un texte (par exemple, un texte littéraire ancien) ayant fait l'objet de multiples analyses et dont les interprétations sont plus ou moins stabilisées. En imaginant qu'un nouveau texte ait été découvert, et que ce dernier illumine un aspect nouveau du texte initial, le sens du texte initial dans la nouvelle société de textes ainsi créée est modifié. Les analyses et interprétations précédentes ne sont cependant pas invalides, elles sont exactes par rapport à l'ancienne société de textes.

La structure englobante de « société de textes » peut alors se ramener à ce que nous avons précédemment appelé l'intertexte. Dans les trois paliers textuels que nous avons évoqués, l'intertexte a été défini comme une structure regroupant les textes que l'utilisateur estime liés à un autre texte, du point de vue de son interprétation. Un texte n'est pas lié à un unique intertexte et peut appartenir à plusieurs intertextes, entraînant ainsi autant d'interprétations différentes[79].

[78] La *Gestalttheorie* est une théorie psychologique prenant place dans le cadre de la psychologie de la forme (le mot allemand *Gestalt* signifiant « forme »). L'être humain est considéré comme un système ouvert interagissant activement avec son environnement. Le principe phare de cette théorie est que le tout est différent de la somme de ses parties. Ainsi, l'air que nous respirons est perçu comme autre chose que l'addition d'azote, d'oxygène, d'argon, de dioxyde de carbone, etc. ; une œuvre musicale est vue comme étant plus qu'une succession rythmée de notes ; le langage ne peut se réduire à l'empilement syntagmatique d'unités discrètes, etc.

[79] Dans [Thlivitis, 1998], l'auteur énonce qu'il y a autant d'interprétations possibles d'un même texte qu'il y a d'« intertextualisations » de ce texte.

Du point de vue de l'utilisateur et plus généralement d'un sujet interprétant, tous les intertextes forment un univers de textes construit par chacun souvent de façon inconsciente. Théodore Thlivitis appelle cet univers l'*anagnose* [Thlivitis, 1998, page 41]. Dans l'objectif que nous avons de positionner l'utilisateur au centre de la tâche d'accès au contenu d'ensembles documentaires, l'anagnose est idéalement ce qu'il faudrait formaliser afin de représenter l'utilisateur et ses différentes expériences. Malheureusement, il n'est pas du tout évident que l'anagnose rassemblant le passé textuel d'un individu, sa culture, sa société soit aisément formalisable.

Dans notre perspective d'accès personnalisé au contenu d'un ensemble documentaire, nous cherchons à prendre le plus possible en considération la cohérence accordée par les utilisateurs aux textes qui le constituent. Nous proposons de représenter l'intertexte *via* l'ensemble documentaire[80]. Nous cherchons à mettre tout particulièrement en évidence les liens intertextuels entre les textes à l'intérieur de l'ensemble documentaire. De tels liens, positionnant les textes par rapport aux autres, facilitent grandement leur interprétation, et en même temps l'accès à leur contenu.

Par cette représentation de l'intertexte, des éléments liés au contexte extérieur à l'ensemble documentaire ne sont pas pris en considération. Pourtant, de tels éléments sont indispensables à l'interprétation des textes de l'ensemble. La partie suivante met en évidence la façon dont nous prenons en considération ce contexte *via* des ressources lexicales représentant les domaines d'intérêt de l'utilisateur au moment de la tâche d'accès au contenu de l'ensemble documentaire.

Des ressources lexicales personnelles comme éléments de représentation du contexte

Le contexte extralinguistique est, tout comme l'intertexte, l'un des trois paliers de la contextualisation du sens. Par définition, il regroupe les conditions pragmatiques d'interprétation d'un texte. À l'instar de l'intertexte, le contexte extralinguistique est difficile à délimiter puisqu'il fait référence à la situation de discours. Dans notre approche centrée-utilisateur pour l'accès au contenu d'ensembles documentaires, la « situation de discours » se limite à l'utilisateur et à l'accès à l'ensemble documentaire qu'il souhaite obtenir *via* la machine. Afin d'apporter des éléments de représentation à cette situation, nous proposons à l'utilisateur d'exprimer ses domaines d'intérêt sur sa tâche.

Ces domaines d'intérêt vont permettre d'apporter des informations sur le contexte dans lequel l'utilisateur souhaite accéder à son ensemble documentaire. Nous avons choisi de représenter ces domaines à l'aide de ressources termino-ontologiques (RTO) (nous renvoyons à [Aussenac-Gilles *et al.*, 2007] pour plus de détails sur ce type de ressources). Les RTO regroupent des terminologies aux structures plus ou moins complexes, comme des réseaux lexicaux, des bases de données lexicales et sémantiques, etc. Le choix d'utiliser des RTO pour représenter les domaines d'intérêt de l'utilisateur est lié à la relative facilité pour un utilisateur d'expliciter des termes, des lexies, des graphies, décrivant son point de vue, plutôt que d'isoler des concepts qui le caractérisent [Paquin, 1990].

La notion de RTO est assez large puisqu'elle regroupe toutes données terminologiques augmentées d'une structure. Le type de RTO choisi est celui proposé par le modèle *LUCIA* de Vincent Perlerin [Perlerin, 2004], abordé précédemment. Le choix de ce modèle a grandement été motivé par les forts emprunts qu'il réalise à la SI et surtout par la place centrale qu'il laisse à l'utilisateur : c'est ce dernier qui a la liberté de construire et de décrire les domaines de son

[80] Une telle représentation rejoint celle décrite dans [Rastier, 1998, note de bas de page n°17], où l'auteur exprime, à propos du corpus, qu'il n'est pas pour autant une simplification de l'intertexte, c'est, selon lui, une objectivation : *Le corpus est la seule objectivation possible (philologique) de l'intertexte, qui sinon demeure une notion des plus vagues.*

choix à l'aide des lexies qu'il souhaite.

Le modèle *LUCIA*, d'inspiration saussurienne, est un modèle différentiel qui part du constat que pour désigner les choses dont on veut parler, pour établir leur valeur sémiotique, on les décrit juste assez pour les différencier des choses avec lesquelles elles pourraient être confondues. L'idée principale est d'exprimer des connaissances et un point de vue sur la terminologie d'un domaine en organisant des lexies selon deux principaux critères :

– des regroupements par similarité, témoignant de la proximité de certaines lexies ;
– des oppositions locales, précisant les différences entre lexies proches.

Nous renvoyons à [Perlerin, 2004, pages 69 à 110] pour une description détaillée du modèle *LUCIA* et des principes qui le régissent.

Dans la pratique, l'utilisateur peut définir un *dispositif* pour chaque domaine qu'il souhaite représenter. Un dispositif contient un ensemble de *tables* en relation, chaque table contenant des lexies d'une même catégorie sémantique selon le point de vue de l'utilisateur. Au sein de chaque table, l'utilisateur doit expliciter des différences entre unités lexicales à l'aide de couples *attribut : valeur*, qui sont l'expression de sèmes. Dans le cadre de la tâche visée par l'utilisateur, plusieurs dispositifs peuvent être utilisés conjointement afin de représenter l'ensemble de ses domaines d'intérêt. Ce regroupement de dispositifs est appelé une *session*.

Afin d'avoir, dès à présent, une vision globale et concrète sur les principes fondamentaux du modèle *LUCIA* et sur la façon dont une RTO respectant ce modèle est construite, nous proposons un exemple de dispositif *LUCIA* (également appelé RTO *LUCIA*). Le dispositif présenté ici porte sur le domaine du cinéma. Le premier travail pour construire une RTO *LUCIA* décrivant le domaine du cinéma a été de rassembler des lexies (et des formes graphiques associées) s'y rapportant selon le point de vue de l'utilisateur. La liste suivante illustre de telle lexies : *acteur*, *comédien*, *réalisateur*, *cameraman*, *producteur*, *scénariste*, *monteur*, *figurant*, *preneur du son*, *Jean-Pierre Jeunet*, *Steven Spielberg*, *Georges Lucas*, *Alfred Hitchcock*, *John Woo*, etc.

À partir de ces lexies, des tables *LUCIA* ont été construites pour les regrouper et les différencier. Dans cette étude, les deux tables de la figure 2.2 ont été construites.

Équipe de tournage
acteur, *comédien*, *réalisateur*, *cameraman*, *producteur*, *scénariste*, *monteur*, *figurant*, *preneur du son*

Réalisateur
Jean-Pierre Jeunet, *Steven Spielberg*, *Georges Lucas*, *Alfred Hitchcock*, *John Woo*

FIG. 2.2 – Création de tables de lexies.

Le modèle *LUCIA* propose de caractériser chacune des tables à l'aide d'un ou plusieurs attributs et de différencier les lexies au sein d'une même table à l'aide de valeurs d'attributs opposées. Ainsi, dans la table précédente Équipe de tournage, l'utilisateur a choisi de faire intervenir l'attribut *rôle* (attribut qui sera donc possédé par toutes les lexies de cette table) avec des valeurs opposées de cet attribut *acteur/dirigeant/technicien*. L'attribut *professionnel*, avec les valeurs opposées *oui/non*, intervient également dans cette table. Pour la table Réalisateur, l'attribut *nationalité* a été choisi avec les valeurs *américaine/française/anglaise/chinoise*. Les deux tables deviennent alors celles présentées en figure 2.3.

Équipe de tournage	*professionnel*	*rôle*
acteur, *comédien*	*oui*	*acteur*
réalisateur, *producteur*, *scénariste*	*oui*	*dirigeant*
cameraman, *monteur*, *preneur du son*	*oui*	*technicien*
figurant	*non*	*acteur*
∅	*non*	*dirigeant*
∅	*non*	*technicien*

Réalisateur	*nationalité*
Steven Spielberg, *Georges Lucas*	*américaine*
Jean-Pierre Jeunet	*française*
Alfred Hitchcock	*anglaise*
John Woo	*chinoise*

FIG. 2.3 – Différenciation des lexies au sein de chaque table.

Une fois l'ensemble des tables créées, le modèle propose de les regrouper et de les lier entre elles au sein d'un même dispositif. Les liens entre tables sont appelés des liens d'héritage. Dans l'exemple présenté ici, il est considéré que la table Réalisateur est reliée à la ligne *professionnel : oui* et *rôle : dirigeant* de la table Équipe de tournage. Ainsi, que chaque lexie de la table Réalisateur hérite du couple *professionnel : oui* et *rôle : dirigeant*. Les différents dispositifs utilisés dans les expérimentations détaillées dans cette thèse sont présentés en annexe C de ce manuscrit. Ils sont également tous disponibles à l'adresse suivante : `http://users.info.unicaen.fr/~troy/dispositifs` (page consultée la 18 juin 2007).

Selon la tâche visée et la précision qu'il souhaite obtenir dans la description de ses domaines d'intérêt, l'utilisateur peut faire intervenir un nombre variable de tables, caractérisées par un nombre plus ou moins important de couples *attribut : valeur*. Si la description des domaines envisagée par l'utilisateur se veut assez simple, le modèle *LUCIA* lui propose de représenter les domaines de son choix sous la forme de simples listes de lexies et formes graphiques associées. Cette représentation « allégée » peut être également vue comme une première description des domaines d'intérêt de l'utilisateur, description pouvant être complétée par la suite par une structuration en tables décrites à l'aide de couples *attribut : valeur*. Dans cette version simple, le domaine du cinéma est décrit par la liste des lexies que l'utilisateur a choisi de faire intervenir. Cette liste est ensuite étiquetée par le nom du domaine que décrivent les lexies, comme l'illustre la figure 2.4.

Cinéma
acteur, *comédien*, *réalisateur*, *cameraman*, *producteur*, *scénariste*, *monteur*, *figurant*, *preneur du son*, *Jean-Pierre Jeunet*, *Steven Spielberg*, *Georges Lucas*, *Alfred Hitchcock*, *John Woo*

FIG. 2.4 – Représentation simple d'un domaine en ensemble de lexies.

Plusieurs utilisations de RTO *LUCIA* sont présentées au chapitre 4 de cette thèse. Des assistances logicielles sont, par ailleurs, fournies aux utilisateurs afin de les aider à construire et à faire évoluer des RTO *LUCIA*. De telles assistances sont présentées dans le chapitre 3 de cette thèse.

2.2.3 La détection des isotopies intra et inter-textuelles comme élément de base aux traitements

Après avoir représenté l'intertexte et le contexte, le troisième et dernier palier de contextualisation du sens est le co-texte. Comme nous l'avons défini précédemment, le co-texte d'une unité linguistique est son « entourage » dans le texte et correspond à une zone de localité sémantique pertinente autour de cette unité, appelée *période*. La notion d'isotopie (ou plutôt, les notions d'isotopies, comme nous le verrons par la suite) joue un rôle important dans l'exploitation du contexte et de l'intertexte pour l'interprétation de l'ensemble documentaire. Un travail au niveau du co-texte mais également au niveau plus global de l'intertexte, exploitant le contexte instauré par l'utilisateur, doit être réalisé afin de faire ressortir différents éléments de signification utiles à l'utilisateur pour son accès au contenu d'un ensemble documentaire.

Les notions d'isotopies intra et inter-textuelles

Nous avons vu précédemment qu'une isotopie est un effet de récurrence d'un même sème au sein d'un texte. Une telle isotopie est *intra-textuelle*, puisqu'elle résume la récurrence d'un même sème au niveau du texte. Cependant, et comme le souligne Théodore Thlivitis [Thlivitis, 1998], une isotopie intra-textuelle ne se construit jamais de manière autonome. Le texte ne se « suffit » pas à lui-même pour permettre à un lecteur une identification des isotopies qu'il contient. L'auteur donne pour exemple différents éléments venant orienter l'identification des isotopies dans le texte : des notes bibliographiques, des commentaires laissés par l'auteur lui-même ou un autre lecteur, des dictionnaires ou glossaires, etc. Selon lui, ces différents éléments sont des composants de l'intertexte qui viennent influencer les isotopies à l'intérieur du texte[81].

Des récurrences d'éléments de signification, portées par des éléments de l'intertexte, vont orienter la présence et la nature des isotopies au niveau du texte. Pour désigner de telles récurrences, on parle d'isotopies *inter-textuelles*, définies de la façon suivante par Théodore Thlivitis [Thlivitis, 1998] :

> [une isotopie inter-textuelle est constituée par] *la récurrence de traits sémantiques qui caractérisent des textes entiers mais au sein d'un intertexte.*

Comme nous l'avons évoqué précédemment, de telles isotopies inter-textuelles ont une influence sur les isotopies intra-textuelles. La réciproque est également vraie puisque c'est à partir des isotopies intra-textuelles que sont constituées les isotopies inter-textuelles : des récurrences de sèmes au niveau local du texte sont « remontées » au niveau plus global d'un sous-ensemble de textes de l'intertexte. Cette double influence des isotopies intra et inter-textuelles se réalise de façon itérative, autant de fois que nécessaire jusqu'à une certaine stabilisation permettant l'interprétation de l'unité textuelle considérée. La figure 2.5 illustre ces notions d'isotopies au niveau du texte et de l'ensemble de textes.

[81] Les éléments de l'intertexte proposés par l'auteur ne constituent pas une liste exhaustive de tous les éléments pouvant influencer les isotopies présentes dans le texte. Plus simplement, un autre texte de l'intertexte peut également être considéré comme un élément venant influencer localement les isotopies d'un autre texte du même intertexte.

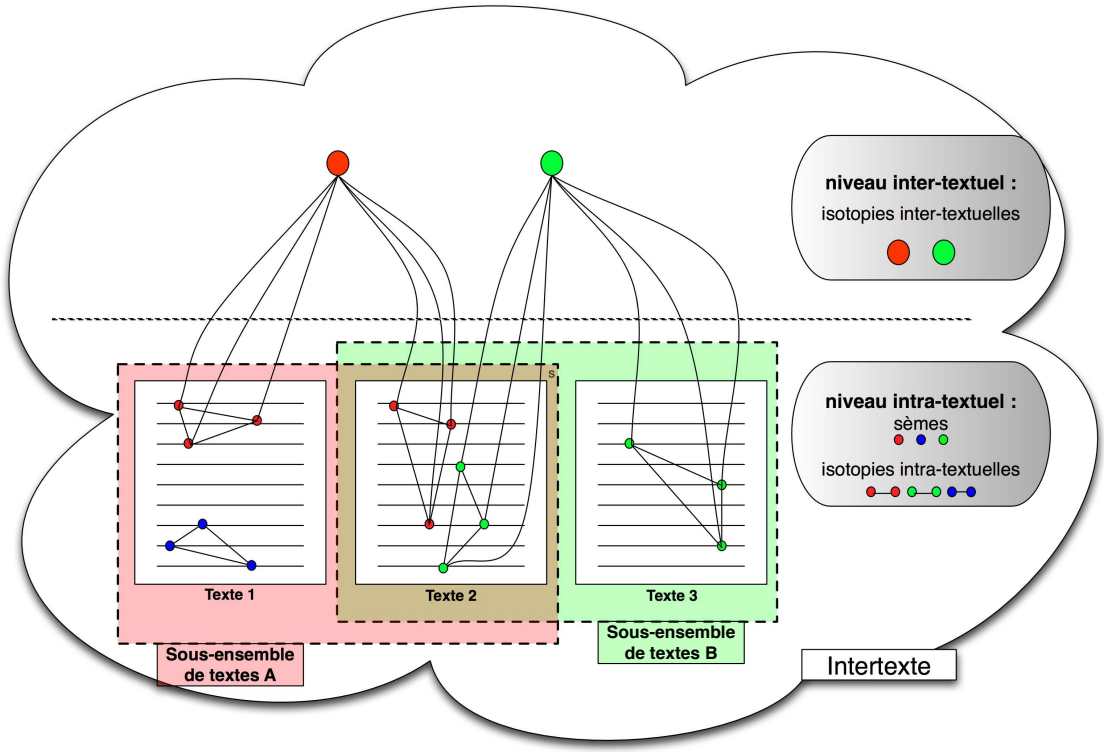


FIG. 2.5 – Isotopies intra et inter-textuelles parcourant textes et sous-ensembles de textes d'un intertexte. L'isotopie inter-textuelle représentée par la couleur rouge parcourt les textes 1 et 2, délimitant ainsi le sous-ensemble de textes A.

Trois niveaux textuels se dégagent alors :

1. **l'ensemble documentaire** dans sa globalité, qui constitue l'objet d'étude de la tâche que nous visons ;
2. **le sous-ensemble de textes** délimité par le parcours des isotopies inter-textuelles ;
3. et **le texte**.

De tels niveaux textuels respectent alors la relation suivante :

texte ⊂ sous-ensemble de textes ⊂ ensemble documentaire

Nous pouvons remarquer que le niveau du sous-ensemble de textes est à la fois le niveau global du texte et le niveau local de l'ensemble documentaire. Dans les vues que nous proposerons par la suite sur les ensembles documentaires, nous chercherons à rendre compte de ces trois niveaux.

La notion d'isotopie inter-textuelle nous oblige à étendre la notion de période, désignant traditionnellement en SI la portée d'une isotopie intra-textuelle à l'intérieur d'un même texte[82]. La portée d'une isotopie inter-textuelle dépasse le cadre du texte pour atteindre celui d'un sous-ensemble de textes, tous issus du même intertexte[83]. Ce sous-ensemble de textes possède alors une forte cohérence : les textes d'un même ensemble partagent des isotopies inter-textuelles, et donc des éléments de signification liés à leur contenu. La mise en évidence d'un ensemble de textes parcouru par une ou plusieurs isotopies inter-textuelles peut être particulièrement informative pour l'utilisateur dans sa tâche d'accès au contenu d'un ensemble documentaire : les textes de l'ensemble contiennent des éléments de signification communs et donc des éléments de contenu similaire. Le paragraphe suivant revient sur les différents types de sèmes constituant de telles isotopies.

Typologie de sèmes et RTO LUCIA

Pour représenter les domaines d'intérêt de l'utilisateur dans le cadre de sa tâche d'accès au contenu d'un ensemble documentaire, nous avons proposé précédemment d'utiliser des RTO respectant le modèle *LUCIA*. Selon sa tâche, l'utilisateur va rassembler un certain nombre de dispositifs au sein d'une session. Cette session est vue, elle-même, comme une table *LUCIA* avec comme seul attribut *Appartenance au domaine* et comme valeurs les différents domaines de l'utilisateur. Chaque ligne de la table est ainsi héritée dans les différents dispositifs, comme l'illustre la figure 2.6. Les lexies de chaque dispositif se voient alors qualifiées par le couple *attribut : valeur*, *Appartenance au domaine : nom du domaine*. Ce trait qualifiant les lexies d'un même dispositif est un sème générique selon la terminologie de la SI et même un sème *macro-générique*, selon Michel Ballabriga dans [Ballabriga, 2005], puisque que c'est un sème de la plus grande généralité possible, du moins dans le cadre de la session de l'utilisateur. Par la suite, nous utiliserons le terme de « sème macro-générique » pour désigner ces sèmes et les opposer aux autres sèmes génériques et spécifiques contenus à l'intérieur des dispositifs (une isotopie macro-générique sera alors la récurrence d'un sème macro-générique).

[82] La notion de période n'est pas propre à la SI. Par exemple, Michel Charolles dans [Charolles, 1988] définit les périodes comme des *unités d'énonciation dont les membres ou composants (phrastiques) entretiennent des rapports de dépendance.*

[83] Nous pouvons alors parler de « période intertextuelle » pour désigner la zone de portée d'une isotopie inter-textuelle.

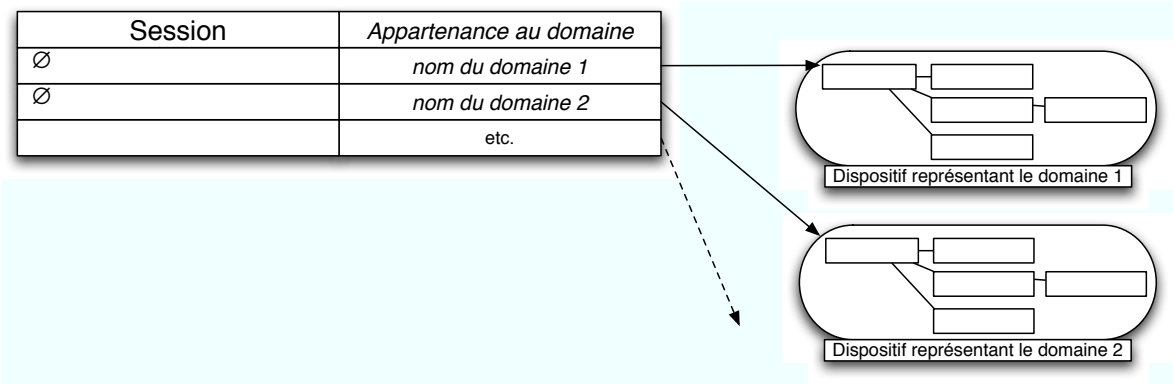


FIG. 2.6 – Héritage des traits d'appartenance aux domaines dans les dispositifs d'une même session.

Au sein d'un même dispositif, les lexies sont organisées en tables et décrites à l'aide de couples *attribut : valeur* (du moins dans le cadre de la représentation « avancée » dépassant la simple liste de lexies). Des sèmes spécifiques et génériques peuvent alors être identifiés comme l'illustre la figure 2.7. Ainsi, les couples *rôle : dirigeant* et *professionnel : oui* sont des sèmes génériques pour les lexies de la table Réalisateur, alors que les couples *nationalité : américaine*, *nationalité : française*, etc. sont des sèmes spécifiques pour les lexies de cette même table.

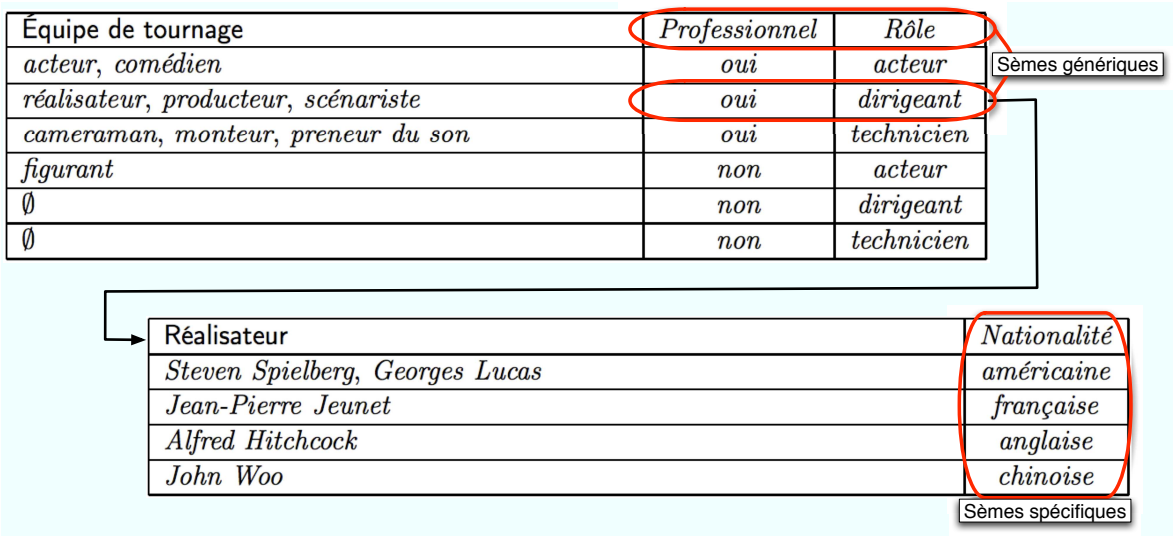


FIG. 2.7 – Sèmes génériques et spécifiques au sein d'un dispositif.

Aux notions de sèmes macro-génériques, génériques et spécifiques, nous ajoutons les notions de sèmes *partagés* et de sèmes *propres*[84]. Le fait qu'un sème soit macro-générique, générique ou spécifique est lié à la structuration des dispositifs dans lesquels il intervient. Le fait qu'un sème soit propre ou partagé n'est pas lié aux dispositifs le contenant, mais est lié à la façon dont il est actualisé dans l'ensemble documentaire. Un sème est propre à un texte (ou à un sous-ensemble de textes de l'ensemble documentaire) s'il y apparaît de façon significativement plus importante

[84] On parlera respectivement d'isotopies propres et partagées pour désigner les récurrences de sèmes propres et de sèmes partagés.

que dans les autres textes de l'ensemble (ou de l'ensemble documentaire). Un tel sème peut alors servir à distinguer un texte ou un sous-ensemble de textes. Au contraire, un sème est partagé par un ensemble documentaire (ou par un ensemble de textes) s'il apparaît de façon équivalente dans tous les textes de l'ensemble documentaire (ou du sous-ensemble de textes).

Dans la section suivante, nous présentons nos propositions pour détecter de tels éléments de signification dans les ensembles documentaires et nous revenons sur l'identification des sèmes propres et partagés, et des isotopies associées.

Détection d'éléments de signification dans des ensembles documentaires

Une projection de RTO *LUCIA* à l'intérieur de chaque texte de l'ensemble documentaire à analyser permet de mettre en évidence des récurrences d'attributs et des récurrences de couples *attribut : valeur*, récurrences assimilables à des isotopies. Une fois l'identification de ces isotopies effectuée, il est possible de passer à un niveau textuel supérieur, afin de mettre en évidence des isotopies inter-textuelles. En partant ainsi d'éléments locaux, les isotopies intra-textuelles, nous proposons de les faire « remonter » au niveau global de l'ensemble documentaire. Si ces isotopies intra-textuelles sont présentes dans plusieurs textes de l'ensemble, alors elles sont considérées comme étant des isotopies inter-textuelles.

Ainsi, un ensemble de textes parcourus par de mêmes isotopies inter-textuelles macro-génériques et génériques est pertinent pour l'utilisateur, du fait qu'il contient des textes partageant certains thèmes génériques[85]. Au sein de cet ensemble de textes, les isotopies inter-textuelles spécifiques le parcourant peuvent compléter l'information thématique ainsi donnée, en explicitant ses particularités sémantiques et en mettant en évidence des thèmes spécifiques. Ce « passage » du niveau local du texte au global de l'intertexte permet de mettre en évidence et de qualifier des sous-ensembles de textes où de mêmes isotopies inter-textuelles sont présentes.

Dans la section précédente, nous avons mis en évidence l'influence de l'intertexte sur des éléments locaux, et donc l'importance du passage du niveau global de l'intertexte au niveau local du texte. Pour représenter cette influence, nous avons choisi d'instaurer une « pondération » sur des éléments locaux (les isotopies intra-textuelles dans un texte t) en fonction d'éléments plus globaux (les isotopies inter-textuelles parcourant un ensemble de textes contenant le texte t). Le principe de cette pondération est lié à la tâche d'accès au contenu que nous nous sommes fixée. Selon nous, pour identifier les éléments de signification pertinents dans un texte, il faut localiser ceux qu'il partage avec son ensemble documentaire, et ceux qui lui sont propres. Par exemple, si une isotopie intra-textuelle est proportionnellement plus forte dans un texte que dans son ensemble documentaire, alors il est intéressant de la mettre en avant, car elle permet de particulariser le texte par rapport à son sur-ensemble[86]. Une telle isotopie est, pour nous, propre au texte. Au contraire, si une isotopie intra-textuelle est présente dans un texte dans les mêmes proportions que dans son intertexte, alors son intérêt est limité, le contenu sémantique porté par l'isotopie partagée étant déjà « affiché » par l'ensemble documentaire.

En appliquant ce principe aux différents niveaux textuels que nous étudions, cela revient à partir de l'ensemble documentaire dans sa globalité, puis à mettre en évidence dans cet ensemble les isotopies inter-textuelles le parcourant. À chacune de ces isotopies est associé un poids donné par la formule ci-dessous. La fonction *poids* prend deux arguments, l'isotopie recherchée et l'ensemble documentaire dans lequel elle se situe. Le résultat retourné est un pourcentage

[85] Le regroupement de textes parcourus par de mêmes isotopies inter-textuelles est réalisé en utilisant des méthodes statistiques que nous présentons dans le chapitre 3 de cette thèse.

[86] Une telle approche s'apparente à celle sous-jacente au TF-IDF (*Term Frequency / Inverse Document Frequency*) décrite par exemple dans [Salton, 1991].

correspondant à la part des récurrences (c'est à dire le nombre de lexies porteuses de l'isotopie) occupées par l'isotopie dans l'ensemble documentaire donné :

$$poids(isotopie,\ ensemble\ documentaire) = \frac{nombre\ de\ recurrences\ dans\ l'ensemble\ documentaire\ de\ traits\ formant\ l'isotopie}{nombre\ total\ de\ recurrences\ dans\ l'ensemble\ documentaire\ de\ traits\ definis\ par\ l'utilisateur} \cdot 100$$

Une fois le poids de chaque isotopie inter-textuelle déterminé au niveau de l'ensemble documentaire, nous proposons de « descendre » au niveau textuel inférieur des sous-ensembles de textes délimités par les isotopies inter-textuelles les parcourant. Le poids associé à une isotopie inter-textuelle dans un sous-ensemble de textes est le suivant :

$$poids(isotopie,\ sous\ ensemble\ de\ textes) = \frac{nombre\ de\ recurrences\ dans\ le\ sous\ ensemble\ de\ textes\ de\ traits\ formant\ l'isotopie}{nombre\ total\ de\ recurrences\ dans\ le\ sousensemble\ de\ textes\ de\ traits\ definis\ par\ l'utilisateur} \cdot 100$$

Pour prendre en considération le niveau textuel global de l'ensemble documentaire à ce niveau plus local, nous proposons alors de définir une fonction *score* qui, à partir d'une isotopie et d'un contexte donnés, fait ressortir les particularités de l'ensemble de textes par rapport à son intertexte :

$$score(isotopie,\ ensemble\ de\ textes) = poids(isotopie,\ ensemble\ de\ textes) - poids(isotopie,\ intertexte)$$

Selon le signe positif ou négatif de la valeur obtenue, un « excès » ou un « déficit » de l'isotopie est observé au niveau local par rapport au niveau global. Une valeur très proche de 0 indique que l'isotopie considérée parcourt les niveaux local et global dans les mêmes proportions, c'est donc une isotopie partagée. De telles informations apportent, dans tous les cas, une idée sur le positionnement du niveau local par rapport au niveau global. Il est ainsi possible de faire ressortir à l'utilisateur les isotopies les plus en excès. De telles isotopies sont ce que nous avons appelé précédemment des isotopies propres. Elles mettent en évidence ce qu'apporte le sous-ensemble de textes par rapport à l'ensemble documentaire dans sa globalité, et peuvent ainsi servir à caractériser et à distinguer le sous-ensemble.

Ces opérations, effectuées ici entre l'ensemble documentaire et le sous-ensemble de textes, se transposent aux niveaux textuels inférieurs. Le sous-ensemble de textes devient alors le contexte global dans lequel nous allons positionner le niveau local du texte. Le score associé à une isotopie intra-textuelle est toujours donné par la fonction *score* :

$$score(isotopie,\ texte) = poids(isotopie,\ texte) - poids(isotopie,\ ensemble\ de\ textes)$$

La pondération des isotopies macro-génériques et génériques peut ainsi mettre en évidence les principales particularités thématiques propres à des textes ou à des sous-ensembles de textes. Par exemple, si une très grande partie de l'intertexte est parcourue de façon uniforme par une même isotopie inter-textuelle macro-générique ou générique, marquant une profonde appartenance à un thème particulier, la pondération permet de faire ressortir au niveau du texte et du sous-ensemble de textes de nouvelles isotopies macro-génériques ou génériques liées, par exemple, à des thèmes un peu plus secondaires. La pondération des isotopies spécifiques dans des textes ou des sous-ensembles de textes va, quant à elle, permettre de les décrire de façon plus fine et plus discriminante. Ainsi, si une isotopie spécifique correspondant, par exemple, à la récurrence d'une évaluation négative est uniformément présente dans l'intertexte, alors la pondération permettra de faire ressortir des isotopies spécifiques différentes et ainsi apporter des informations complémentaires permettant de distinguer de nouveaux éléments de signification.

De tels calculs de pondération sont également abordés dans [Roy *et al.*, 2007] et différents exemples sont présentés. Nous reprenons l'un de ces exemples en figure 2.1 afin d'illustrer ces calculs (exemples traduits de l'anglais). Le tableau 2.1 fait ressortir les principales isotopies inter-textuelles parcourant un groupe de textes délimité au sein d'un ensemble documentaire. Les scores des isotopies avec pondération, comme expliqué précédemment avec la fonction *score*, sont présents dans la colonne de droite du tableau. Les scores sans pondération sont présentés en colonne centrale. Ils prennent les valeurs retournées par la fonction *poids*, également présentée précédemment. Les classements impliqués par ces deux scores sont indiqués.

Isotopies inter-textuelles	Score sans pondération (fonction *poids*)	Score avec pondération (fonction *score*)
rapport au domaine : objet	70.5% (position 1)	14.9% (position 1)
évaluation : mal	2.5% (position 3)	1.0% (pos : 2)
état : gaz	2.0% (position 4)	0.7% (pos : 3)
type d'agent : organisation	1.4% (position 5)	-0.02% (pos : 4)
type d'objet : matériel	16.0% (position 2)	-0.06% (pos : 5)
type d'activité : professionnelle	0.02% (position 6)	-0.08% (pos : 6)

TAB. 2.1 – Exemple de classements avec et sans pondération des isotopies inter-textuelles présentes dans un ensemble de textes.

Le classement réalisé à partir des scores pondérés déclasse l'isotopie *type d'objet : matériel* de la deuxième position à la cinquième position. Les isotopies *évaluation : mal*, *état : gaz* et *type d'agent : organisation* gagnent une place dans le classement grâce à la pondération. Les isotopies *rapport au domaine : objet* et *type d'activité : professionnelle* restent respectivement en première et dernière position du classement. Une analyse des textes de l'ensemble considéré avait révélé que les textes abordaient des problèmes liés à la pollution, au développement durable et au rejet de gaz à effets de serre. La pondération a été efficace puisqu'elle a permis de faire remonter dans le classement les isotopies inter-textuelles *évaluation : mal* et *état : gaz*, exprimant bien le contenu des textes de l'ensemble étudié. Au contraire, l'isotopie *type d'objet : matériel* a été déclassée du fait de son important partage dans l'ensemble documentaire. Ce déclassement s'est lui aussi révélé pertinent.

Une fois les différentes isotopies déterminées, il faut transmettre les informations qu'elles expriment à l'utilisateur. Comme nous l'avons déjà évoqué précédemment, nous avons choisi de proposer aux utilisateurs des visualisations interactives sur l'ensemble documentaire. Des telles visualisations se doivent alors de prendre en considération les différents éléments présentés dans cette section afin de faire ressortir aux utilisateurs, des informations adaptées sur le contenu de l'ensemble documentaire considéré.

2.2.4 Des visualisations cartographiques interactives comme pour l'accès personnalisé au contenu d'ensembles documentaires

Visualisation personnalisée et interactive de données

Dans [Pednault, 2000], l'auteur propose un certain nombre de recommandations pour la visualisation personnalisée de données. Nous proposons de les reprendre dans notre problématique. Tout d'abord, l'accent est respectivement mis sur la *simplicité* et la *flexibilié* des vues. Il faut ainsi représenter à l'utilisateur uniquement ce qui lui est nécessaire pour résoudre ou assister sa tâche. Il ne faut pas être trop restrictif ou trop précis dans la représentation, mais il faut plutôt

proposer des vues à différents paliers (dans notre cas, les trois niveaux textuels). La représentation doit également être *fluide* et *riche*. La notion de fluidité fait ici référence à la valeur ajoutée apportée par l'interface et ses possibilités d'interactions. La richesse de la représentation est, quant à elle, liée à l'intelligibilité des informations représentées. L'adéquation entre les données fournies et celles qui apparaissent dans les représentations doit pouvoir être facilement perçue par l'utilisateur.

Les cartes d'un ensemble documentaire que nous proposons reprennent ces différentes idées. Dans les cartes proposées, l'accent est mis tout particulièrement sur l'interaction. De telles cartes sont des supports de visualisation interactive d'ensembles documentaires, supports tenant compte des domaines d'intérêt de l'utilisateur que nous avons abordés précédemment. Notre objectif à travers de tels supports est de permettre à l'utilisateur d'identifier les différentes isotopies pertinentes de son point de vue en parcourant son ensemble documentaire. L'identification de ces isotopies se fait par la visualisation et la manipulation des cartes. Par la manipulation de la carte, l'utilisateur identifie, au fil du parcours impliqué, les différentes isotopies structurant et décrivant les éléments qu'il observe (ensemble documentaire dans sa globalité, sous-ensemble de textes ou texte). Cette visualisation « isotopique » interactive oriente ainsi les parcours interprétatifs de l'utilisateur sur l'ensemble documentaire. De telles cartes, que nous décrivons dans le paragraphe suivant, se posent alors comme de véritables médias entre l'utilisateur et son ensemble documentaire, guidant ses parcours interprétatifs et facilitant ainsi son accès au contenu de l'ensemble documentaire.

Propositions de visualisations multi-échelles personnalisées et interactives d'ensembles documentaires

Dans le chapitre précédent, nous avons présenté différentes techniques de visualisation de données textuelles et nous avons souligné l'intérêt que nous portons aux techniques de cartographie. Afin que l'utilisateur puisse avoir un accès personnalisé et interactif au contenu de son ensemble documentaire, nous proposons de construire différents types de cartes sur cet ensemble prenant en considération ses domaines d'intérêt.

Nous proposons, tout d'abord, une vue cartographique globale sur les textes de l'ensemble. Une telle vue globale est élaborée à partir des isotopies intra-textuelles macro-génériques parcourant les textes. Le fait de considérer les isotopies macro-génériques permet d'isoler, dans un premier temps, des informations globales sur les domaines présents dans l'ensemble documentaire. Sur la vue cartographique, chaque texte est représenté par un élément (par exemple, un point) positionné par rapport aux autres textes de façon plus ou moins proche selon les isotopies intra-textuelles macro-génériques qu'ils partagent. Dans le chapitre 3 de cette thèse, nous présentons en détail les méthodes statistiques utilisées pour positionner les textes sur la vue cartographique. De façon plus générale, les visualisations proposées ici sont mises en œuvre dans le prochain chapitre de cette thèse.

Une vue cartographique globale sur les sous-ensembles de textes de l'ensemble documentaire est également proposée. Cette vue fournit à l'utilisateur une visualisation des sous-ensembles de textes délimités par les isotopies inter-textuelles macro-génériques les parcourant. Tout comme pour les textes, le positionnement des sous-ensembles de textes se fait selon leurs isotopies inter-textuelles macro-génériques.

Des vues plus locales sont ensuite proposées. Ces vues mettent en évidence les particularités du texte et de l'ensemble de textes en affichant respectivement, par exemple sous forme de listes, les isotopies intra et inter-textuelles macro-génériques, génériques et spécifiques les parcourant. Par ces vues, l'utilisateur doit pouvoir positionner le texte et le sous-ensemble de textes dans leur

contexte global, respectivement celui du sous-ensemble de textes et de l'ensemble documentaire. Pour cela, une mise en parallèle de la description du niveau local avec une description du niveau global doit être proposée, par exemple en mettant en parallèle les isotopies présentes dans le local (comme le texte) avec celles présentes dans le global (l'ensemble de textes contenant le texte).

Un retour au texte est également particulièrement important. Ainsi, une visualisation du texte, augmentée par un coloriage des différentes lexies des domaines de l'utilisateur, peut aider l'utilisateur dans sa tâche d'accès au contenu de l'ensemble documentaire.

Les vues cartographiques précédentes vont refléter la dimension synchronique de l'ensemble documentaire. Les textes de l'ensemble sont tous considérés sur un même plan temporel. Dans certains cas, l'ensemble documentaire est fortement dépendant de la dimension temporelle, par exemple, lorsqu'il contient des articles de presse ou des messages de forums de discussion. L'ancrage de certains ensembles documentaires avec la dimension temporelle est donc à prendre en considération afin d'interroger la dimension diachronique. Pour cela, nous proposons de rendre dynamiques les vues cartographiques globales de l'ensemble documentaire. Ce dynamisme représente l'axe temporel, les modifications sur les vues au fil du temps illustrant l'évolution des isotopies au sein des textes de l'ensemble documentaire sur la période temporelle considérée.

Les vues décrites précédemment proposent un grand nombre d'interactions. La navigation sur les vues cartographiques est la principale interaction que nous proposons. L'utilisateur doit pouvoir se déplacer sur les vues et effectuer des zooms sur certaines zones. Le passage d'un niveau textuel à un autre est une interaction primordiale. L'utilisateur doit ainsi pouvoir faire des aller-retours entre une vue globale sur l'ensemble documentaire et une vue locale sur le texte, en passant par une vue sur le sous-ensemble de textes. Pour chaque niveau, l'utilisateur doit pouvoir parcourir interactivement des informations qui lui sont propres (par exemple, en parcourant des listes de lexies et d'isotopies présentes au niveau textuel considéré). Sur les vues dynamiques, l'utilisateur doit pouvoir, en plus des interactions énoncées précédemment, contrôler la dimension temporelle en faisant des pauses quand il le souhaite.

Le chapitre suivant de cette thèse illustre ces différentes propositions en présentant les différentes vues construites à partir des domaines d'intérêt de l'utilisateur et de son ensemble documentaire.

2.2.5 Utilisations du modèle *AIdED*

Dans les paragraphes précédents, nous avons présenté les différentes caractéristiques de notre modèle. En nous basant sur des représentations approchées de l'intertexte par l'ensemble documentaire et du contexte par les domaines d'intérêt de l'utilisateur, nous avons mis en évidence comment détecter des éléments de signification et comment exploiter ces derniers pour proposer des visualisations interactives et multi-échelles sur l'ensemble documentaire. Les propositions mises en place dans notre modèle peuvent alors se résumer en quatre étapes :

1. Sélection par l'utilisateur de l'ensemble documentaire à analyser et construction des dispositifs *LUCIA* représentant ses domaines d'intérêt ;
2. Projection des dispositifs *LUCIA* sur l'ensemble documentaire afin de mettre en évidence les isotopies liés à ces domaines d'intérêt ;
3. Construction de différentes vues sur l'ensemble documentaire prenant en considération les isotopies détectées et pondérées à l'étape précédente ;
4. Navigation et manipulation des vues par l'utilisateur afin d'assister son interprétation de l'ensemble documentaire.

La quatrième et dernière étape proposée ci-dessus n'est pas une étape terminale. Après avoir visualisé interactivement son ensemble documentaire, l'utilisateur peut ressentir deux types de besoins :

- *Le besoin de modifier et de faire évoluer ses domaines d'intérêt.*
  Si l'utilisateur constate qu'un domaine est peu ou pas représenté sur les visualisations qui lui sont retournées, il peut décider de le supprimer ou, au contraire, de le compléter. Il peut également décider d'ajouter de nouveaux domaines si des vues au niveau local du texte lui en font ressentir le besoin.
- *Le besoin de modifier et de faire évoluer son ensemble documentaire.*
  Les vues retournées à l'utilisateur vont lui permettre d'observer la répartition de ses domaines d'intérêt sur l'ensemble documentaire, d'appréhender des regroupements de textes en sous-ensembles, etc. L'utilisateur peut alors vouloir ajouter de nouveaux textes à l'ensemble documentaire afin, par exemple, de visualiser comment ils se répartissent sur les vues. Le retrait d'un ou plusieurs textes peut également être réalisé par l'utilisateur, par exemple, si ce dernier observe un sous-ensemble de textes dont les caractéristiques ne lui paraissent pas pertinentes dans sa tâche d'accès au contenu.

Après ces modifications de l'utilisateur sur l'ensemble documentaire et sur ses domaines d'intérêt, une nouvelle projection des ressources sur le nouvel ensemble documentaire doit être réalisée. Ainsi, de la quatrième étape abordée précédemment, un retour vers la première étape est effectué. La boucle d'utilisations ainsi instaurée par notre modèle permet alors de prendre en considération les enrichissements et l'expérience acquise des précédentes analyses. Les RTO *LUCIA*, utilisées pour représenter les domaines d'intérêt de l'utilisateur, ne sont donc jamais dans un état final, chacune de leurs utilisations, de leurs projections en ensembles documentaires, entraînant un retour sur les RTO et donc de potentielles modifications.

Une telle approche interactive et itérative nous éloigne encore un peu plus de celles défendues dans le cadre du Web Sémantique. Là où l'on peut considérer que le Web Sémantique cherche à rendre le plus possible partagées de vastes ontologies terminales qui synthétisent une connaissance pensée comme objective et devant convenir à tous les utilisateurs, nous préférons exploiter des ressources propres à un utilisateur. Il découle de notre démarche une *légèreté* des ressources et des traitements au sens de [Perlerin, 2004]. Ainsi, les ressources utilisées ne représentent que ce qui est important du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple, une centaine de termes) ce qui les rend moins complexes à construire, à maintenir dans le temps, à enrichir et surtout à exploiter à travers des projections dans des ensembles documentaires[87]. Dans le cadre de ce travail de thèse, nous avons illustré dans [Roy et Beust, 2007], à travers différentes expérimentations, comment une telle boucle d'interaction facilitait tout particulièrement la maintenance et l'évolution de RTO (nous revenons sur ce point dans la suite de cette thèse). Notre démarche s'inscrit ainsi dans un processus de recherche et de développement en aller-retours entre des logiciels, des ensembles documentaires et des utilisateurs, les uns étant conditionnés par les autres. Les RTO, mais aussi les ensembles documentaires, que nous considérons évoluent de façon endogène dans une boucle d'interactions entre la machine, l'utilisateur, ses domaines et son ensemble documentaire où chaque pôle est déterminant.

[87] Cette démarche nous paraît être une réponse au constat que dressent Didier Bourigault et Nathalie Aussenac-Gilles dans [Bourigault et Aussenac-Gilles, 2003] sur 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.*

Conclusion : une nouvelle façon de percevoir l'accès au contenu d'ensembles documentaires

Les différentes propositions réalisées dans ce chapitre ont pour objectif de poser un nouveau cadre pour l'accès personnalisé au contenu d'ensembles documentaires. Ces propositions, formant le modèle *AIdED*, permettent de donner à l'utilisateur la place centrale qui lui est peu souvent attribuée dans cette problématique. En se fondant sur le principe de détermination du local par le global, nous avons proposé de prendre en considération les différents paliers textuels et leurs influences mutuelles. Cette prise en considération a donné lieu à deux publications [Beust et Roy, 2006a, Beust et Roy, 2006b], où nous avons résumé et illustré les principes de cette proposition. La mise au point de visualisations cartographiques interactives de l'ensemble documentaire, permettant à l'utilisateur de naviguer entre les différents paliers textuels, a ensuite été proposée. Des telles cartes, résultant de la projection des domaines d'intérêt de l'utilisateur sur l'ensemble documentaire, font ressortir les différentes isotopies parcourant l'ensemble. Ces cartes constituent alors de véritables guides aux parcours interprétatifs de l'utilisateur, facilitant au mieux son accès personnalisé au contenu de l'ensemble documentaire. Nous illustrons la valeur ajoutée de nos propositions, pour des tâches variées d'accès au contenu d'ensembles documentaires, dans le dernier chapitre de cette thèse à travers différentes expérimentations. Avant cela, la mise en œuvre logicielle des différentes propositions du modèle *AIdED* est détaillée dans le chapitre suivant.

Chapitre 3

Instrumentation logicielle du modèle

Sommaire

Introduction 70
3.1 Généricité visée de la mise en outils et en instruments 71
3.1.1 Indépendance par rapport aux langues 71
3.1.2 Indépendance par rapport aux codages des caractères 72
3.1.3 Indépendance par rapport aux formats des ensembles documentaires . . 73
3.1.4 Propriétés des ressources textuelles et des transformations produites . . 73
3.2 Aides logicielles à la construction de RTO personnelles 76
3.2.1 Une approche centrée-utilisateur de la construction et la description de ressources lexicales 76
3.2.2 Généralités sur les outils proposés pour l'extraction et la description de lexies 77
3.2.3 *Memlabor* et *ThemeEditor* : outils existants pour l'extraction et la structuration de lexies 79
3.2.4 *VisualLuciaBuilder* : construction interactive de dispositifs *LUCIA* . . . 83
3.2.5 *FlexiSemContext* : outil pour la mise en contexte de lexies et de sèmes . 89
3.3 *ProxiDocs* : projections de RTO personnelles dans des ensembles documentaires 93
3.3.1 Présentation et objectifs 93
3.3.2 Traitements mis en œuvre 95
3.3.3 Méthodes de représentation d'ensembles documentaires à partir de RTO personnelles 95
3.3.4 Projection et classification de l'ensemble documentaire 97
3.3.5 Construction des supports de visualisation interactive de l'ensemble documentaire 99
3.3.6 Mise en œuvre logicielle au sein de la plate-forme *ProxiDocs* 108
3.3.7 Intégration et utilisation de *ProxiDocs* au sein de projets d'enseignement et de recherche 112
Conclusion : mise en instruments du modèle d'analyse à travers des logiciels interactifs 112

Introduction

Dans le chapitre précédent, nous avons détaillé le modèle *AIdED* et les différentes propositions qu'il réalise, aussi bien au niveau de la représentation des domaines d'intérêt de l'utilisateur qu'au niveau de leur projection visuelle et interactive en ensembles documentaires. Ces différentes propositions vont nécessiter une instrumentation informatique afin de les rendre opératoires. Les logiciels présentés dans ce chapitre sont ce que nous appelons des *logiciels d'étude*. Ce sont des logiciels de recherche qui permettent d'étudier un phénomène complexe.

Au sens d'Anne Nicolle dans [Nicolle, 1996], les logiciels d'étude sont conçus dans le but de vérifier des hypothèses sur les langues en les expérimentant sur du matériau textuel attesté, c'est-à-dire du matériau textuel produit par des locuteurs, des auteurs dans le cadre d'une pratique. On oppose tout particulièrement les matériaux textuels attestés aux phrases et textes construits artificiellement par l'expérimentateur dans le but de mettre en évidence un phénomène prévu.

Anne Nicolle précise dans [Nicolle, 2002] que les logiciels d'étude ne sont pas des logiciels dont la durée de vie et la diffusion sont aussi grandes que des logiciels commerciaux. Cette affirmation est toutefois à nuancer aujourd'hui avec le développement du logiciel libre. Dans le domaine du traitement automatique des langues, de nombreuses plates-formes logicielles comme *GATE*[88], *NLTK*[89], *Unitex*[90], *LinguaStream*[91], etc. sont accessibles gratuitement, avec leur documentation et sont, le plus souvent, *open-source*.

Selon Benoît Habert dans [Habert, 2005], l'*outillage* de la linguistique fait intervenir trois types d'objets, les deux premiers étant assimilables à des logiciels d'étude :

- les *instruments* prenant en entrée des données langagières et produisant en sortie des représentations transformées de ces données langagières ;
- les *outils*, objets informatiques plus génériques, polyvalents, pouvant manipuler des données langagières mais pas uniquement ;
- et les *ressources* qui sont les données langagières transformées par les instruments et dans une moindre mesure, manipulées par des outils.

Les logiciels d'étude présentés dans ce chapitre ont tous pour objectif d'apporter une aide à l'utilisateur dans des tâches bien distinctes.

Le premier type de tâche visée est l'aide à la constitution de ressources termino-ontologiques (RTO) décrivant les domaines d'intérêt de l'utilisateur. Une telle tâche nous a amené à développer un certain nombre d'*outils* informatiques, assez généralistes, assistant l'utilisateur dans des tâches de constitution et de caractérisation de RTO décrivant des domaines (tels *Memlabor*, *ThemeEditor*, *VisualLuciaBuilder* et *FlexiSemContext*).

Le second type de tâche que nous visons est la production de vues sur l'ensemble documentaire tenant compte de ces domaines. Dans une telle tâche, nous avons développé un *instrument* (la plate-forme *ProxiDocs*) permettant d'obtenir différentes représentations visuelles et interactives mettant en évidence la projection des domaines d'intérêt dans des ensembles documentaires. Ces outils et instruments manipulent des *ressources* de plusieurs types : à la fois des ensembles documentaires, mais aussi l'expression des domaines pertinents pour l'utilisateur, ces domaines étant représentés par des données langagières.

Dans ce chapitre, nous mettons d'abord en évidence la généricité visée pour l'implémentation de notre modèle d'analyse, généricité aussi bien en langues qu'en formats d'entrées ou de sorties. La mise en œuvre des outils logiciels dédiés à l'aide à la construction des RTO personnelles est

[88] `http://gate.ac.uk/` (page consultée le 21 juillet 2007).
[89] `http://nltk.sourceforge.net` (page consultée le 21 juillet 2007).
[90] `http://www-igm.univ-mlv.fr/~unitex` (page consultée le 21 juillet 2007).
[91] `http://www.linguastream.org/home.html` (page consultée le 21 juillet 2007).

ensuite détaillée. La partie suivante présente les instruments logiciels dédiés à la projection des ressources de l'utilisateur sur des ensembles documentaires. Enfin, nous concluons ce chapitre sur les différentes utilisations que nous envisageons de ces outils et instruments afin de mettre en évidence leur valeur ajoutée.

3.1 Généricité visée de la mise en outils et en instruments

Cette partie va présenter en quoi l'implémentation de nos outils et de nos instruments est générique et à quels niveaux se place une telle généricité.

3.1.1 Indépendance par rapport aux langues

Le modèle présenté dans la partie précédente est considéré comme *alingue* [Vergne, 2004]. L'utilisateur explicite les domaines de son choix dans la langue qu'il souhaite, ces domaines sont ensuite projetés sur un ensemble documentaire toujours choisi par l'utilisateur dans la langue de son choix (et en toute vraisemblance, dans la langue des ressources construites).

Les traitements mis en œuvre afin d'implémenter notre modèle d'analyse doivent être indépendants des langues : aucun traitement d'identification de langue n'est réalisé, les traitements sont indépendants des langues puisqu'aucune ressource linguistique, hormis les RTO personnelles construites par l'utilisateur, ne sont exploitées.

Néanmois, nous avons, à plusieurs moments pendant cette thèse, pensé à intégrer des outils d'analyse morphologique, voire morpho-syntaxique, afin de permettre en quelque sorte une désambiguïsation de certains termes. Le cas le plus souvent rencontré est celui d'un nom commun correspondant également à un verbe à l'infinitif, comme c'est le cas avec *pouvoir* ou *devoir*. Dans un tel cas, un étiqueteur morphologique comme le *TreeTagger* [Schmidt, 1994] permettrait de déterminer (à un taux d'erreur près) la catégorie du terme et ainsi d'éviter toute confusion.

Afin de mettre en pratique ce genre d'analyses, l'utilisateur devrait préciser, pour chaque terme de ses RTO, la catégorie morphologique qu'il considère. Cette tâche supplémentaire serait assez lourde pour l'utilisateur et nous avons donc choisi de gagner en légèreté en ne permettant pas une telle opération. De plus, un outil comme le *TreeTagger* utilise des ressources propres à chaque langue pour effectuer les analyses. Il n'est alors possible de traiter que des langues pour lesquelles on possède des ressources, ceci rendant les analyses dépendantes des langues.

De tels cas d'ambiguïté morphologique restent rares, comme nous avons pu l'expérimenter dans le projet *IsoMeta*, présenté dans le chapitre suivant de cette thèse. Au cours de nos expériences dans ce projet, concernant l'étude de trois métaphores conceptuelles (la météorologie boursière, la santé financière et la guerre économique) dans un corpus d'articles boursiers, nous n'avons observé aucune ambiguïté liée à la morphologie des lexies exploitées[92].

Dans nos analyses, la redondance nous permet de pallier ce genre d'ambiguïtés morphologiques. Le nom commun et l'auxiliaire *été* peuvent bien évidemment être confondus mais l'environnement sémantique dans des ensembles documentaires de ces deux termes sera très différent. Le nom commun *été* pourra par exemple être entouré des lexies *estivale*, *vacances*, *saison*, etc. actualisant des isotopies très différentes de ce que pourrait actualiser l'auxiliaire *été*. Les ambiguïtés observées dans le projet *IsoMeta* étaient principalement d'ordre sémantique, avec de mêmes

[92] Par exemple, une lexie dont l'utilisation visée était en tant que nom commun ne s'est jamais retrouvée employée en corpus dans une autre catégorie morphologique entraînant une interprétation différente (telle *je* souris et *la* souris ou encore *il* lit et *le* lit où les graphies sont soit des verbes, soit des noms, les deux cas entraînant des interprétations très différentes).

lexies décrivant des entités différentes (homonynes). Comme par exemple, *dépression* décrivant aussi bien une dépression nerveuse qu'une dépression météorologique.

Les versions futures des implémentations du modèle pourraient cependant intégrer des outils d'analyse morphologique et même morpho-syntaxique eux-mêmes alingues. Les auteurs de [Houben et Rioult, 2006] proposent une méthode apportant un début de réalisation d'un étiqueteur morpho-syntaxique alingue, se basant uniquement sur des propriétés très générales des langues et exploitant des méthodes issues des domaines de la fouille de données et de l'apprentissage supervisé. Les résultats présentés dans cet article sont assez encourageants et laissent penser que de telles approches, sans ressource linguistique, ni règle symbolique, seront utilisables dans un futur proche[93]. Une telle approche correspond beaucoup plus à nos attentes que des étiqueteurs basés sur des ressources et des règles propres à chaque langue. Malgré cela, l'utilisateur aura toujours à indiquer dans ses RTO personnelles des informations liées à la catégorie morpho-syntaxique des lexies qu'il y fait intervenir, avec les différentes contraintes que cela entraîne.

3.1.2 Indépendance par rapport aux codages des caractères

Être indépendant des langues signifie également être indépendant des codages des caractères. Le développement du Web et des différents moyens de communication a mis en évidence un grand nombre de problèmes liés à l'échange de fichiers textuels. Ainsi, chaque pays codait ses fichiers texte dans un jeu de caractères propre à sa langue, une telle opération limitait, voire rendait impossible, la lecture de tels fichiers sur des ordinateurs ne possédant pas le jeu de caractères exploité à l'origine.

Unicode[94] a été développé dans le but de remplacer l'utilisation de jeux de caractères nationaux. Cette norme, développée par le Consortium Unicode[95] vise à donner à tout caractère, de n'importe quel système d'écriture de langue, un nom et un identifiant numérique, et ce de manière unifiée, quelle que soit la plate-forme informatique ou le logiciel. Des données textuelles respectant la norme Unicode sont donc lisibles sur n'importe quelle machine disposant d'un support adéquat. Nous pouvons citer [Andries, 2002] pour une présentation détaillée d'Unicode et de son histoire.

Dans nos développements logiciels, nous avons choisi d'intégrer cette norme. Le schéma d'algorithme suivant illustre une telle intégration :

Entrée : Données textuelles dans n'importe quel jeu de caractères
Sortie : Données textuelles dans un encodage Unicode

1.1 *Détection du jeu de caractères utilisé dans les données textuelles prises en entrée*;
1.2 *Conversion des ressources en Unicode* ;
1.3 *Traitements internes réalisés sur les données en Unicode* ;
1.4 *Production des données textuelles de sortie en Unicode* ;

Algorithme 1 : Schéma d'algorithme de tout traitement multilingue.

De cette manière, il est possible à partir de ressources textuelles prises dans n'importe quelle

[93] Tout dépendra alors du degré de finesse souhaité pour l'analyse morphologique. Une méthode sans ressource pourra être moins efficace mais sera applicable sur un très grand nombre de langues alors qu'une méthode avec ressources et règles pourra être plus efficace sur une langue donnée mais ne sera utilisable que sur cette langue.

[94] La première version d'Unicode remonte à octobre 1991, la version la plus récente à l'heure actuelle est la version 5.0.0 [The-Unicode-Consortium, 2006].

[95] `http://www.unicode.org` (page consultée le 14 juillet 2007).

écriture (et donc, éventuellement dans un jeu de caractères local différent d'Unicode) de produire des sorties en Unicode, lisibles de tous. Cette « lecture universel » est, tout de même, à nuancer même à l'heure actuelle. L'utilisation d'Unicode n'est pas encore partagée par tous et il est encore très (trop ?) fréquent que des problèmes de lisibilité subsistent, par exemple, sur des pages de sites Internet ou sur des courriers électroniques. Nous renvoyons à [Giguet et Lucas, 2002] où les auteurs présentent le développement d'un outil de recherche sur Internet exploitant Unicode et les nombreux problèmes liés à cette utilisation. Plusieurs années après l'écriture de cet article, de tels problèmes sont encore bien présents, l'intégration de la norme Unicode dans les différents systèmes et logiciels utilisés prenant du temps.

3.1.3 Indépendance par rapport aux formats des ensembles documentaires

Les données textuelles sont disponibles sur Internet dans un très grand nombre de formats, tous différents les uns des autres. Parmi ces formats, nous pouvons citer le format texte brut, le format HTML, le format DOC proposé par *Microsoft Word* et sa version *open-source* ODT proposée par *OpenOffice*, le format PDF d'*Adobe*, etc. Chacun de ces formats a une utilisation privilégiée, par exemple, le format HTML est préféré lorsque qu'une consultation en ligne est souhaitée, un document au format PDF est préféré pour une impression du document, un document au format XML peut être choisi lorsque plusieurs personnes doivent manipuler et transformer le document, etc.

Afin de prendre en considération cette variété de formats de documents textuels, nous avons choisi d'inclure ou de développer différents convertisseurs permettant de transformer ces différents formats vers un format texte. Une telle transformation entraîne une perte d'information sur la structure et la mise en forme des documents. Nos travaux actuels ne traitant pas de telles caractéristiques, nous pouvons nous contenter d'une conversion vers un format textuel brut. À l'avenir, il serait, par contre, tout à fait envisageable de tenir compte de la structure des documents dans nos analyses. Dans ce cas, les différents convertisseurs utilisés devront être revus afin de transformer les fichiers textuels de différents formats vers un format permettant de représenter la structure des documents (par exemple, un format XML). Le format OpenDocument proposé initialement par *OpenOffice* pourrait ainsi servir, c'est un format de données ouvert pour les applications bureautiques et basé sur XML. Bien évidemment, les différents traitements réalisés par la suite devront être modifiés en conséquence afin de traiter utilement ces nouvelles informations.

3.1.4 Propriétés des ressources textuelles et des transformations produites

Les ressources prises en entrée ainsi que les transformations de ces ressources produites par nos instruments respectent certaines propriétés, que nous détaillons dans cette partie.

Légèreté

Comme nous l'avons énoncé dans le chapitre précédent, les RTO *LUCIA* que nous utilisons sont *légères* au sens de Vincent Perlerin [Perlerin, 2004], du fait qu'elles ne représentent que ce qui est pertinent du point de vue de l'utilisateur et restent ainsi de taille raisonnable (par exemple, une centaine de termes). De telles ressources sont moins complexes à construire, à maintenir, à enrichir et à projeter dans des ensembles documentaires. En conséquence, des traitements basés sur des ressources légères seront également *légers*, c'est-à-dire réalisables dans un temps raisonnable et d'une complexité informatique raisonnable.

Nous ajoutons ici, qu'en plus de ressources légères, nous souhaitons également obtenir des transformations « légères ». Autrement dit, nous souhaitons que ces transformations ne représentent que des choses simples, pertinentes et facilement accessibles à l'utilisateur, reprenant ainsi les recommandations de [Pednault, 2000] énoncées au chapitre précédent. Nous nous situons donc assez loin des systèmes proposant des représentations très complexes et pour lesquels une longue période d'apprentissage est nécessaire.

Cette notion de légèreté touche également les ensembles documentaires étudiés. Dans nos travaux, nous ne supposons pas que de tels ensembles puissent être représentatifs d'un phénomène langagier quelconque. Benoît Habert montre, par exemple, dans [Habert, 2004] que des « méga-corpus » [Kennedy, 1998][96] ne sont pas forcément représentatifs d'une langue. L'auteur montre ainsi que sur les 1 345 emplois du verbe *vendre* dans un corpus constitué de cinq années du journal *Le Monde*, aucun n'exprime une trahison.

En opposion à l'utilisation de tels méga-corpus, les ensembles documentaires utilisés dans nos travaux regroupent des textes pertinents aux yeux d'un utilisateur pour une tâche donnée à un moment précis. La cardinalité de ces ensembles doit donc être assez réduite et, en aucun cas, ne vouloir représenter autre chose qu'un point de vue très précis sur une tâche donnée pour un individu.

Réutilisabilité et portabilité

Lorsqu'un utilisateur passe du temps à travailler sur des RTO dans un format électronique, ou à utiliser un logiciel produisant des transformations de ces ressources, il faut bien évidemment que les ressources et les transformations construites soient facilement réutilisables par l'utilisateur et même par n'importe quel autre utilisateur estimant en avoir besoin. Pour cela les ressources et les transformations doivent être accessibles en « clair », dans un format libre et non propriétaire.

Un autre point important et complémentaire au précédent est la portabilité des ressources et des transformations. Une fois les ressources construites et/ou les transformations produites, il est important de pouvoir les utiliser à différents endroits, sur différentes machines et différents systèmes d'exploitation, et dans le cadre de différentes tâches.

Le format XML [W3C, 2006a] permet de répondre à ces exigences de réutilisabilité et de portabilité. Des fichiers XML sont des fichiers contenant du texte directement lisible par n'importe quel éditeur de texte, de n'importe quelle machine, et sous n'importe quel système d'exploitation. Si des informations sur la mise forme du fichier XML sont fournies (par exemple, à l'aide d'une feuille de transformation XSL [W3C, 2006b]), la quasi-totalité des navigateurs Internet permet de visualiser le fichier XML mis en forme.

Nous produisons en sortie, non pas des transformations finies des données textuelles prises en entrée, à la manière d'un analyseur syntaxique produisant en sortie une annotation un texte étiqueté, mais des supports dédiés à la visualisation et à la manipulation interactives des données textuelles prises en entrée. Pour fournir à l'utilisateur des supports visuels d'interaction, nous avons choisi d'utiliser le format de représentation graphique vectorielle *Scalable Vector Graphics* (SVG) [W3C, 2003].

En SVG, les objets graphiques et leurs caractéristiques sont codés sous forme alpha-numérique dans un document XML. Le code source d'un fichier SVG est donc directement lisible dans un éditeur de texte. Son formalisme XML lui permet d'être manipulé et enrichi par d'autres composants logiciels (feuilles de transformations, applications, etc.). Le format SVG gère les formes géométriques de base comme des rectangles, des ellipses, etc., mais aussi des chemins

[96] C'est-à-dire des corpus contenant une centaine de millions de mots, ce qui représente environ un millier de romans ou encore cinq années des productions d'un quotidien comme *Le Monde*.

qui utilisent les courbes de Bézier et qui permettent ainsi d'obtenir n'importe quelle forme. Le remplissage peut se faire à l'aide de couleurs, de dégradés de couleurs, de filtres, etc.

Le SVG peut être visualisé nativement avec certains navigateurs Internet, comme *Konqueror*, *Opera*, et *Mozilla Firefox*, ou à l'aide du plug-in proposé par *Adobe* pour d'autres, comme *Internet Explorer* ou *Mozilla* (version originale encore développée aujourd'hui par *Mozilla* en parallèle à *Mozilla Firefox*)[97] Les composants graphiques ainsi produits sont visualisables directement dans un navigateur sous des systèmes d'exploitation de type Windows, Apple Mac OS ou Linux. L'inclusion de ces composants dans des pages HTML est tout à fait réalisable.

Interactivité

Le choix du langage SVG pour les sorties visuelles que nous produisons est lié aux nombreuses possibilités d'interactions qu'il permet d'offrir aux utilisateurs. Des interactions simples, tels des zooms avant et arrière, ainsi que des déplacements sur les objets graphiques, sont directement accessibles en visualisant le document SVG avec un navigateur Internet. Le format SVG permet également l'intégration d'animations de type XML/SMIL [W3C, 2001] et la manipulation des objets graphiques par des scripts de type ECMAScript [ECMA-International, 1999] modifiant dynamiquement les attributs des éléments du fichier XML[98].

Les différents points abordés précédemment (et en particulier sa lisibilité, sa portabilité, sa réutilisabilité) nous ont fait préférer l'utilisation d'objets graphiques au format SVG, plutôt que des objets prenant place dans le langage Flash. Cette technologie est cependant majoritaire sur Internet pour mettre en œuvre des animations et des interactions. Son principal avantage par rapport au SVG est sa relative simplicité de mise en œuvre *via* un logiciel dédié[99]. Ses principaux inconvénients sont alors liés : il faut utiliser un logiciel particulier pour faire du Flash, rendant par la même occasion très difficile la production automatique d'objets graphiques Flash par des programmes informatiques. Une animation Flash est un fichier dans un format propriétaire (format `swf`), alors qu'un simple éditeur de texte suffit pour produire et modifier du SVG.

Soutenu par la *Free Software Fundation*[100], le projet *Gnash*[101] (« grincement de dents ») vise, depuis 2005, à créer une alternative libre (sous licence GNU GPL) au lecteur Flash d'*Adobe*. Il fait suite à de nombreux projets en la matière n'ayant pas vraiment abouti. Celui-ci propose d'ores et déjà un plug-in Firefox et une version autonome, reconnaissant les fichiers Flash 7. Il respecte également le système d'échanges XML de la spécification officielle. Ce projet est l'une des six campagnes prioritaires[102] de la *Free Software Fondation*, avec *OpenOffice* ou le projet *Classpath*. Si, comme nous le pensons, une telle initiative aboutit, il serait alors tout à fait envisageable de produire des supports d'interactions dans ce format « Open Source Flash ».

[97] Le plugin proposé par Adobe est accessible à l'adresse suivante : `http://www.adobe.com/svg/viewer/install/mainframed.html` (page consultée le 26 janvier 2007). Devant le développement de navigateurs intégrant nativement le formalisme SVG (mais aussi pour des raisons commerciales), *Adobe* a décidé d'arrêter le développement de ce plugin : `http://www.adobe.com/svg/pdfs/ASV_EOL_FAQ.pdf` (page consultée le 26 janvier 2007).

[98] De nombreux sites Web proposent des exemples d'images, d'animations, de jeux au format SVG, le site `http://croczilla.com/svg/samples` (page consultée le 26 janvier 2007) passe en revue ces différentes possibilités à travers différents exemples.

[99] Par exemple, le logiciel d'*Adobe* : `http://www.adobe.com/fr/products/flash/flashpro` (page consultée le 26 janvier 2007).

[100] `http://www.fsf.org` (page consultée le 26 janvier 2007).

[101] Description du projet disponible à l'adresse suivante : `http://www.gnu.org/software/gnash` (page consultée le 25 janvier 2007).

[102] Campagnes prioritaires présentées à l'adresse suivante : `http://www.fsf.org/campaigns/priority.html` (page consultée le 25 janvier 2007).

3.2 Aides logicielles à la construction de RTO personnelles

L'objectif de notre travail est de proposer aux utilisateurs des vues personnelles et interactives sur des ensembles documentaires. Comme nous avons pu le voir dans le chapitre précédent, la construction de telles vues passe par la description, à l'aide de lexies et de formes graphiques associées, d'attributs et de valeurs d'attributs, des domaines pertinents aux yeux de l'utilisateur dans le cadre de sa tâche.

Afin d'exprimer son point de vue sur des domaines de son choix, nous proposons à l'utilisateur différents outils l'assistant durant les phases d'extraction de lexies d'un ensemble documentaire et de description de ces lexies selon le modèle *LUCIA*.

3.2.1 Une approche centrée-utilisateur de la construction et la description de ressources lexicales

La construction et la description de RTO se basent sur ce que nous appelons des ensembles documentaires d'observation, désignant aussi bien des corpus et des collections d'observation (se reporter au chapitre 1 de cette thèse pour les définitions de ces termes adoptées). Des ensembles documentaires d'observation doivent donc être constitués de textes attestés, n'ayant pas été créés artificiellement par un utilisateur pour une tâche donnée. Les textes placés dans ces ensembles d'observation doivent être en rapport très étroit avec la pratique de l'utilisateur et la tâche qu'il vise.

Précédemment, nous avons proposé une représentation approchée de l'intertexte par l'ensemble documentaire que l'utilisateur souhaite analyser. Ce choix a été grandement motivé par la place centrale que nous donnons à l'ensemble documentaire et aux liens existant entre les textes qu'il regroupe. Dans une telle démarche, il nous semble tout à fait cohérent de se placer dans le cadre d'une observation empirique de phénomènes langagiers et plus particulièrement d'entités lexicales (lexies et graphies) dans des ensembles documentaires. À partir de ces observations, attestant l'utilisation de telles lexies ou permettant la mise en évidence de nouvelles lexies (une graphie présente dans l'ensemble documentaire peut inciter l'utilisateur à créer une ou plusieurs lexies liées à la graphie observée), l'utilisateur peut isoler celles lui paraissant les plus pertinentes pour la description de ses domaines d'intérêt pour la tâche qu'il vise. Une telle démarche rejoint complètement celle de [Bourigault et Slodzian, 1999], où les auteurs affirment :

> *La tâche d'analyse terminologique vise alors avant tout la construction d'une description des structures lexicales à l'œuvre dans un corpus textuel à partir d'une analyse réglée de ce corpus. (...) Pour chaque unité choisie, l'analyste construit une signification (type) à partir des sens (occurrences) attestés dans le corpus.*

La construction de terminologies à partir d'ensembles documentaires est une tâche particulièrement délicate et constitue un champ de recherche à part entière (se reporter, par exemple, à [Jacquemin et Zweigenbaum, 2000] et [Nazarenko et Hamon, 2002]). Contrairement à des approches entièrement automatiques, basées, par exemple, sur des règles d'extraction morpho-syntaxiques [Smadja et McKeown, 1990], sur des statistiques lexicales [Lebart *et al.*, 1998] ou sur des approches combinant les deux précédentes [Bourigault, 1994], notre particularité est, encore une fois, la place que nous laissons à l'utilisateur dans cette tâche. Les principes de cette extraction supervisée par l'utilisateur de lexies à partir d'ensembles documentaires pour construire des RTO *LUCIA* ont déjà été présentés dans [Perlerin, 2004, pages 127-129], nous y renvoyons pour plus de détails. Différents travaux en TAL, proches de l'extraction terminologique, exploitent une telle approche supervisée, laissant ainsi l'utilisateur exprimer son point de vue sur l'appartenance d'un candidat-terme à une classe thématique [Rossignol et Sébillot, 2003], sur la qualité d'une

relation sujet-verbe [Claveau et Sébillot, 2004] ou encore sur la pertinence de couples de termes antonymes [Schwab *et al.*, 2005]. Cependant, le coté « supervisé » de tels travaux ne rend pas pour autant des résultats personnalisés pour un utilisateur et une tâche donnée, il s'agit juste ici de valider ou non des propositions du système. Dans notre approche, c'est l'utilisateur qui choisit et décrit les termes de son choix pour la tâche qu'il vise.

Quand l'utilisateur a extrait les lexies pertinentes pour la description des domaines de son choix, nous lui proposons de rassembler ces lexies en catégories, puis de structurer ces catégories en dispositifs *LUCIA*. Cette structuration peut se faire *via* une interface logicielle, comme celle détaillée dans [Perlerin, 2004, pages 151-159]. Différentes plates-formes logicielles existent afin de proposer une aide dans la structuration de terminologie. Nous pouvons citer, par exemple, *Contexto* de Jean-Luc Minel [Minel, 2001] ou encore *Terminae* de Brigitte Biébow et Sylvie Szulman [Biébow et Szulman, 2000]. De tels outils sont particulièrement intéressants dans le sens où ils laissent à leurs utilisateurs les moyens d'interagir sur les termes en les mettant en relation et en les décrivant. Cependant, ils sont plutôt dédiés à des spécialistes et non à des utilisateurs lambda. Au contraire, les différentes aides logicielles que nous proposons se veulent accessibles au plus grand nombre, aussi bien pour des descriptions de domaines très simples en ensembles de lexies, que pour des descriptions fines en dispositifs *LUCIA*.

3.2.2 Généralités sur les outils proposés pour l'extraction et la description de lexies

Les différentes aides logicielles que nous proposons pour l'extraction et la structuration de RTO *LUCIA*, ont donc, chacune, la particularité de donner à l'utilisateur et à son ensemble documentaire une place centrale. Avant de détailler chacun de ces outils, et d'insister tout particulièrement sur les outils *VisualLuciaBuilder* et *FlexiSemContext* développés durant cette thèse, nous donnons quelques informations générales sur ces derniers, leur fonctionnement et leurs interactions. Le schéma de la figure 3.1 illustre cette position centrale et présente les outils abordés dans cette section.

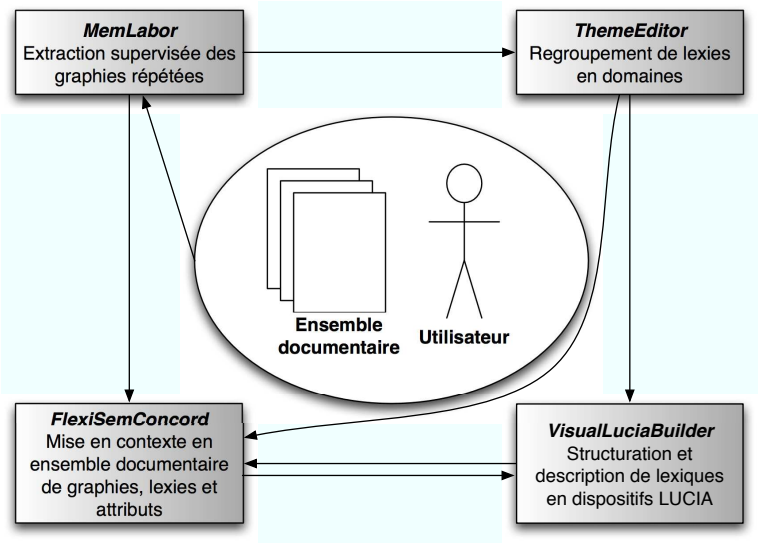


FIG. 3.1 – Les différents outils logiciels proposés pour l'extraction et la caractérisation de lexies ; leurs interactions sont représentées par des flèches.

Ces outils sont donc utilisés de la façon suivante :

- *Memlabor* permet une extraction supervisée des graphies répétées dans l'ensemble documentaire d'observation de l'utilisateur, l'outil propose également une aide dans la phase d'assemblage de graphies en lexies ;
- *ThemeEditor* assiste l'utilisateur dans le regroupement de lexies relevant d'un même domaine en permettant un coloriage interactif d'ensembles de lexies des textes d'un ensemble documentaire ;
- *VisualLuciaBuilder* permet alors de décrire interactivement chaque domaine sous forme d'un dispositif *LUCIA* ;
- *FlexiSemContext* est un concordancier pouvant être utilisé en complément de chacun des logiciels précédents afin de mettre en contexte dans l'ensemble documentaire d'observation, des graphies, des lexies, des attributs, des valeurs d'attributs et des couples *attribut : valeur*.

L'utilisation de ces différents logiciels se fait en « aller-retours ». A tout moment, il est possible de revenir à l'étape précédente afin de réviser, de revoir un choix réalisé auparavant. Par exemple, lors de la création d'un dispositif *LUCIA* avec *VisualLuciaBuilder*, il est possible de vouloir remettre une lexie en contexte dans l'ensemble documentaire avec *FlexiSemContext*. L'utilisation de ces outils est « en spirale » et non finalisée. Les ressources sont utilisables et projetables à tout moment du cycle avec différents niveaux de description et de révision. Contrairement à la majorité des outils existants en TAL, les ressources prises en entrée ne sont ici jamais considérées comme finalisées. Elles sont stabilisées pour un moment et une tâche donnés. Cette utilisation cyclique des outils entraîne un enrichissement constant des ressources au fil du temps et des utilisations.

Ces outils *Memlabor*, *ThemeEditor* et *VisualLuciaBuilder* échangent des données *via* des fichiers XML. L'outil *FlexiSemContext* n'intègre pas de formalisme XML d'échange avec les autres outils, son rôle est de proposer à l'utilisateur de saisir une graphie, une lexie, un attribut, etc. *via* une interface Web reliée à des ensembles documentaires afin de mettre en contexte les éléments donnés. La figure 3.2 illustre ces échanges.

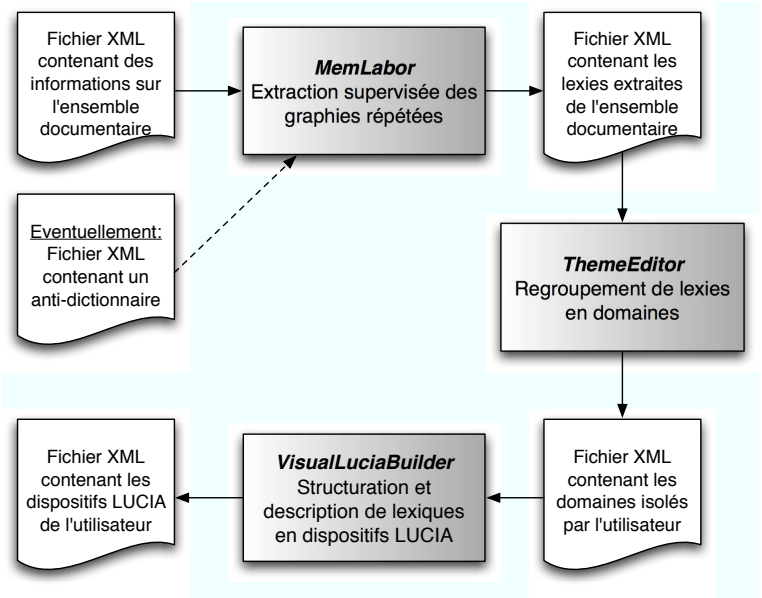


FIG. 3.2 – Les différents échanges entre les logiciels *Memlabor*, *ThemeEditor* et *VisualLuciaBuilder*.

Ces échanges de données au format XML entre des composants logiciels font fortement penser à des plates-formes logicielles dédiées au TAL. Par exemple, la plate-forme *LinguaStream* [Bilhaut et Widlöcher, 2006] (figure 3.3), développée au Laboratoire GREYC de l'Université de Caen depuis plusieurs années, propose ce type d'échanges entre des « briques » logicielles à l'intérieur de chaînes de traitements.

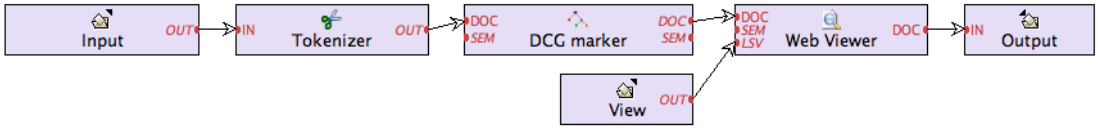


FIG. 3.3 – Un exemple de chaîne de traitement *LinguaStream*.

Chacune de ces briques est dédiée à une tâche très spécialisée, comme la tokenisation, l'étiquetage morpho-syntaxique, etc. L'enchaînement automatique de plusieurs de ces briques au sein d'une même chaîne de traitement permet de réaliser des tâches complexes (comme des analyses thématiques [Bilhaut, 2006] et rhétoriques [Widlöcher, 2006] de textes), mais ne laisse à aucun moment une place aux interactions de l'utilisateur, comme nos outils peuvent le faire *via* leurs interfaces respectives.

Les quatre outils abordés dans cette section sont détaillés dans les parties suivantes autour d'un exemple concret d'utilisation. Ces outils sont tous *open-source*, développés dans les langages de programmation Java [Gosling *et al.*, 2005] (pour *Memlabor*, *ThemeEditor* et *VisualLuciaBuilder*) et PHP [Achour *et al.*, 2007] (pour *FlexiSemContext*) et disponibles sur Internet avec leur documentation à l'adresse suivante : `http://www.greyc.unicaen.fr/island/logiciel` (page consultée le 25 janvier 2007).

3.2.3 *Memlabor* et *ThemeEditor* : outils existants pour l'extraction et la structuration de lexies

L'outil *Memlabor* a été développé par Vincent Perlerin [Perlerin, 2002] afin de fournir à l'utilisateur une assistance dans l'extraction des lexies présentes dans un ensemble documentaire. Nous avons réutilisé cet outil pour deux tâches principales :

- le calcul des graphies répétées avec un possible filtrage à l'aide d'un anti-dictionnaire ;
- et l'assemblage de graphies en lexies.

En exploitant le principe de cohésion lexicale, *Memlabor* se fonde sur l'hypothèse que plus une graphie (hors mots de l'anti-dictionnaire) est répétée dans un ensemble de textes, plus elle est susceptible d'être associée à l'un des thèmes présents dans l'ensemble considéré [Perlerin, 2004, page 141]. En présentant aux utilisateurs une liste des graphies classées par ordre décroissant de fréquence d'apparition, le logiciel permet une première assistance à l'extraction de mots intéressants pour la tâche à partir d'un ensemble documentaire. À partir de cette liste de graphies répétées, l'utilisateur peut construire la liste des lexies qui lui semblent pertinentes par rapport aux domaines qu'il souhaite décrire dans le cadre de sa tâche.

L'application de l'outil sur un ensemble documentaire constitué de 789 articles du journal *Le Monde* de 1989 choisis aléatoirement parmi l'ensemble des articles de l'année[103] permet d'obtenir dans un premier temps l'écran présenté en figure 3.4.

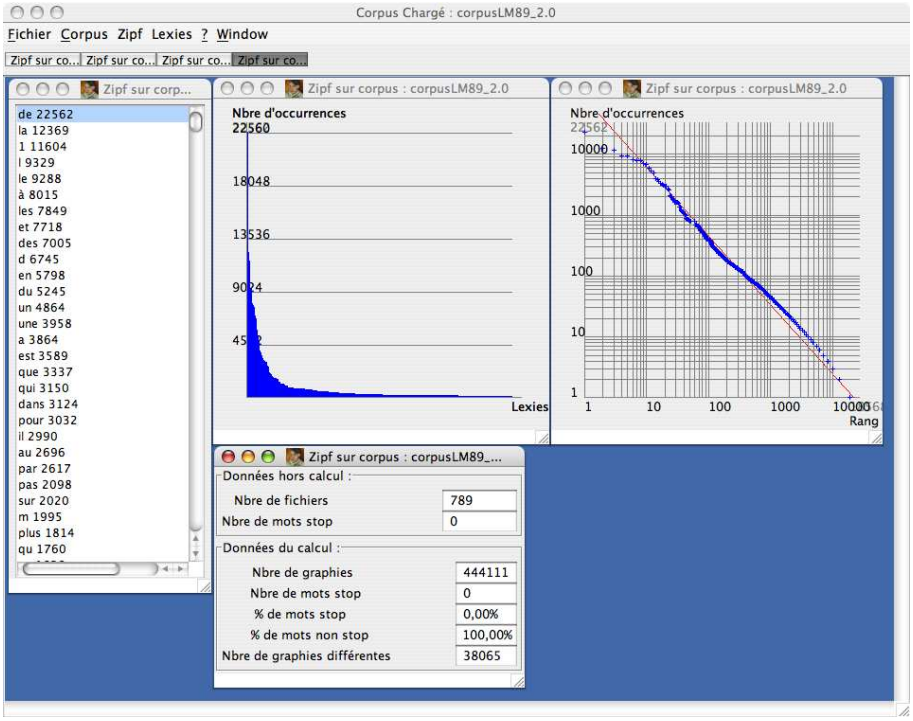


FIG. 3.4 – *Memlabor* : Écran présentant les graphies répétées dans l'ensemble documentaire triées par ordre décroissant de fréquence. L'interface présente également deux graphiques mettant respectivement en évidence les graphies en fonction de leur fréquence et les rangs des graphies en fonction de la fréquence.

Memlabor propose un calcul de type « Zipf »[104]. L'utilisation linguistique de cette loi de Zipf est celle qui nous intéresse le plus. Emmanuel Giguet dans [Giguet, 1998, page 168] explique que la liste des graphies issues d'un calcul de type Zipf amène à repérer, de façon approximative, deux zones distinctes et contigües[105]. La première de ces zones, contenant les graphies les plus fréquentes, regroupe majoritairement des mots grammaticaux, alors que la seconde zone contient des termes « lexicaux thématisants ». Ces zones restent cependant difficilement repérables automatiquement. Malgré tout, retourner à l'utilisateur une telle liste des graphies répétées dans son ensemble documentaire est un bon moyen pour lui permettre d'appréhender le contenu de cet ensemble documentaire. Si, en plus, ce dernier peut interagir avec cette liste en sélectionnant

[103] Cet ensemble documentaire a été construit dans le but d'étudier les différentes thématiques abordées dans le journal au cours de cette année. Le choix aléatoire des articles parmi l'ensemble des 39 290 articles produits sur l'année est lié à la légéreté des traitements que nous voulons garder, c'est-à-dire à leur relative rapidité, au moins dans le cadre de l'expérience présentée dans ce chapitre afin d'illustrer les différents outils. L'ensemble documentaire ainsi constitué contient 2% des articles de l'année.

[104] En 1935, le linguiste Harvard Georges Kingsley Zipf vérifie à la main que pour un corpus donné (l'œuvre *Ulysse* de James Joyce), la fréquence d'un terme est inversement proportionnelle à son rang [Zipf, 1949].Une telle loi a alors entraîné un grand nombre de travaux dans des domaines très différents (compression informatique [Caron *et al.*, 2003], démographie [Hill, 1970], etc.).

[105] Emmanuel Giguet propose de réaliser en ligne, *via* une *applet* Java, de tels calculs sur du texte choisi par l'utilisateur : `http://users.info.unicaen.fr/~giguet/java/zipf.html` (page consultée le 25 janvier 2007).

certaines graphies qu'il juge pertinentes, alors la liste devient un très bon premier support pour l'interaction entre l'utilisateur et l'ensemble documentaire.

Pour permettre à l'utilisateur de mettre en évidence ces entités lexicales thématisantes, nous lui proposons de parcourir la liste des graphies répétées dans l'ensemble documentaire. Pour chaque graphie de la liste, un clic droit sur la graphie permet d'afficher un menu contextuel proposant :

- soit d'ajouter la graphie à un anti-dictionnaire[106] ; une fois un anti-dictionnaire constitué ou réutilisé, il est alors possible de réaliser un nouveau comptage en filtrant sur les éléments qu'il contient ;
- soit d'ajouter ces graphies dans une nouvelle zone de l'interface dédiée à la description de lexies comme la figure 3.5 l'illustre.

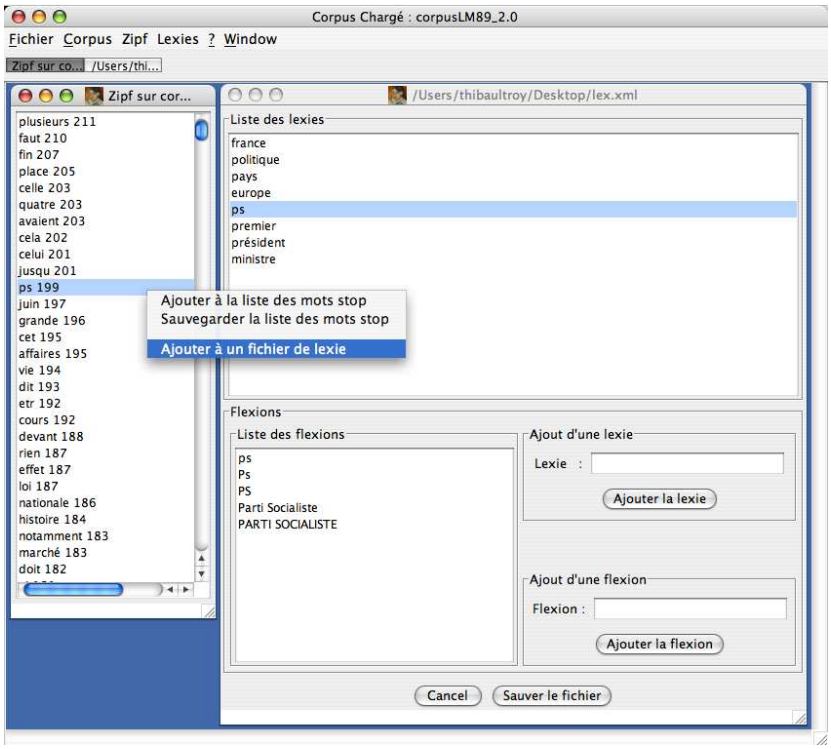


FIG. 3.5 – *Memlabor* : Écran mettant en évidence la constitution d'une liste de lexies à partir de la liste des graphies répétées dans l'ensemble documentaire.

Par cette interface, l'utilisateur peut ajouter pour une lexie donnée les formes fléchies pertinentes à ses yeux. Dans la figure 3.5, on peut, par exemple, voir que la graphie *PS* a été associée à la lexie *Parti Socialiste*. De même, à la lexie *société*, sa forme plurielle *sociétés* peut être associée[107]. Il est également possible de composer et d'enrichir plusieurs graphies en une lexie, comme les graphies *premier* et *ministre* pouvant former la lexie *premier ministre*.

[106] L'utilisateur peut constituer son propre anti-dictionnaire, par exemple avec les principaux mots grammaticaux présents dans son ensemble documentaire, ou bien réutiliser l'un des nombreux anti-dictionnaires existant sur Internet, comme ceux que l'on peut trouver sur le site de l'Institut Interfacultaire d'Informatique de l'Université de Neuchâtel : `http://www.unine.ch/Info/clef` (page consultée le 8 février 2007) pour 15 langues (anglais, français, allemand, arabe, etc.). Se reporter à [Savoy et Berger, 2005] pour une exploitation de ces anti-dictionnaires.

[107] Ce n'est pas une obligation de préciser les différents accords en genre et en nombre des lexies, nous verrons par la suite qu'il est possible d'aller chercher automatiquement de telles formes fléchies dans une base de données lexicales (uniquement pour le français).

Dans l'ensemble documentaire pris en exemple, un premier comptage sans anti-dictionnaire a permis de dénombrer au total 444 111 graphies dont 38 065 différentes. Un second comptage, prenant en considération un anti-dictionnaire de 463 mots (celui abordé précédemment, proposé par l'Université de Neuchâtel), a permis de filtrer 195 776 des 444 111 occurrences globales pour ne laisser « que » 37 747 graphies différentes. À partir de la liste de ces graphies triées par l'outil dans l'ordre décroissant de leur fréquence, une liste de 644 lexies a été construite. La constitution de cette liste a été faite dans l'objectif de représenter les principales thématiques présentes dans l'ensemble documentaire.

Une fois une telle liste de lexies construite par l'utilisateur, ce dernier peut, s'il le souhaite, réaliser un comptage de ces lexies dans l'ensemble documentaire (comptage tenant compte de la forme graphique de chaque lexie ainsi que de leurs formes fléchies). Un enregistrement de la liste de lexies dans un format XML est alors proposé à l'utilisateur afin de simplifier leur utilisation future par d'autres outils et en particulier par l'outil *ThemeEditor* permettant de regrouper des lexies en domaines révélant les thématiques illustrées par ces lexies.

Après une première sélection de lexies à l'aide de *Memlabor*, l'étape suivante consiste donc à les rassembler en domaines. L'outil *ThemeEditor*, élaboré par Pierre Beust [Beust, 2002], propose une aide pour une telle tâche. L'utilisateur est invité par l'outil à nommer les domaines qui l'intéressent. Une fois cette opération réalisée, il est proposé à l'utilisateur de placer des lexies (comme celles extraites avec *Memlabor*, mais ce dernier peut également ajouter d'autres lexies de son choix) au sein de ces différents domaines. Ce placement peut se faire de façon non exclusive, une lexie pouvant être associée à plusieurs domaines.

Les ressources ainsi constituées sont alors projetées sur l'ensemble documentaire initial. Une couleur est associée à chaque domaine. La couleur peut être choisie par l'utilisateur, si ce n'est pas le cas, une couleur par défaut, non utilisée, est attribuée. Le principe de coloriage des lexies selon le domaine auquel elles appartiennent permet de mettre en évidence la répartition, l'alternance et les enchaînements au long d'un texte des sujets liés aux domaines ainsi créés. Une difficulté rencontrée lors de cette étape de coloriage est qu'une lexie peut appartenir à plusieurs domaines et donc avoir plusieurs étiquettes. Ce serait par exemple le cas de la lexie *avocat*, que l'on pourrait aussi bien affecter au domaine des aliments qu'à celui de la justice. Dans un tel cas, nous avons choisi d'attribuer la lexie au domaine le plus représenté dans le texte. Ceci revient en quelque sorte à prolonger les isotopies génériques (cf. chapitre précédent) du texte et donc à favoriser la redondance thématique.

Cette interface de lecture de l'ensemble documentaire permet à la fois une évaluation qualitative et quantitative des données (cf. figure 3.6). La lecture des extraits de textes où apparaissent des lexies coloriées peut alors amener :

- à des modifications d'association entre un mot et un domaine ;
- au repérage de mots peu redondants dans l'ensemble documentaire mais susceptibles de présenter un intérêt pour la description des domaines.

Après plusieurs utilisations itératives de *ThemeEditor*, plusieurs domaines peuvent avoir été décrits. Ils sont stockés en machine sous la forme de listes au format XML, format illustré en annexe A de cette thèse.

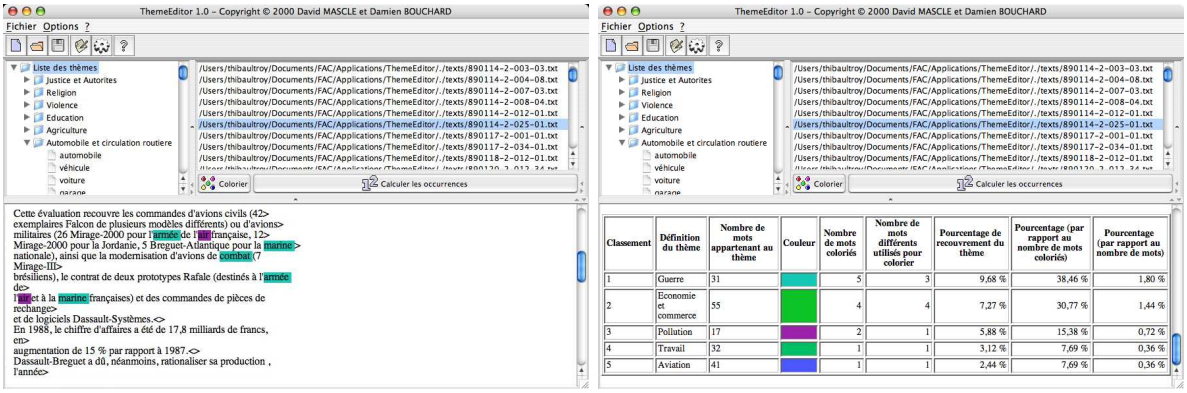


FIG. 3.6 – *ThemeEditor* : l'écran de gauche présente un texte de l'ensemble documentaire colorié à l'aide des domaines. L'écran de droite propose des statistiques sur les domaines présents dans le texte (nombre de lexies coloriées, taux de recouvrement du domaine, etc.).

Les 644 lexies extraites avec *Memlabor* de l'ensemble documentaire d'observation ont pu être regroupées en domaines et coloriées dans l'ensemble documentaire à l'aide de *ThemeEditor*. Plusieurs utilisations itératives de l'outil ont permis de mettre en évidence 18 domaines[108]. Certaines des lexies extraites avec *Memlabor* n'ont pas été positionnées dans les domaines, le coloriage ne s'étant, par exemple, pas révélé pertinent. D'autres lexies ont pu être placées dans plusieurs domaines (c'est le cas de la lexie « combat » que l'on a placée dans les domaines de la violence, de la guerre et du sport). Enfin, de nouvelles lexies ont pu apparaître du fait de leur proximité avec les lexies coloriées dans l'ensemble documentaire. Au total, 646 lexies constituent les 18 domaines mis en évidence.

Pour le français, *ThemeEditor* permet à l'utilisateur d'associer des formes fléchies aux lexies choisies, en utilisant la base de données lexicale *BDLex* [de Calmès et Pérennou, 1998][109]. Nous avons déjà abordé précédemment le fait que l'utilisation de telles bases lexicales s'opposait à notre approche alingue. Pourtant, dans certains cas et pour des langues pour lesquelles nous avons des ressources, il peut être intéressant de permettre à l'utilisateur d'obtenir automatiquement des formes fléchies des lexies de son intérêt. Dans de tels cas, les formes fléchies sont stockées dans le fichier XML produit par *ThemeEditor*.

3.2.4 *VisualLuciaBuilder* : construction interactive de dispositifs *LUCIA*

Après avoir extrait avec *Memlabor*, des lexies d'un ensemble documentaire et les avoir regroupées en domaines avec *ThemeEditor*, nous proposons à l'utilisateur de décrire chacun de ces domaines selon le modèle *LUCIA*.

[108] Ces domaines sont : la justice (24 lexies), la religion (29 lexies), la violence (21 lexies), l'éducation (29 lexies), l'agriculture (36 lexies), la circulation routière (46 lexies), l'aviation (41 lexies), la mer (40 lexies), le dopage (24 lexies), l'économie (56 lexies), la politique (57 lexies), l'espace (43 lexies), la guerre (31 lexies), l'informatique (48 lexies), la pollution (17 lexies), le sport (34 lexies), la télévision (37 lexies) et le travail (33 lexies).

[109] Une base de données *MySQL* a été utilisée pour stocker et consulter la base *BDLex*.

Pour cela, nous proposons l'outil interactif *VisualLuciaBuilder* permettant un grand nombre d'opérations de création et de révision de dispositifs[110]. À la suite de l'outil *LUCIABuilder* [Perlerin, 2004, pages 151 à 159] (figure 3.7) développé par Vincent Perlerin, nous avons développé l'outil *VisualLuciaBuilder* dans un objectif similaire : celui d'assister l'utilisateur dans la construction de dispositifs *LUCIA*. Les motivations de ce développement sont liées aux limites constatées sur l'outil original.

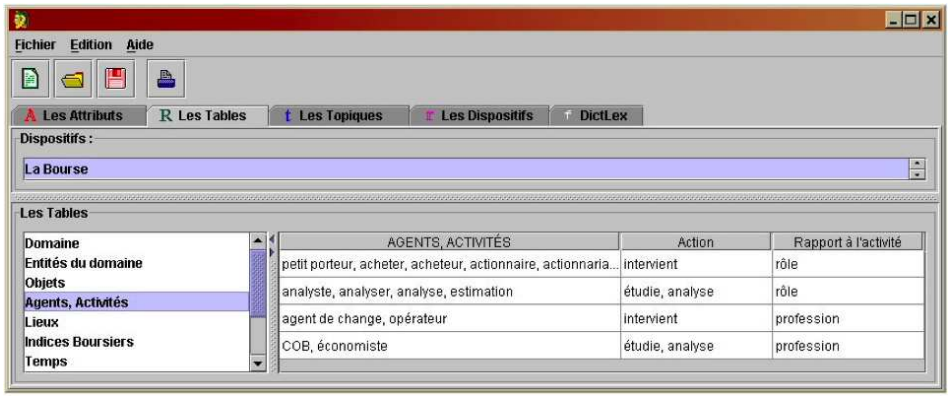


FIG. 3.7 – *LUCIABuilder* : Interface montrant la phase de remplissage d'une table d'un dispositif.

Tout d'abord, *LUCIABuilder* oblige l'utilisateur à une élaboration séquentielle de ses dispositifs, en définissant, tout d'abord les attributs et valeurs d'attributs, puis les tables, puis l'organisation des tables entre elles au sein du dispositif. De nombreux « aller-retours » sont pourtant nécessaires dans l'élaboration des dispositifs. Par exemple, il faut pouvoir revenir à tout moment sur les attributs et les valeurs définies, en ajoutant ou en supprimant un attribut et/ou une valeur. De telles modifications sont ensuite répercutées sur les tables possédant les attributs concernés, comme nous le verrons par la suite à travers un exemple. Hormis cette vue séquentielle de la phase de construction d'un dispositif, *LUCIABuilder* donne une vision assez locale sur les tables. Ainsi, il n'est pas possible pour l'utilisateur de visualiser à la fois le contenu de tables du dispositif et les liens entre les tables. Pourtant, une telle vue conjointe sur le contenu des tables et leurs liens serait particulièrement utile afin de mieux appréhender les descriptions déjà réalisées.

La réalisation de l'outil *VisualLuciaBuilder* a donc été principalement guidée par ces idées d'offrir à l'utilisateur à la fois plus d'interactivité et une vue globale durant la phase de construction des dispositifs. Le développement de ce logiciel a commencé dans le cadre d'un projet d'étudiants en Master première année d'informatique à l'Université de Caen durant l'année universitaire 2005-2006. Ce projet, mené par Kahina Hamadache et Matthieu Vernier, a conduit à l'élaboration d'une première version utilisable de l'outil [Hamadache et Vernier, 2006]. Après avoir poursuivi ce développement afin de corriger les dernières imperfections et d'intégrer quelques améliorations, nous avons pu stabiliser une version de l'outil *VisualLuciaBuilder*, outil présenté en figure 3.8.

[110] À noter, tout de même, que ce dernier peut être utilisé directement. L'utilisation des outils précédents permet d'amorcer très utilement la phase de construction de dispositifs *LUCIA*, cependant l'utilisation de *Memlabor* et de *ThemeEditor* n'est pas une obligation pour l'utilisateur, l'outil *VisualLuciaBuilder* pouvant s'utiliser de façon autonome (il en est de même pour tous les outils que nous proposons).

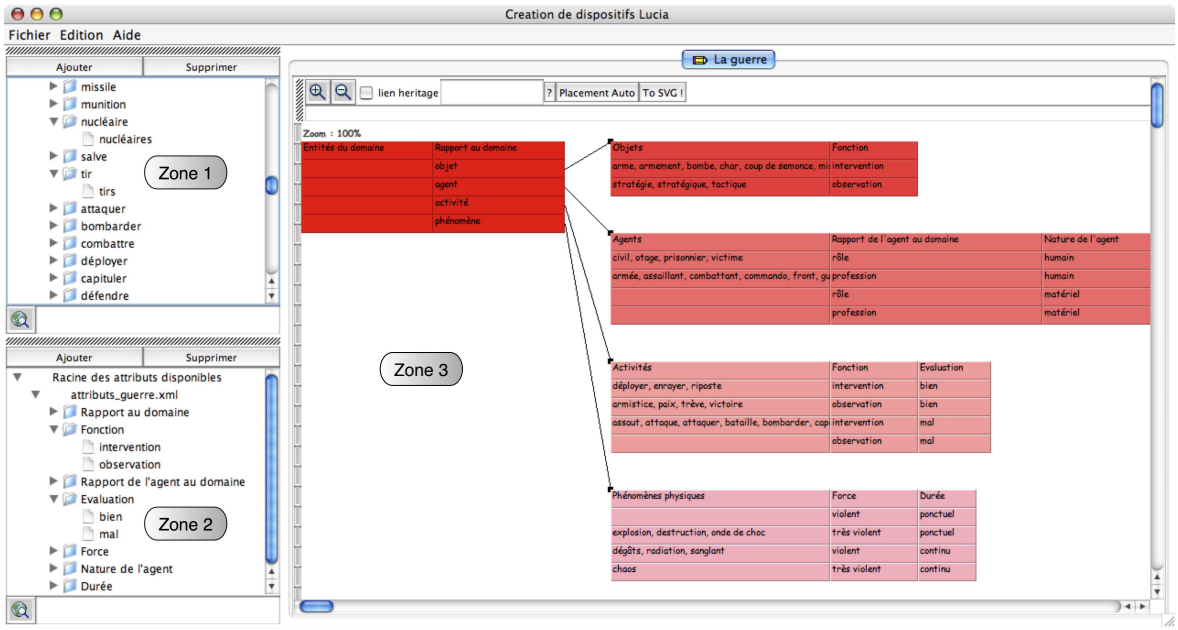


FIG. 3.8 – *VisualLuciaBuilder* : Interface illustrant les principales zones numérotées manuellement de 1 à 3. Le dispositif sélectionné représente le domaine de la guerre. Il décrit et met en relation les lexies d'une liste constituée à l'aide des outils *MemLabor* et *ThemeEditor* à partir de l'ensemble documentaire abordé précédemment.

L'interface de l'outil comporte des zones distinctes marquées de 1 à 3 sur la figure précédente.

La **zone 1** contient une ou plusieurs listes de lexies sélectionnées par l'utilisateur dans le cadre d'études précédentes. L'outil permet de charger aussi bien des fichiers XML de lexies produits par *Memlabor* que des fichiers XML de domaines produits par *ThemeEditor*, dans ce cas les catégories sont bien évidemment conservées dans l'affichage du lexique qui se fait sous forme d'arbre. L'utilisateur peut ainsi ajouter de nouvelles lexies, modifier et supprimer des lexies existantes. Également, il est possible de partir de zéro en saisissant les lexies à la main dans la zone dédiée. L'interface permet de saisir des formes fléchies de ces lexies pertinentes pour l'utilisateur. Si ces formes sont déjà présentes (par exemple, si l'utilisateur a utilisé la fonctionnalité de *ThemeEditor* associant automatiquement, à chaque lexie, ses formes fléchies associées présentes dans la base *BDLex*), ce dernier peut modifier et/ou compléter les flexions existantes. Pour chaque lexie de la zone, un clic droit avec la souris ouvre un menu contextuel offrant la possibilité d'aller voir la définition de la lexie proposée par *Wikipédia*[111], de mettre en contexte la lexie dans un ensemble documentaire *via FlexiSemContext* (voir partie suivante) et d'obtenir des synonymes avec le dictionnaire des synonymes de l'Université de Caen[112]. Le but recherché, en proposant à l'utilisateur ces différents liens, est de lui permettre de relier sa ressource à un espace hypertextuel (encyclopédie et dictionnaire de synonymes) et à un espace intertextuel (concordancier). En reliant la ressource de l'utilisateur à une encyclopédie et à un dictionnaire de synonymes, nous lui proposons en quelque sorte de reprendre à son compte certains éléménts de « consensus » proposés par les auteurs de l'encyclopédie et du dictionnaire. Le but visé est ainsi de permettre à l'utilisateur un éventuel complément de sa propre ressource avec de nouveaux éléments pertinents de son point de vue auxquels il n'aurait pas encore fait référence.

[111] `http://www.wikipedia.fr` (page consultée le 25 janvier 2007).
[112] `http://elsap1.unicaen.fr/cgi-bin/cherches.cgi` (page consultée le 25 janvier 2007).

Ces différentes fonctionnalités peuvent permettre d'enrichir les listes avec des lexies auxquelles l'utilisateur n'aurait pas pensé jusqu'à présent. Un moteur de recherche dans le lexique est également disponible, afin de rechercher si une lexie ou l'une de ses formes fléchies est bien présente dans les listes. Une fois les listes de lexies stabilisées, l'utilisateur peut les sauvegarder dans un fichier au format XML afin de permettre une exploitation future par l'outil. La figure 3.9 illustre différentes fonctionalités offertes par *VisualLuciaBuilder* lors de la phase de chargement du lexique.

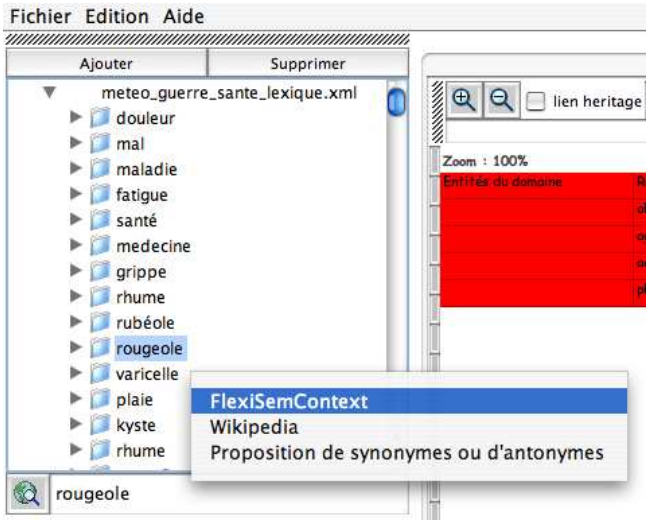


FIG. 3.9 – *VisualLuciaBuilder* : Recherche de la lexie « rougeole » dans les listes. Une fois la lexie trouvée, elle est automatiquement sélectionnée. Un clic droit sur la lexie permet d'obtenir différentes informations par l'intermédiaire de *Wikipédia*, d'un dictionnaire des synonymes ou encore d'un outil de mise en contexte dans l'ensemble documentaire.

La **zone 2** représente une ou plusieurs listes d'attributs et de valeurs d'attributs définis par l'utilisateur. Là encore, l'utilisateur peut soit charger et modifier une liste d'attributs construite lors d'une précédente utilisation de l'outil, ou bien partir d'une liste vierge et saisir de nouveaux attributs et de nouvelles valeurs. Tout comme pour la zone 1, un module de recherche est disponible afin de retrouver et de sélectionner directement un attribut ou une valeur d'attribut donné. Des valeurs d'attributs prenant la forme d'images ou d'icônes sont définissables. L'application *VisualLuciaBuilder* contient un ensemble d'images (répertoire « images » à la racine du répertoire contenant l'application) pré-sélectionnées. Il est bien sûr possible d'ajouter ses propres images. Ceci n'est pas qu'une fonctionnalité supplémentaire de l'interface mais cela constitue un premier pas vers une sémiotique différente du lexical, utilisée dans un même but : pouvoir exprimer des différences. La figure 3.10 illustre l'usage de telles images comme valeur d'un attribut. La sauvegarde dans un fichier au format XML du dictionnaire d'attributs est proposée à l'utilisateur quand ce dernier a stabilisé la liste des attributs à faire intervenir dans la description de ses domaines.

Enfin, la **zone 3** est une zone où l'utilisateur « dessine » son dispositif. Il crée de nouvelles tables par un simple clic droit dans la zone de dessin, puis nomme les tables comme il le souhaite. L'utilisateur peut alors glisser interactivement au sein des tables les attributs (zone 2) qu'il estime décrire les tables concernées. Un tel ajout entraîne la création dans la table d'autant de colonnes qu'il y a d'attributs, et d'autant de lignes qu'il y a de combinaisons des différentes valeurs des attributs choisis. Ensuite, l'utilisateur peut remplir les tables en glissant les lexies de la zone 1

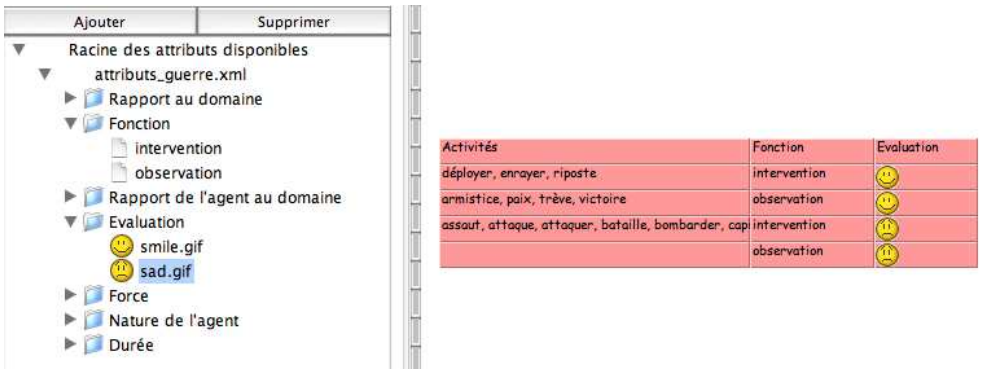


FIG. 3.10 – *VisualLuciaBuilder* : Des images (*smileys*) ont été utilisés pour définir l'attribut « évaluation ». De telles images se retrouvent ensuite dans les tables concernées dans le dispositif (dans l'exemple, il s'agit d'un dispositif sur le domaine de la guerre).

dans les cases des tables qu'il juge pertinentes. L'utilisateur dessine également les liens d'héritage entre tables en traçant tout simplement un trait entre la ligne de la table parente et le nom de la table fille. Les différentes tables sont déplaçables dans la zone de dessin, mais l'utilisateur peut utiliser la fonction de placement automatique pour disposer le mieux possible les tables dans la zone de dessin.

Une fonction de zoom est disponible dans la zone de dessin afin, par exemple, de mettre l'accent sur une table donnée ou, au contraire, pour garder une vue très globale sur le dispositif durant son élaboration. La création de plusieurs dispositifs en parallèle est également possible, un onglet de la zone de dessin est alors associé à chaque dispositif. L'application intègre un module de recherche dans la zone de dessin. Ce module permet de localiser une lexie dans un dispositif, mais aussi de faire ressortir les différents attributs et valeurs d'attributs portés par la lexie (cf. figure 3.11).

Une opération un peu plus complexe de modification d'un dispositif est la suppression d'un attribut (i.e. une colonne d'une table). Une telle suppression implique le regroupement des lexies séparées par l'une des valeurs de l'attribut (cf. 3.12 pour une illustration de cette opération). De la même manière, nous avons décidé que lors de la suppression de lexies et d'attributs dans leurs dictionnaires respectifs, ces suppressions se répercutent dans le dispositif : les lexies supprimées dans le lexique disparaissent des tables dans lesquelles elles étaient positionnées. Les attributs et valeurs d'attributs retirés du dictionnaire d'attributs sont supprimés des tables de la même manière que pour le retrait de lexies.

Dès que l'utilisateur considère ses dispositifs stabilisés, du moins momentanément, leur enregistrement sous forme de fichiers XML est réalisé. Un fichier XML est alors associé à chaque dispositif. Ce fichier reprend les identifiants des lexies et des attributs présents dans les fichiers XML décrivant ces listes respectives. L'enregistrement d'un ensemble de dispositifs, appelé une session, est également possible. Un fichier XML est alors utilisé pour décrire ces éléments. Plus concrètement, ce fichier contient des pointeurs vers les différents fichiers XML des dispositifs, des dictionnaires d'attributs et des lexiques utilisés pour décrire les domaines de l'utilisateur (des extraits de tels fichiers XML sont présentés en annexe A de cette thèse).

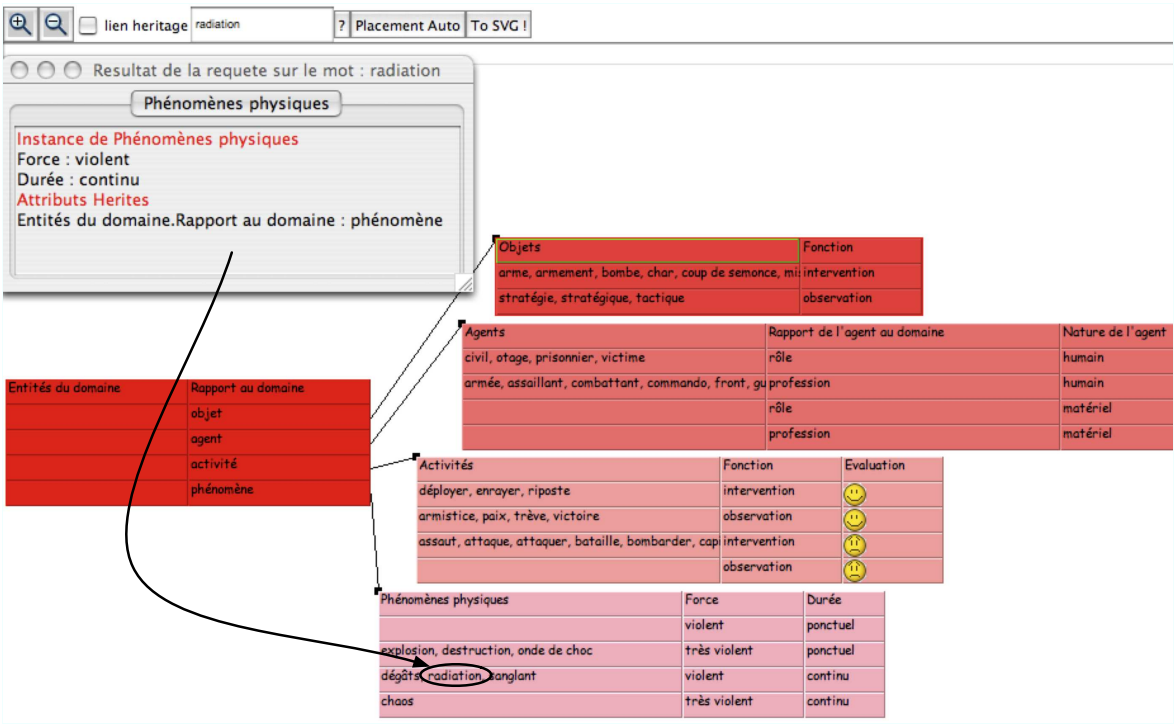


FIG. 3.11 – *VisualLuciaBuilder* : Résultat de la recherche de la lexie *radiation* dans le dispositif représentant le domaine de la guerre. La lexie a été entourée et reliée à ses attributs manuellement sur la figure.

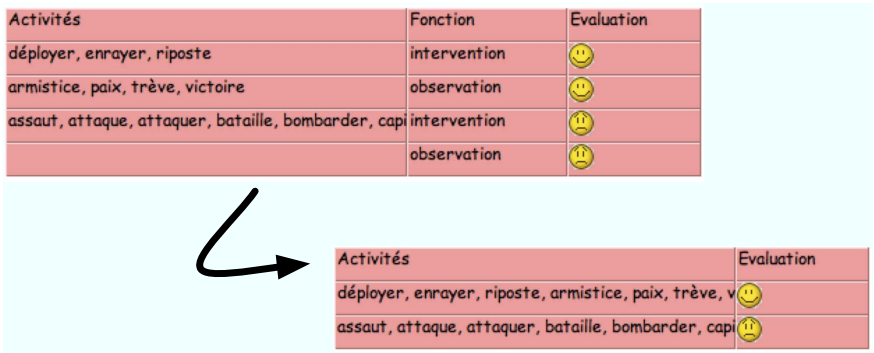


FIG. 3.12 – *VisualLuciaBuilder* : Suppression de l'attribut *fonction* de la table des activités du domaine de la guerre.

Des sorties SVG des dispositifs sont également retournées par *VisualLuciaBuilder*, de telles sorties permettent de visualiser les dispositifs de façon interactive en dehors d'une utilisation du logiciel, par exemple, dans une page d'un site Internet[113].

Dans notre exemple d'utilisation sur un ensemble documentaire constitué d'articles de presse, nous avions obtenu à l'issue de l'utilisation de *ThemeEditor*, 18 domaines, chacun étant décrit en moyenne par une trentaine de lexies. Ce premier regroupement en domaines constitue, non seulement un préliminaire particulièrement intéressant à l'élaboration de dispositifs *LUCIA*, mais permet aussi d'obtenir des ensembles de lexies directement projetables dans l'ensemble documentaire par la plate-forme *ProxiDocs*, comme nous le verrons par la suite. À partir de 4 de ces 18 domaines, nous avons élaboré des dispositifs *LUCIA* qui pourront, par exemple, nous servir à une étude plus poussée du contenu de l'ensemble documentaire par rapport aux domaines ainsi représentés (ces quatre domaines sont les suivants : la télévision, la guerre, l'espace et la justice). Ces dispositifs pourront également être projetés dans l'ensemble documentaire par *ProxiDocs*, les projections donnant, dans ce cas, plus d'informations à l'utilisateur.

3.2.5 *FlexiSemContext* : outil pour la mise en contexte de lexies et de sèmes

Les trois outils présentés précédemment apportent une aide à l'utilisateur pour extraire et structurer des lexies en domaines. Pour apporter à l'utilisateur une aide supplémentaire dans la sélection de lexies pertinentes, mais aussi un retour pertinent sur les contextes d'utilisation des lexies choisies, ce dernier peut utiliser l'outil *FlexiSemContext*. Cet outil permet l'observation des contextes d'apparition de lexies dans un ensemble documentaire, ainsi que des contextes d'actualisation d'attributs et de valeurs d'attributs attribués à ces lexies.

Ce besoin de retour aux ensembles documentaires dans une phase de construction de RTO personnelles est apparu lors de l'atelier formation du CNRS « Variation, construction et instrumentation du sens » où une première expérimentation du modèle *LUCIA* a été effectuée (se reporter à [Perlerin et Beust, 2003] et au chapitre 4 de cette thèse pour plus de détails sur cette expérimentation). Durant cet atelier, il était envisagé de 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) à partir d'une liste de lexies données extraites d'un corpus d'article de *Le Monde sur CD-ROM*. À l'issue de cette expérience, un manque de retour au corpus a été ressenti comme un fort handicap par les participants, principalement à cause de l'impossibilité de revenir sur un texte faisant intervenir les lexies proposées.

C'est tout particulièrement pour pallier une telle absence de retour dans les ensembles documentaires que nous avons développé l'outil *FlexiSemContext*. À partir d'une entité lexicale donnée, l'utilisateur peut visualiser, à l'aide de l'outil, les contextes d'apparition de cette entité et de ses formes fléchies (s'il le désire, pour cela nous intégrons encore une fois la base de données lexicale *BDLex*) au sein des textes de l'ensemble. Également, à partir d'un attribut ou d'un couple *attribut : valeur*, ce dernier peut visualiser les contextes d'actualisation de ces éléments en ensembles documentaires. L'utilisateur peut alors juger la pertinence d'une lexie, d'un attribut, d'une valeur d'attribut, dans le cadre de sa tâche et les faire ou non intervenir dans les représentations lexicales des domaines de son choix.

[113] Un grand nombre de dispositifs au format SVG sont accessibles à l'adresse suivante : `http://www.info.unicaen.fr/~troy/dispositifs` (page consultée le 20 juillet 2007).

Cet outil est disponible sur Internet à l'adresse suivante : `http://www.info.unicaen.fr/~troy/flexisemcontext`. À partir d'ensembles documentaires chargés sur le serveur, il est alors possible de mettre en contexte dans cet ensemble une lexie choisie par l'utilisateur *via* l'interface de l'outil. Une copie de l'ensemble documentaire doit ainsi être présente sur le serveur[114].

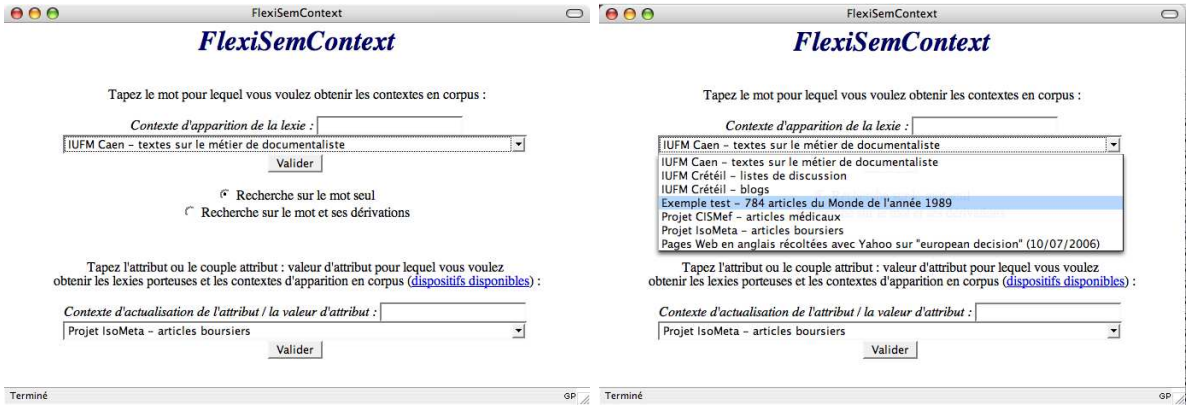


FIG. 3.13 – *FlexiSemContext* : Interface d'interrogation de l'application.

La figure 3.13 illustre l'interface d'accueil de l'application. Cette interface invite, dans sa partie supérieure, à saisir une lexie à mettre en contexte dans un ensemble documentaire choisi parmi la liste des ensembles documentaires présents sur le serveur, cette liste est mise en évidence dans l'écran de droite. L'utilisateur a le choix de réaliser son interrogation sur la lexie seule ou sur la lexie et ses fléxions associées dans la base lexicale *BDLex*. La partie inférieure propose les mêmes traitements mais à partir d'un attribut, d'une valeur d'attribut ou d'un couple *attribut : valeur*. Ainsi, il est possible de mettre en évidence les contextes d'actualisation de ces éléments dans l'ensemble documentaire choisi (cet ensemble documentaire doit avoir été au préalable étiqueté selon les dispositifs de l'utilisateur, cet étiquetage est réalisé par la plate-forme *ProxiDocs* que nous verrons dans la partie suivante).

Une fois une lexie et un ensemble documentaire choisis par l'utilisateur, l'application retourne l'écran présenté en figure 3.14. La lexie recherchée, et éventuellement ses flexions, sont alors mises en évidence dans les textes de l'ensemble choisi. Les textes sont classés dans l'ordre décroissant du nombre d'occurrences de l'élément demandé. Les extraits des textes laissent apparaître une fenêtre d'un certain nombre de caractères autour de la lexie demandée. Ces contextes d'apparition permettent alors d'obtenir des informations sur le ou les emplois de la lexie dans l'ensemble documentaire. Pour chaque extrait, il est également possible de revenir sur le texte original dans son intégralité à l'aide d'un lien hypertexte.

[114] La mise sur le serveur d'un ensemble documentaire pour son interrogation par *FlexiSemContext* se fait sur une simple demande à `thibault.roy@info.unicaen.fr`, de même pour obtenir une version du logiciel. Dans le cadre d'une consultation de l'ensemble documentaire accessible par tous *via* Internet, l'ensemble documentaire doit être au préalable libre de droit.

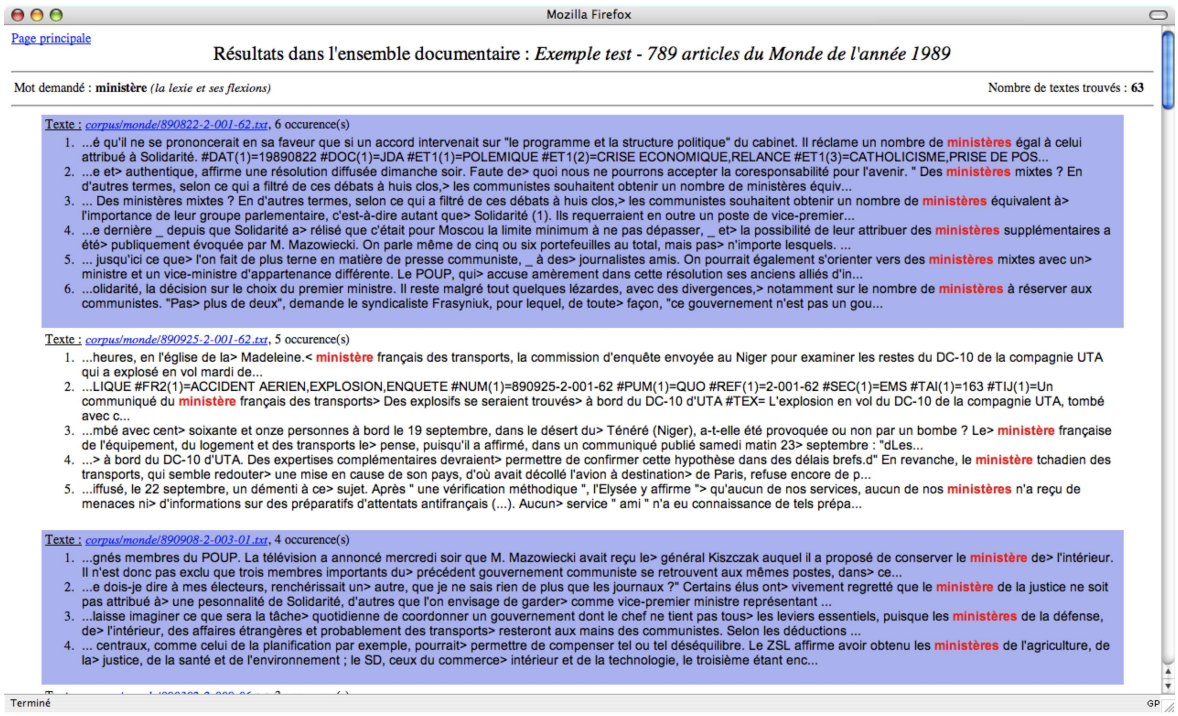


FIG. 3.14 – *FlexiSemContext* : Mise en contexte de la lexie « ministère » dans un ensemble documentaire.

L'outil *FlexiSemContext* fait partie des outils d'analyse d'ensembles documentaires appelés des concordanciers. De tels outils, couramment utilisés en linguistique de corpus, permettent d'observer les contextes d'apparition d'entités linguistiques. Plus formellement, ces outils permettent d'isoler les concordances d'une expression linguistique donnée, appelée le pivot, dans un contexte d'une taille donnée. Dans [Pincemin *et al.*, 2006], les auteurs définissent une concordance de la façon suivante :

> *Un corpus étant fixé, une concordance est la liste de toutes les occurrences d'un pivot, alignées verticalement en colonne (nous dirons « empilées »), entourées de part et d'autre par leur contexte, et triées selon un critère pertinent pour l'analyse.*

Selon eux, un concordancier doit permettre un alignement strict de l'expression pivot entre les extraits du corpus dans lesquels elle apparaît, ainsi que des opérations de tri et de sélection sur les contextes gauche et droit de l'expression pivot.

Toute comme *FlexiSemContext*, différents outils de mise en contexte de termes sont accessibles sur Internet pour l'étude de différents ensembles documentaires. Par exemple, Jean Véronis propose un outil pour naviguer dans la constitution européenne : `http://aixtal.blogspot.com/2005/04/texte-naviguez-dans-la-constitution.html` (page consultée le 21 février 2007). Également, Tom Cobb de l'Université du Québec à Montréal propose sur le site `http://132.208.224.131/Francord.htm` (page consultée le 21 février 2007) un concordancier en ligne permettant de travailler sur différents ensembles documentaires (articles du journal Le Monde, textes de Guy de Maupassant, etc.).

D'autres outils beaucoup plus évolués (mais non accessibles en ligne) permettent de réaliser des analyses de concordances plus poussées. Nous pouvons, par exemple, citer les outils

Contextes[115], *WConcord*[116], *WordSmith*[117] et *XSARA*[118], ou encore ceux inclus dans les plates-formes de TAL *Unitex*[119], *Intex*[120] et *Nooj*[121]. De tels outils proposent, par exemple, des possibilités de définition d'expressions pivot par des expressions régulières ; des fonctions de tri et de sélection évolués sur les contextes gauches et droits ; des traitements plus efficaces, intégrant une indexation pour plus de rapidité de traitement (très utiles pour les très grands ensembles documentaires) ; etc.

Notre outil *FlexiSemContext* intègre une possibilité originale de mise en contexte : la mise en contexte de lexies servant de supports à un attribut ou un couple *attribut : valeur*. La figure 3.15 illustre une telle mise en contexte pour le couple *évaluation : mal* dans un corpus d'articles boursiers. Les articles du corpus sont présents sur le serveur, ils ont été préalablement étiquetés à l'aide de la plate-forme *ProxiDocs* (décrite dans la section suivante) avec trois dispositifs définis par des experts dans le cadre d'une expérience précise (l'expérience réalisée dans le cadre du projet *IsoMeta* décrite au chapitre suivant). Les lexies porteuses de cet élément de signification sont précisées en tête du document. Ce sont ces lexies qui sont alors mises en contexte dans l'ensemble documentaire.

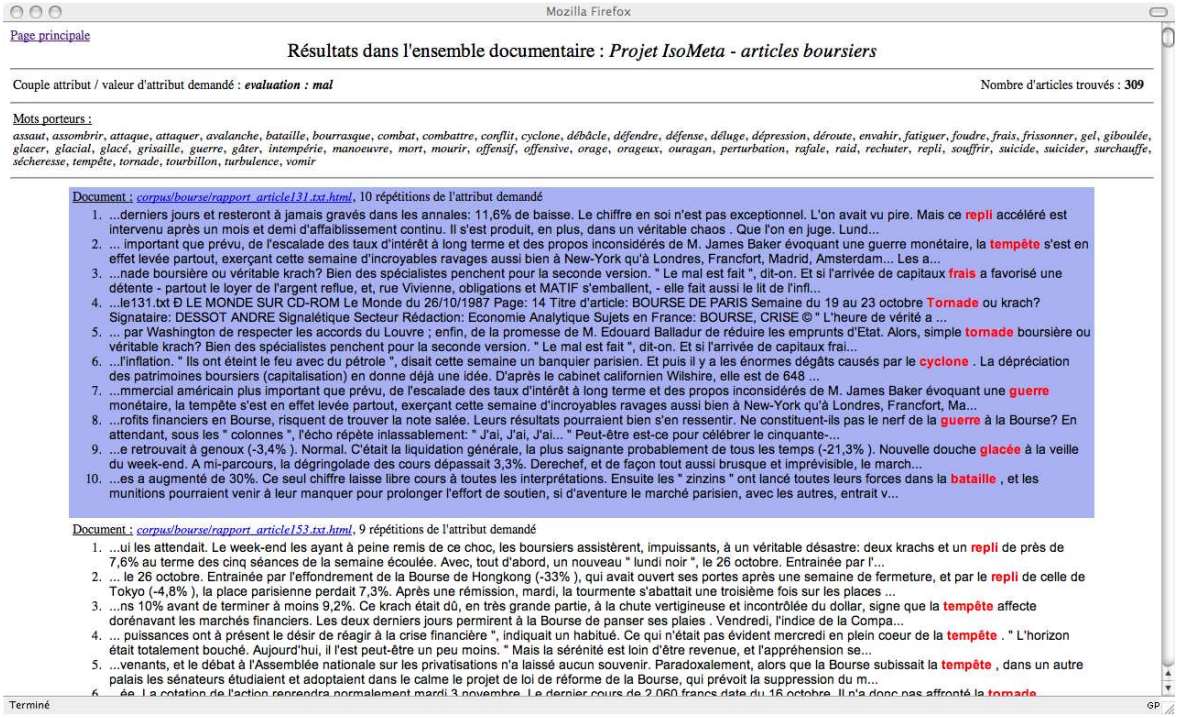


FIG. 3.15 – *FlexiSemContext* : Mise en contexte d'un couple *attribut : valeur*.

Contrairement aux logiciels *Memlabor*, *ThemeEditor* et *VisualLuciaBuilder* présentés précédemment qui sont développés en Java, *FlexiSemContext* est développé en PHP qui est un langage dédié aux application Internet et à l'interfaçage avec des bases de données. Ce choix est lié à

[115] `http://www.up.univ-mrs.fr/~veronis/logiciels/Contextes` (page consultée le 21 février 2007).
[116] `http://www1.ujaen.es/~aalcaraz/HEL/wconcord.zip` (page consultée le 21 février 2007).
[117] `http://www.lexically.net/wordsmith/index.html` (page consultée le 21 février 2007).
[118] `http://users.ox.ac.uk/~lou/papers/xsara.xml` (page consultée le 21 février 2007).
[119] `http://www-igm.univ-mlv.fr/~unitex` (page consultée le 21 juillet 2007).
[120] `http://intex.univ-fcomte.fr` (page consultée le 25 septembre 2007).
[121] `http://www.nooj4nlp.net` (page consultée le 25 septembre 2007).

la facilité de consultation que nous voulions donner au logiciel : pas besoin d'installation, un simple navigateur Internet suffit. De plus, il est très facile, comme nous le verrons dans la section suivante dédiée à la projection de RTO dans des ensembles documentaires, d'inclure des liens vers *FlexiSemContext* par de simples liens hypertexte.

3.3 *ProxiDocs* : projections de RTO personnelles dans des ensembles documentaires

3.3.1 Présentation et objectifs

Les outils présentés précédemment apportent une aide à l'utilisateur dans une tâche de construction de RTO *LUCIA*. Ils permettent à l'utilisateur d'extraire, d'observer, de mettre en contexte et de décrire des lexies pouvant lui permettre d'exprimer son point de vue sur des domaines de son intérêt. Une fois que l'utilisateur a exprimé un tel point de vue par l'intermédiaire de RTO dans un format simple (ensembles de lexies) ou avancé (dispositifs *LUCIA*), le principal intérêt pour ce dernier réside dans l'utilisation et la projection de ses RTO personnelles dans des ensembles documentaires.

Afin de proposer à l'utilisateur différents moyens pour projeter ses RTO dans des ensembles documentaires, nous avons développé la plate-forme logicielle *ProxiDocs*. Nous développons cette application depuis 2003 au sein de l'équipe ISLanD du laboratoire GREYC de l'Université de Caen. Le langage de programmation Java a été utilisé, les sorties de l'application utilisent les formats HTML, SVG et XML. Nous qualifions *ProxiDocs* d'*instrument* car elle construit des représentations transformées de données langagières (des RTO personnelles et des ensembles documentaires). Nous la qualifions également de *plate-forme* car l'utilisateur a la possibilité de réaliser et de combiner différents types d'analyses (différents types de projection, de classification, de visualisation, etc.).

La fonctionnalité première de *ProxiDocs* est de permettre à l'utilisateur de visualiser interactivement la répartition de ses RTO personnelles dans des ensembles documentaires. Une telle appréhension globale et interactive de l'ensemble documentaire, prenant en considération les domaines d'intérêt de l'utilisateur, est, de notre point de vue, un moyen très efficace pour appréhender le contenu de l'ensemble analysé. La plate-forme permet de construire différentes cartes, différents rapports d'analyse, des ensembles documentaires selon les ressources de l'utilisateur, ce qui est illustré en figure 3.16.

L'interaction entre les sorties de la plate-forme *ProxiDocs* et l'utilisateur est très forte. Notre objectif n'est pas de retourner une vue statique et terminale sur un ensemble documentaire à partir de RTO personnelles, mais de fournir aux utilisateurs des supports d'interactions sur leurs ensembles documentaires leur permettant de visualiser la répartition de RTO. Cette visualisation interactive permet d'avoir un retour, aussi bien sur l'ensemble documentaire analysé, que sur ses RTO, et ainsi de poursuivre l'enrichissement et l'évolution des ressources, continuant le cycle d'interaction des logiciels instrumentant le modèle *AIdED*, comme l'illustre la figure 3.17.

Comme nous le verrons dans le chapitre suivant, ces visualisations interactives permettent de réaliser différentes tâches, portant sur des domaines différents comme la recherche documentaire, l'analyse linguistique de corpus ou encore l'analyse de forums de discussion. Avant d'aborder la valeur ajoutée de notre plate-forme pour des tâches d'accès au contenu d'ensembles documentaires, nous présentons, dans les sous-sections suivantes, les différents traitements mis en œuvre dans *ProxiDocs*.

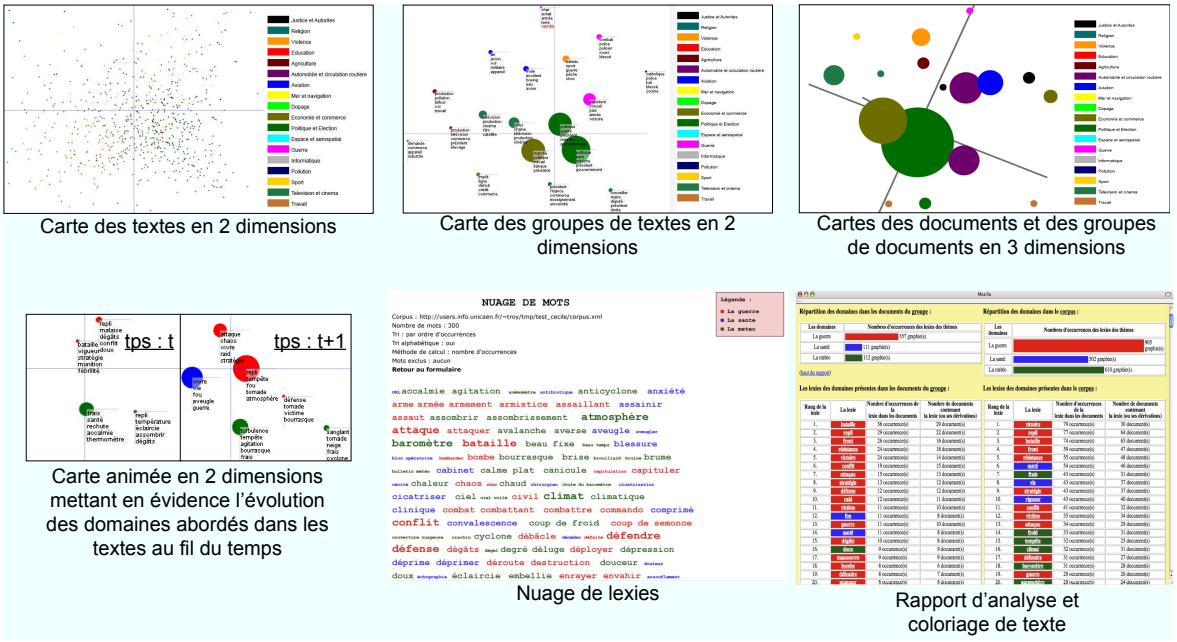


FIG. 3.16 – Illustration des principales sorties de la plate-forme *ProxiDocs*.

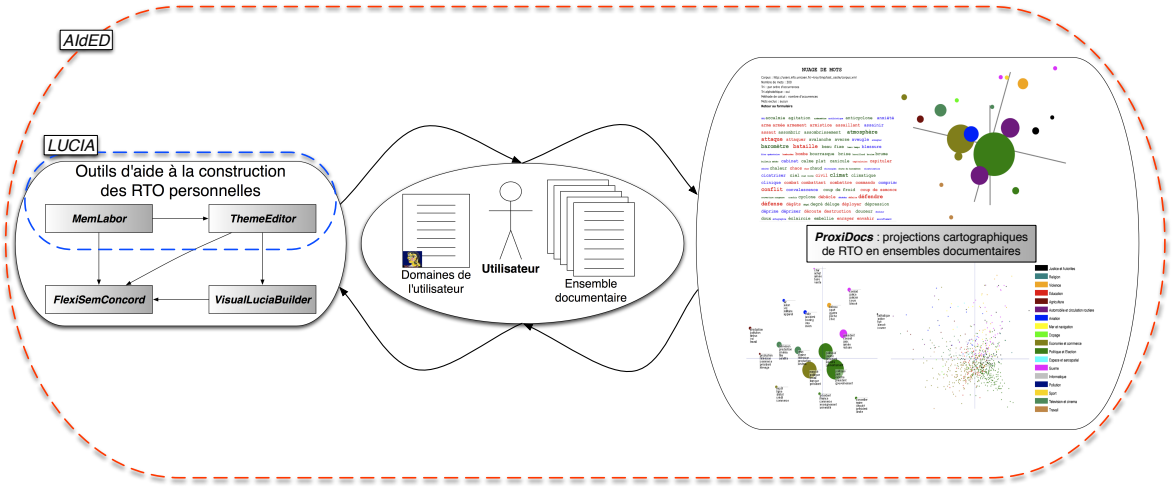


FIG. 3.17 – Cycle d'interactions entre les outils logiciels d'aide à la construction de RTO personnelles et *ProxiDocs*. Ces différents logiciels constituent l'instrumentation du modèle *AIdED*.

3.3.2 Traitements mis en œuvre

Les traitements mis en œuvre dans *ProxiDocs* fonctionnent d'une manière chaînée, c'est-à-dire que chaque entité va prendre en entrée des données d'une entité précédente, effectuer des calculs sur ces données, puis fournir les résultats obtenus après ces calculs à l'entité suivante.

Ce chaînage facilite la maintenance et l'évolution de la plate-forme. Les échanges entre composants de la chaîne de traitements sont soit des données textuelles respectant un formalisme XML, soit des objets du langage de programmation passés directement en paramètres. L'utilisation d'une structure de chaîne de traitements permet assez facilement la modification, l'ajout ou la suppression d'un composant, de telles opérations n'affectant que le composant impliqué. Il arrive cependant que des « ajustements » des entrées/sorties des autres composants soient nécessaires lors de changements un peu trop lourds impliquant de nouveaux traitements en amont et en aval.

Les traitements réalisés par *ProxiDocs* sont réalisés de manière chaînée, cependant, et comme nous l'avons déjà évoqué dans la partie précédente, les sorties produites à l'issue de ces traitements ne constituent pas des « finalités » mais sont des supports de visualisation et d'interaction sur un ensemble documentaire, supports faisant émerger de nouvelles informations, de nouveaux besoins, et entraînant de nouvelles utilisations de la plate-forme. Les différentes tâches de la chaîne de traitements de notre application interviennent dans l'ordre suivant :

1. **La prise en compte des RTO personnelles dans l'ensemble documentaire.** Cette étape consiste à compter les RTO personnelles de l'utilisateur dans l'ensemble de textes. Pour cela, nous verrons dans la sous-section suivante les différents comptages que nous réalisons au niveau des lexies et des couples *attribut : valeur* ;
2. **La projection des textes de l'ensemble documentaire.** Une fois réalisée l'étape de prise en considération des RTO personnelles de l'utilisateur dans l'ensemble documentaire, une structure numérique est associée à l'ensemble documentaire. Cette structure est ensuite projetée sur un espace à 2 ou 3 dimensions afin d'en permettre des visualisations ;
3. **La classification de l'ensemble documentaire.** L'étape suivante consiste à proposer des regroupements automatiques entre textes de l'ensemble documentaire. Pour cela, nous présentons par la suite les différentes méthodes numériques utilisées ;
4. La dernière étape consiste en **la construction des supports de visualisation en 2 ou 3 dimensions de l'ensemble documentaire**. Ces supports sont interactifs et proposent des accès à différents rapports d'analyse et visualisations.

3.3.3 Méthodes de représentation d'ensembles documentaires à partir de RTO personnelles

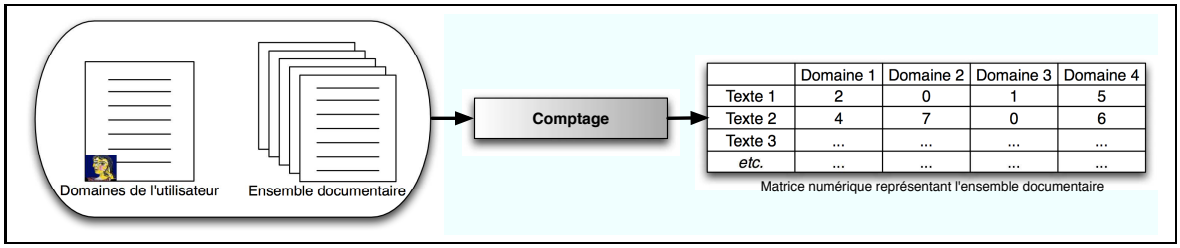


FIG. 3.18 – Étape de comptage des domaines de l'utilisateur dans l'ensemble documentaire.

Comme nous l'avons énoncé précédemment, la première étape réalisée par *ProxiDocs* consiste à projeter dans l'ensemble documentaire les RTO personnelles de l'utilisateur. Ces RTO prennent la forme de fichiers XML dans des formats que nous proposons selon le niveau de description souhaité par l'utilisateur[122]. Ces informations sont stockées dans une table de hachage, où la clé est le nom du domaine, et les valeurs sont les lexies décrivant ce domaine pour l'utilisateur. Dans le cas d'une représentation des domaines par des dispositifs *LUCIA*, chaque lexie se voit qualifiée par les tables dans lesquelles elle apparaît, ainsi que par les attributs et valeurs d'attributs qu'elle possède ou dont elle hérite[123].

Le tableau suivant définit quatre domaines avec quelques lexies associées, ces différents domaines nous serviront dans la présentation des méthodes de comptage.

Éducation	école	professeur	cours	diplôme
Agriculture	cultiver	terre	élevage	vache
Automobile	véhicule	garagiste	Fiat	panne
Politique	président	ministère	élections	chancelier

Après avoir parcouru le ou les fichiers XML décrivant les domaines de l'utilisateur, *ProxiDocs* va procéder au comptage des lexies de chaque domaine dans chaque texte de l'ensemble.

La tâche suivante consiste à compter, pour chaque texte de l'ensemble analysé, le nombre d'occurrences des lexies de chaque domaine qu'il contient. Une telle tâche revient à mettre en évidence le nombre de récurrences de sèmes macro-génériques marquant l'appartenance des lexies aux domaines. Prenons, par exemple, l'extrait **texte_1** suivant :

> *En transformant le* ***ministère*** *de l'agriculture en un* ***ministère*** *de la protection des consommateurs, le* ***chancelier*** *Gerhard Schroeder a renversé la vapeur, alors que le pays sombrait dans l'hystérie depuis la découverte du premier cas de vache folle en novembre 2000.* (Le Monde, 29 janvier 2001 - Arnaud Leparmentier)

Les mots soulignés (pour le domaine de l'agriculture) et en gras (pour le domaine de la politique) dans le texte ci-dessus sont des mots présents dans le tableau des domaines décrit précédemment. On aura donc les occurrences des mots des domaines de ce tableau pour **doc_1** :

	Éducation	Agriculture	Automobile	**Politique**
texte_1	0	2	0	3

Ce traitement doit être réalisé pour chaque document de l'ensemble documentaire, afin de construire un tableau où chaque ligne correspond aux nombres d'occurrences des domaines pour un texte de l'ensemble. Par exemple, le tableau suivant pourra alors être retourné :

	Éducation	**Agriculture**	**Automobile**	**Politique**
doc_1	0	2	0	3
doc_2	1	0	3	0
doc_3	0	0	2	1
...	...	...	...	...

Nous qualifierons cette méthode de comptage d'« absolue », par opposition à la méthode suivante dite « relative ». Cette dernière réalise les mêmes opérations que la méthode absolue, mais au lieu d'attribuer aux textes le nombre de lexies de chaque domaine qu'ils contiennent,

[122] Formats présentés en annexe A en figure A.1 et A.2.

[123] Une table de hachage est également utilisée avec pour clés, les différentes lexies, et pour valeurs, les descriptions de ces lexies en termes de tables où elles figurent et de couples *attribut : valeur* qu'elles possèdent.

nous faisons intervenir la taille de ces documents, en divisant le nombre trouvé par la méthode absolue par le nombre de graphies que chaque document contient. La méthode relative permet de ne pas attribuer le même poids à un document de 100 tokens possédant 10 lexies du domaine « politique » qu'à un document de 1000 tokens possédant également 10 lexies du même domaine, contrairement à la méthode absolue. Cette méthode de comptage est particulièrement utile dans les cas où la taille des documents de l'ensemble documentaire varie énormément.

Une fonctionnalité assez simple, laissée à l'utilisateur, est le filtrage des textes de l'ensemble documentaire contenant un nombre d'occurrences de lexies des domaines inférieur ou égal à une valeur donnée (appelons cette valeur $k_{restriction}$). Ainsi, si l'utilisateur choisit d'utiliser cette fonctionnalité et de prendre une valeur $k_{restriction}$ égale à 2, les textes contenant un nombre d'occurrences d'éléments des domaines inférieur ou égal à 2 (tout domaine confondu) ne seront pas pris en considération dans le comptage et donc dans les étapes suivantes.

Un autre facteur sur lequel nous nous sommes penché est celui de la prise en considération de la taille des ressources dans la phase de représentation de l'ensemble documentaire. Nous sommes partis du constat que certains domaines d'une même session pouvaient être décrits avec beaucoup plus de lexies que d'autres. Une réflexion s'est alors portée sur la prise en considération de la taille des domaines dans le comptage, par exemple, en minorant les domaines décrits avec beaucoup de lexies (ayant donc plus de « chance » d'apparaître dans un texte) et au contraire en majorant les domaines décrits avec peu de lexies. Le problème est que si l'utilisateur a choisi de décrire plus finement un domaine (par exemple, avec une plus grande quantité de lexies ou avec plus de couples *attribut : valeur*), c'est qu'il a choisi de mettre l'accent sur ce dernier. Il nous paraît alors que diminuer le poids de domaines plus décrits (en quantité) ou mieux décrits (en qualité) irait à l'encontre du choix de l'utilisateur et des principes du modèle lui laissant une place centrale et une liberté d'expression de son point de vue.

Le choix et la configuration de la méthode de comptage sont bien évidemment laissés à l'utilisateur. Lors de nos différentes expérimentations, nous avons pu toutefois remarquer que la méthode de comptage absolue était plus adaptée à l'analyse d'ensembles documentaires où les textes étaient d'une taille équivalente et que dans le cas contraire, la méthode relative pouvait être utilisée. Le filtrage sur les textes contenant un petit nombre de lexies des ressources permet d'éviter, en quelque sorte, le « bruit » causé dans les résultats d'analyse par un trop grand nombre de textes contenant peu ou pas d'éléments des ressources de l'utilisateur. Les méthodes de comptage présentées ici attribuent un espace numérique, représenté par une matrice, à un ensemble documentaire. À partir de ces matrices, nous réalisons différents traitements afin d'en proposer des visualisations. C'est ce que nous développons dans la sous-section suivante.

3.3.4 Projection et classification de l'ensemble documentaire

Projection de l'espace numérique représentant l'ensemble documentaire

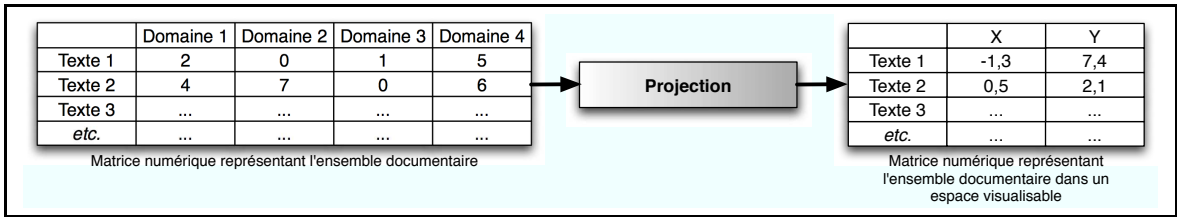


FIG. 3.19 – Étape de projection de la matrice numérique de grande dimension représentant l'ensemble documentaire vers un espace visualisable

Les espaces représentés dans la partie précédente possèdent un nombre de dimensions élevé, correspondant au nombre de domaines de l'utilisateur. Pour visualiser de tels espaces numériques, de telles matrices, et donc avoir un regard global sur l'espace documentaire considéré, nous avons choisi de les projeter sur un plan ou dans un espace en 3 dimensions. Pour cela, des méthodes issues des statistiques et de l'analyse des données ont été utilisées et développées. Ces méthodes permettent de « réduire » un espace numérique de grande dimension en un espace numérique de plus faible dimension.

Ainsi, des méthodes ayant déjà fait leur preuve dans le domaine de l'analyse des données et des statistiques, telles la méthode de Sammon [Sammon, 1969], la méthode de l'Analyse en Composantes Principales (ACP) [Bouroche et Saporta, 1980] et la méthode de l'Analyse Factorielle des Correspondances (AFC) [Benzécri, 1980], ont été utilisées. Des méthodes de projection que nous avons développées, plus simples, ont également été utilisées afin de proposer aux utilisateurs un large choix pour leurs analyses, d'une complexité et d'une finesse variable selon leurs objectifs.

Dans l'annexe B de cette thèse, nous proposons des descriptions des différentes méthodes de projection énoncées précédemment, nous expliquons leur implémentation *via* différents algorithmes, et nous comparons les visualisations obtenues par chaque méthode sur un exemple.

Classification de l'espace documentaire visualisé

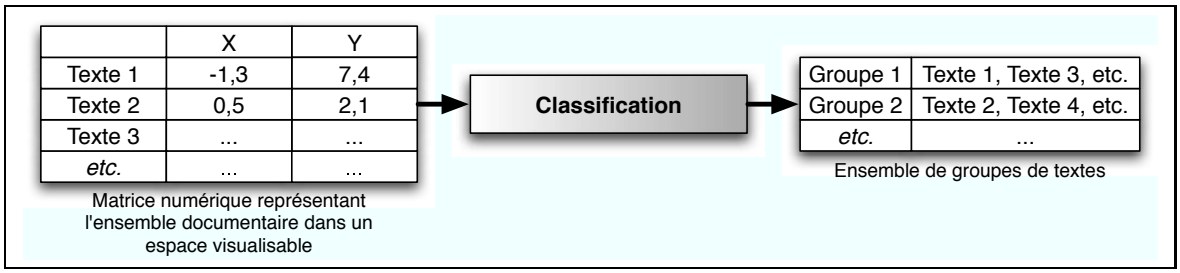


FIG. 3.20 – Étape de classification de l'ensemble documentaire.

Après application de l'une des méthodes de projection précédentes, chaque texte est représenté par un point contenu dans un espace à 2 ou 3 dimensions. Une carte de l'ensemble documentaire représentant les textes par un nuage de points peut être construite. Si l'ensemble étudié contient un grand nombre de textes (par exemple, plusieurs centaines) alors le nuage de points les représentant peut être très dense et/ou très étendu. Dans le but de faciliter la visualisation d'un tel ensemble, nous avons choisi d'intégrer des méthodes de classification afin d'assister l'utilisateur dans l'analyse des cartes.

Les méthodes choisies sont des méthodes de classification d'ensembles d'éléments, elles ont donc pour objectif de proposer des regroupements entre éléments « proches » de l'ensemble de départ. Ces méthodes (appelées méthodes non-supervisées) sont à opposer aux méthodes supervisées de classification. Elles partent d'un ensemble d'éléments et d'un ensemble de classes (c'est-à-dire de regroupements potentiels étiquetés par l'utilisateur et initialement vides) et cherchent à « remplir » ces classes, à positionner les éléments dans les groupes prévus initialement.

Les méthodes de classification implantées dans *ProxiDocs* permettent donc de regrouper des textes proches sur la projection en 2 ou 3 dimensions. Le choix de regrouper les textes après projection est lié à la visualisation des groupes que nous voulons proposer à l'utilisateur. Une classification réalisée avant projection puis visualisée après projection sur une carte à 2 ou 3 dimensions pourrait révéler des regroupements surprenants pour les utilisateurs, avec par exemple des éléments de mêmes groupes distants sur la carte. Cependant, il serait tout à fait concevable dans les prochaines versions de la plate-forme *ProxiDocs* de laisser à l'utilisateur la telle possibilité de choisir le « moment » qu'il estime idéal pour la phase de classification.

Ainsi, il est mis en évidence des groupes de textes partageant de mêmes récurrences de sèmes macro-génériques marquant l'appartenance aux domaines, et partageant ainsi de mêmes isotopies macro-génériques. Deux méthodes de classification ont été implantées dans la plate-forme *ProxiDocs* : la méthode de la classification hiérarchique ascendante (CHA) [Bouroche et Saporta, 1980] et la méthode des K-Means [MacQueen, 1967]. L'utilisateur peut alors soit choisir un nombre de groupes à mettre en évidence, soit ne rien préciser et dans ce cas, uniquement réalisable avec la CHA, un nombre de groupes jugé « optimal » est déterminé de façon automatique (cf. section B.3.3). Ces méthodes proposent des regroupements exclusifs des textes, un texte n'appartient qu'à un seul groupe[124]. Comme pour les méthodes de projection, les principes et l'implémentation de ces méthodes de classification sont décrits et illustrés en annexe B de cette thèse.

3.3.5 Construction des supports de visualisation interactive de l'ensemble documentaire

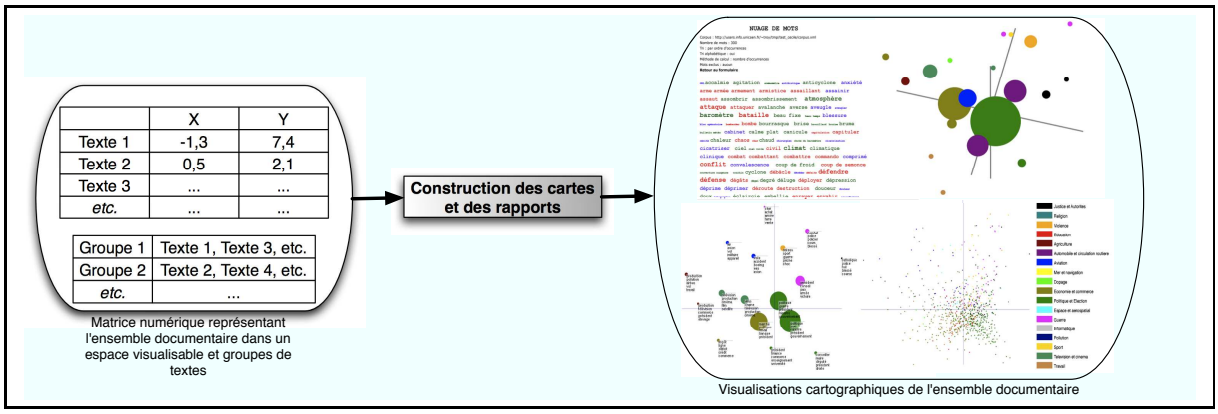


FIG. 3.21 – Étape de construction des cartes de l'ensemble documentaire.

Une fois les traitements précédents de comptage, de projection et de classification réalisés, toutes les informations nécessaires à la construction de visualisations cartographiques et de rapports d'analyse de l'ensemble documentaire, à partir des domaines de l'utilisateur, sont disponibles.

[124] Des méthodes de classification permettant des chevauchements entre plusieurs groupes pourront être cependant ajoutées dans les évolutions futures de *ProxiDocs*, telle la méthode ECCLAT de [Durand et Crémilleux, 2002].

Carte des textes en 2 dimensions

Fonctionnalité visée par la carte : première visualisation de la répartition des domaines de l'utilisateur dans un ensemble documentaire au grain « texte ».

La première visualisation que nous proposons (figure 3.22) est une carte en 2 dimensions représentant chaque texte de l'ensemble documentaire. Cette carte met directement en évidence les résultats de la méthode de projection permettant le passage de l'espace numérique initial de grande dimension à un espace numérique en 2 dimensions, en représentant chaque texte de l'ensemble documentaire. Chaque texte est représenté sur la carte par un point, dont la couleur correspond au domaine majoritairement présent dans le texte. La légende des couleurs associées aux domaines est placée en partie droite de la carte. Le langage SVG est utilisé pour construire les différentes cartes que nous proposons. Afin de proposer certaines interactions à l'utilisateur, le langage Javascript est associé.

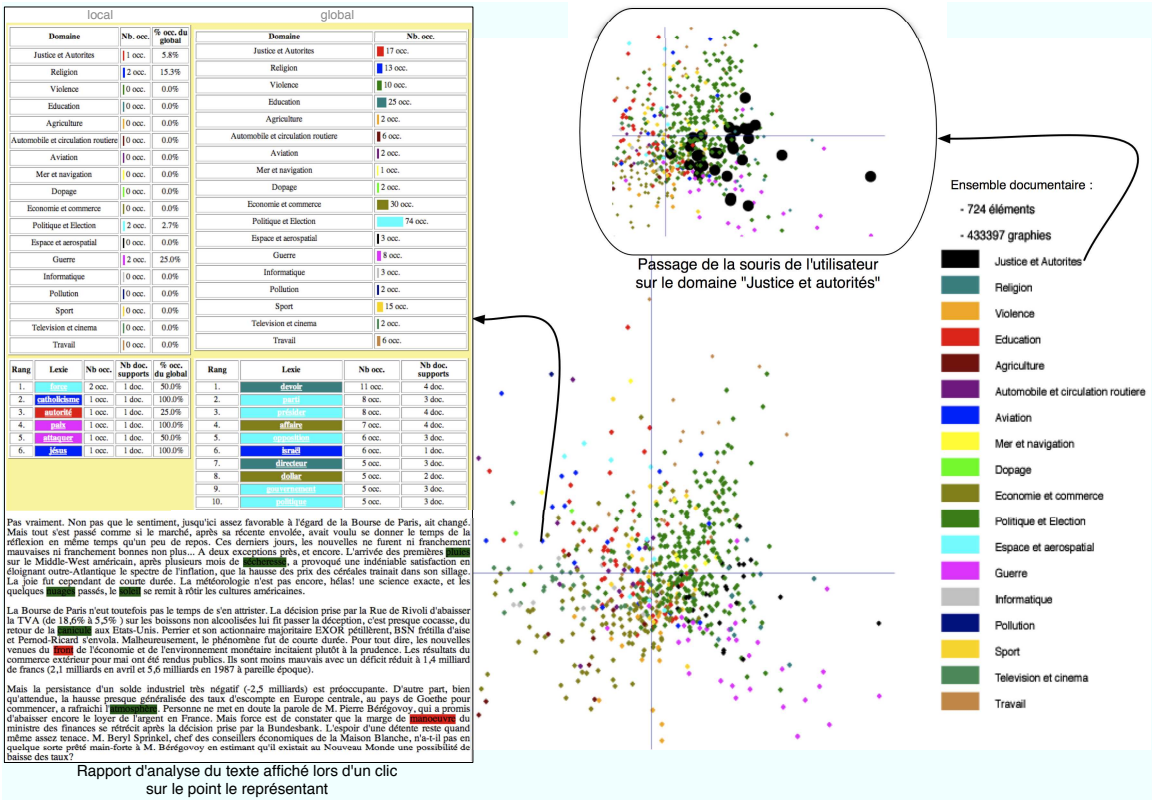


FIG. 3.22 – Carte des textes en 2 dimensions et les possibilités d'interactions offertes.

Parmi ces interactions, décrites également en figure 3.22, le passage de la souris sur le nom d'un domaine dans la légende provoque un agrandissement des points représentant des textes majoritairement de ce domaine. Également, chaque point sur la carte est un lien vers un rapport d'analyse du texte. Ce rapport met en évidence :

- pour chaque domaine, le nombre de lexies présentes dans le texte ;
- une liste des lexies des domaines présentes dans le texte ; ces lexies sont triées par ordre décroissant de leurs occurrences ;
- dans le cas d'une analyse basée sur des dispositifs *LUCIA*, une liste des isotopies intra-textuelles présentes dans le texte, triées par ordre décroissant de leurs scores (avec pondération, le score sans pondération est également indiqué) ;

– un texte colorié (au format HTML, présentant un coloriage similaire à celui effectué par *ThemeEditor*) où chaque occurrence de lexie des domaines est coloriée selon la couleur attribuée à son domaine, dans le cas d'une analyse basée sur des dispositifs *LUCIA*, les couples *attribut : valeur* associés aux lexies coloriées sont accessibles lors d'un passage de la souris sur la lexie.

Pour chaque élément du rapport d'analyse, un parallèle est réalisé entre le texte et son sur-ensemble (le groupe de textes, déterminé à l'issue de l'étape de classification, dans lequel le texte apparaît) afin de mettre en évidence comment le texte se positionne dans son contexte global. Dans ce cas, les différents calculs d'occurrences de lexies sont réalisés au niveau de l'ensemble des textes du groupe. La liste des isotopies inter-textuelles présentes dans le groupe est ainsi mise en parallèle avec les isotopies intra-textuelles du texte.

Carte des groupes de textes en 2 dimensions

Fonctionnalité visée par la carte : visualisation plus détaillée de l'ensemble documentaire au grain « groupe de textes », mise en évidence et description de regroupements entre textes selon les domaines de l'utilisateur.

Le second type de visualisation proposée (figure 3.23) est une carte mettant en évidence des groupes de textes en 2 dimensions. Cette carte se base sur les résultats de la projection présentée dans la carte précédente et propose de regrouper des textes jugés proches sur cette carte à l'aide d'une méthode de classification[125].

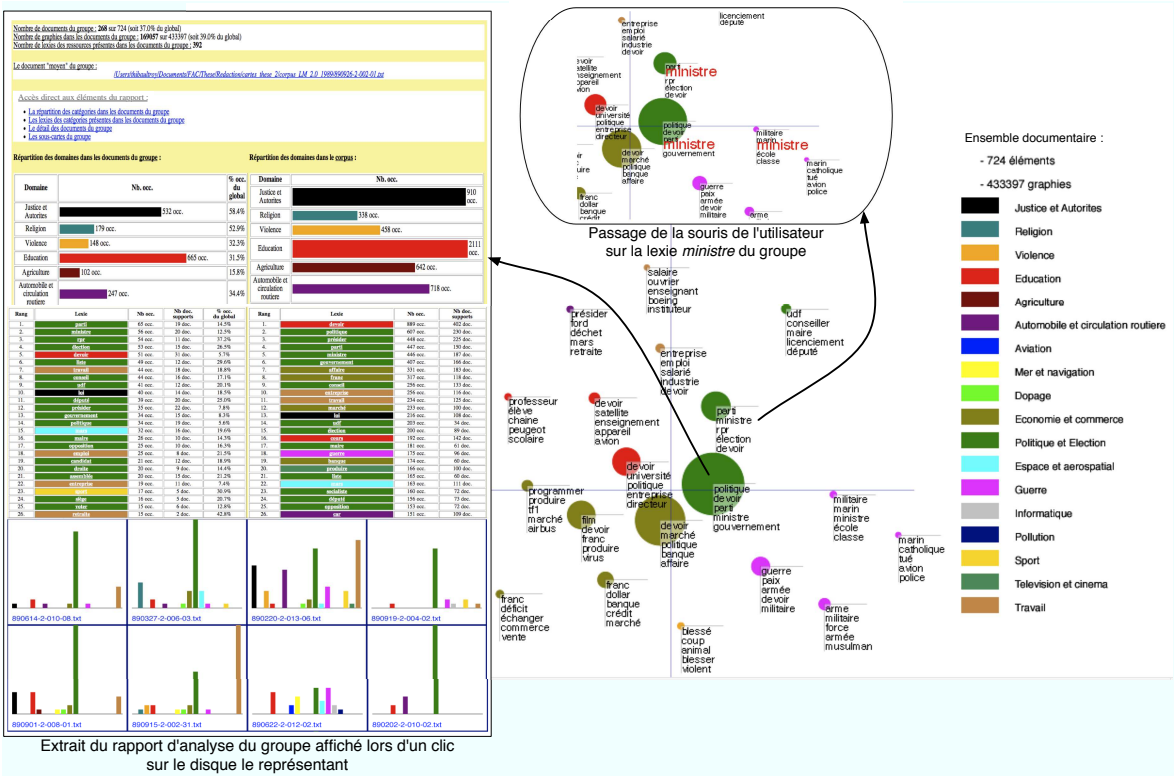


FIG. 3.23 – Cartes des groupes de textes en 2 dimensions et les possibilités d'interactions offertes.

[125] Nous renvoyons en section B.3 de cette thèse pour une présentation et une motivation des méthodes de classification proposées aux utilisateurs pour la construction des cartes de groupes de textes.

La carte obtenue présente des groupes de textes, ces groupes étant représentés par des disques plus ou moins grands selon leur cardinalité[126]. Chaque disque a la couleur du domaine majoritairement présent dans les textes du groupe, c'est-à-dire le domaine représenté par le plus d'occurrences de ses lexies. Le disque est positionné sur le centre de gravité de l'ensemble des points qu'il représente sur la carte. Les disques sont étiquetés par les cinq lexies les plus fréquentes des domaines de l'utilisateur présents dans les textes du groupe.

Différentes interactions sont proposées sur cette carte. Par exemple, lorsqu'un utilisateur place sa souris sur une lexie étiquetant un groupe, cette lexie est mise en évidence dans les différents groupes qu'elle étiquette. Cette action permet alors d'identifier les groupes de textes partageant cette lexie en nombre assez important. C'est également une façon d'identifier une autre proximité entre textes que celle proposée dans l'espace métrique de la carte. De manière similaire à la carte précédente, chaque disque est un lien vers un rapport d'analyse du groupe qu'il représente.

Ces rapports d'analyse de groupe contiennent différents éléments dont certains sont semblables à ceux des rapports d'analyse de texte. Le parallèle entre le contexte local (ici, le groupe) et le contexte global (ici, l'ensemble documentaire) est également réalisé. Les rapports contiennent les informations suivantes :

- le nombre de lexies de chaque domaine dans le groupe et dans l'ensemble documentaire ;
- une liste des lexies des domaines triées par ordre décroissant de leurs nombres d'occurrences dans les textes du groupe et dans l'ensemble documentaire ;
- dans le cas où les domaines sont des dispositifs *LUCIA*, une liste des isotopies inter-textuelles triées par ordre décroissant de leur score dans le groupe et dans l'ensemble documentaire, les attributs non répétés dans le groupe sont également marqués ;

De nouveaux éléments, propres à l'analyse de groupes de textes, sont également accessibles dans les rapports d'analyse :

- un lien vers le texte le plus proche du centre de gravité du groupe, ce texte peut être utile pour aider l'utilisateur dans la caractérisation du groupe ;
- chaque lexie du groupe ou de l'ensemble documentaire est un lien vers *FlexiSemContext* afin de mettre en évidence l'élément considéré dans son contexte (le groupe ou l'ensemble documentaire) ;
- également, chaque isotopie inter-textuelle présentée dans le rapport d'analyse est un lien vers *FlexiSemContext* permettant de visualiser en contexte les lexies du groupe ou de l'ensemble documentaire supportant cette isotopie ;
- un accès aux textes du groupe et à leur rapport d'analyse est proposé *via* un histogramme mettant en évidence la répartition des domaines dans chaque texte ;

Deux autres possibilités majeures d'analyse sont également offertes aux utilisateurs. La première consiste à proposer à l'utilisateur une « sous-cartographie » de chaque groupe de textes présent sur la carte (interaction illustrée en figure 3.24).

Cette sous-cartographie propose d'itérer les différents traitements réalisés par *ProxiDocs* au niveau des textes de chaque groupe. Chacune de ces sous-cartes, construites en même temps que la carte principale, met en évidence un nombre de sous-groupes égal à la moitié du nombre de groupes de la carte principale[127]. La sous-cartographie d'un groupe, accessible *via* son rapport d'analyse, est particulièrement utile si ce dernier contient un grand nombre de textes. Elle permet de déterminer plus finement les caractéristiques propres à un groupe, en mettant, par exemple, en évidence son homogénéité et son hétérogénéité par rapport aux domaines de l'utilisateur.

[126] Le diamètre D en pixels d'un disque répond à la formule : diamètre$(disque_{groupe}) = 2.\sqrt{|groupe|}$.
[127] Choix réalisé empiriquement afin de ne pas surcharger la sous-carte.

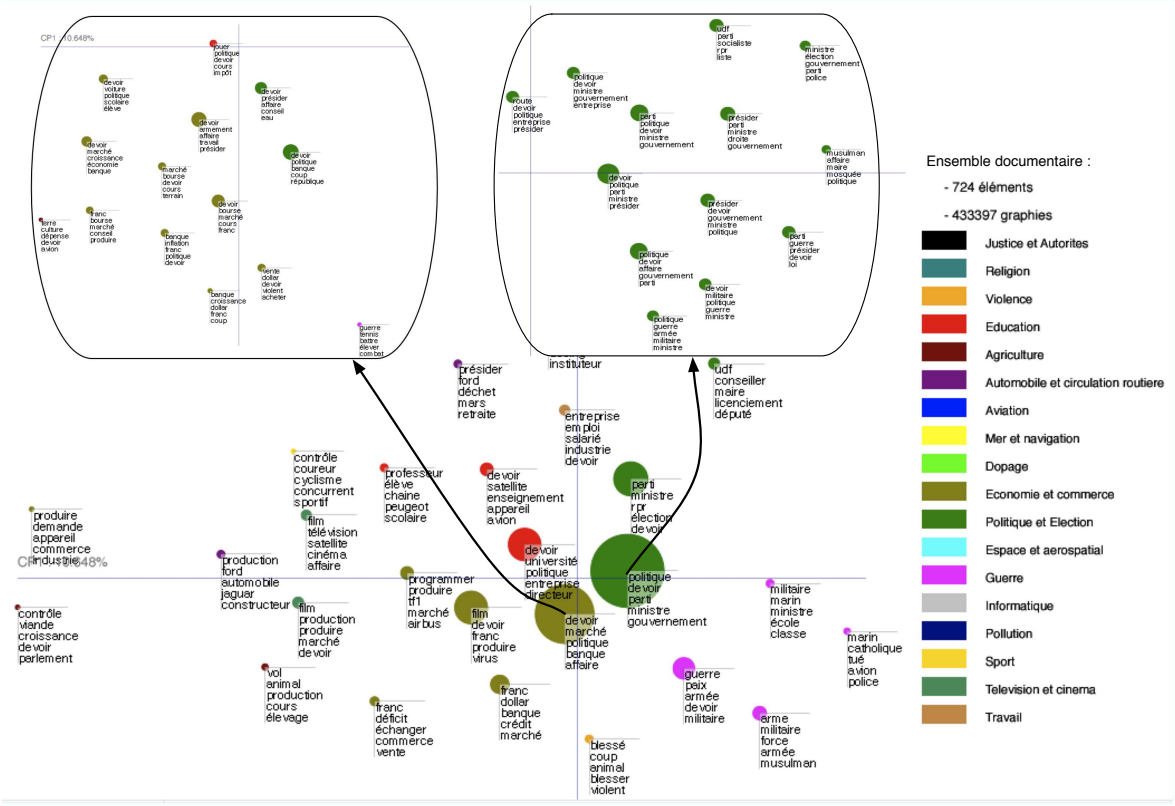


FIG. 3.24 – Mise en évidence des sous-cartographies de deux groupes de textes.

Une dernière possibilité offerte à l'utilisateur consiste en la suppression d'un groupe de textes sur la carte. Si l'utilisateur juge qu'un groupe de textes présenté sur la carte n'est pas ou plus pertinent dans son analyse, ce dernier peut relancer la construction d'une cartographie *via* ProxiDocs en ne tenant plus compte des textes de ce groupe. Une telle carte fait ressortir les particularités de l'ensemble documentaire, privé des textes d'un groupe, et ainsi, permet de réaliser de nouvelles observations sur le contenu du reste de l'ensemble documentaire.

La carte des groupes de textes permet à l'utilisateur de poursuivre l'analyse initiée par la carte des textes. Sur cette nouvelle carte, il peut appréhender plus finement le contenu de l'ensemble en visualisant des classes regroupant des textes partageant de mêmes isotopies macro-génériques. Au sein de chaque classe, l'utilisateur peut encore affiner son analyse en observant les domaines abordés, les lexies présentes, les isotopies inter-textuelles calculées, les sous-cartographies construites, etc. Le parallèle réalisé entre le groupe de textes et son contexte global, l'ensemble documentaire, permet à l'utilisateur de voir comment se positionne le groupe par rapport à l'ensemble documentaire. Ceci peut permettre d'observer un certain « héritage » des domaines majoritairement abordés dans l'ensemble documentaire sur les textes du groupe, par exemple, si les répartitions des domaines observées au niveau du groupe sont très proches de celles observées au niveau de l'ensemble documentaire. Au contraire, l'influence des textes du groupe sur l'ensemble documentaire peut également être mise en évidence, par exemple, si la répartition des différents éléments observés dans le groupe est très éloignée de celle observée dans l'ensemble documentaire.

Cartes des textes et des groupes de textes en 3 dimensions

Fonctionnalité visée : visualisations complémentaires de la répartition des domaines de l'utilisateur dans un ensemble documentaire aux grains « texte » et « groupe de textes ».

Les types de cartes présentés précédemment prennent place sur un plan. Ce choix de proposer à l'utilisateur des cartes dans un espace en deux dimensions est principalement lié à leur plus grande facilité d'accès. Pourtant, une fois que l'utilisateur est familiarisé avec des cartes en deux dimensions, il peut être particulièrement intéressant de lui proposer des cartes dans un espace en trois dimensions. Comme nous l'avons évoqué dans des parties précédentes de ce chapitre et en annexe B de cette thèse, les différentes méthodes de projection implémentées dans la plate-forme *ProxiDocs* proposent de réduire, de résumer, un espace de grande dimension vers un espace à 2 ou 3 dimensions, la réduction vers un espace à 2 dimensions entraînant une plus grande perte d'informations que celle vers un espace en 3 dimensions.

Des cartes en 3 dimensions (exemples donnés en figure 3.25) reflètent potentiellement mieux l'espace d'origine que les cartes en 2 dimensions. Leur principal intérêt est de permettre à l'utilisateur d'observer plus finement les proximités pouvant exister entre plusieurs textes et groupes de textes et ainsi d'en déduire des informations sur leur contenu. De telles cartes en 3 dimensions sont également particulièrement utiles pour des utilisateurs « surpris » des résultats révélés par leurs cartes en 2 dimensions. Il peut arriver que plusieurs textes ou groupes de textes soient situés à proximité les uns des autres sur une carte en 2 dimensions et que l'utilisateur ne comprenne pas de telles proximités. Des cartes en 3 dimensions peuvent ainsi permettre, soit de vérifier les proximités présentes sur la carte initiale, soit de les contredire, expliquant, par exemple, une perte importante d'informations lors de l'étape de projection.

L'utilisateur est libre de faire pivoter, à l'aide de sa souris, l'espace qui lui est proposé, et ainsi d'avoir différents angles de vue sur l'ensemble documentaire analysé. L'ajout de cette troisième dimension aux cartes d'ensembles documentaires entraîne, bien évidemment, une plus grande complexité dans l'analyse des cartes. Cependant, les observations peuvent révéler de nouvelles informations utiles, absentes des cartes en deux dimensions. Il est donc tout à fait recommander de

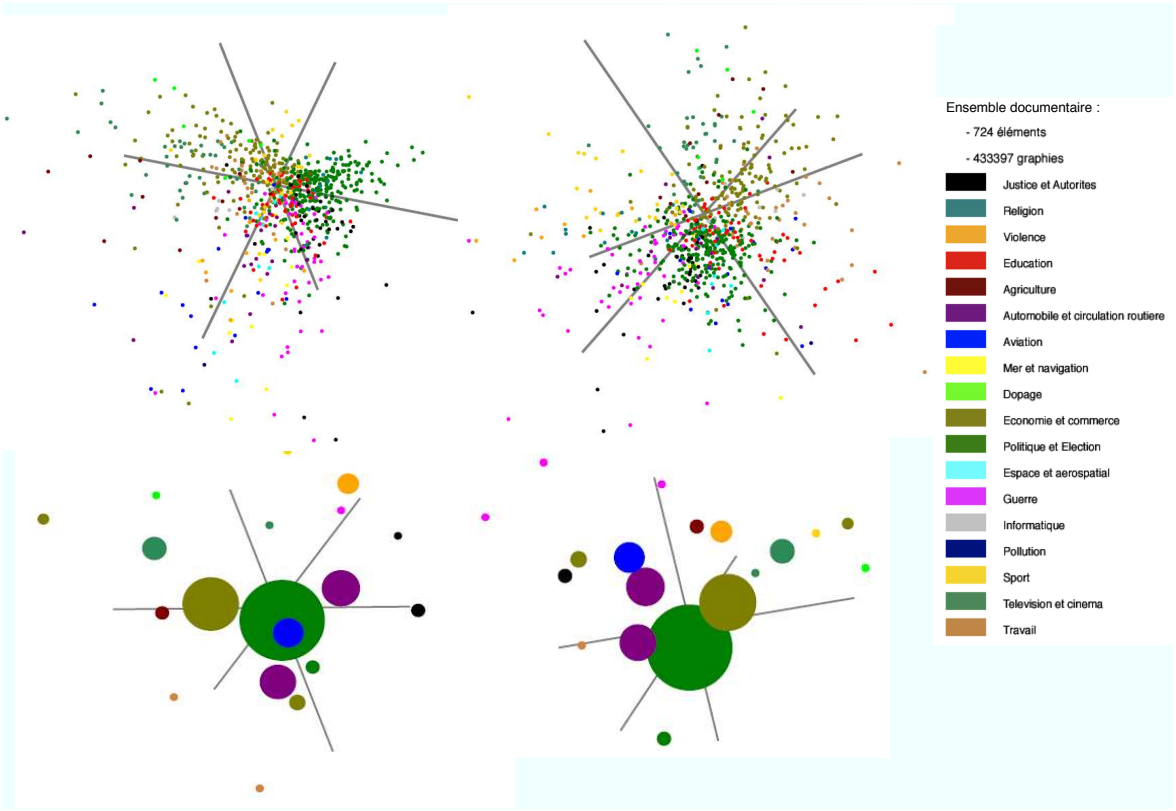


FIG. 3.25 – Exemples de cartes des textes (partie supérieure) et de cartes de groupes de textes (partie inférieure) en 3 dimensions.

proposer à un utilisateur habitué à analyser des cartes en deux dimensions, d'affiner ses analyses en ajoutant une troisième dimension, afin d'obtenir de nouvelles informations sur l'ensemble documentaire et les domaines étudiés.

Cartes temporelles des textes et des groupes de textes en 2 dimensions

Fonctionnalité visée par ces cartes : visualisation chronologique des domaines de l'utilisateur dans un ensemble documentaire aux grains « texte » et « groupe de textes ».

Les cartes précédentes proposent aux utilisateurs des vues globales sur les ensembles documentaires. Ces cartes ne prennent pas en considération la dimension temporelle des textes. De telles vues permettent l'étude de la dimension synchronique de l'ensemble documentaire mais ignorent sa dimension diachronique. Dans le chapitre 2, nous avons indiqué l'importance de l'ancrage des ensembles documentaires dans la dimension temporelle, dans une certaine actualité. Afin de permettre aux utilisateurs d'interroger un tel ancrage, nous proposons des cartes temporelles des textes et des groupes de textes, animées, mettant dynamiquement en évidence l'évolution des domaines présents dans l'ensemble documentaire au fil du temps.

Pour prendre en compte le temps dans nos analyses, il faut que les textes constituant l'ensemble documentaire soient datés (avec par exemple, la date de rédaction ou de publication des textes). Nous construisons ensuite des cartes des textes et des groupes de textes à partir de l'ensemble documentaire et des domaines de l'utilisateur sur différentes périodes. L'« enchaînement » automatique de ces cartes permet à ce dernier d'observer l'évolution des domaines présents au fil du temps. La figure 3.26 illustre un tel enchaînement.

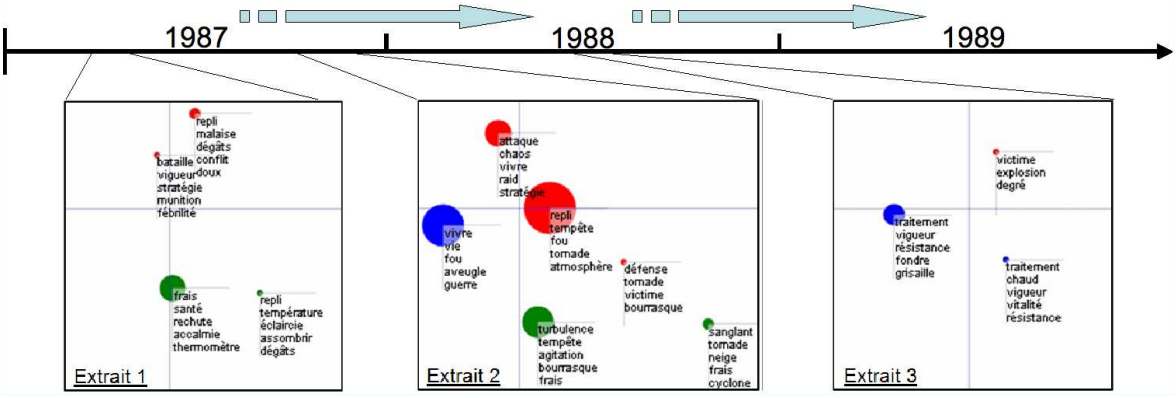


FIG. 3.26 – Illustration de l'enchaînement dynamique proposé par une carte temporelle. Trois extraits capturés à des moments différents sont présentés.

L'utilisateur doit tout d'abord choisir la fenêtre temporelle que les cartes doivent mettre en évidence. Cette fenêtre correspond à la durée maximale pouvant exister entre deux textes présents sur une même carte. L'utilisateur doit ensuite préciser sur quelle unité de temps cette fenêtre doit se déplacer sur l'axe temporel. Si ce dernier a choisi une fenêtre temporelle égale à un mois, alors il peut, par exemple, choisir une unité de temps égale à un jour. De cette manière, si l'on considère le texte le plus « ancien » de l'ensemble documentaire de l'utilisateur et soit D la date de ce texte, alors, la première carte est construite en considérant les textes compris entre D et (D + 1 mois), la seconde carte est construite en considérant les textes compris entre (D + 1 jour) et (D + 1 mois + 1 jour), etc. Le choix de la fenêtre temporelle et de l'unité de temps dépend grandement du type d'analyse que l'utilisateur souhaite réaliser.

Il est ensuite retourné à l'utilisateur une carte globale contenant les différentes cartes construites sur les différentes périodes. Cette carte globale proposera un enchaînement automatique de ces cartes. L'utilisateur doit définir l'équivalence entre le temps réel et l'unité de temps choisie, c'est-à-dire le délai entre l'enchaînement de deux cartes. Par exemple, si ce dernier considère qu'une seconde dans le temps réel est équivalente à une unité de temps, alors, une carte succédera à l'autre toutes les secondes sur la carte dynamique globale. Cet enchaînement des cartes permet à l'utilisateur d'observer les changements thématiques présents dans les textes de son ensemble documentaire au fil du temps.

Nuages et anti-nuages de lexies de l'ensemble documentaire

Fonctionnalité visée par ces sorties : autres visualisations de la répartition des domaines de l'utilisateur dans un ensemble documentaire.

Les différentes cartes, en 2 ou 3 dimensions, statiques ou animées, qui ont été présentées dans les parties précédentes, constituent les supports de visualisation les plus complets, les plus détaillés, les plus interactifs, offerts aux utilisateurs par l'application *ProxiDocs* pour leur interprétation d'ensembles documentaires. Ces supports de visualisations peuvent cependant être utilement complétés, ou même introduits, par des supports de visualisation plus simples, proposant moins d'informations à l'utilisateur.

Parmi ces supports plus simples, nous proposons ce que nous appelons des *nuages* et des *anti-nuages* de lexies[128]. Ces derniers mettent respectivement en évidence les lexies des domaines de l'utilisateur les plus fréquentes et les moins fréquentes dans l'ensemble documentaire (plus une lexie est grande dans le nuage, plus elle est occurrente dans l'ensemble documentaire, plus une lexie est grande dans l'anti-nuage, moins elle est présente dans l'ensemble documentaire).

La figure 3.27 présente de tels nuages à partir d'un domaine portant sur la guerre construit par un utilisateur. En partie gauche de la figure, le nuage nous permet d'observer que les lexies *bataille*, *conflit*, *front*, *repli*, *résistance*, *stratégie* et *victoire* sont les plus fréquentes du domaine dans l'ensemble documentaire. Au contraire, la partie droite, représentant un anti-nuage de lexies, permet d'observer que les lexies *bombarder*, *capitulation*, *char*, *défaite* et *hostilité* sont très rares voire absentes de l'ensemble documentaire. Ces nuages sont interactifs, un survol de la souris de l'utilisateur sur une lexie entraîne l'affichage d'informations sur cette lexie : son nombre d'occurrences dans l'ensemble documentaire et le nombre de textes de l'ensemble la contenant.

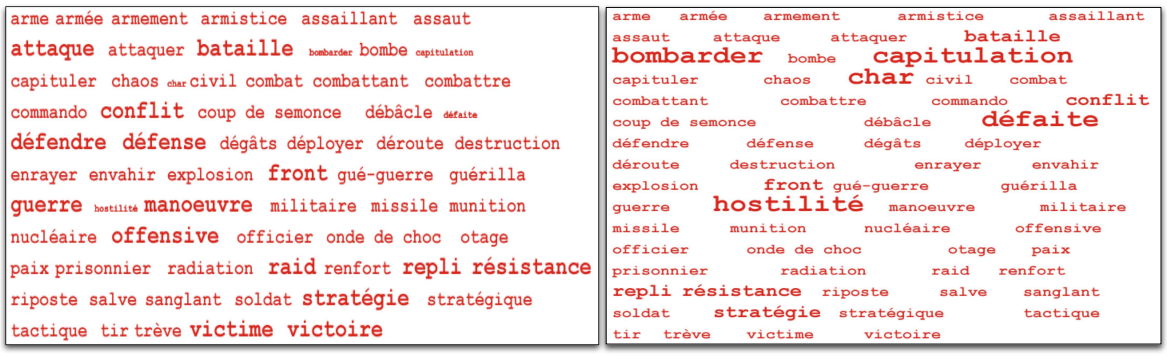


FIG. 3.27 – Exemples d'un nuage (à gauche) et d'un anti-nuage (à droite) de lexies d'un domaine (la guerre) dans un ensemble documentaire.

[128] Visualisations en nuages de mots inspirées, entre autres, du site *TagCloud* `http://www.tagcloud.com` (page consultée le 26 mai 2007) proposant des visualisations de contenu de *blogs* sur Internet.

Ces nuages permettent à l'utilisateur de visualiser de premières informations sur le contenu de son ensemble documentaire selon les domaines qu'il a définis. Ils lui permettent de s'interroger sur la trop grande généricité des lexies très fréquentes, ainsi que sur la cohérence de la présence dans les domaines de lexies absentes ou trop peu présentes dans un ensemble documentaire.

3.3.6 Mise en œuvre logicielle au sein de la plate-forme *ProxiDocs*

Les différents traitements détaillés précédemment ont été implémentés dans la plate-forme *ProxiDocs* afin de permettre à l'utilisateur d'obtenir les différentes cartes et visualisations de son ensemble documentaire. Le langage de programmation Java a été utilisé afin de mettre en œuvre les différents traitements. La machine virtuelle Java doit donc être préalablement installée sur la machine de l'utilisateur pour exécuter la plate-forme[129]. Le langage SVG, couplé avec le langage Javascript, a été utilisé afin de produire les cartes et les nuages interactifs de l'ensemble documentaire. De même, les langages SVG et HTML ont été utilisés afin de produire les différents rapports d'analyses des textes et des groupes de textes. Le schéma présenté en figure 3.28 résume les différents traitements abordés dans les parties précédentes et mis en œuvre par l'application.

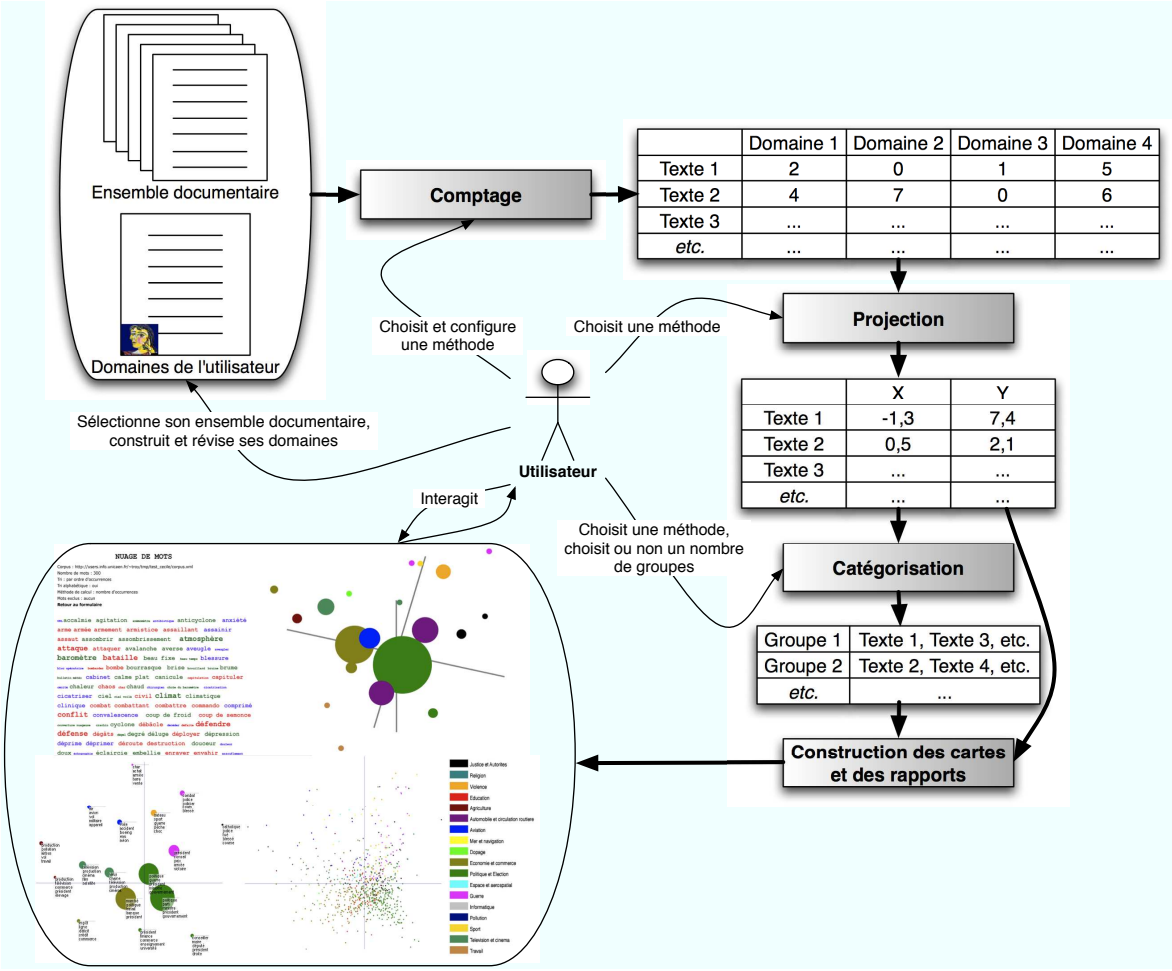


FIG. 3.28 – Enchaînement des traitements réalisés par *ProxiDocs*.

[129] Machine virtuelle disponible : `http://java.sun.com` (page consultée le 14 juillet 2007).

Sans entrer dans les détails de l'implémentation de ces traitements dans la plate-forme[130], de nombreuses classes ont été développées, regroupées en un certain nombre de *packages*. Nous avons le plus possible réalisé nos développements logiciels afin de faciliter les ajouts de nouveaux composants. La structuration en *packages* et classes que nous avons choisie pour la plate-forme *ProxiDocs* permet de tels ajouts et modifications. Cette structuration est décrite succinctement au tableau 3.1.

Nom du package	Classes contenues
`parseurs`	Classes permettant de lire les fichiers XML contenant les domaines de l'utilisateur.
`comptage`	Classes décrivant les méthodes de comptage des domaines de l'utilisateur en ensemble documentaire.
`projection`	Classes implémentant les méthodes de projection de l'ensemble documentaire sur un plan ou un espace en 3 dimensions.
`classification`	Classes mettant en œuvre les méthodes de classification de l'ensemble documentaire.
`cartographie`	Classes permettant de produire les cartes interactives de l'ensemble documentaire.
`gui`	Classes décrivant l'interface graphique de la plate-forme *ProxiDocs*.
`rapport`	Classes permettant la construction des différents rapports d'analyse des textes et des groupes de textes, des nuages et anti-nuages de lexies, des histogrammes, etc.
`Jampack`	Ce *package*, que nous réutilisons, contient des classes permettant de représenter des matrices numériques et de réaliser différents calculs sur ces matrices, calculs intervenant en particulier dans les méthodes de projection (plus d'informations sur ce *package* : `ftp://math.nist.gov/pub/Jampack/Jampack/AboutJampack.html` (page consultée le 21 mai 2007)).

TAB. 3.1 – *Packages* de la plate-forme *ProxiDocs*.

Les différentes classes contenues dans ces *packages* s'articulent les unes avec les autres. Le diagramme de classes, présenté en figure 3.29, met en évidence les principaux liens et articulations entre les classes. Ainsi, pour ajouter une méthode de comptage, de projection ou de classification, il suffit de développer une nouvelle classe héritant respectivement de MethodeComptage, de MethodeProjection ou de MethodeCategorisation. De la même manière, pour construire un nouveau type de carte interactive, il suffit d'étendre la classe Carte.

[130] Pour cela, nous renvoyons au site Internet dédié à *ProxiDocs*, permettant de télécharger la plate-forme et de consulter un grand nombre de cartes : `http://www.info.unicaen.fr/~troy/proxidocs` (page consultée le 14 juillet 2007).

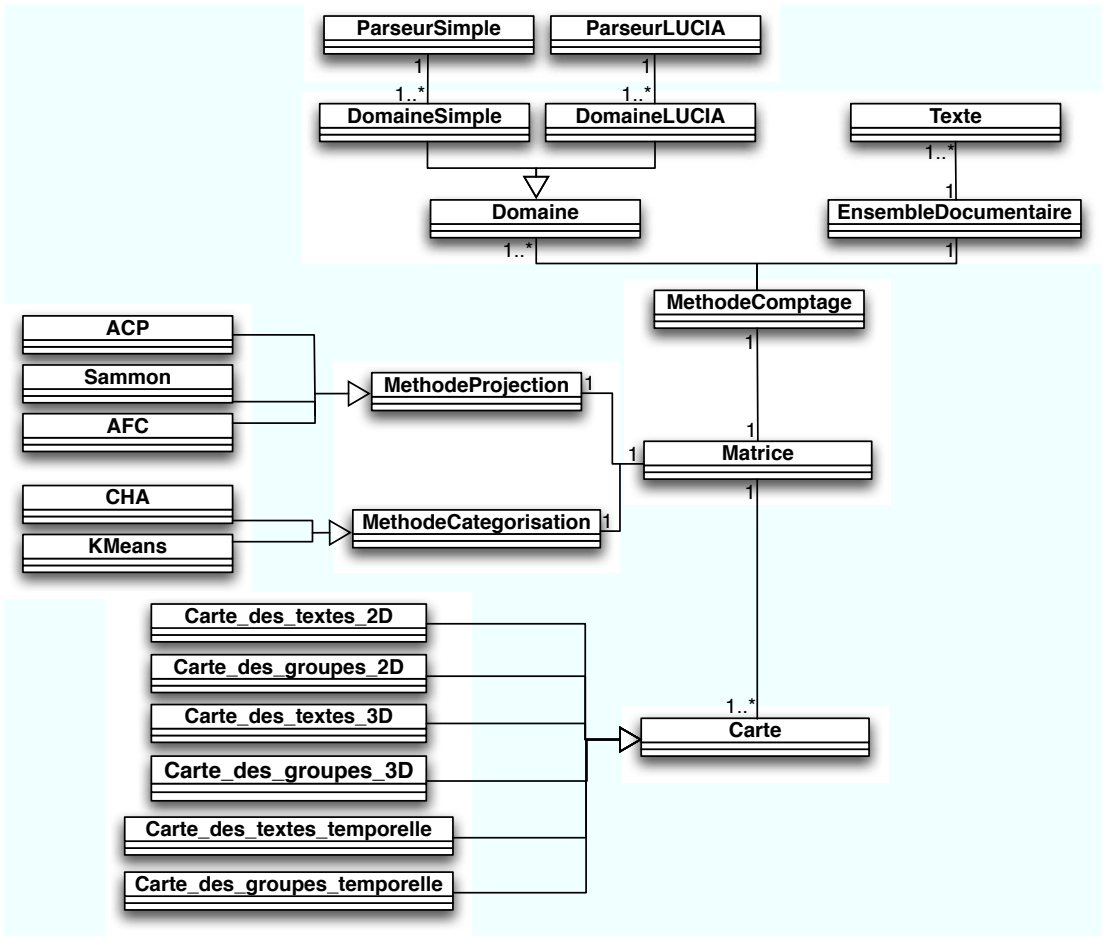


FIG. 3.29 – Diagramme de classes mettant en évidence les principaux éléments logiciels composant la plate-forme *ProxiDocs*.

Afin de permettre à l'utilisateur de créer facilement des cartes d'ensemble documentaire à partir de domaines de son choix, la plate-forme *ProxiDocs* propose l'interface présentée en figure 3.30.

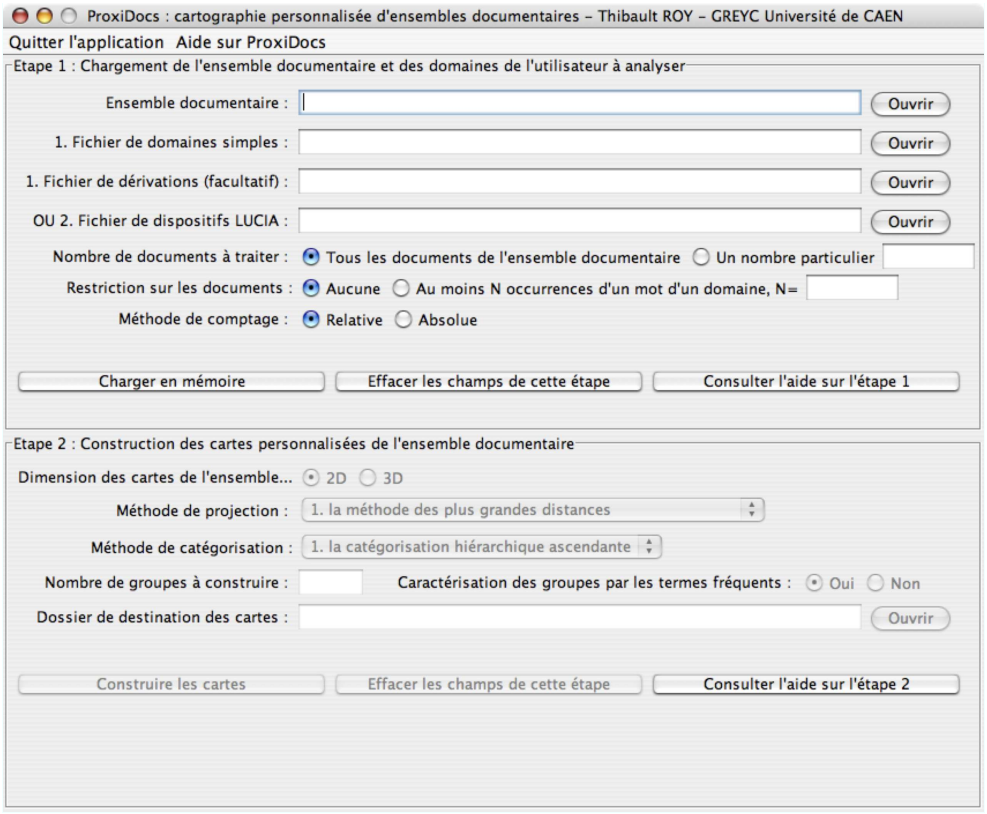


FIG. 3.30 – Interface de la plate-forme *ProxiDocs*.

Cette interface propose de décomposer la construction des cartes en 2 étapes :

1. La première étape est l'étape de comptage. L'utilisateur doit choisir sur son disque dur le répertoire contenant l'ensemble documentaire et le fichier XML contenant les domaines. Une fois la méthode de comptage configurée, l'utilisateur peut lancer l'exécution de cette première étape. Une instance d'un objet décrivant l'ensemble documentaire (la matrice numérique et un certain nombre d'éléments nécessaires pour la construction des rapports et des cartes) est alors créée sur le disque dur[131] de l'utilisateur, le temps de la session d'utilisation de la plate-forme.
2. La seconde étape consiste à appliquer les traitements de projection, de classification et de construction des cartes. L'utilisateur doit alors choisir et configurer les méthodes qu'il souhaite utiliser.

Cette séparation en deux étapes permet à l'utilisateur de ne réaliser qu'une seule fois l'étape de comptage[132], puis d'appliquer autant de méthodes de projection, de classification, de cartographie qu'il le souhaite, avec différents paramètres.

[131] Toutes nos classes Java implémentent l'interface *Serializable* afin de permettre leur représentation par une suite d'octets stockable dans un fichier.

[132] L'étape de comptage peut être assez longue selon le nombre d'éléments de l'ensemble documentaire et le nombre de lexies dans les domaines de l'utilisateur. Éviter de répéter cette étape nous semble indispensable pour une utilisation optimale de la plate-forme.

3.3.7 Intégration et utilisation de *ProxiDocs* au sein de projets d'enseignement et de recherche

Le développement de la plate-forme *ProxiDocs* a débuté en 2003 dans le cadre d'un projet de Maîtrise d'Informatique de l'Université [Roy et Sagit, 2003]. Son développement a donné lieu à plusieurs publications scientifiques. Ainsi, dans [Roy et Beust, 2004], nous avons décrit les méthodes statistiques exploitées dans la plate-forme. Dans [Roy, 2005, Roy et Beust, 2005], nous avons présenté en détail les différents types de cartes pouvant être construits par la plate-forme.

Différentes versions de la plate-forme se sont ainsi succédées, avec l'ajout fréquent de nouvelles fonctionnalités. *ProxiDocs* est à la fois le résultat d'un projet de recherche et d'enseignement, mais est également à l'initiative de plusieurs autres projets. Au cours de cette thèse, différents étudiants de Master d'Informatique de l'Université de Caen ont travaillé sur des projets portant, aussi bien, sur l'ajout de nouvelles fonctionnalités à la plate-forme, que sur des développements gravitant autour de la thématique des RTO personnelles.

Ainsi, en 2005, Samuel Authesserre et Jérémy Courvalet ont développé un métamoteur de recherche utilisant les fonctionnalités de la plate-forme *ProxiDocs* pour proposer une cartographie personnalisée des résultats d'un moteur de recherche [Authesserre et Courvalet, 2005]. En 2006, Kahina Hamadache et Matthieu Vernier ont réalisé la première version de l'outil *Visual-LuciaBuilder* [Hamadache et Vernier, 2006]. Cécile Raffy [Raffy, 2006] a intégré à la plate-forme un module logiciel permettant la construction des nuages et des anti-nuages de lexies. En 2007, François Malherbe et Jérôme Le Moulec ont travaillé sur des méthodes de visualisation permettant de suivre le rapport à l'actualité de ressources terminologiques personnelles telles que des RTO *LUCIA* [Malherbe et Moulec, 2007].

La plupart de ces projets d'étudiants nous ont permis de mener des développements venant compléter différentes expérimentations, notamment dans le cadre de différents projets de recherche. Ces expérimentations ont été impliquées par différentes collaborations, entraînant des utilisations de la plate-forme *ProxiDocs* par différents chercheurs, dans différents contextes et avec des objectifs variés. Ces expérimentations ont constitué une véritable mise à l'épreuve pour nos propositions, et nous les détaillons chacune dans le chapitre suivant de cette thèse.

Conclusion : mise en instruments du modèle d'analyse à travers des logiciels interactifs

Ce chapitre a permis de présenter les composants logiciels proposant des instrumentations des différentes propositions constituant notre modèle. Pour concrétiser la place centrale que nous donnons aux utilisateurs, les instrumentations proposées sont particulièrement orientées vers ces derniers : ce sont eux qui construisent et manipulent les RTO à la base des traitements, et ce sont également eux qui naviguent sur les cartes des ensembles documentaires.

Ces réalisations logicielles ont pour objectif de rendre le plus accessible possible des tâches de construction de RTO personnelles et de projections cartographiques de ces RTO en ensembles documentaires. Ces réalisations nous ont permis de mettre en évidence les besoins de généricité (par exemple, en langues, en codages), de robustesse, d'accessibilité, mais surtout d'interactivité. En tant que logiciels d'étude, nos outils et instruments restent toujours modifiables, et ne sont donc pas, par principe, dans un état finalisé. En fonction de telle ou telle hypothèse expérimentale, ces logiciels seront mis à l'épreuve et seront amenés à évoluer selon les attentes des utilisateurs.

Afin de mettre en évidence la valeur ajoutée de notre modèle et de son implémentation dans des tâches variées d'accès au contenu, nous proposons dans le chapitre suivant différentes expérimentations. Ces expérimentations, réalisées dans le cadre de collaborations et de contextes très différents, ont impliqué la mise en place de différents *dispositifs expérimentaux* selon [Habert, 2005], où un dispositif expérimental est *un montage d'instruments, d'outils et de ressources servant à produire des « faits » dont la reproductibilité et le statut (l'interprétation) font l'objet de controverses.* La pertinence accordée par les utilisateurs aux dispositifs mis en place nous permettra, alors, de mesurer la valeur ajoutée de nos propositions.

Chapitre 4

Expérimentations et évaluations du modèle *AIdED*

Sommaire

Introduction 116
4.1 La problématique de l'évaluation en TAL 116
4.1.1 Méthodes d'évaluation traditionnelles 117
4.1.2 Vers d'autres méthodes d'évaluation pour des systèmes interactifs et/ou centrés-utilisateur 118
4.1.3 Évaluer des systèmes centrés sur leurs utilisateurs 120
4.2 Recherche documentaire personnalisée 122
4.2.1 Assistance dans une recherche d'information généraliste sur Internet . . 122
4.2.2 Assistance dans une recherche d'information médicale 128
4.3 Analyse d'expressions métaphoriques 136
4.3.1 Étude de métaphores conceptuelles 136
4.3.2 Ressources et corpus 138
4.3.3 Analyses réalisées 139
4.3.4 Vers une nouvelle phase du projet IsoMeta 152
4.4 Étude de forums de discussion pédagogiques 154
4.4.1 Observation de l'acquisition de l'identité professionnelle 155
4.4.2 Observation des usages d'une terminologie professionnelle 160
4.4.3 Apports de vues globales et interactives dans l'accès au contenu de forums pédagogiques 166
4.5 Mesurer les premiers regards portés sur des cartes d'ensembles documentaires 167
4.5.1 Motivations et contexte 167
4.5.2 Cadre expérimental 169
4.5.3 Analyse des résultats et retour sur les cartes d'ensembles documentaires 172
Conclusion : valeur ajoutée et flexibilité du modèle *AIdED* 177

Introduction

Dans ce chapitre, nous proposons d'expérimenter et d'évaluer les différentes propositions que nous faisons à travers le modèle *AIdED* et son instrumentation. La problématique de l'évaluation en TAL est particulièrement active. Différentes campagnes, différents protocoles, différentes mesures sont proposés afin de mettre en évidence la pertinence de systèmes pour des tâches précises. Évaluer un système d'accès personnalisé au contenu d'ensembles documentaires comme le nôtre est plus complexe, l'utilisateur y projetant son point de vue, et donc une forte part de subjectivité. Cependant, une telle évaluation n'est pas impossible, elle ne pourra peut-être pas s'effectuer avec un protocole classique, à l'aide, par exemple, des mesures de rappel et de précision, mais plutôt par la mise en place de différentes expérimentations auprès d'utilisateurs visant des tâches d'accès au contenu d'ensembles documentaires bien précises.

Ainsi, la section suivante de ce chapitre aborde la problématique de l'évaluation en TAL et plus particulièrement les moyens d'évaluer des systèmes centrés sur leurs utilisateurs. Les sections 2 à 4 présentent des expérimentations très variées prenant place dans des contextes précis. Nos propositions seront respectivement mises à l'épreuve dans des tâches de recherche documentaire, d'étude d'expressions métaphoriques et d'analyse de forums de discussion. Dans chacune de ces tâches, des utilisateurs « experts » ont été impliqués afin d'avoir un retour le plus pertinent possible sur nos propositions par rapport à la tâche visée. La partie 5 de ce chapitre est consacrée à une expérimentation un peu différente où l'interface des supports d'interaction que nous proposons est tout particulièrement interrogée, notamment avec un dispositif de suivi du regard. Enfin, nous conclurons sur ces différentes mises à l'épreuve de notre modèle afin d'en évaluer sa valeur ajoutée et son adaptabilité dans différentes tâches d'accès personnalisé au contenu d'ensembles documentaires.

4.1 La problématique de l'évaluation en TAL

Dès les débuts de l'informatique, la problématique de l'évaluation a été abordée dans le but de mesurer l'efficacité, la rapidité de programmes. Il est ainsi tout à fait souhaitable d'avoir des programmes dont le temps de calcul et l'espace mémoire occupé soient les plus réduits possibles. Mesurer la qualité d'un programme et des algorithmes sous-jacents du point de vue de sa complexité constitue le domaine de recherche de l'algorithmique. Ce domaine cherche ainsi à définir formellement la complexité temporelle (tel le nombre d'itérations nécessaires à son exécution) et spatiale (telle la taille des structures de données nécessaires) associée à un algorithme et donc aux systèmes l'implémentant.

Cette discipline met alors particulièrement l'accent sur le fonctionnement interne de programmes, sur leur temps d'exécution, sur la place occupée en mémoire, etc. Dans le cadre de systèmes de TAL, l'étude du fonctionnement interne est certes importante à prendre en considération mais n'est pas le facteur permettant de dire si un système est pertinent ou non pour une tâche visée[133]. L'élément principal à évaluer sont les sorties produites par le système et si ces dernières répondent à la tâche visée. Si des entrées sont nécessaires au système (corpus de textes, ressources lexicales, etc.), le temps nécessaire à leur élaboration doit également être pris en considération. Également, si les systèmes sont interactifs, les différentes fonctionnalités proposées aux utilisateurs doivent être considérées.

[133]Même si la problématique du « temps réel » en TAL est abordée depuis un certain temps dans le cadre de systèmes de reconnaissance et de synthèse de la parole, des systèmes de traduction, des systèmes de veille, etc. où des réponses rapides doivent être obtenues. Nous pouvons par exemple citer dans cette problématique la thèse de Leila Zouari [Zouari, 2007] proposant un système de transcription automatique de parole en temps réel.

Dans la suite de cette section, nous illustrons la problématique de l'évaluation en TAL à travers différentes actions menées dans le cadre de différentes tâches. Différentes propositions et métriques pour l'évaluation en TAL sont ainsi présentées.

4.1.1 Méthodes d'évaluation traditionnelles

Dès les débuts du TAL, la question de l'évaluation des systèmes a été posée, notamment avec des travaux portant sur la traduction automatique de [Bar-Hillel, 1960, Alpac, 1966]. La problématique de l'évaluation en TAL constitue même depuis quelques années un champ de recherche à part entière, avec des conférences, des campagnes de recherches, des organisations.

Ainsi, les conférences TREC[134] ont été mises en place aux États-Unis depuis 1992 dans le but d'évaluer des systèmes de recherche documentaire. Entre 1987 et 1998, sept conférences MUC[135] ont été organisées afin de proposer des évaluations de systèmes d'extraction d'information. Plus récemment, les conférences LREC[136] ou encore SENSEVAL[137] traitent plus particulièrement de tâches respectivement liées à l'adéquation de ressources lexicales dans différentes tâches de TAL et à la désambiguïsation sémantique.

La plupart des organisations traitant du TAL se positionnent autour de la problématique de l'évaluation. C'est par exemple le cas de l'association ELRA[138] ou encore de l'association LDC[139].

Différentes campagnes d'évaluation sur les différents thèmes du TAL naissent également à intervalles réguliers. Récemment, le programme Technolangue[140] a été mis en place en France. L'un de ses objectifs est de proposer des méthodes d'évaluation de différents outils de TAL, tels des analyseurs syntaxiques, des systèmes de questions / réponses ou encore des outils de traductions automatiques[141].

Plusieurs indicateurs sont traditionnellement utilisés afin d'évaluer de tels systèmes de TAL. Les plus fréquents sont les mesures de *rappel* et de *précision*[142]. D'autres mesures peuvent alors être déduites des précédentes, tel le *bruit* (proportion de fausses réponses parmi les résultats du système), le *silence* (proportion de bonnes réponses absentes des résultats retournés par le système) ou encore la *f-mesure* [Rijsbergen, 1979] synthétisant les mesures de rappel et de précision. Selon les domaines visés par l'évaluation, de nouvelles mesures sont proposées. C'est par exemple le cas des mesures BLEU (*Bilingual Evaluation Understudy*) [Papineni *et al.*, 2002] et NIST (mesure du *National Institute of Standards and Technology*) [Doddington, 2002] pour la traduction automatique, ou encore de la mesure ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) [Lin et Hovy, 2003] pour le résumé automatique.

Ces mesures permettent principalement d'établir des classements entre systèmes. Un positionnement des systèmes par rapport à une mesure plancher (appelée encore *baseline*) est souvent

[134] Text Retrieval Conference - `http://trec.nist.gov` (page consultée le 26 juin 2007).

[135] Message Understanding Conferences - `http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html` (page consultée le 26 juin 2007).

[136] Language and Ressources Evaluation Conferences - `http://www.lrec-conf.org` (page consultée le 26 juin 2007).

[137] `http://www.senseval.org` (page consultée le 26 juin 2007).

[138] European Language Resources Association - `http://www.elra.info` (page consultée le 26 juin 2007).

[139] Linguistic Data Consortium - `http://www.ldc.upenn.edu` (page consultée le 26 juin 2007).

[140] `http://www.technolangue.net` (page consultée le 26 juin 2007).

[141] Respectivement avec les campagnes *EASY* (Évaluation des Analyseurs Syntaxiques du français), *EQueR* (Évaluation en Question/Réponse) et *CESTA* (Campagne d'Evaluation de Systèmes de Traduction Automatique).

[142] Respectivement définies de la manière suivante : *rappel = pourcentage de réponses pertinentes proposées par le système évalué dans l'ensemble des réponses existantes dans le corpus de base* et *précision = pourcentage de réponses pertinentes dans l'ensemble des réponses proposées par le système évalué.*

réalisé. Des valeurs élevées ou basses de ces mesures permettent également de mettre en évidence les forces et les faiblesses des systèmes étudiés. Par exemple, un analyseur syntaxique ayant un fort taux de précision mais un faible taux de rappel retourne de bons étiquetages mais en nombre très limité, largement inférieur au nombre d'étiquettes attendues.

Des évaluations utilisant de tels indicateurs nécessitent aux « évaluateurs » humains de s'accorder sur un grand nombre de points (tels par exemple un ou plusieurs corpus de tests, une ou plusieurs séries de requêtes, d'alignements multi-lingues, etc.), ce qui peut être particulièrement délicat et nécessiter un important travail. Ce type de travail a alors pour but de borner le champ d'application des outils à évaluer et d'attribuer une certaine pertinence à des résultats pouvant être potentiellement retournés par ces outils. Dans un tel cas, la pertinence est alors considérée comme une décision binaire et objective. Ceci peut être valable pour certaines tâches où il est assez facile de s'accorder sur une décision, par exemple sur une réponse devant être retournée par un système de question / réponse à une interrogation portant sur une date, ou encore sur une étiquette morphosyntaxique. Par contre, pour des systèmes interrogeant le contenu d'ensembles documentaires tels des systèmes d'aides à la navigation ou des systèmes de recherche et de veille documentaire, il est beaucoup plus difficile de se mettre d'accord sur le bon résultat à retourner à telle requête ou manipulation de l'utilisateur.

Pour nous, et comme nous avons pu l'aborder précédemment, la pertinence est une notion dépendante de l'utilisateur, et donc forcément subjective. Plusieurs utilisateurs n'auront pas forcément le même point de vue des résultats à obtenir pour une tâche donnée. Un même utilisateur pourra juger différemment la pertinence de quelque chose selon l'instant du jugement. C'est l'utilisateur qui est au final le seul juge de la pertinence d'un résultat selon ses besoins, cette place prépondérante est pour l'instant trop souvent ignorée dans les protocoles d'évaluation actuels.

Nous renvoyons à [Beust, 2005] où l'auteur dresse un panorama des principales limites des différentes méthodes et métriques traditionnellement proposées pour l'évaluation en TAL, et principalement dans le cadre de systèmes interactifs et personnalisés d'accès au contenu d'ensembles documentaires. Nous pouvons également citer [Chaudiron, 2004] où il est proposé un état de l'existant très complet des méthodes d'évaluation de différents systèmes de TAL.

Les paragraphes suivants de ce chapitre prennent en compte de telles limites pour mettre en avant différentes propositions prenant en considération l'utilisateur dans l'activité d'évaluation.

4.1.2 Vers d'autres méthodes d'évaluation pour des systèmes interactifs et/ou centrés-utilisateur

Pour prendre en considération le point de vue de l'utilisateur sur la pertinence d'un système pour une tâche donnée, il faut donc l'intégrer d'une certaine façon dans l'évaluation du système. Karen Spark-Jones et Julia R. Galliers [Spark Jones et Galliers, 1995] ont proposé de faire intervenir deux « facettes » dans l'évaluation : l'évaluation *intrinsèque* et l'évaluation *extrinsèque*. L'évaluation intrinsèque mesure les propriétés concernant la nature du sujet à évaluer et son objectif, alors que l'évaluation extrinsèque mesure les aspects concernant les impacts et les effets de sa fonction. Ainsi, des critères liés à la fonction propre du système (coté intrinsèque) sont pris en considération, par exemple à l'aide de mesures présentées précédemment, mais également des critères liés à l'usage du système dans son environnement (coté extrinsèque), par exemple à l'aide de questionnaires de satisfaction adressés aux utilisateurs.

En pratique, et plus particulièrement dans une tâche de production automatique de résumés de textes, il est proposé dans [Farzindar et Lapalme, 2005] que l'évaluation intrinsèque du système soit la mesure ROUGE énoncée précédemment. Pour l'évaluation extrinsèque, il a été demandé à des utilisateurs typiques du système de juger la qualité des résumés produits. La prise

en considération de ces deux éléments dans le cadre de l'évaluation globale du système permet alors d'avoir une véritable information sur la pertinence du système pour une tâche donnée.

Toujours pour donner plus de place à l'utilisateur dans l'évaluation de systèmes informatiques, Daniel Luzzati propose dans [Luzzati, 1996] de définir un taux de compétence et un taux d'efficacité. Ces taux consistent principalement en la mesure du nombre de corrections devant être apportées par un utilisateur dans un système interactif de dialogue homme-machine. Moins il y a de corrections réalisées par l'utilisateur, plus le système est jugé pertinent pour la tâche visée.

Certains travaux mettent l'accent plus particulièrement sur l'ergonomie, comme le propose Laurence Bellies dans [Bellies, 2002] dans le cadre de l'évaluation d'un système informatique de commandes. De plus en plus, les travaux s'intéressent ainsi à la façon dont les usagers perçoivent l'information qui leur est présentée. Ainsi, des grandes marques et organisations ont réalisé des analyses du suivi du regard sur la page d'accueil de leur site Internet afin de mettre en évidence les éléments les plus visualisés. Certaines marques ont même modifié leur site à l'issue de telles analyses afin de mettre le plus possible en avant les éléments qui leur semblent importants (nom de la marque ou de l'organisation, produits phares, événements importants, etc.). La figure 4.1 illustre ce suivi du regard sur une page d'un site Internet. L'image de gauche correspond aux regards portés sur l'ancienne version de la page d'accueil du site de la police de San Francisco, celle de droite correspond aux regards portés sur une version de cette page d'accueil corrigée afin de faire ressortir des éléments jugés importants et non visualisés dans la première version du site[143].

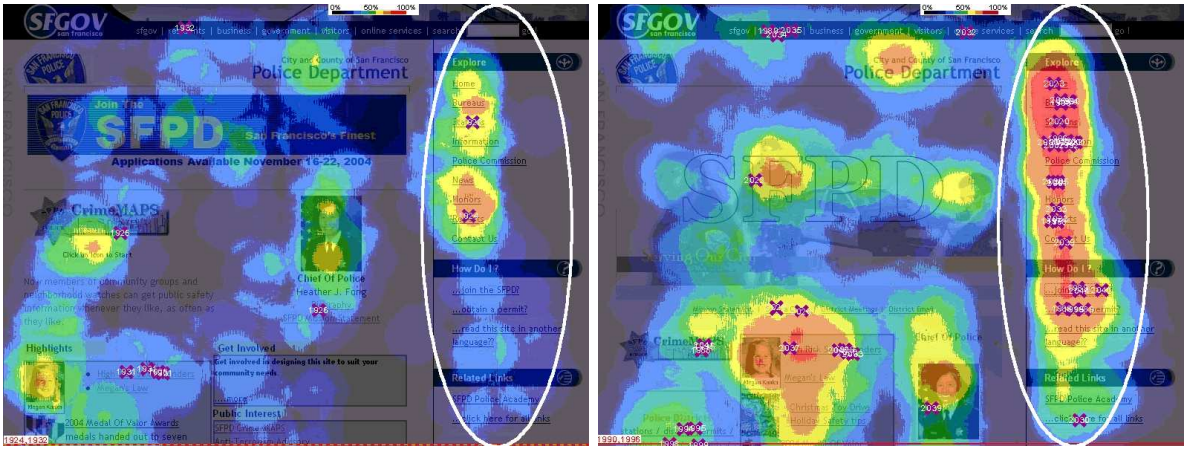


FIG. 4.1 – Illustration des regards portés sur deux versions d'une page d'accueil d'un site Internet : l'originale (à gauche) et la version « optimisée » (à droite).

Ces travaux permettent de mieux tenir compte des particularités des systèmes et de leurs interactions avec les utilisateurs dans les contextes « naturels » de leurs utilisations. Dans le cadre des systèmes centrés sur leurs utilisateurs dans lequel nous situons nos travaux, une évaluation doit donc prendre en considération les différents éléments abordés dans cette section. Ainsi, l'aspect extrinsèque de l'évaluation est selon au moins tout aussi important que l'aspect intrinsèque. Pour prendre en considération un tel aspect extrinsèque dans l'évaluation d'un système, il est par exemple possible d'interroger la satisfaction de ses usagers, de mesurer l'adéquation à la tâche des interfaces et des interactions proposées, d'isoler des zones des interfaces particulièrement

[143] D'autres expériences du même type sont disponibles à l'adresse suivante : `http://blog.eyetools.net/eyetools_research` (page consultée le 2 juillet 2007).

visualisées, etc. Dans la section suivante de ce chapitre, nous traçons les grandes lignes que nous nous sommes fixées pour l'évaluation de notre modèle.

4.1.3 Évaluer des systèmes centrés sur leurs utilisateurs

Le modèle *AIdED* propose de centrer sur l'utilisateur les tâches d'accès au contenu d'ensembles documentaires. Cette place donnée à l'utilisateur a entraîné la conception de logiciels centrés sur ce dernier ou encore individu-centrés[144]. Des tels logiciels sont individus-centrés car ils produisent des sorties d'analyses (dans notre cas, des supports interactifs de visualisation cartographique) dont la création est intégralement dépendante de ressources prises en entrée décrivant le point de vue de l'utilisateur. Pierre Beust dans [Beust, 2005] définit de tels systèmes et les oppose aux systèmes « technocentrés » :

> *Les systèmes individu-centrés s'opposent aux systèmes dits technocentrés où l'utilisateur n'a principalement qu'un rôle d'interprétation des résultats de la machine sans qu'il puisse déterminer la façon de les produire. Pour qu'un système soit effectivement individu-centré, il faut que ses traitements soient déterminés par l'expression d'un point de vue particulier, celui de son utilisateur, sur une tâche particulière. II ne s'agit pas simplement que 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.*

Dans les systèmes individu-centrés, les sorties retournés à l'utilisateur ne constituent pas une finalité mais elles doivent plutôt être considérées comme des regards possibles sur la tâche visée à un moment donné. De tels regards évolueront en même temps que le point de vue de l'utilisateur sur sa tâche.

Proposer des éléments d'évaluation de tels systèmes individu-centrés entraîne donc de prendre en considération différents aspects qui leur sont propres. Nous proposons de décomposer l'évaluation de systèmes individu-centrés, et plus particulièrement de l'instrumentation logicielle d'*AIdED*, selon les trois étapes suivantes :

1. *Évaluation de la phase de construction des ressources décrivant le point de vue de l'utilisateur :*
   Durant cette étape, il faudra évaluer si la construction des ressources décrivant le point de vue de l'utilisateur sur la tâche qu'il souhaite accomplir est facilement réalisable dans un temps raisonnable. Il faudra également mettre en évidence si ce dernier est correctement assisté durant cette phase.
2. *Évaluation de la phase d'exécution du logiciel :*
   Dans cette étape, il faudra s'intéresser au délai nécessaire à l'exécution du logiciel à partir des entrées spécifiées par l'utilisateur. Il sera également utile de mesurer l'espace mémoire nécessaire à l'exécution du logiciel.
3. *Évaluation des sorties produites par le logiciel :*
   Durant cette phase de l'évaluation, il faudra demander à chaque utilisateur si les résultats retournés sont pertinents pour lui et en relation étroite avec le point de vue qu'il a exprimé précédemment. Par la suite, il faudra évaluer si les méthodes de visualisation utilisées pour présenter les résultats sont facilement exploitables par l'utilisateur. Enfin, il faudra évaluer comment les résultats obtenus seront appropriés par l'utilisateur et comment ils pourront être exploités par ce dernier afin de faire évoluer correctement la représentation de son point de vue sur cette tâche.

[144] Théodore Thlivitis parle également dans [Thlivitis, 1998] de systèmes « anthropocentrés ».

Au cours d'un travail précédent, s'intéressant à la première étape définie ci-dessus, une évaluation de la phase de construction des ressources caractérisant le point de vue de l'utilisateur a été menée. L'atelier de formation du CNRS « Variation, construction et instrumentation du sens » (juillet 2002, île de Tatihou, Manche) a ainsi été l'occasion de mettre en place une expérimentation portant sur la création de RTO *LUCIA* (cf. [Perlerin et Beust, 2003] pour plus de détails sur cette expérimentation). L'objectif était de tester la capacité d'utilisateurs novices à s'approprier les principes généraux du modèle *LUCIA* (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, il a été 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 type Zipf sur l'ensemble des articles traitant de la bourse et de l'économie de laquelle les éléments non verbaux et non substantivaux avaient été enlevés (cette liste contenait par exemple des lexies comme *action*, *back office*, *dévaluation*, *OPA* ou encore *palais Brogniart*). 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).

À l'issue des deux séances d'expérience, tous les participants ont au moins proposé des groupes de lexies, précisé les différences qu'ils considéraient comme 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 les participants, il a tout d'abord été estimé 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.

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, il a ainsi été mis en évidence que la méthode de construction des RTO *LUCIA* s'acquiert rapidement et que les principes qui la régissent sont facilement assimilables.

Dans cette première phase d'évaluation, les nouveaux outils logiciels pour l'aide à la construction de RTO *LUCIA* que nous avons présentés au chapitre 3 doivent être pris en considération. De même, des éléments d'évaluation portant sur la phase d'exécution des logiciels (étape 2) et les sorties produites pour nos logiciels (étape 3) doivent être apportés. Selon nous, différentes expérimentations de notre modèle doivent être menées afin de contribuer à son évaluation pour chacune des trois étapes énoncées précédemment. Dans chaque expérimentation, des ressources représentant les domaines d'intérêt des utilisateurs sur la tâche visée seront construites puis projetées dans l'ensemble documentaire considéré. Les sorties d'analyses produites seront ensuite manipulées par les utilisateurs qui exprimeront des opinions sur leur pertinence.

Dans les sections suivantes de ce chapitre, nous abordons un certain nombre d'expérimentations ayant toutes pour objectif de mettre à l'épreuve le modèle *AIdED* et de mettre ou non en évidence sa valeur ajoutée. Dans un premier temps, nous présentons différentes utilisations de notre instrumentation logicielle dans le but d'interroger la pertinence des informations et des interactions proposées pour l'accès au contenu d'ensembles documentaires. Dans un second temps, nous interrogerons plus particulièrement la lisibilité des interfaces cartographiques que nous proposons auprès d'un plus large panel de sujets. Un retour sur ces différentes expérimen-

tations est finalement présenté afin de regrouper les différents éléments permettant de mesurer la valeur ajoutée par le modèle *AIdED* et son instrumentation dans des tâches d'accès personnalisé au contenu d'ensemble documentaire.

4.2 Recherche documentaire personnalisée

Les premières expérimentations que nous présentons prennent place dans le domaine de la recherche documentaire. La première a un objectif d'assistance dans une recherche d'information « traditionnelle » sur Internet alors que la seconde a des visées beaucoup plus spécifiques en assistant une recherche d'information dans des documents médicaux.

4.2.1 Assistance dans une recherche d'information généraliste sur Internet

Contexte de l'expérimentation

La tâche que nous visons dans cette expérience est une tâche de recherche d'information sur Internet. À la manière des métamoteurs de recherche *KartOO* et *MapStan*, présentés au chapitre 1 de cette thèse, nous proposons à l'utilisateur d'avoir un regard global sur les résultats d'un ou plusieurs moteurs de recherche pour une requête donnée. Comme nous l'avons déjà évoqué précédemment, un tel regard permet une appréhension globale des différents éléments d'un ensemble documentaire et facilite ainsi l'accès à son contenu. Nous proposons que ce regard global prenne la forme de cartes construites à partir de l'ensemble documentaire formé par les pages retournées par un ou plusieurs moteurs de recherche selon une requête donnée. Ces cartes seront alors élaborées à l'aide des domaines d'intérêt de l'utilisateur sur sa tâche de recherche d'information

La plate-forme *ProxiDocs*, telle que nous l'avons détaillée au chapitre 3, permet l'analyse d'ensembles documentaires présents localement sur la machine. Au cours de ce travail de thèse, un projet d'étudiants en Informatique à l'Université de Caen a permis de réaliser une application permettant d'interroger différents moteurs de recherche pour une requête donnée et de cartographier, *via* des composants de *ProxiDocs*, les N premières pages retournées selon des domaines choisis par l'utilisateur [Authesserre et Courvalet, 2005]. Cette application, prenant la forme de servlets *Java Server Pages*, invite l'utilisateur à saisir les différentes informations décrivant sa tâche de recherche documentaire :

- le ou les mots-clés de la requête ;
- le ou les moteurs de recherche (*Google*, *Tiscali*, *Yahoo*, *Altavista*, *Msn Search* et *Lycos*)[145] ;
- la langue (pour l'instant, uniquement le français ou l'anglais) ;
- le nombre N de pages souhaitées par l'utilisateur ;
- ainsi que la configuration traditionnelle de la plate-forme *ProxiDocs* (domaines d'intérêt, méthode de projection, méthode de classification, emplacement des cartes produites, etc.).

Nous ne donnons pas ici les détails techniques liés à la réalisation de cette application, nous renvoyons à [Roy et Beust, 2006] pour plus d'informations.

L'objectif que nous nous sommes fixé ici est une recherche d'information sur Internet dans le large contexte des décisions européennes. Une telle recherche a été effectuée selon différents domaines d'intérêt que nous avons construits dans le cadre de cette expérimentation. Nous avons

[145] Dans le cas d'une interrogation portant sur plusieurs moteurs de recherche, la plate-forme distribue la recherche sur les différents moteurs et uniformise les liens obtenus afin d'éliminer la redondance et d'obtenir le nombre de pages souhaité par l'utilisateur.

choisi d'effectuer cette recherche sur des documents de langue anglaise. Les RTO *LUCIA* représentant les domaines seront donc également dans cette langue. Les domaines que nous avons choisis sont les suivants : *computer science* (informatique), *farming* (agriculture), *pollution* (pollution), *road safety* (sécurité routière), *space* (espace) et *sport* (sport). Un dispositif *LUCIA* est associé à chacun de ces domaines. Ces dispositifs sont constitués de 3 à 5 tables et regroupent de 30 à 60 graphies choisies manuellement, sans l'aide d'outil logiciel. La construction des dispositifs s'est fait à l'aide de *VisualLuciaBuilder*. Certains attributs sont communs à plusieurs dispositifs, comme c'est le cas pour l'attribut *Link with domain* (rapport au domaine) avec les valeurs *Object* (objet), *Agent* (agent) et *Phenomenon* (phénomène) ou encore pour l'attribut *Evaluation* (évaluation) avec les valeurs *Good* (bien) et *Bad* (mauvais).

L'ensemble des dispositifs exploités durant cette expérimentation sont décrits en annexe C de cette thèse. La figure 4.2 présente plus particulièrement le dispositif associé au domaine de l'informatique utilisé durant l'expérimentation.

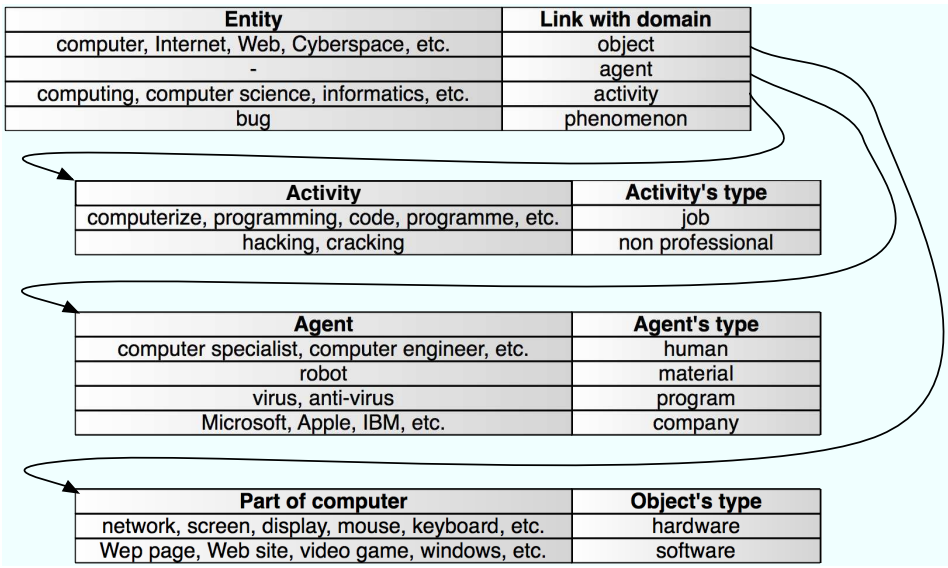


FIG. 4.2 – Le dispositif de l'informatique utilisé durant cette expérimentation.

Quatre tables sont utilisées : la table principale nommée *Entity* (entité) et trois autres tables, respectivement nommées *Part of Computer* (élément d'un ordinateur), *Agent* (agent) et *Activity* (activité). Chacune de ces trois tables est reliée à la table principale. Par exemple, la table *Part of Computer* est reliée à la première ligne de la table *Entity*. De cette manière, toutes les graphies de *Part of Computer* héritent de l'attribut *Link with domain* (rapport au domaine) avec la valeur *Object*.

Afin de constituer l'ensemble documentaire à analyser, le terme-clef *European decision* (décision européenne) a été choisi pour constituer l'objet de la recherche. Le moteur de recherche *Yahoo* a été sélectionné dans l'application ainsi que les résultats en anglais. Le nombre de pages à prendre en considération a été fixé à 150. Seuls les documents au format HTML, PDF et DOC ont été retenus. Pour ces documents, seules les parties textuelles ont été analysées. L'ensemble documentaire ainsi constitué contient donc 150 textes dont la taille varie entre 1 000 et 50 000 tokens. Les différentes vues de cet ensemble documentaire prenant en considération les domaines définis précédemment sont détaillées dans la suite de cette sous-section.

Vues globales sur l'ensemble documentaire

La figure 4.3 contient un nuage des graphies des dispositifs présentes dans l'ensemble documentaire. Ici, les graphies du domaine de l'informatique sont particulièrement présentes avec *programme*[146] (programmer), *network* (réseau), *Microsoft*, *software* (logiciel), *etc.* Certaines graphies du domaine de la pollution et de l'agriculture sont également mises en évidence par le nuage.

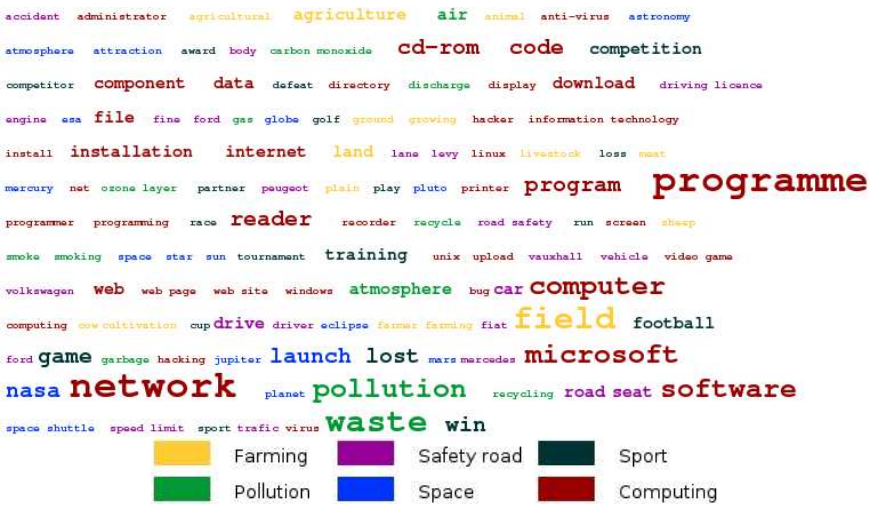


FIG. 4.3 – Nuage de graphies de l'ensemble documentaire.

Ce nuage constitue une première analyse de l'ensemble documentaire et permet d'avoir un premier aperçu des domaines les plus présents. Pour poursuivre cette analyse, nous proposons en figure 4.4 une carte des textes de l'ensemble documentaire. Cette carte est le résultat d'une ACP où les deux premières composantes ont été choisies, la méthode de comptage relative ayant été choisie auparavant. Le choix de cette méthode de projection est lié à la meilleure dispersion des points qu'elle a pu mettre en évidence sur la carte. Le taux d'inertie associé à cette ACP est proche de 40%, ce qui implique que 40% de l'information comprise dans l'espace de départ est représentée dans l'espace d'arrivée présenté par la carte. La carte révèle que chaque domaine est représenté dans l'ensemble documentaire avec une légère prédominance du domaine de l'informatique dans le quadrant inférieur droit. À noter également que le domaine de la pollution n'est majoritaire que dans quatre textes de l'ensemble.

La figure 4.5 met en évidence le résultat d'une CHA en 18 groupes (nombre choisi empiriquement). La présence majoritaire du domaine de l'informatique se confirme sur cette carte. Le plus grand disque (manuellement annoté « groupe 1 », de même pour les groupes 2 et 3) sur la carte représente un groupe de textes majoritairement dans ce domaine. Le groupe 2 aborde plus particulièrement le domaine de la pollution alors que le groupe 3, comme le groupe 1, contient des textes traitant plutôt du domaine de l'informatique. Le fait que le groupe 2 soit majoritairement du domaine de la pollution est assez surprenant. Comme nous l'avons évoqué précédemment, seuls quatre textes sont principalement dans ce domaine. Le rassemblement d'un plus grand nombre de textes dans le groupe 2 a fait remonter au premier plan le domaine de la pollution, pourtant non majoritaire dans la plupart des textes qu'il contient. Cette remontée

[146]Nous avons fait « cohabiter » dans le dispositif de l'informatique les graphies *programme* (orthographe britannique) et *program* (orthographe américaine) pour désigner le verbe français *programmer*. À noter également que le nom français *programme* a pu être désigné par ses équivalents anglais *programme* et *program*.

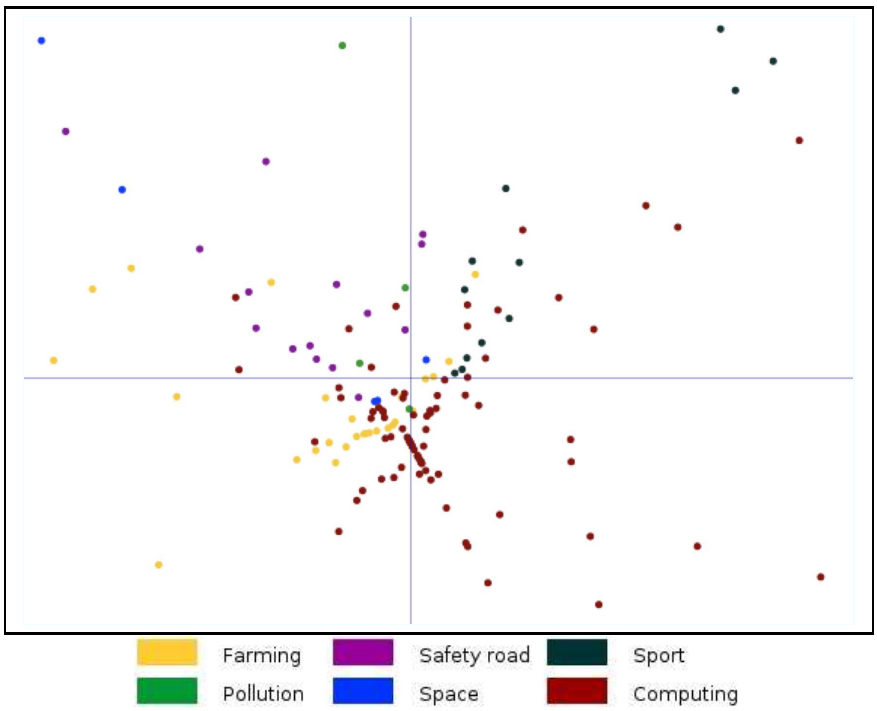


FIG. 4.4 – Carte des textes de l'ensemble documentaire.

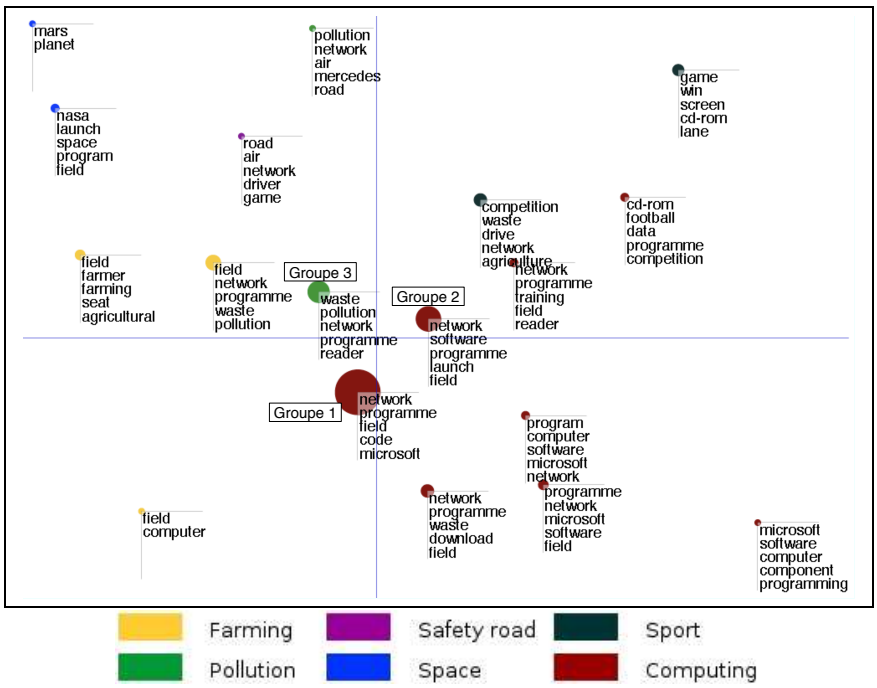


FIG. 4.5 – Carte des groupes de textes de l'ensemble documentaire.

se justifie par la présence significative du domaine de la pollution dans chacun des textes du groupe 2. La « somme » de ces présences a dépassé celles des autres domaines, abordés pourtant majoritairement dans certains textes, mais absents dans d'autres.

Le tableau 4.1 décrit les particularités de ces trois groupes (les informations du tableau sont données dans les rapports de groupes accessibles sur la carte). Les différentes informations contenues dans ce tableau orientent l'interprétation des différents groupes de textes.

Groupe	Nombre de textes	Nombre d'occurrences total de graphies					
		Farming	Pollution	Road Safety	Space	Sport	Computing
Groupe 1	67	100	24	17	19	20	401
Principales graphies	1. *network*, 2. *programme*, 3. *field*						
Principales isotopies inter-textuelles	1. *Object's type* , 2. *Activity's type* (valeurs respectives les plus fréquentes : *produced* et *job*)						
Groupe 2	15	80	443	116	21	58	213
Principales graphies	1. *waste*, 2. *pollution*, 3. *network*						
Principales isotopies inter-textuelles	1. *Link with domain*, 2. *Evaluation* (valeurs respectives les plus fréquentes : *agent* et *bad*)						
Groupe 3	20	37	14	24	39	79	392
Principales graphies	1. *network*, 2. *software*, 3. *programme*						
Principales isotopies inter-textuelles	1. *Object's type*, 2. *Link with domain* (valeurs respectives les plus fréquentes : *used* et *activity*)						

TAB. 4.1 – Détails des trois groupes marquées de 1 à 3 sur la carte des groupes de documents de l'ensemble documentaire présentée en figure 4.5.

Ainsi, le groupe 1, très majoritairement du domaine de l'informatique, a comme graphies les plus fréquentes *network*, *programme*, *field*. En parcourant les textes du groupe, nous avons observé que l'informatique était abordée en relation avec de nombreux domaines non représentés par les dispositifs, tels la santé, la politique, la diffusion de l'information, etc. L'informatique est ici plutôt un moyen de communication entre différentes personnes dans différentes activités liées à la Communauté Européenne, comme l'illustre l'extrait suivant :

> *Network ETUE-net II is helping to develop the use of computer-based communication within the European trade unions.* (Le réseau ETUE-net II aide actuellement au développement de l'utilisation de moyen de communication informatisé à l'intérieur des syndicats européens de salariés.) - `http://education.etui-rehs.org/en/projects/etue-net_II` (page consultée le 4 juillet 2007)

Les principales isotopies inter-textuelles détectées (*Object's type* et *Activity's type* avec respectivement comme valeurs les plus fréquentes *produced* et *job*) vont dans ce sens. La notion d'objet est présente ainsi que la notion de travail. Ces notions sont par contre assez générales et n'orientent pas de façon très fine l'interprétation.

Le groupe 2 est majoritairement du domaine de la pollution mais les domaines de l'informatique et de la sécurité routière y sont également abordés. Les deux principales isotopies inter-textuelles du groupe sont *Link with domain* et *Evaluation* avec comme valeurs les plus fréquemment associées *agent* et *bad*. Dans ce cas, les isotopies orientent bien l'interprétation des textes du groupe. Les différents textes du groupe traitent des décisions de la Communauté Européenne sur le développement durable ou plutôt du besoin de décisions sur ce sujet. L'extrait suivant illustre cette idée :

> *Over the last decade there has been a considerable increase in research activities that claim to address sustainable development.* (À travers la dernière décennie, il y a eu une augmentation considérable des activités de recherche prétendant s'intéresser au développement durable.) - `http://www.edis.sk/ekes/eur20389en.pdf` (page consultée le 4 juillet 2007)

L'isotopie inter-textuelle classée en troisième position du groupe (*State* avec comme valeur la plus fréquente *gas*) oriente encore un peu plus l'interprétation, les rejets de gaz à effet de serre étant significativement abordés dans les textes. Dans ce groupe, la pondération des isotopies inter-textuelles a été particulièrement utile, puisqu'elle a permis de faire remonter dans le classement les isotopies *Evaluation* (de la troisième à la deuxième position) et *State* (de la sixième à la troisième position) tout en déclassant des isotopies plus génériques moins appropriées, comme *Object's type* ou *Agent's type*.

Comme le groupe 1, le groupe 3 est très majoritairement dans le domaine de l'informatique. Contrairement au premier, les textes qu'il contient n'abordent pas le domaine de l'informatique de façon secondaire. Ces textes traitent de l'informatique comme un domaine principal, et abordent plus particulièrement les décisions de la Communauté Européenne en réponse à la situation de monopole imposée par *Microsoft*, comme dans l'extrait suivant :

> *The European Commission's Microsoft antitrust decision, released earlier today, is potentially important to enterprise customers because of what it says about server-to-server interoperability.* (La décision de la Commission Européenne sur la situation de monopole de Microsoft, parue plus tôt dans la journée, est potentiellement importante pour les clients d'entreprise sur ce qu'elle annonce pour l'interopérabilité entre serveurs.) - `http://esto.jrc.es/docs/AnnexDInfoMembership.doc` (page consultée le 4 juillet 2007)

Les principales isotopies inter-textuelles du groupe (*Object's type* et *Link with domain* avec comme valeurs respectives les plus fréquentes *used* et *activity*) orientent légèrement l'interprétation des documents du groupe mais restent tout de même générales et très liées aux isotopies inter-textuelles parcourant l'ensemble documentaire dans sa globalité.

Vers une recherche documentaire personnalisée et interactive

Cette expérimentation avait pour objectif d'illustrer l'intérêt de notre modèle dans une tâche usuelle de recherche d'information. Les différentes cartes et analyses présentées ont permis d'isoler des groupes de textes se positionnant dans des domaines particuliers et abordant certains sujets. Ces guides à l'interprétation apportent ainsi une valeur ajoutée importante en personnalisant la tâche et en proposant différents niveaux de visualisation[147]. Dans le cadre de recherches répétées sur Internet (l'application de notre modèle d'analyse ne nous semble pas adaptée pour des tâches de recherche documentaire très ponctuelles, principalement à cause de la construction des RTO *LUCIA* qui demande un certain effort), le modèle *AIdED* peut ainsi être efficacement appliqué pour accéder au contenu de l'ensemble documentaire retourné par le ou les moteurs de recherche.

Cette expérimentation prend place dans une problématique très proche des travaux de Vincent Perlerin sur la recherche documentaire. Ainsi, [Perlerin, 2001] propose d'utiliser des RTO personnelles différentielles afin d'effectuer un filtrage et un réordonnancement des résultats d'un moteur de recherche généraliste. Une telle idée est intéressante et pourrait être exploitée dans les vues

[147] L'idée de proposer une vue globale sur les résultats de moteurs de recherche n'est pas neuve, mais nous pouvons tout de même remarquer que la plupart des outils (tels *Kartoo* et *Mapstan*) proposent des vues assez rapidement limitées en nombre d'éléments (avec pas plus de 30 éléments par vue) et également en interactions.

que nous proposons : au lieu de prendre les N premiers résultats du moteur de recherche, nous pourrions considérer les N premiers résultats du moteur de recherche jugés pertinents du point de vue des RTO *LUCIA* de l'utilisateur.

Le schéma d'évaluation de notre modèle donne les éléments suivants pour cette expérimentation :

- *Étape 1 : phase de constitution des ressources*
  Les dispositifs utilisés dans cette expérimentation sont assez simples. Le temps nécessaire à leur construction a été assez réduit, de l'ordre d'une dizaine de minutes par dispositif. Les graphies placées dans les dispositifs ont été choisies selon qu'elles se rapportaient, de notre point de vue, aux domaines considérés.
- *Étape 2 : phase d'exécution des outils*
  La phase d'exécution des outils a été assez variable, l'interrogation du ou des moteurs de recherche sélectionnés entraînant des délais plus ou moins longs. Dans notre expérimentation, moins d'une minute a été nécessaire pour constituer l'ensemble documentaire avec les 150 premières pages retournées par le moteur de recherche *Yahoo*. Ce délai est d'autant plus long que le nombre de pages désirées par l'utilisateur est important. La construction des cartes de l'ensemble documentaire selon les six dispositifs est également de l'ordre de la minute. Là également, ce délai s'allonge avec la taille de l'ensemble documentaire.
- *Étape 3 : phase d'évaluation des sorties produites*
  La phase d'analyse des cartes a été assez courte, de l'ordre d'une vingtaine de minutes. Les cartes en 2 dimensions ont été principalement analysées et elles ont permis d'isoler assez rapidement les principaux sujets abordés dans les textes de l'ensemble documentaire.

Dans la section suivante, nous poursuivons l'application du modèle *AIdED* dans le domaine de la recherche documentaire. Cette fois-ci, la recherche n'est pas généraliste et est fortement ciblée sur la problématique de l'accès à des documents médicaux.

4.2.2 Assistance dans une recherche d'information médicale

L'expérimentation précédente prenait place dans le cadre d'une tâche habituelle de recherche documentaire sur Internet, où nous étions à la fois les concepteurs et les expérimentateurs. Ce double rôle concepteur/expérimentateur nous a permis d'avoir un premier retour sur nos propositions et leurs implémentations, mais a forcément introduit un biais dans l'évaluation. Pour remédier à cela, les prochaines expérimentations présentées prennent toutes place dans des contextes d'utilisations « réelles », définis dans le cadre de collaborations avec des utilisateurs extérieurs ayant des besoins particuliers en accès au contenu d'ensembles documentaires.

La première de ces expérimentations en contexte réel a été réalisée en collaboration avec Aurélie Névéol[148], travaillant sur la recherche d'information dans le domaine biomédical. Dans son travail de thèse [Névéol, 2005], Aurélie Névéol s'est intéressée à l'automatisation et à la répartition entre l'homme et la machine de tâches d'accès au contenu d'un catalogue de documents médicaux. L'objectif de cette collaboration, et de l'expérimentation qu'elle a impliquée, était alors principalement de mettre en avant l'intérêt de vues cartographiques sur des ensembles de documents médicaux en prenant en considération certaines particularités de ce domaine.

[148] Aurélie Névéol a réalisé sa thèse de doctorat en Informatique à l'INSA de Rouen, elle est actuellement en post-doctorat à la *National Library of Medicine* (Bethesda, États-Unis) dans l'équipe *Indexing Initiative*.

Contexte

Comme dans la plupart des disciplines, les documents scientifiques dans le domaine de la santé ne sont donc pas épargnés par l'essor du numérique. Plusieurs projets se donnent pour objectif de guider les utilisateurs dans leur recherche d'information en santé. Ainsi, la fondation Suisse HON[149] propose un portail vers une information de santé de qualité dans plusieurs langues européennes. La base documentaire MEDLINE[150] recense une grande partie des publications scientifiques dans le domaine biomédical. Depuis 1995, le Catalogue et Index des Sites Médicaux Francophones[151] recense des documents de santé institutionnels à l'usage des professionnels de santé, des étudiants en médecine et du grand public.

Afin de retrouver des informations pertinentes dans de tels ensembles documentaires, les méthodes traditionnelles interrogent des bases documentaires à l'aide de mots-clés. Comme nous l'avons déjà évoqué précédemment, l'objectif à travers cette expérimentation est de proposer à des experts de la santé (médecins, documentalistes, chercheurs, etc.) des vues globales sur des ensembles de documents médicaux. De telles vues sont construites à partir de domaines, ou plutôt de spécialités médicales et biologiques propres au domaine de la santé.

Corpus d'étude et domaines d'intérêt

Pour cette étude, nous avons travaillé avec un corpus de 70 documents en langue française extraits aléatoirement du catalogue CISMeF. Chaque document du corpus comporte une indexation à l'aide de descripteurs du thésaurus MeSH (Medical Subject Headings), nous renvoyons à [Névéol, 2005, pages 105 à 113] pour plus de détails sur l'indexation MeSH. Une telle indexation se présente sous la forme d'une liste pondérée de mots-clés ou de paires mot-clé / qualificatif issus du MeSH, elle permet de caractériser assez finement les concepts abordés dans chaque document considéré de manière isolée. La pondération « majeur » dénote les thèmes traités en profondeur dans le document, et la pondération « mineur » signale les thèmes traités plus succinctement. La figure 4.6 donne un exemple de document et de son indexation MeSH.

L'idée de ce travail est de proposer un accès plutôt d'ordre thématique au corpus en tenant compte de ce qui est appelé « métaterme » dans la terminologie CISMeF[152]. La caractérisation des documents du corpus en terme de métatermes est effectuée grâce à un outil bibliométrique [Darmoni *et al.*, 2005] utilisant récursivement l'algorithme de classification décrit dans [Névéol *et al.*, 2004]. Cet algorithme est fondé sur l'indexation MeSH des documents en listes pondérées de mots-clés ou de paires mot-clé / qualificatif, et exploite les liens sémantiques existant entre les mots-clés MeSH et les métatermes d'une part, les qualificatifs MeSH et les métatermes d'autre part. Ainsi, chaque descripteur MeSH attribué à un document permet de le caractériser avec le ou les métatermes auxquels renvoient le descripteur. Par exemple, un document indexé avec le mot-clé *<diabète>* relève du métaterme « endocrinologie ». Le score attribué à « endocrinologie » sera de 100 si *<diabète>* est un thème majeur pour le document et de 1 si c'est un thème mineur[153].

[149] Health On the Net - `http://www.hon.ch` (page consultée le 29 juin 2007).

[150] Page descriptive de MEDLINE : `http://www.nlm.nih.gov/pubs/factsheets/medline.html` (page consultée le 29 juin 2007). Le moteur de recherche PubMeb permet d'effectuer des recherches dans la base MEDLINE : `http://www.pubmed.gov` (page consultée le 29 juin 2007).

[151] CISMeF - `http://www.cismef.org` (page consultée le 29 juin 2007).

[152] Nous renvoyons à l'adresse suivante pour une description des différents métatermes proposés : `http://www.chu-rouen.fr/ssf/santspe.html` (page consultée le 29 juin 2007).

[153] Le terme « score » utilisé dans cette expérimentation reprend la terminologie CISMeF et désigne l'importance d'un métaterme dans un document ou dans un ensemble de documents. Il ne correspond donc pas au terme « score » utilisé pour désigner l'importance d'une isotopie intra ou inter-textuelle.

(...)
La prophylaxie antibiotique chez les enfants
Comité des maladies infectieuses et d'immunisation, Société canadienne de pédiatrie (SCP)
(...)
Le présent énoncé remplace l'énoncé sur la prophylaxie antibiotique publié par la Société canadienne de pédiatrie en 1982 (1). Il vise à orienter les pédiatres généralistes et les médecins de famille. Il traite de la plupart des situations dans lesquelles on fait appel à la prophylaxie antibiotique pour traiter les enfants, mais non de toutes. Il exclut les prophylactiques antiviraux ou antiparasitaires. Depuis la publication du dernier énoncé, d'importants changements se sont produits dans le domaine de la prophylaxie antibiotique. Certains s'expliquent par les résultats d'essais cliniques, tandis que d'autres proviennent des préoccupations relatives à l'évolution et à la propagation des bactéries antibiorésistantes. (...)

```
<PubmedArticle>
  <CISMeF_ID>64</CISMeF_ID>
  <MeshHeadingList>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N"> adolescent
</DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N"> child
</DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N"> child,
preschool </DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N"> infant
</DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="N"> infant,
newborn </DescriptorName>
    </MeshHeading>
    <MeshHeading>
      <DescriptorName MajorTopicYN="Y"> antibiotic
prophylaxis </DescriptorName>
    </MeshHeading>
  </MeshHeadingList>
</PubmedArticle>
```

FIG. 4.6 – En partie gauche de la figure, un exemple de document constituant le corpus, en partie droite, son indexation MeSH au format XML PubMed.

La figure 4.7 présente les 20 métatermes ayant obtenu les scores les plus élevés dans le corpus d'étude. Au total, 78 métatermes sont apparus dans le corpus parmi les 115 définis dans CISMeF. À partir du classement des métatermes établi avec la méthode précédente sur le corpus d'étude, nous obtenons une information globale sur cet ensemble. Le même traitement peut alors être réalisé, non plus au niveau global du corpus, mais au niveau local du document afin de caractériser ce dernier à l'aide de scores associés aux métatermes qu'il contient.

Specialities	Score	Majors		Minors		PubMed link	CISMeF link
		Keywords	Qualifiers	Keywords	Qualifiers		
therapeutics	6620	9	56	21	99	PubMed	CISMeF
infectiology	5203	51		103		PubMed	CISMeF
virology	4652	46		52		PubMed	CISMeF
diagnosis	2858	17	11	31	27	PubMed	CISMeF
epidemiology	2760	10	17	32	28	PubMed	CISMeF
statistics	2659	8	18	28	31	PubMed	CISMeF
bacteriology	2550	25		50		PubMed	CISMeF
preventive medicine	2227	1	21	5	22	PubMed	CISMeF
neurology	2139	21		39		PubMed	CISMeF
pharmacology	2075	5	15	21	54	PubMed	CISMeF
allergology and immunolgy	1717	16	1	17		PubMed	CISMeF
rheumatology	1621	16		21		PubMed	CISMeF
gastroenterology	1520	15		20		PubMed	CISMeF
hepatology	1309	13		9		PubMed	CISMeF
oncology	1113	11		12	1	PubMed	CISMeF
pulmonary disease (specialty)	1013	10		13		PubMed	CISMeF
surgery	1010	6	4	9	1	PubMed	CISMeF
physiology	921	3	6	12	9	PubMed	CISMeF
gynecology	723	7		23		PubMed	CISMeF
obstetrics	723	7		23		PubMed	CISMeF

FIG. 4.7 – Les 20 métatermes les plus présents dans l'ensemble de notre corpus

Contrairement aux autres expérimentations que nous avons pu réaliser, nous avons dû ici adapter très légèrement nos logiciels afin de prendre en considération, non pas les lexies et leurs formes graphiques décrivant un domaine, mais les scores associés à chaque métaterme. Cette adaptation ne nous sort cependant pas de notre problématique de recherche, les métatermes constituant, en quelque sorte, des domaines dans le milieu de la santé et c'est l'ensemble de ces domaines qui a été jugé pertinent par Aurélie Névéol dans la tâche visée. De cette manière, un document est représenté de la façon suivante, sous la forme d'un vecteur de dimension égale aux nombres de métatermes détectés dans le corpus :

$$vecteur_{document} = (score_{therapeutics}(document), score_{infectiology}(document), score_{virology}(document), ...)$$

Si des métatermes apparaissant dans le corpus ne sont pas présents dans le document, des valeurs nulles sont placées aux coordonnées correspondantes dans le vecteur. Ce processus est répété pour chaque document de l'ensemble étudié. Ainsi, un espace à 78 dimensions où les documents prennent place a pu être construit. Les différentes cartes que nous avons pu construire, à partir de cet espace, sont présentées au paragraphe suivant.

Cartographie du corpus d'étude

Une première cartographie du corpus a été dressée avec *ProxiDocs*, nous la présentons en figure 4.8. Cette figure présente une carte des groupes de documents. La méthode de projection de Sammon a été utilisée et une CHA en 12 groupes (nombre choisi empiriquement) a été réalisée. La méthode de Sammon a été choisie du fait qu'elle permettait d'obtenir une carte se répartissant beaucoup mieux dans l'espace, contrairement aux autres méthodes retournant des cartes où les points et disques étaient plus rapprochés les uns des autres. Une telle méthode semble bien adaptée à la projection d'espace de grande dimension (ici, un espace de 78 dimensions) contrairement aux méthodes de l'ACP et de l'AFC semblant moins efficaces sur de tels espaces.

La couleur attribuée à chaque groupe correspond à son métaterme majoritaire, c'est-à-dire celui ayant le score le plus élevé dans les documents du groupe. Chaque groupe est étiqueté par ses cinq métatermes de score le plus élevé. Pour des raisons de lisibilité, nous avons choisi de faire figurer une légende attribuant une couleur aux 15 métatermes majoritaires dans le corpus (et non à l'ensemble des 78 métatermes pris en considération). Cette légende est disponible sur la partie droite de la figure. Cette carte des groupes de documents en 2 dimensions a été jugée plus pertinente et plus lisible par Aurélie Névéol que la carte des documents ou les cartes en 3 dimensions. La classification a permis de mettre en évidence des « familles » de documents partageant de mêmes métatermes.

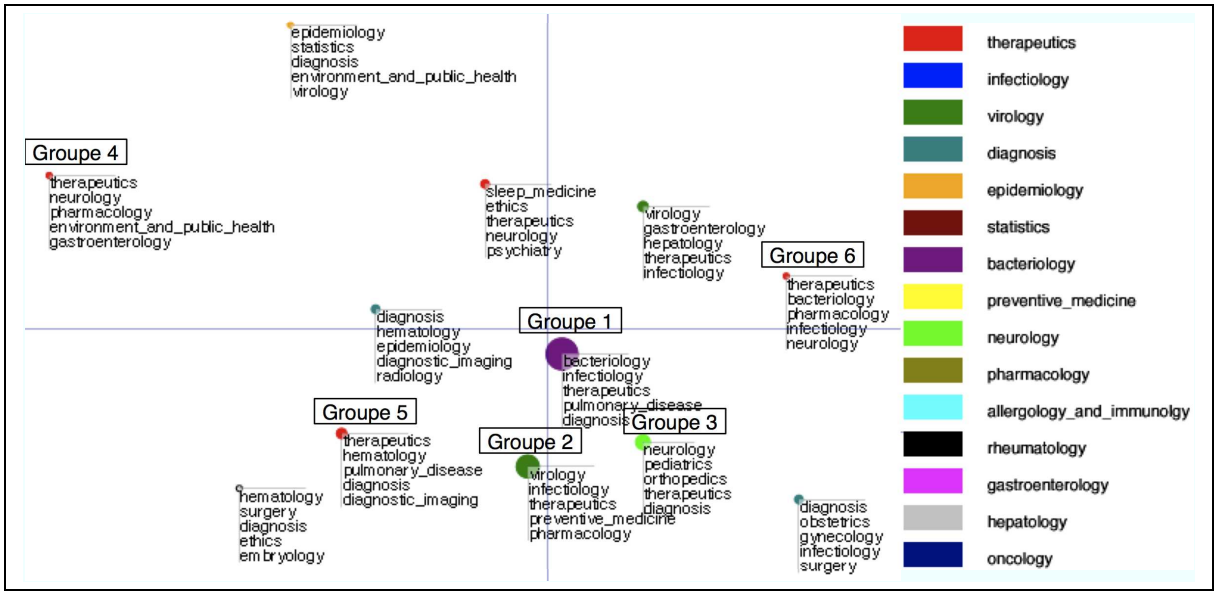


Fig. 4.8 – Carte des groupes de documents obtenus à partir des 78 métatermes présents dans le corpus. Les groupes de 1 à 6 ont été marqués manuellement sur la carte afin d'en faciliter l'analyse.

La répartition des métatermes dans le corpus présentée par la carte est très différente de celle présentée en figure 4.7. Par exemple, le métaterme *therapeutics*, qui possède de façon assez nette le score le plus élevé de l'analyse globale du corpus, apparaît très peu de façon majoritaire sur la carte. Seuls les groupes numérotés de 4 à 6 sont dominés par ce métaterme. La carte permet également de différencier des groupes de documents de même métaterme principal mais ayant des métatermes secondaires très différents. Ainsi, les groupes 4 à 6 ont pour métaterme principal *therapeutics*, mais sont pourtant distants les uns des autres sur la carte. Cet éloignement se justifie par des métatermes secondaires très différents : *environment and public health* et *gastroenterology*

pour le groupe 4, *hematology* et *pulmonary disease* pour le groupe 5 et *bacteriology* et *infectiology* pour le groupe 6. En visualisant les contenus des groupes représentés *via* leurs rapports d'analyse accessibles sur la carte, les différences de sujets abordés dans les documents de chaque groupe semblent tout à fait justifier les différences de métatermes secondaires et la distance entre les groupes sur la carte. En parcourant l'ensemble des groupes de documents de la carte, il semble que le métaterme *therapeutics* constitue une sorte de « trame de fond » : il figure dans les cinq premiers métatermes pour huit groupes, et dans les dix premiers pour trois autres groupes.

Afin d'approfondir l'analyse de la carte, le tableau 4.2 détaille les thématiques traitées par les groupes de plus grande cardinalité (groupes marqués de 1 à 3 sur la figure 4.8).

	Nombre de documents	Métatermes les plus importants du groupe (les scores sont précisés entre parenthèses, le score d'un métaterme correspond à la somme des scores de ce métaterme dans les documents du groupe)
Groupe 1	29	*bacteriology* (1538), *infectiology* (1469), *therapeutics* (739), *pulmonary disease* (406), *diagnosis* (343)
Groupe 2	15	*virology* (2540), *infectiology* (1528), *therapeutics* (1522), *preventive medicine* (905), *pharmacology* (519)
Groupe 3	6	*neurology* (507), *pediatrics* (409), *orthopedics* (404), *infectiology* (306), *diagnosis* (302)

TAB. 4.2 – Description des groupes 1 à 3 de la carte présentée en figure 4.8.

Le groupe 1 est constitué de 29 documents avec comme métaterme majoritaire *bacteriology*, pourtant placé en septième position dans le classement global. Ce métaterme semble principalement employé dans les documents de ce groupe (les deux tiers de son score dans le score global y sont liés). Le métaterme *infectiology* est en deuxième position dans le classement des métatermes du groupe. Le score de ce métaterme est également élevé dans le groupe 2 et le groupe 3. Les scores des métatermes de ce groupe sont assez faibles proportionnellement au nombre de documents qu'il contient. En parcourant ces documents, nous pouvons observer qu'ils abordent des sujets assez disparates, avec des liens entre eux très généraux.

Au contraire, le groupe 2 possède des scores de métatermes très élevés compte tenu de sa cardinalité (ce groupe est constitué de 15 documents). Le métaterme *virology* arrive en tête de classement, suivi des métatermes *infectiology* et *therapeutics* ayant également des scores très élevés. En parcourant les documents du groupe, nous observons qu'ils traitent, à une grande majorité, de vaccination (13 sur 15). Cette thématique est liée aux métatermes obtenus en tête de classement et semble constituer un lien entre les documents du groupe.

Le troisième et dernier groupe détaillé contient 6 documents. Les métatermes en tête de ce dernier figurent très bas dans le classement global des métatermes sur le corpus (*neurology*, premier métaterme du groupe, n'apparaît qu'en neuvième position dans le classement global, les métatermes *pediatrics* et *orthopedics*, respectivement deuxième et troisième du groupe, sont situés en dehors des 20 premiers métatermes du classement global). La lecture des documents de ce groupe révèle qu'elles traitent principalement de médecine pédiatrique dans différents aspects (troubles neurologiques chez l'enfant, chirurgie orthopédique pour l'enfant, etc.).

Une telle carte a été jugée dans un premier temps par Aurélie Névéol comme une bonne illustration du corpus. Les différents regroupements présentés sur la carte lui ont d'abord permis d'avoir un appréhension globale des métatermes partagés entre les documents. Certains groupes (les groupes 2 et 3, par exemple), rassemblant des documents particulièrement proches du point de vue des métatermes abordés, ont été jugés comme très pertinents dans le cadre de regroupements

thématiques de documents. Des métatermes particulièrement partagés par différents groupes ont aussi été identifiés. Ces métatermes ont été qualifiés par Aurélie Névéol de « métatermes transversaux » (tel le métaterme *therapeutics*).

Dans le cadre de l'analyse de ce corpus, ces métatermes ont été jugés comme étant assez peu pertinents pour décrire les documents et pour les caractériser et les distinguer les uns des autres. Quinze métatermes transversaux ont ainsi été identifiés : *anatomy*, *diagnosis*, *diagnostic imaging*, *disease transmission*, *economics*, *education*, *epidemiology*, *ethics*, *history of medicine*, *histology*, *organization and administration*, *pathology*, *patient*, *statistics* et *therapeutics*. Un métaterme transversal ne constitue pas un « tout », il ne correspond pas vraiment un domaine médical à part entière et bien délimité, mais intervient plutôt en association avec un grand nombre d'autres métatermes. De tels métatermes constituent en quelque sorte des fonds sémantiques communs aux documents du corpus. Dans la suite de l'expérimentation, il a été ressenti un besoin de construire une nouvelle carte du corpus ne prenant pas en considération les métatermes tansversaux. La figure 4.9 représente cette carte.

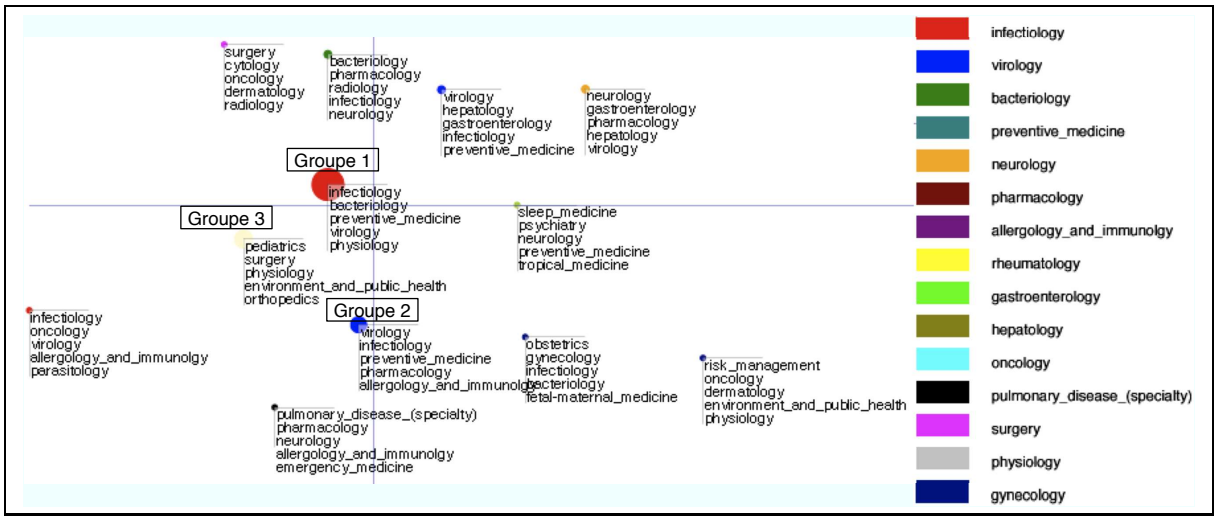


FIG. 4.9 – Carte des groupes de ressources obtenues à partir des métatermes non transversaux présents dans le corpus. Les groupes de 1 à 3 ont été numérotés manuellement sur la carte afin d'en faciliter l'analyse.

Sur cette carte, construite également avec la méthode de Sammon et une CHA en 12 groupes, le groupe 1 possède 36 documents, les trois métatermes majoritaires dans ce groupe sont *infectiology*, *bacteriology* et *preventive medecine*. Un parcours rapide des documents de ce groupe révèle qu'ils abordent des thématiques assez variées, mais cependant liées aux métatermes principaux assez génériques étiquetant le groupe. Le groupe 2 contient 11 documents et possède comme métatermes majoritaires *virology*, *infectiology* et *preventive medecine*. Il regroupe des documents tous étroitement liées au domaine de la virologie, comme des documents traitant du virus de la grippe et des différents vaccins existants contre ce virus. Le groupe 3 contient également 11 documents, ses trois métatermes principaux sont : *pediatrics*, *surgery*, *physiology*. Les documents de ce groupe sont tous étroitement liés à ces métatermes et plus particulièrement à la médecine pédiatrique et ses différents aspects. Ce lien étroit entre les thématiques abordées dans les documents des groupes et leurs métatermes majoritaires se retrouvent pour la quasi-totalité des regroupements présentés sur la carte.

Cette seconde carte du corpus reste tout de même assez proche de la première carte que nous

avons présentée. Par exemple, il existe des similitudes entre les groupes n°2 des deux cartes et également entre les groupes n°3. D'après Aurélie Névéol, une caractérisation un peu plus fine des groupes est proposée sur la seconde carte. Cette meilleure caractérisation est principalement liée à une meilleure différenciation des groupes. Ne pas avoir considéré les métatermes transversaux dans cette seconde carte a permis de limiter ce phénomène de partage des métatermes entre un grand nombre de groupes.

*Apport d'*AIdED *dans une recherche documentaire médicale*

Cette expérimentation a permis de faire ressortir l'intérêt de prendre en considération la dimension globale du corpus afin de caractériser finement des sous-ensembles de textes. Les cartes ont ainsi révélé une répartition des métatermes très différente de celle obtenue lors de l'analyse globale du corpus présentant les informations sous forme de listes. Des groupes de documents caractérisés par des métatermes majoritaires assez « enfouis » dans le classement global du corpus ont pu être mis en évidence. Ces cartes ont été considérées comme de bonnes illustrations du corpus, permettant un accès aux différents thèmes abordés dans des documents.

Les ressources lexicales exploitées dans cette expérimentation, les métatermes décrivant les documents, étaient de nature assez différente de celles que nous proposons, principalement par la généricité visée. Cependant, le besoin d'adaptation à la tâche et, d'une certaine manière, de personnalisation de ressources lexicales a rapidement été ressenti par Aurélie Névéol. Ainsi, un filtrage de certains métatermes a été réalisé afin d'affiner l'accès au contenu des textes du corpus. Ceci illustre bien l'utilisation itérative de notre modèle et l'enrichissement qui en découle : deux « allers-retours » ont étés réalisés sur le corpus, dans un premier temps avec un ensemble généraliste de métatermes qui a été affiné dans un second temps à l'issue de la première analyse.

Si nous reprenons le schéma d'évaluation de notre modèle en trois étapes que nous proposons, nous obtenons alors pour cette expérimentation :

- *Étape 1 : phase de constitution des ressources*
  Dans cette expérimentation, il n'y pas vraiment eu de construction des ressources. L'outil décrit dans [Darmoni *et al.*, 2005] a été appliqué afin d'extraire l'ensemble des métatermes présents dans chaque document du corpus. Ce sont ces métatermes, et les valeurs numériques leur étant associées par l'outil dans chaque document, qui nous ont servi dans la construction des cartes du corpus, à la manière de cartes construites à partir de domaines représentés par des ensembles de lexies. Le temps d'exécution de cet outil a été assez court sur notre corpus, de l'ordre de la minute.
- *Étape 2 : phase d'exécution des outils*
  La construction des cartes du corpus a été réalisée par la plate-forme *ProxiDocs* dans un temps assez réduit grâce au faible nombre de documents (moins de deux minutes pour la construction des différentes cartes des documents et des groupes de documents).
- *Étape 3 : phase d'évaluation des sorties produites*
  L'analyse par Aurélie Névéol des différentes cartes a tout d'abord commencé par un « rejet » assez rapide de la carte des documents en 2 dimensions et des cartes en 3D, rejet lié principalement à la trop grande quantité d'informations perçues par Aurélie Névéol sur les cartes. Une analyse plus poussée de cartes des groupes a ensuite été réalisée, c'est cette analyse qui a été commentée précédemment.

L'expérimentation détaillée dans cette section a donné lieu à une publication, nous y renvoyons pour d'éventuels compléments [Roy et Névéol, 2006].

Les perspectives offertes par cette expérimentation sont assez nombreuses. La première d'entre elles consiste à proposer un accès cartographique à un catalogue de santé. Un tel catalogue re-

groupe un grand nombre de documents, il serait par exemple très utile de proposer une projection cartographique d'une partie ou de l'ensemble de ces documents à partir de leurs différents métatermes (transversaux ou non).

L'identification et la prise en considération de métatermes transversaux pourraient également être poursuivies. L'objectif de ce travail consisterait à s'intéresser à l'indexation des documents médicaux en retenant des candidats-index qui sont peu ou pas liés à des métatermes transversaux dans l'ensemble documentaire. Un tel travail pourrait permettre d'affiner les éléments indexant des documents selon le contexte global formé par l'ensemble documentaire dans lequel ils se situent.

Une troisième perspective de recherche consiste à se situer plutôt dans un contexte de veille, et à proposer des analyses cartographiques évolutives, toujours en tenant compte des métatermes, sur des documents dans l'actualité médicale[154]. Un tel travail permettrait de mettre en évidence les métatermes abordés au fil du temps.

Cette expérimentation et surtout la collaboration qu'elle a impliquée ont commencé mi-2005 et les dernières avancées ont eu lieu en fin 2006. Les perspectives évoquées précédemment pourront cependant donner lieu à une poursuite de cette expérimentation.

Dans la section suivante de ce chapitre, nous abordons une expérimentation également guidée par une collaboration entre chercheurs. L'objet de cette expérimentation est, par contre, très différent de celui étudié dans cette partie puisque nous nous intéresserons à certaines expressions métaphoriques pouvant apparaître dans des textes. Les visualisations cartographiques seront utilisées pour observer et caractériser leurs usages en corpus.

4.3 Analyse d'expressions métaphoriques

Toujours dans l'objectif de mettre à l'épreuve notre modèle dans des tâches variées d'accès au contenu d'ensembles documentaires, nous nous sommes intéressés à la façon dont ce modèle nous permettait d'accéder dans un ensemble documentaire à certains types d'expressions, et plus particulièrement à des expressions métaphoriques. Nous détaillons dans cette section une telle expérimentation basée sur un travail collaboratif entre différents chercheurs.

4.3.1 Étude de métaphores conceptuelles

Les travaux présentés ici s'inscrivent dans le cadre du projet *IsoMeta*[155], projet inter-équipes au sein du laboratoire GREYC de l'Université de Caen initié par Pierre Beust, Vincent Perlerin (équipe ISLanD[156]) et Stéphane Ferrari (équipe DoDoLa[157]). Le projet *IsoMeta*, commencé en 2001, a pour objectif de mener une analyse de métaphores conceptuelles fondée sur une représentation sémique. Ce projet est soutenu par le pôle pluridisciplinaire « Modélisation en Sciences Cognitives (ModeSCo) » de la Maison de la Recherche en Sciences Humaines (MRSH) de l'Université de Caen / Basse Normandie.

L'objet d'étude concerne le phénomène autant cognitif que linguistique introduit sous le nom de « métaphore conceptuelle » par [Lakoff et Johnson, 1980]. Une métaphore conceptuelle peut être vue comme une projection d'un domaine source sur un domaine cible mettant en saillance certains éléments et en effaçant d'autres. Le principe central est l'existence d'images récurrentes

[154] Tels les documents de la section « Quoi de neuf » de CISMeF : `http://www.chu-rouen.fr/documed/neuf.html` (page consultée le 29 juin 2007).

[155] *IsoMeta* pour « Isotopie et Métaphores ».

[156] Interaction Sémiotique Langage Diagrammes.

[157] Document Données Langue.

donnant lieu à de nombreux emplois métaphoriques dans la langue. Ces emplois ne sont pas les seuls, des comparaisons explicites peuvent les accompagner et le caractère conventionnel de nombreuses métaphores conceptuelles étudiées par Lakoff est aussi souvent lié à l'existence d'emplois figés, de catachrèses. Nous renvoyons à [Ferrari, 2006] pour une présentation plus détaillée de la notion et de travaux qui y sont liés.

Indépendamment de la nature précise des figures de rhétorique sous lesquelles se réalise une métaphore conceptuelle, une telle métaphore offre avant tout l'intérêt de fournir une connaissance *a priori* exploitable en TAL. En effet, il existe un lien privilégié entre le domaine source et le domaine cible d'une métaphore conceptuelle. L'exploitation de ce lien entre domaines source et cible est désormais récurrente pour coder des connaissances préalables sur les métaphores conceptuelles, par exemple dans des lexiques sémantiques de grande taille [Chibout *et al.*, 2001], pour l'enrichissement de la version italienne d'EuroWordNet [Alonge et Castelli, 2003] ou encore [Guazzini *et al.*, 2004], pour l'analogie [Gentner, 1988] et [Veale, 2003], etc.

Dans la première phase d'*IsoMeta*, les instigateurs du projet se sont intéressés à une analyse locale d'une seule métaphore conceptuelle afin de dégager quelle dynamique sémique est à l'œuvre dans les emplois lexicaux associés. Il a été mené une étude sur corpus d'une métaphore conceptuelle conventionnelle spécifique : la « météorologie boursière ». Ces travaux ont permis de dégager une caractérisation des sèmes transportés dans les emplois métaphoriques, en relation avec les domaines source et cible de la métaphore conceptuelle sous-jacente (se reporter à [Perlerin *et al.*, 2005] pour plus de détails sur ces travaux initiaux).

Dans la poursuite de ces premiers travaux limités à l'étude locale d'emplois métaphoriques, l'objectif principal visé par cette expérimentation est de dégager des régularités à l'échelle d'un corpus de textes, en vue de caractériser la dimension intertextuelle des emplois de métaphores conceptuelles. Pour ce faire, nous[158] exploitons le même corpus, constitué d'articles boursiers, présenté en détail dans les lignes suivantes. Nous étudions conjointement dans ce corpus plusieurs métaphores conceptuelles ayant des domaines sources distincts, la météorologie, la santé et la guerre, mais un domaine cible commun, la bourse. De nombreux emplois de ces trois métaphores conceptuelles ont été observés en corpus, de même que des utilisations non métaphoriques des lexiques des trois domaines sources qui y conservent donc une grande partie de leur polysémie. Les exemples suivants l'illustrent sur deux emplois du mot *orage*, respectivement métaphorique et non métaphorique :

> *Paris laisse passer l'orage et campe sur des positions relativement solides.*
> *(...) à la suite du violent orage qui avait éclaté dans la nuit de lundi à mardi derniers, les logiciels, pieds dans l'eau, ont refusé de fonctionner.*

Nos objectifs sont multiples dans cette expérimentation. Tout d'abord, nous cherchons à observer comment se répartissent en corpus les emplois de différentes métaphores conceptuelles ayant un même domaine cible : si les métaphores s'excluent les unes les autres ou si elles apparaissent conjointement dans les textes. Ainsi, nous cherchons à exploiter les vues cartographiques du corpus d'étude afin d'observer les rapports entre textes contenant des emplois métaphoriques et ceux contenant des emplois non métaphoriques du lexique d'un domaine. Enfin, nous proposons d'étudier les liens éventuels entre la présence de certaines métaphores conceptuelles et des faits d'actualité. Tous ces éléments pourront nous aider à interroger la dimension intertextuelle des métaphores conceptuelles étudiées.

[158] Les chercheurs ayant participé à cette deuxième phase du projet *IsoMeta* sont Stéphane Ferrari, Pierre Beust et moi-même.

4.3.2 Ressources et corpus

Afin de représenter les trois domaines sources des métaphores conceptuelles étudiées, nous avons choisi dans un premier temps de regrouper des lexies désignant des entités étroitement liées à chacun de ces domaines. Cette représentation des domaines selon le modèle *LUCIA* simplifié a été exploitée dans un premier temps pour amorcer l'étape de construction des différentes ressources. Les lexies constituant les domaines ont été extraites de notre corpus après avoir attesté leur appartenance à l'un des domaines sources étudiés. Une telle représentation simple des domaines en ensembles de lexies a rapidement été complétée par une représentation des domaines en dispositifs *LUCIA*[159]. Ainsi, au domaine de la santé sont, par exemple, associées les lexies suivantes : *grippe*, *essouflement*, *blessure*, *déprimer*, *guérison*, etc. Les attributs *type de pathologie* avec les valeurs *maladie, infection*, *conséquence d'infection*, *trouble psychologique* et *trouble physique* et *évaluation* avec les valeurs *bien* et *mal* ont entre autres été associés à des lexies de ce domaine dans le dispositif (le premier attribut présenté ici est utilisé uniquement dans le dispositif de la santé alors que le second intervient dans les dispositifs des trois domaines). Le nombre de lexies rassemblées dans les dispositifs varie selon les domaines : 64 lexies pour le domaine de la guerre, 112 pour celui de la météorologie et 111 pour celui de la santé[160]. Le nombre d'attributs varie également selon les domaines : 8 pour la météorologie, 7 pour la guerre et 8 pour la santé. Le dispositif de la météorologie utilisé dans cette expérimentation est très proche de celui utilisé dans la première phase du projet. Les dispositifs de la santé et de la guerre ont, par contre, été créés dans cette deuxième phase du projet *IsoMeta*. Nous renvoyons en annexe C pour une vue détaillée sur les dispositifs utilisés dans cette expérimentation.

À noter que même dans le cadre de cette étude où les domaines sont assez « fermés », la polysémie entre les domaines est encore présente. C'est en effet le cas avec la lexie *dépression* qui peut être associée aussi bien au domaine de la météorologie qu'à celui de la santé. Nous avons choisi dans cette étude de placer cette lexie dans les deux dispositifs. Dans celui de la santé, le couple *attribut : valeur type de pathologie* : *trouble psychologique* est associé à la lexie, alors que dans le domaine de la météorologie, les couples *évaluation* : *mal* et *axe* : *pression* lui sont associés. De telles lexies ambiguës restent cependant très rares, seule la lexie *dépression* appartient à plusieurs des domaines. Ses apparitions en corpus dans 9 articles (1 occurrence de la lexie par article) ne représentent qu'environ 0,5% des occurrences des lexies des domaines étudiés. Les occurrences observées sont de plus assez facilement interprétables du fait des lexies des autres domaines les entourant dans le texte, formant ainsi des isotopies intra-textuelles.

Pour mener nos observations, nous avons constitué un corpus de 303 articles de taille variable (de 200 à 2 000 graphies) issus du journal *Le Monde sur CD-ROM*, tous relatifs au domaine de la bourse et totalisant environ 250 000 graphies. Ce corpus, appelé `Corpus_Travail`, couvre la période 1987 - 1989 et contient 46% de bilans boursiers et 54% de dépêches portant sur l'actualité boursière des trois années. Les bilans sont reconnaissables par leurs titres contenant soit le mot *Bilan*, soit une chaîne *Semaine du X au Y Mois* où *X*, *Y* et *Mois*, sont le jour du début, le jour de fin et le mois de la période considérée. Ce corpus est un sous-ensemble du corpus `Corpus_Total` de 594 articles utilisé dans la phase initiale du projet IsoMeta [Perlerin *et al.*, 2005]. À partir de ce corpus initial, nous avons sélectionné les articles contenant au moins deux occurrences de lexies ou de formes fléchies de ces lexies de nos domaines d'étude (tous domaines confondus),

[159]Les outils *Memlabor*, *FlexiSemContext* et *VisualLuciaBuilder* ont été utilisés pour constituer ces dispositifs. Le premier a permis d'extraire les graphies répétées du corpus, le second a permis une mise en contexte des lexies « candidates » aux dispositifs et le troisième a été utilisé pour construire et réviser les dispositifs.

[160]Dans nos analyses, nous avons également associé aux différentes lexies des dispositifs leurs formes fléchies à l'aide de la base de données lexicale *BDLex*.

soit 303 articles.

Par cette sélection, nous faisons l'hypothèse que nous limitons dans les analyses portant sur `Corpus_Travail` le possible « bruit » causé par des articles contenant très peu d'occurrences des domaines étudiés. Les exemples suivants illustrent des emplois de chacune des trois métaphores conceptuelles dans `Corpus_Travail`[161] :

> *Le dénouement dans la* ***bataille*** *autour de la première banque commerciale privée du pays a eu peu d'effet sur les cours.* (27/02/1989)
> *Une véritable* ***tempête*** *de hausses, alimentée par une marée de capitaux, étrangers pour partie, en quête de placement.* (03/08/1987)
> *(...) la* ***plaie*** *ouverte par le krach est presque* ***cicatrisée****.* (27/06/1988)

La partie suivante de cette section présente les différentes analyses réalisées à partir des trois dispositifs *LUCIA* et de `Corpus_Travail` selon les propositions du modèle *AIdED*.

4.3.3 Analyses réalisées

Préliminaires statistiques

Dans un premier temps, nous avons réalisé une première analyse quantitative de la répartition des dispositifs *LUCIA* représentant les domaines sources dans le corpus étudié. Pour mettre en évidence le « passage » du corpus exploité initialement dans le projet *IsoMeta* (`Corpus_Total`) à notre corpus d'étude actuel (`Corpus_Travail`), nous proposons le tableau 4.3 mettant en évidence des informations sur les occurrences des lexies des domaines dans les corpus considérés.

Domaine	Nb de lexies	Nombre total de lexies et de leurs flexions		Nombre d'articles de `Corpus_Total` contenant au moins	
		`Corpus_Total`	`Corpus_Travail`	une lexie	deux lexies
Guerre	64	916	831	320 articles -> (- 32%) -> 216 articles	
Météo	112	627	569	252 articles -> (- 36%) -> 160 articles	
Santé	111	567	485	282 articles -> (- 46%) -> 151 articles	

TAB. 4.3 – Tableau présentant une première répartition des domaines dans `Corpus_Total` et `Corpus_Travail`.

Un nombre plus important de lexies du domaine de la guerre dans les articles est observé, malgré la taille réduite de la RTO utilisée. Il est également intéressant de remarquer que des proportions très importantes d'articles ne contiennent qu'une seule occurrence de lexies des domaines d'étude. C'est à cause de ces fortes proportions que nous avons constitué le corpus `Corpus_Travail` afin de favoriser le plus possible la redondance des domaines au sein des articles.

L'influence de l'actualité boursière est très forte sur les articles constituant notre corpus d'étude. Pour permettre une première appréhension de cette influence, nous proposons en figure 4.10 deux graphiques mettant en évidence la répartition des articles du corpus selon le mois de l'année et le nombre moyen par article de lexies de chaque domaine selon le mois de l'année.

Le graphique de gauche de la figure 4.10 montre que le nombre d'articles du corpus varie fortement selon la période de l'année. On peut ainsi observer qu'un nombre important d'articles prend place entre octobre 1987 et février 1988, ceci à cause d'une période très agitée dans le monde boursier. Le graphique de droite de cette même figure permet d'observer la répartition des domaines d'étude tout au long de l'année. Il se confirme que le domaine de la guerre est

[161] Les mots en gras dans les extraits font partie des dispositifs représentant les domaines sources.

plus présent que les autres domaines avec une valeur particulièrement importante en avril 1987. Une première exploration du corpus met en évidence que les articles de cette période traitent principalement d'un conflit armé en Afrique ayant des répercutions sur la bourse, le domaine de la guerre y semble plutôt employé de façon non métaphorique même si de nombreux emplois métaphoriques sont également présents.

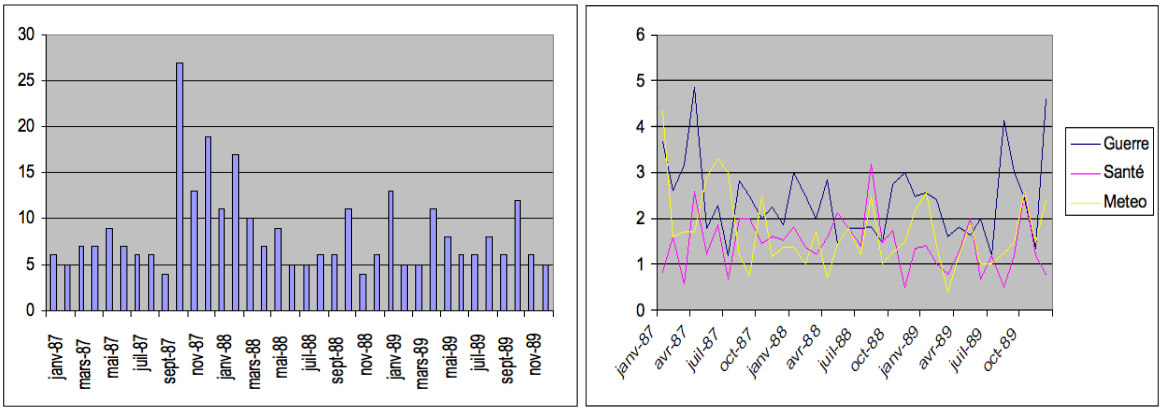


FIG. 4.10 – Le graphique de gauche présente le nombre d'articles de `Corpus_Travail` selon le mois de l'année, celui de droite, le nombre moyen par article d'occurrences de graphies des domaines étudiés selon le mois de l'année.

Afin de favoriser la présence des domaines sources dans les articles, nous considérons qu'un article est étiqueté `Guerre` s'il possède au moins deux occurrences du domaine de la guerre (procédure identique pour les deux autres domaines). Un article possède donc au moins une étiquette (par constitution de `Corpus_Travail`) et peut en avoir plusieurs s'il contient plus de deux occurrences de plusieurs domaines. La figure 4.11 met en évidence les éventuels « liens » entre les trois domaines dans les articles ainsi étiquetés.

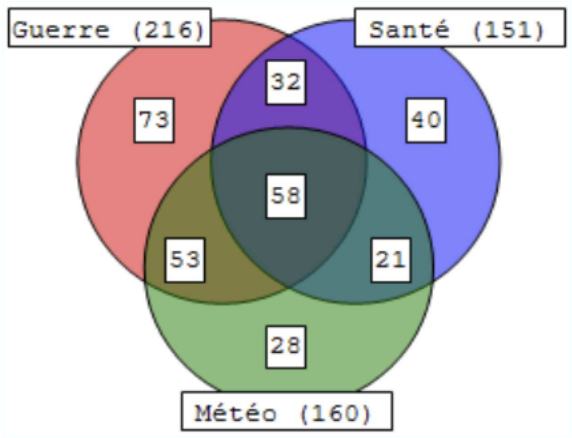


FIG. 4.11 – Représentation schématique des ensembles d'articles de `Corpus_Travail` étiquetés `Guerre`, `Météo` et `Santé` et de leurs intersections, les valeurs numériques indiquées sont les cardinalités des différents ensembles résultant des différentes intersections et exclusions des anneaux.

Nous pouvons observer que le domaine de la météorologie est assez lié avec les autres domaines : sur les 160 articles étiquetés `Météo` seulement 28 possèdent cette seule étiquette alors que 73 et 40 articles sont respectivement étiquetés uniquement par `Guerre` et `Santé`.

Les différentes données présentées précédemment ont été obtenues à l'issue de simples comptages de lexies et de leurs flexions en corpus. Ces comptages ont été réalisés à l'aide de certains composants logiciels de la plate-forme *ProxiDocs*. Pour terminer cette section sur les premières analyses de la répartition des domaines en corpus, nous présentons un extrait du nuage et un extrait d'un anti-nuage (figure 4.12) de lexies des dispositifs dans le corpus.

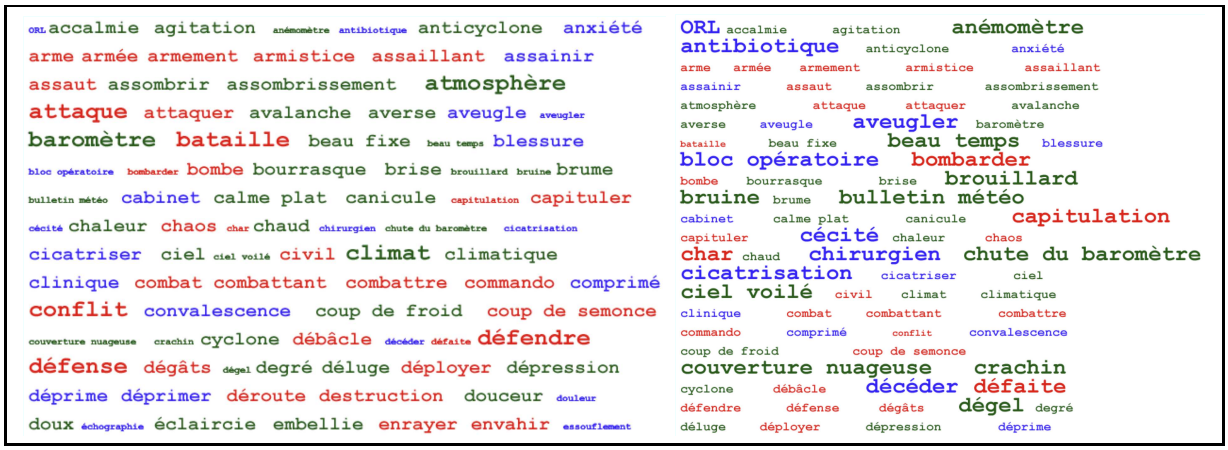


FIG. 4.12 – En partie gauche de la figure, extrait du nuage des lexies des trois dispositifs en corpus, en partie droite, extrait de l'anti-nuage des lexies des trois dispositifs en corpus.

Ces nuages mettent alors en évidence les lexies des dispositifs particulièrement présentes dans le corpus, comme par exemple les lexies *atmosphère*, *attaque*, *baromètre*, *bataille* et *défendre*, et les lexies des dispositifs très peu présentes voire absentes, telles les lexies *anémomètre*, *antibiotique*, *aveugler*, *bombarder* ou encore *défaite*. Nous complétons les premières observations de la répartition des domaines sources en corpus réalisées dans cette section, par différentes cartes que nous présentons dans les paragraphes suivants de cette section.

Cartes de documents et des groupes de documents en 2 et 3 dimensions

Afin d'avoir un regard plus précis sur le corpus et les domaines sources de métaphores conceptuelles étudiées, nous avons utilisé la plate-forme *ProxiDocs* afin de produire différentes représentations cartographiques du corpus selon les domaines. La première carte, présentée en figure 4.13 a été construite par notre plate-forme en faisant intervenir la méthode de comptage relative et une ACP. Les deux premières composantes issues de l'ACP ont été choisies pour construire la carte, l'inertie associée à chacune est respectivement de 41,07% pour la première et de 37,14% pour la deuxième, le taux d'inertie global est donc de 78,21%. Un tel taux indique une représentation assez fidèle des données de l'espace de départ par la carte (ce taux s'expliquant par la relative simplicité du passage d'un espace à 3 dimensions vers un espace à 2 dimensions).

Le cercle des corrélations associé à cette ACP, également déterminé par *ProxiDocs*, est présenté en partie inférieure gauche de la figure 4.13, il apporte une information sur la présence et le lien entre les domaines étudiés dans le corpus. Par définition, le cercle des corrélations attribue un caractère plus exprimé aux éléments éloignés de son centre. Nous pouvons ainsi remarquer que les domaines de la météorologie et de la guerre sont fortement exprimés, c'est-à-dire très présents dans un certain nombre de textes du corpus. Le domaine de la santé est par contre moins

exprimé, car plus proche du centre. Un tel caractère nous indique que le domaine de la santé est moins présent en corpus et de façon mieux répartie que les deux autres domaines. Sa position sur le cercle, entre les domaines de la guerre et de la météorologie, indique également des liens entre ce domaine et les deux autres. Ces liens se traduisent par des utilisations privilégiées du domaine de la santé dans des textes où le domaine majoritaire est soit celui de la météorologie, soit celui de la guerre, donnant ainsi une vision différente de celle proposée en figure 4.11.

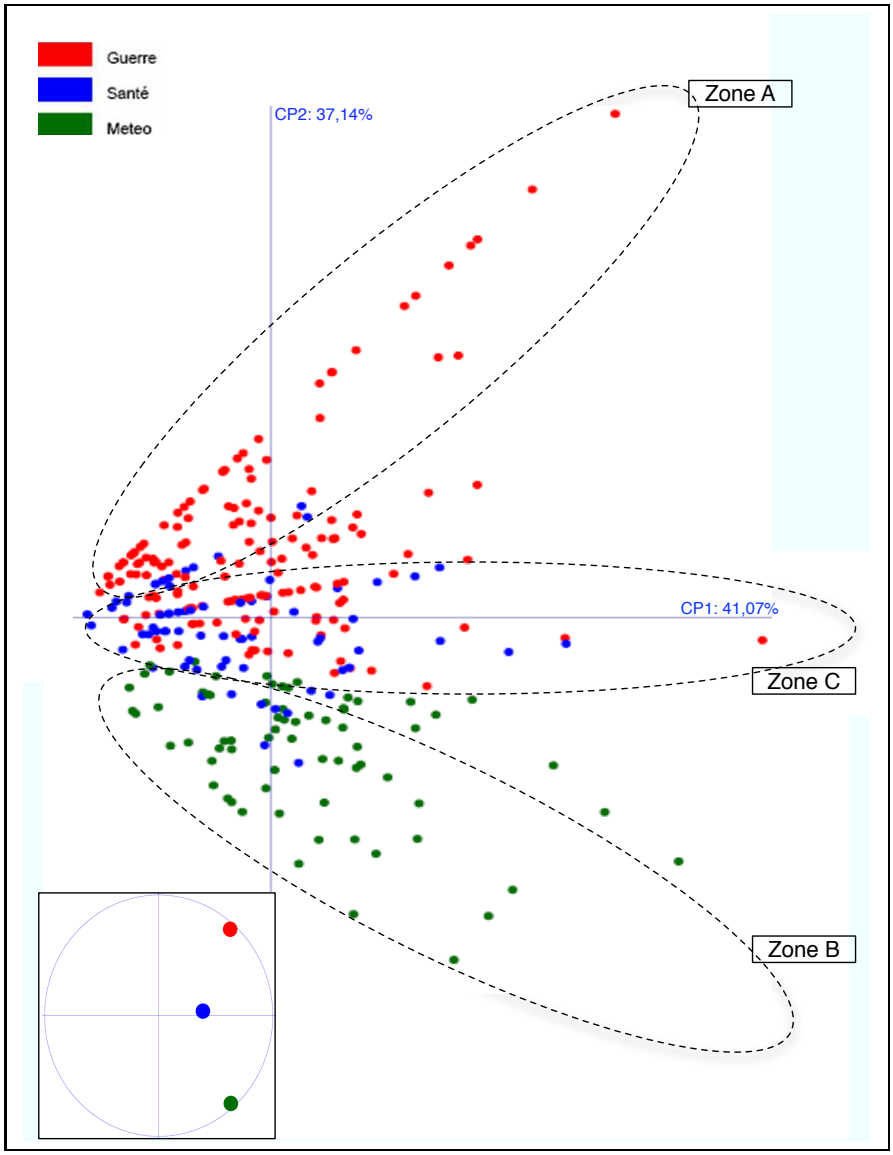


FIG. 4.13 – Carte en 2 dimensions des textes du corpus. Des zones ont été marquées manuellement sur la carte afin de faciliter son analyse.

Cette carte des textes représente donc chaque texte du corpus par un point dont la couleur correspond à son domaine majoritaire. La carte des textes révèle que les domaines de la guerre et de la météorologie sont les plus représentés dans les articles du corpus (respectivement en zones A et B). Un parcours des articles de ces deux zones laisse paraître que la zone B (météorologie) contient une majorité de dépêches liées à l'actualité boursière alors que la zone A (guerre) semble contenir autant de bilans que de dépêches. Le domaine de la santé est moins représenté et il est intéressant de voir que les articles contenant majoritairement des lexies de ce domaine se « mélangent » en zone C avec des articles majoritairement du domaine de la guerre. Un parcours rapide des articles de cette zone révèle une proportion importante de bilans.

La carte en 3 dimensions (dont des extraits sont présentés en figure 4.14) révèle également une disposition en 3 zones comme celle présentée précédemment.

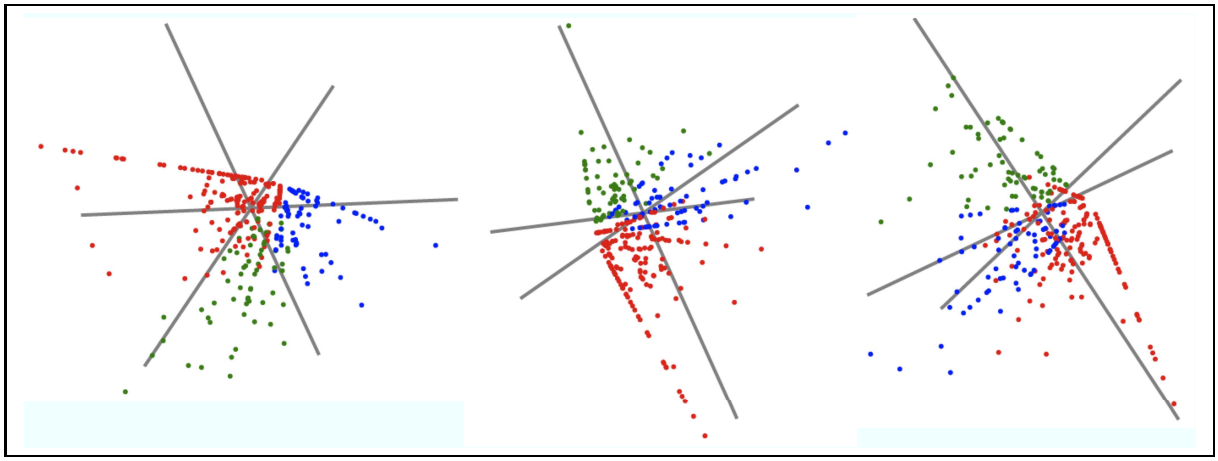


FIG. 4.14 – Extraits de la carte en 3 dimensions des textes du corpus.

La carte des groupes de textes présentée en figure 4.15 est issue de la même ACP que la carte précédente, elle met en plus en évidence le résultat d'une CHA en 12 groupes (nombre choisi empiriquement) de `Corpus_Travail`. Cette carte confirme la prédominance des domaines de la guerre et de la météorologie (sur les 12 groupes d'articles représentés sur la carte, 8 groupes contiennent des articles contenant majoritairement des lexies dans le domaine de la guerre, 4 dans le domaine de la météorologie et aucun dans le domaine de la santé).

Le tableau 4.4 résume les propriétés des groupes et des zones marqués sur la carte (la présentation des groupes et zones dans le tableau se fait de la partie supérieure de la carte vers sa partie inférieure). Les informations du tableau sont accessibles dans les rapports de groupes consultables à partir la carte des groupes[162].

[162] Lorsque les informations portent sur une zone de plusieurs groupes, un parcours manuel des différents groupes a été réalisé afin de rassembler leurs descriptions.

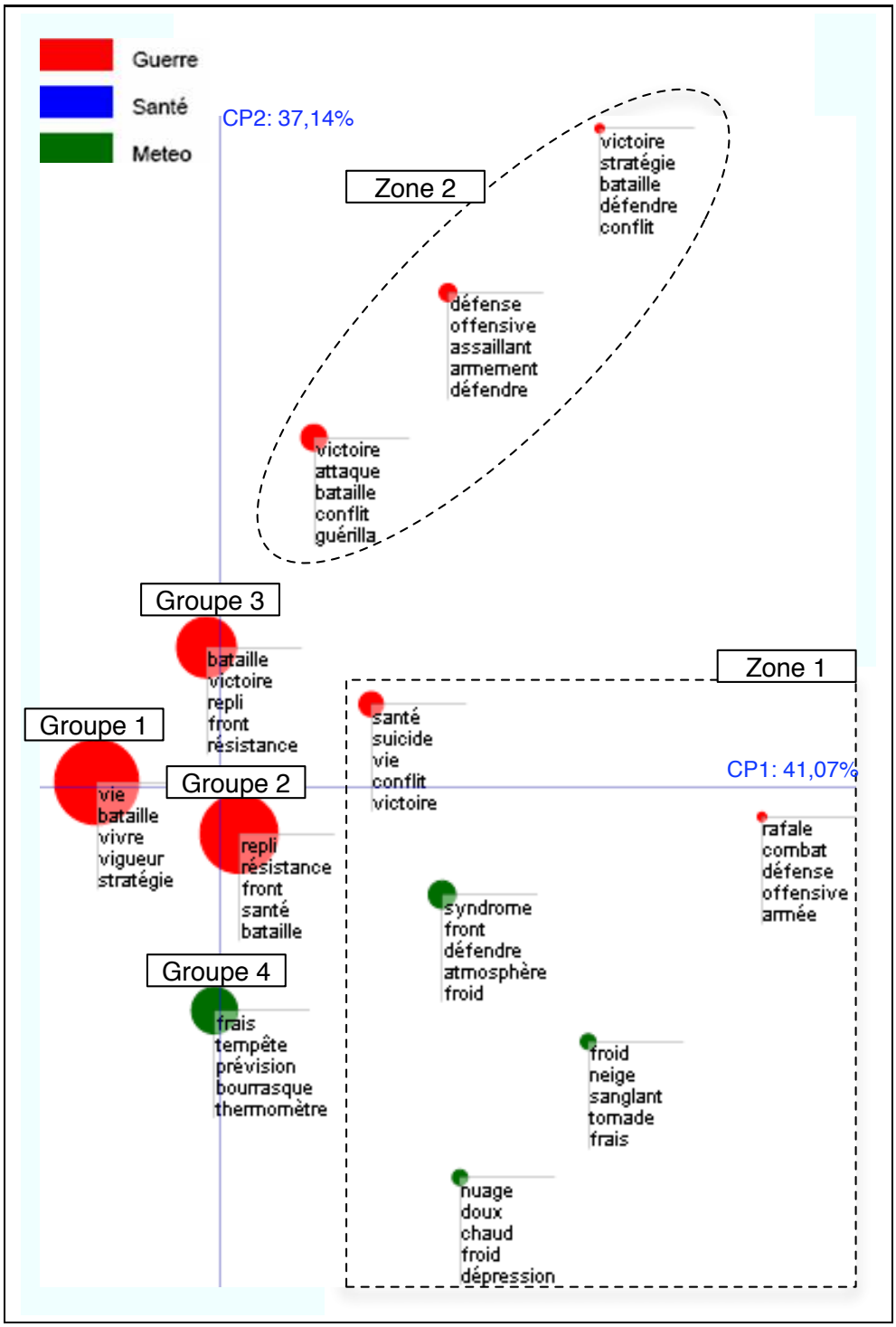


FIG. 4.15 – Cartes en 2 dimensions des groupes de textes du corpus. Des groupes et des zones ont été marqués manuellement sur la carte afin de faciliter son analyse.

Groupe / Zone	Nombre d'articles	Type d'articles		Nombre total de lexies de domaine		
		Bilans	Dépêches	Guerre	Météorologie	Santé
Zone 2 (3 groupes)	14	7%	93%	101	1	8
Principales lexies	1. *victoire*, 2. *défense*, 3. *bataille*					
Principales isotopies inter-textuelles	1. *Fonction*, 2. *Rapport de l'agent au domaine* (valeurs respectives les plus fréquentes : *intervention* et *profession*)					
Groupe 3	49	49%	51%	217	37	56
Principales lexies	1. *bataille*, 2. *victoire*, 3. *repli*					
Principales isotopies inter-textuelles	1. *Fonction*, 2. *Nature de l'agent* (valeurs respectives les plus fréquentes : *intervention* et *humain*)					
Groupe 1	98	47%	53%	162	79	150
Principales lexies	1. *vie*, 2. *bataille*, 3. *vivre*					
Principales isotopies inter-textuelles	1. *Rapport au domaine*, 2. *Durée* (valeurs respectives les plus fréquentes : *objet* et *ponctuel*)					
Groupe 2	84	73%	27%	231	222	185
Principales lexies	1. *repli*, 2. *résistance*, 3. *front*					
Principales isotopies inter-textuelles	1. *Nature de l'agent*, 2. *Évaluation* (valeurs respectives les plus fréquentes : *humain* et *mal*)					
Groupe 4	29	31%	69%	18	110	31
Principales lexies	1. *frais*, 2. *tempête*, 3. *prévision*					
Principales isotopies inter-textuelles	1. *Axe*, 2. *Direction* (valeurs respectives les plus fréquentes : *température* et *monte*)					
Zone 1 (5 groupes)	25	32%	68%	90	105	54
Principales lexies	1. *froid*, 2. *syndrome*, 3. *santé*					
Principales isotopies inter-textuelles	1. *Rapport au domaine*, 2. *Évaluation* (valeurs respectives les plus fréquentes : *objet* et *bien*)					

TAB. 4.4 – Détails des différents groupes et zones de groupes marquées sur la carte de la figure 4.15.

Une lecture de cette carte des groupes de textes complétée à l'aide du tableau permet de remarquer que les groupes 1 et 3 contiennent en grand nombre des lexies du domaine de la guerre. Les proportions de bilans et de dépêches de ces groupes sont très proches des proportions globales du corpus (soit 46% de bilans et 54% de dépêches). Les métaphores conceptuelles remarquées lors d'une lecture des articles de ces deux groupes font intervenir les domaines sources de la guerre et, dans une moindre mesure, de la santé. Toutefois le nombre de ces métaphores est assez faible et les emplois sont plutôt figés.

Le groupe 2 contient une majorité de bilans boursiers. Les métaphores conceptuelles observées font intervenir les trois domaines sources étudiés. Leur nombre est plus important que dans les groupes 1 et 3 et les emplois métaphoriques sont plus variés et originaux.

Le groupe 4 contient en majorité des dépêches. Un nombre important de métaphores conceptuelles faisant intervenir le domaine source de la météorologie y est observé. Tout comme dans les articles du groupe 2, les emplois métaphoriques sont variés et originaux.

La zone 1 marquée sur la carte met en évidence cinq petits groupes de textes (contenant de 1 à 10 articles). Beaucoup de métaphores conceptuelles de domaines sources de la guerre et de la météorologie ont été observées dans les articles de ces groupes.

Au contraire, la zone 2 rassemble trois groupes (contenant de 1 à 9 articles) où un grand nombre de lexies du domaine de la guerre sont présentes. Les emplois des lexies de ce domaine sont non métaphoriques et relèvent des sujets abordés dans ces articles, à savoir, de conflits armés ayant des répercussions sur le marché boursier.

Les différentes isotopies inter-textuelles, et les valeurs qui leur sont le plus fréquemment associées, apportent également une information sur la caractérisation des différents groupes. Par exemple, les deux couples *nature de l'agent* : *humain* et *évaluation* : *mal* sont les plus présents dans le groupe 2. Ils peuvent ainsi aiguiller globalement l'interprétation des textes du groupe en laissant penser que des lexies des domaines sources sont employées pour exprimer des faits négatifs liés à des agents humains. Une telle idée est d'ailleurs confortée par les trois lexies les plus présentes dans le groupe (*repli*, *résistance* et *front*) qui confirment la forte connotation négative.

Nous donnons en figure 4.16 un extrait de la sortie retournée par l'outil *FlexiSemContext* mettant en évidence les contextes d'actualisation du couple *évaluation* : *mal* dans les textes du groupe 2. Une telle sortie est accessible *via* le rapport d'analyse du groupe. Comme nous l'avons fait remarquer précédemment, le groupe 2 contient un nombre important de lexies des domaines sources étudiés. Un parcours plus détaillé des textes du groupe 2 confirme les emplois métaphoriques des lexies de ces trois domaines (les trois domaines sont utilisés de façon à peu près équivalente). Les métaphores conceptuelles présentes expriment principalement des idées négatives liées à des baisses, des incidents sur le cours de la bourse. La présence de telles idées dans les textes du groupe semblent justifier les principales isotopies inter-textuelles qui le caractérisent.

Nous venons d'illustrer que les isotopies inter-textuelles pouvaient permettre d'orienter l'interprétation globale d'un groupe de textes. Elles ont également de fortes retombées sur l'interprétation des éléments plus locaux comme les textes. Par exemple, considérons le groupe situé en partie inférieure gauche de la zone 1 (groupe de domaine majoritaire de la météorologie et étiqueté par les lexies *nuage*, *doux chaud*, *froid* et *dépression*). Ce groupe de 4 documents (que nous appellerons le groupe n°5) contient 5 lexies du domaine de la guerre, 3 de la santé et 33 de la météorologie. Ses isotopies inter-textuelles sont détaillées dans le tableau 4.5, ce tableau reprend des informations données dans le rapport d'analyse du groupe.

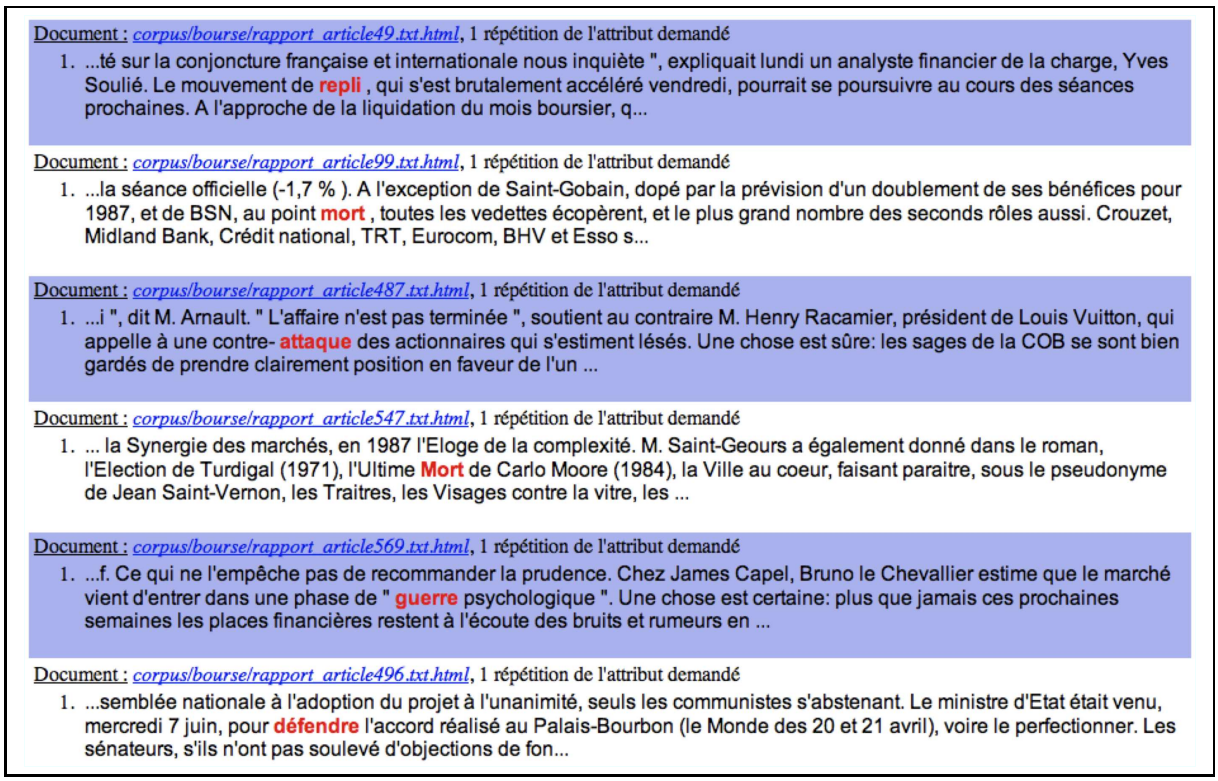


FIG. 4.16 – Mise en évidence des contextes d'actualisation du couple *évaluation* : *mal* dans des textes du groupe 2.

Rang	Attribut répété	Nombre de répétitions	Score de l'isotopie	Nombre de textes supports	Rang sans pondération
1.	*Axe*	29	25.8%-8.8% = **17.0%**	4	3.
	1. *couverture nuageuse* - 12 répétitions (41.3%) (la météorologie, 12) 2. *température* - 9 répétitions (31.0%) (la météorologie, 9) 3. *pression* - 5 répétitions (17.2%) (la météorologie, 5) 4. *agitation* - 3 répétitions (10.3%) (la météorologie, 3)				
2.	*Évaluation*	30	26.7%-21.0% = **5.7%**	4	2.
	1. *mal* - 12 répétitions (57.3%) (la guerre, 4 - la météorologie, 8) 3. *bien* - 8 répétitions (42.6%) (la santé, 1 - la météorologie, 7)				
3.	*Rapport au domaine*	38	33.9%-33.8% = **0.1%**	4	1.
	1. *phénomène* - 23 répétitions (60.5%) (la santé, 1 - la météorologie, 22) 2. *objet* - 10 répétitions (26.3%) (la santé, 1 - la météorologie, 9) 3. *activité* - 4 répétitions (10.5%) (la guerre, 4) 4. *agent* - 1 répétition (2.6%) (la guerre, 1)				
4.	*Direction*	5	4.4%-4.8% = **-0.4%**	4	5.
	1. *monte* - 4 répétitions (80.0%) (la météorologie, 4) 2. *descend* - 1 répétition (20.0%) (la météorologie, 1)				
5.	*Fonction*	10	8.9%-14.7% = **-5.8%**	4	4.
	1. *intervention* - 5 répétitions (50.0%) (la guerre, 4 - la météorologie, 1) 2. *observation* - 5 répétitions (50.0%) (la météorologie, 5)				

TAB. 4.5 – Détails des isotopies inter-textuelles parcourant le groupe 5, informations accessibles dans le rapport d'analyse du groupe. Pour rappel, le nombre de répétitions correspond d'un attribut ou d'une valeur d'attribut correspond au nombre d'occurrences de mots les portant des domaines issus des domaines sources.

La lecture du tableau nous donne un grand nombre d'informations sur les différentes isotopies inter-textuelles présentes dans le groupe de textes considéré. Le rang des isotopies inter-textuelles avec pondération (à gauche) et sans pondération (à droite), l'attribut lié à l'isotopie, son nombre de répétition, son score et le nombre de textes du groupe la supportant[163] sont ainsi présentés. L'isotopie inter-textuelle classée en première position est liée à l'attribut *Axe*. Cet attribut représente la dimension des phénomènes physiques mesurables exprimés dans le domaine source de la météorologie, telles la température, la force du vent, l'épaisseur de la couverture nuageuse, etc. L'usage métaphorique de lexies porteuses de cet attribut a été jugé par les membres du projet *IsoMeta* comme un bon guide pour l'interprétation des textes du groupe en terme de phénomènes mesurables de la Bourse. L'isotopie inter-textuelle classée en troisième position (*Rapport au domaine* avec comme valeur majoritairement associée *phénomène*) renforce l'idée de phénomènes physiques appliqués au monde de la Bourse. L'isotopie inter-textuelle liée à l'attribut *Évaluation* est classée en seconde position, la valeur *mal* étant la plus fréquemment associée à l'attribut répété. Cette isotopie permet également d'orienter l'interprétation des textes en ajoutant une connotation négative à l'idée de phénomènes physiques mesurables appliqués au monde boursier. Les attributs *Direction* et *Fonction* sont respectivement liés aux isotopies inter-textuelles classées dans les deux dernières positions du classement. Contrairement aux isotopies les précédant dans le classement, les scores de ces deux dernières isotopies inter-textuelles sont négatifs. Le groupe ne contribue pas à renforcer de telles isotopies dans le corpus mais au contraire en présente un déficit. Il a ainsi été confirmé que de telles isotopies n'apportaient que peu de nouvelles informations venant guider l'interprétation des textes du groupe.

Dans ce groupe, la pondération des isotopies inter-textuelles que nous proposons *via* la fonction de score a été jugée particulièrement utile. Pour rappel du chapitre 3 de cette thèse (page 63), le calcul du score d'une isotopie se fait en déterminant la différence entre le pourcentage (appelé *poids*) de cette isotopie dans le contexte local (ici, le groupe de textes) et le pourcentage de cette isotopie dans le contexte global (ici, l'ensemble documentaire). Principalement, la « remontée » de l'isotopie inter-textuelle liée à *Axe*[164] a permis de guider rapidement l'interprétation des textes du groupes en termes de phénomènes physiques mesurables. Le déclassement de l'isotopie inter-textuelle liée à *Rapport au domaine* se justifie également. Cet attribut est très présent dans l'ensemble du corpus du fait de sa position en entrée des dispositifs. Cet attribut est alors assez générique et même s'il est très présent dans le groupe (38 répétitions), il en est de même dans le corpus, puisque le score attribué à l'isotopie correspondante est à peine supérieur à la valeur nulle.

Afin de mieux illustrer les « guides » aux parcours interprétatifs fournis par les isotopies inter-textuelles détaillées précédemment, nous avons extrait, de chacun des quatre textes du groupe, des exemples d'usages métaphoriques de lexies du domaine source de la météorologie, majoritairement présent dans ce groupe.

> *Le **chaud** et le **froid** ? Depuis le grand « krach » d'octobre, la Bourse de Paris connait ce genre de phénomène **climatique**. Mais les écarts de **températures** qu'elle a eu à subir au cours de la séance du jeudi 10 décembre pourront figurer dans le grand livre des records. (...) Jusqu'à la clôture à 14 h 30, le **thermomètre** du marché ne cessa de monter pour s'élever de 3,4%.* (12/12/1987)

[163] Nous pouvons déjà remarquer que chaque texte du groupe supporte chaque isotopie inter-textuelle. Nous pouvons dès maintenant en déduire une forte homogénéité des textes, chaque texte contribuant de manière plus ou moins prononcée aux différentes isotopies inter-textuelles du groupe.

[164] Sans pondération, cette isotopie aurait été en troisième position, avec la pondération, elle est passée en première position.

*Le **Baromètre** de la Bourse mis au point par les analystes suédois de Delphi, (...).* (18/05/1987)

***Mois maussade**, juillet n'a pas été un **mois pourri**. Une fois remis de leur déception de n'avoir pas assisté à la « mythique » hausse d'été, les boursiers ont repris, ces jours derniers, quelque espoir de surprendre quelques belles **éclaircies** rue Vivienne. (...) Vendredi, les **nuages** qui **assombrissaient** l'horizon boursier huit jours plus tôt avaient disparu. La tension entre la France et l'Iran continuait de laisser **froids** les professionnels.* (27/07/1987)

*(...), la hausse presque généralisée des taux d'escompte en Europe centrale, au pays de Goethe pour commencer, a **rafraichi** l' **atmosphère**.* (04/07/1988)

Pour conclure notre analyse de la carte des groupes de textes, nous proposons le tableau 4.6 résumant nos différentes observations et donnant, pour chaque groupe, un exemple d'emplois de lexies des domaines sources (les groupes et zones sont abordés du haut vers le bas de leur position sur la carte). Pour chaque groupe et zone, nous faisons figurer le type de métaphores conceptuelles observé : de métaphores variées à des emplois littéraux en passant par des emplois figés.

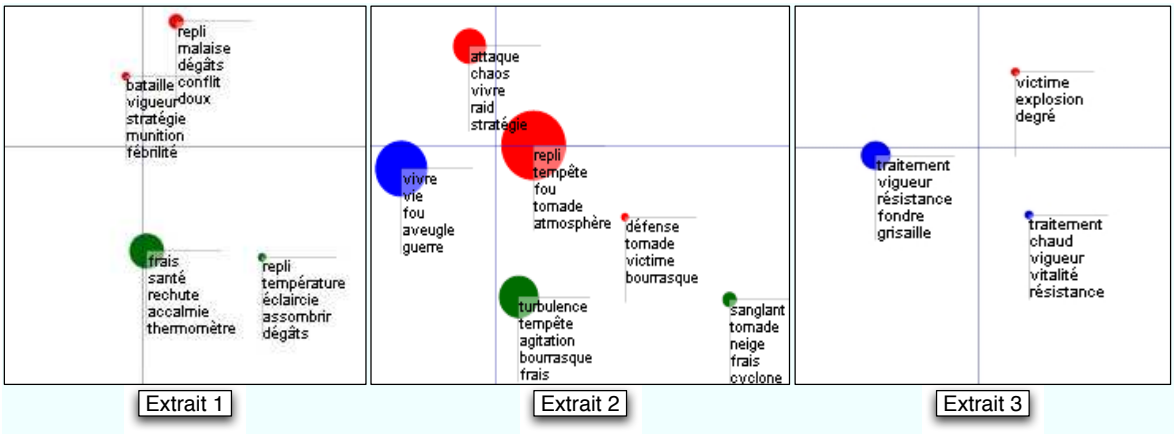
Groupe / Zone	Métaphores conceptuelles	Exemples extraits du corpus
Zone 2	Aucune	*Pour se déplacer (...), les **officiers** de la **guérilla** utilisent des motos récupérées pendant les **attaques**.* (13/04/1987)
Groupe 3	Figées	*Selon le SNUI, qui rappelle que le **conflit** des impôts dure depuis sept mois, (...).* (22/09/1989)
Groupe 1	Figées	*En neuf mois, six firmes sur les trente-trois OPA ont été l'objet de véritables **batailles** boursières.* (26/09/1988)
Groupe 2	Variées	*(...) après avoir **contaminé** New-York et Londres, la **fièvre** des OPA s'est mise (...) à ronger la Bourse.* (08/02/1988)
Groupe 4	Variées	*Un petit **vent frisquet** a soufflé, ces derniers jours rue Vivienne, qui, sans crier gare, s'est soudain éclipsé à la dernière minute pour laisser la place à une **brise** nettement plus **chaude**.* (15/05/1989)
Zone 1	Variées	*Porteur du terrible **virus** de la défiance, il se propage à la vitesse de l'éclair et les tentatives désespérées de **réanimation** (...) sont inopérantes.* (30/10/1987)

TAB. 4.6 – Détails des différents groupes et zones de groupes marquées sur la carte de la figure 4.15.

Nous constatons que la répartition sur la carte des différents groupes et zones de groupes se fait selon les types de métaphores que nous y avons observés : les emplois littéraux ou les métaphores figées sont situés dans la partie supérieure de la carte, tandis que les emplois variés et plus originaux sont globalement situés dans la partie inférieure. Une gradation de cette répartition peut même être observée si l'on remarque que les emplois figés sont situés entre les emplois littéraux et les métaphores plus vives. La carte révèle alors ce que nous avons appelé le degré de *métaphoricité*. Le dégré 0 se situe ainsi en haut de la carte avec les emplois littéraux des lexies des domaines utilisés. Le degré maximum, le degré 1, se situe en bas de la carte avec des emplois métaphoriques variés des lexies. Entre ce degré 0 et ce degré 1, se situent des degrés de métaphoricité intermédiaires, où des emplois métaphoriques plus figés, plus habituels des lexies des domaines sources sont présents dans les textes.

Cartes temporelles des groupes de documents

Les cartes présentées dans la section précédente ont permis d'observer la répartition des domaines d'étude en corpus, de localiser des emplois de certaines métaphores et de mettre en évidence un degré de métaphoricité du corpus, en distinguant des ensembles de textes contenant soit des emplois métaphoriques, soit des emplois non métaphoriques du lexique des domaines étudiés. Étant donné l'influence du temps dans l'apparition des domaines d'étude dans les articles du corpus (cf. graphiques de la figure 4.10), nous proposons de rendre compte de cette dimension temporelle dans une carte temporelle du corpus. Les traitements statistiques réalisés pour la construction de cette carte restent identiques à ceux de l'expérience précédente, seul le mode d'affichage change comme nous avons pu l'expliquer au chapitre 3. Dans l'expérience présentée ici, il nous a paru pertinent de choisir une fenêtre temporelle d'un mois et une unité de déplacement de cette fenêtre d'un jour. Il en résulte une animation de 1065 images dont la figure 4.17 présente trois extraits significatifs.



Extrait	Période	Nombre d'articles	Nb d'occurrences de lexies		
			Guerre	Météo	Santé
1	25/04/87-25/05/87	10 (5 bilans, 5 dépêches)	24	36	16
2	17/10/87-17/11/87	33 (4 bilans, 29 dépêches)	78	73	46
3	13/05/88-13/06/88	7 (5 bilans, 2 dépêches)	9	11	15

FIG. 4.17 – Trois extraits de la carte temporelle du corpus. Le tableau de la partie inférieure détaille chaque extrait.

Les extraits précédents illustrent la répartition temporelle très hétérogène des domaines dans les articles de notre corpus. Le premier extrait présente 4 petits groupes de documents, le domaine majoritaire est celui de la météorologie. En parcourant les articles de cet extrait, nous pouvons remarquer que l'actualité boursière a été calme sur cette période, les seuls faits marquants sont les conditions climatiques, la plupart des articles font d'ailleurs allusion à ces conditions par des emplois métaphoriques ou non (les lexies soulignées dans les extraits d'articles font partie des ressources représentant les domaines d'étude) :

> *La **température** a très brutalement baissé, non seulement à l'extérieur (...) mais rue Vivienne aussi. (...) Le coup de **froid** a été sévère (-5,5%) et d'autant plus inquiétant pour la « végétation mobilière » que le **thermomètre** a chuté de 4% au cours de la seule séance de vendredi.* (18/05/1987)

Le deuxième extrait présente six groupes d'articles. Le domaine de la guerre est majoritaire dans cet extrait, devant le domaine de la météorologie. Cet extrait se distingue par un nombre

de dépêches très important lié à l'actualité boursière très agitée sur la première moitié de cette période, ces dépêches faisant intervenir un grand nombre d'emplois métaphoriques de lexies de nos trois domaines :

> *(...) la pire* ***débâcle*** *qu'ait jamais connue la Bourse de New York (-22,4%).* (22/10/1987)
> *Les monnaies et les marchés sont entrés en* ***turbulence****, un* ***climat*** *qui n'est guère favorable aux affaires.*(22/10/1987)
> *Les deux derniers jours permirent à la Bourse de* ***panser*** *ses* ***plaies****.* (02/11/1987)

Le troisième et dernier extrait met en évidence trois petits groupes d'articles, le domaine de la santé est majoritairement représenté dans les articles. L'actualité boursière de cette période a été particulièrement calme, seule une légère hausse des cours a été mise en évidence :

> *Malgré quelques signes d'****essoufflement****, la Bourse n'a pas manqué de* ***vigueur*** *(...).* (13/06/1988)

Ces exemples illustrent alors la grande variété des métaphores conceptuelles rencontrées et surtout le lien très fort qu'elles entretiennent avec l'actualité. Malgré le choix d'un corpus thématiquement très homogène, il ressort tout de même une influence constante du contexte extérieur aux actualités boursières, telle que nous avons pu la mettre en évidence précédemment.

4.3.4 Vers une nouvelle phase du projet IsoMeta

Au cours de cette expérience, nous avons pu mettre en pratique la plupart des fonctionnalités offertes par la plate-forme *ProxiDocs* : cartes en 2 et 3 dimensions, cartes temporelles, exploration de rapports d'analyse de groupes de documents, mise en contexte de lexies porteuses d'un couple *attribut* : *valeur* particulier dans des textes du corpus, etc. L'utilisation de cet arsenal logiciel nous a permis de mettre visuellement en évidence et de caractériser les usages de trois métaphores conceptuelles. La répartition des lexiques des trois domaines sources a tout d'abord montré des utilisations conjointes des trois métaphores qui ne sont donc pas exclusives l'une de l'autre, et ceci indépendamment de la nature des articles, bilan ou dépêche.

Les différentes vues interactives construites à partir du corpus et des domaines étudiés nous ont permis de définir la notion de degré de métaphoricité. Ce degré, qui reste cependant à mieux caractériser, pourrait se révéler particulièrement utile pour distinguer l'emploi de certaines lexies. Selon le degré de métaphoricité attribué au texte ou au groupe de textes, des lexies présentes, issues de domaines potentiellement sources de métaphores conceptuelles, auraient plus de probabilités d'être employées de façon métaphorique si le degré de métaphoricité est proche de 1, et inversement si le degré est proche de 0. Selon nous, la prochaine phase du projet *IsoMeta* devra approfondir l'étude d'un tel degré de métaphoricité.

Les cartes temporelles du corpus ont mis en évidence le rapport entre les usages des métaphores étudiées et l'actualité. Par exemple, la métaphore conceptuelle de la « guerre boursière » a été très utilisée lors du mini-krach boursier de fin 1987, alors que celle de la « météorologie boursière » a été très employée lors de conditions climatiques remarquables, hiver rude ou forte chaleur. Continuer une étude sur ce lien étroit entre métaphores conceptuelles et actualité nous semble également important. Pour affiner une étude, et plus particulièrement une détection de métaphores conceptuelles, il semble alors tout à fait possible d'intégrer les principaux faits d'actualités et les domaines qui y sont liés, afin de « prévoir », d'« anticiper » la présence de certaines métaphores conceptuelles pour affiner les analyses de contenu d'ensembles documentaires.

Étendre notre étude avec d'autres métaphores conceptuelles semble également pertinent afin de mieux caractériser ce phénomène. Certaines de ces métaphores ont d'ailleurs déjà été observées dans notre corpus de travail, faisant par exemple intervenir les domaines sources du jeu et du cinéma[165].

Ces différentes perspectives, révélant pour la plupart une certaine dimension intertextuelle des métaphores étudiées, vont faire émerger une nouvelle phase du projet *IsoMeta* cherchant ainsi à aller encore plus loin dans une telle description de métaphores conceptuelles.

Le schéma d'évaluation de notre modèle appliqué à cette expérimentation est le suivant :

- *Étape 1 : phase de constitution des ressources*
  Cette phase s'est déroulée en plusieurs allers-retours entre les trois membres du projet *IsoMeta*. Le dispositif *LUCIA* en rapport avec le domaine de la météorologie a été repris de la phase précédente du projet et n'a été que très peu modifié. Les deux autres dispositifs décrivant les domaines de la santé et de la guerre ont été réalisés à partir d'observations faites sur notre corpus de travail. Différents échanges ont eu lieu autour de ces dispositifs, tout d'abord sur leur contenu (initialement, nous avons d'abord rassemblé les lexies nous paraissant liées aux domaines sources dans de simples ensembles), puis sur leur structure (dans un second temps, les listes plates de lexies ont été « transformées » en dispositifs). Cette phase s'est déroulée pendant environ un mois avant d'arriver à une première stabilisation des trois domaines sources (des modifications mineures, consistant en l'ajout ou la suppression de lexies, d'attributs ou de valeurs d'attributs, ont cependant été apportées par la suite).
- *Étape 2 : phase d'exécution des outils*
  Dans cette expérimentation, seul *ProxiDocs* a été utilisé *via* son interface. Les différentes cartes (2 dimensions, 3 dimensions, temporelles) et les rapports d'analyses liés ont alors été construits en quelques dizaines de minutes. Plusieurs exécutions de *ProxiDocs* ont été nécessaires à chaque modification des domaines sources.
- *Étape 3 : phase d'évaluation des sorties produites*
  Les cartes de textes et de groupes de textes en 2 dimensions ont principalement été analysées. Plusieurs allers-retours ont également eu lieu entre les différents membres du projet afin de partager les analyses des différentes cartes. Les premières analyses, et principalement les retours aux textes permis par les cartes, ont révélé certains compléments ou modifications à apporter dans les domaines sources. Cette phase a sans doute été la plus longue, puisqu'elle s'est déroulée sur plusieurs mois avec des échanges réguliers entre les membres du projets (en moyenne, un échange par semaine).

Nous avons caractérisé cette expérimentation comme la deuxième étape du projet *IsoMeta* après la première phase détaillée dans [Perlerin *et al.*, 2005]. Cette expérimentation a donné lieu à plusieurs publications. Ainsi, dans [Roy *et al.*, 2005], nous avons présenté et commenté les premières cartes obtenues à partir du corpus de travail. Dans [Roy *et al.*, 2006], nous avons poursuivi cette analyse et nous l'avons complétée par une analyse diachronique.

Dans cette expérimentation, nous avons illustré l'intérêt de notre modèle pour étudier trois métaphores conceptuelles dans un corpus d'articles boursiers. Dans l'expérimentation suivante, le matériau textuel étudié est très différent puisqu'il s'agit de forums de discussion.

[165] Des exemples de ces deux métaphores conceptuelles, respectivement de domaine source du jeu et du cinéma : *Sous les lambris du palais Brongniart, les professionnels se demandent combien de temps encore va durer ce jeu de l'oie.* (15/06/1987) ; *(...), le film lamentable de l'effritement a été projeté jusqu'au redressement final, (...).* (21/03/1988).

4.4 Étude de forums de discussion pédagogiques

Dans le cadre de ce travail de thèse, une participation à une Équipe de Recherche Technologique éducation (ERTé) est réalisée depuis juin 2005 sur la thématique des communautés d'apprentissage en ligne. Cette ERté, intitulée CALICO (Communautés d'Apprentissage en Ligne, Instrumentation, COllaboration), a 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.

Un triple objectif est ainsi visé :

- un objectif de recherche fondamentale : rendre intelligibles les dynamiques d'interaction entre les participants à des activités finalisées dans un contexte de formation et notamment l'activité du formateur-tuteur ;
- un objectif de recherche-développement : fournir des instruments aux communautés d'apprentissage (favoriser l'instrumentation des différents rôles : concepteurs, organisateurs, tuteurs, apprenants, ...) notamment par la visualisation des traces d'activité ;
- un objectif de formation : aider à la mise en place de formations collaboratives par la diffusion des artefacts conçus dans les activités de recherche et la mise en place d'actions spécifiques en direction de formateurs.

Cette ERTé réunit plusieurs laboratoires de sciences de l'éducation et d'informatique, ainsi que plusieurs Instituts Universitaires de Formation des Maîtres (IUFM). Plus de détails sont disponibles à l'adresse suivante : `http://calico.inrp.fr/CALICO` (page consultée le 4 juillet 2007). L'objectif de notre participation à cette équipe est d'étudier comment notre modèle d'accès au contenu d'ensembles documentaires peut apporter une valeur ajoutée dans des analyses ciblées de forums de discussion à visée pédagogique.

Dans le cadre de formations à distance en IUFM, les forums de discussion sont souvent considérés comme d'importants moyens d'échange [Bruillard, 2007]. La définition du forum la plus communément admise est celle le décrivant comme un lieu d'échange asynchrone de messages textuels [Henri et Lundgren-Cayrol, 2001]. Un forum est organisé en un ou plusieurs fils de discussion, chaque fil correspondant à un sujet choisi par les participants ou les modérateurs du forum. Les forums considérés ici contiennent des échanges entre formateurs et stagiaires d'IUFM et entre stagiaires, autour d'objectifs précis fixés par les formateurs.

Notre apport dans une telle problématique consiste à proposer aux formateurs un accès informatisé au contenu de forums de discussion, cet accès étant ciblé sur des domaines d'intérêt définis par les formateurs. Deux collaborations ont débuté avec des membres de l'ERTé CALICO autour de deux tâches différentes d'accès informatisé au contenu de forums. Ces deux tâches sont les suivantes :

- *L'acquisition de l'identité professionnelle* en collaboration avec Georges Ferone[166] : l'objectif est ici d'observer les échanges sur des forums de discussion entre stagiaires d'IUFM réalisant leurs premiers enseignements sur le terrain. Ces observations se font dans le but de mettre en évidence comment les stagiaires découvrent et s'approprient le métier d'enseignant.
- *L'appropriation et la maîtrise d'une terminologie professionnelle* en collaboration avec Nicole Clouet et Marie-Laure Compant-Lafontaine[167] : cette seconde tâche consiste à visualiser, dans des forums de discussion, l'usage d'une terminologie professionnelle pour « vérifier » sa bonne compréhension et sa bonne utilisation par les stagiaires.

[166] Équipe Coditexte, IUFM de Créteil et Équipe ESCOL, Université de Paris 8.
[167] IUFM de Caen, Université de Caen - Basse-Normandie.

Nous détaillons chacune de ces deux collaborations dans les sections suivantes.

4.4.1 Observation de l'acquisition de l'identité professionnelle

Construction des domaines d'intérêt

La première tâche présentée ici a été réalisée en collaboration avec George Ferone, formateur à IUFM de Créteil intervenant dans la formation des professeurs des écoles. Afin d'interroger la façon dont les stagiaires échangent sur des forums de discussion dédiés lors de leurs premières semaines d'enseignement, ce dernier a construit six domaines d'intérêt à l'aide des outils *Memlabor* et *ThemeEditor*. La représentation des domaines utilisée est celle des ensembles de graphies. 222 graphies ont été sélectionnées (ce nombre regroupe des formes lemmatisées et des formes fléchies de lexies, George Ferone ayant sélectionné toutes les formes graphiques qu'il jugeait pertinentes). Ces domaines font intervenir ce que George Ferone appelle [Ferone, 2006] la *dimension collaborative* (comment et sur quoi les stagiaires s'entraident et questionnent leurs formateurs), la *dimension réflexive* (comment et sur quoi les stagiaires échangent à propos de leur formation) et la *vie du groupe* (comment les stagiaires se soutiennent, s'encouragent, plaisantent entre eux). Les domaines retenus sont les suivants[168] :

1. *Information* (dimension collaborative) - 13 graphies liées à des demandes d'information d'ordre général : *demander*, *envoyer*, *info*, *information*, *informations*, *infos*, etc.
2. *Ressources* (dimension collaborative) - 34 graphies liées à des demandes de ressources pédagogiques : *aide*, *besoin*, *demande*, *doc*, *docs*, *document*, etc.
3. *Pratique* (dimension réflexive) - 64 graphies liées aux activités de la classe : *activité*, *activités*, *album*, *apprendre*, *apprentissage*, *apprentissages*, etc.
4. *IUFM* (dimension réflexive) - 39 graphies liées à la formation dispensée à l'IUFM, à l'environnement de l'école : *APP*[169], *bilan*, *CDI*, *commission*, *conférence*, *cours*, etc.
5. *Collectif* (vie du groupe) - 41 graphies liées à la vie en société, la soutien entre stagiaire : *aime*, *ambiance*, *amis*, *anniversaire*, *bises*, *bisou*, etc.
6. *Personnel* (vie du groupe) - 31 graphies liées aux difficultés et aux sentiments personnels sur la formation : *aider*, *bonheur*, *choix*, *contente*, *difficile*, *difficiles*, etc.

Ces domaines ont été projetés sur quatre forums correspondant aux échanges de deux groupes de stagiaires professeurs des écoles de septembre 2002 à juin 2003 et de deux autres groupes de stagiaires de septembre 2003 à juin 2004. Chaque groupe est constitué d'environ 30 stagiaires. Les échanges sur les forums sont libres, aucune consigne n'a été donnée aux stagiaires pour leurs échanges sur les forums. Les forums sont accessibles à chaque stagiaire du groupe ainsi qu'à leurs formateurs. Pour abréger, nous appelons $E1_{2002-2003}$, le forum du premier groupe de l'année 2002-2003 (417 messages totalisant 45 124 graphies) et de la même manière, nous avons les forums $E2_{2002-2003}$ (365 messages - 24 956 graphies), $E1_{2003-2004}$ (219 messages - 23 687 graphies) et $E2_{2003-2004}$ (487 messages - 52 111 graphies).

Les différentes cartes construites par notre plate-forme sur chacun de ces forums à partir des domaines d'intérêt sont décrites dans la suite de cette sous-section.

Analyses des cartes des forums obtenues

Nous présentons en figure 4.18 les cartes de groupes de messages des différents forums étudiés. Afin d'analyser des forums de discussion, la plate-forme *ProxiDocs* intègre un module permettant

[168] La description complète de ces domaines est donnée en annexe C de cette thèse.
[169] Atelier de pédagogie personnalisée.

l'extraction des messages d'un forum au format XMLForum, format mis en place durant l'ERTé CALICO afin de permettre le partage et la diffusion de forums de discussion[170]. De cette manière, un ensemble documentaire est constitué avec les différents messages du forum. Nous rejoignons ainsi l'approche de Nadine Lucas dans [Lucas, 2005] où il est considéré qu'un forum possède une certaine cohérence linguistique et constitue un « tout ». Dans ses travaux, l'auteur exploite cette cohérence pour mener des analyses thématiques automatiques de forums.

Pour obtenir les cartes de forums de discussion présentées en figure 4.18, la méthode de comptage relative a été utilisée, des ACP ont été réalisées (taux d'inertie entre 50% et 65%) ainsi que des CHA automatiques (sans nombre de groupes précisé).

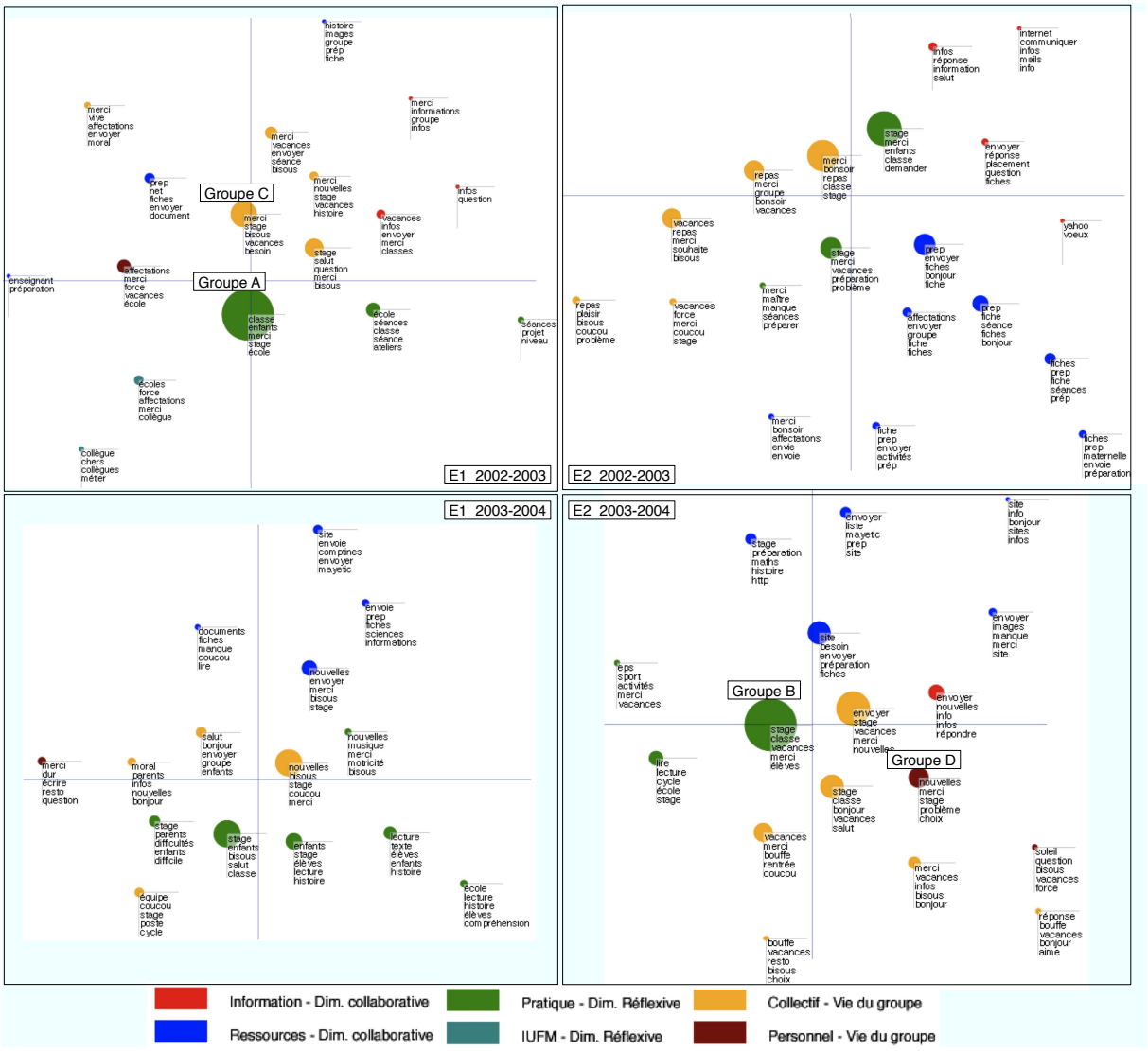


FIG. 4.18 – Cartes des groupes des différents forums étudiés.

[170]Description de ce format : `http://calico.inrp.fr/CALICO/groupspace.2007-03-07.9748726835/preparation-des-corpus/benjamin_huynh_kim_bang_annexe_xmlforum.rtf/view` (page consultée le 5 juillet 2007).

Les cartes ainsi présentées sont assez semblables d'un forum à l'autre. Les graphies étiquetant les groupes marquent bien l'appartenance des forums à un « genre » très particulier avec des formes très fréquentes comme *salut*, *merci*, *bisous*, *bonjour*, etc. Le domaine *Pratique* est le plus abordé dans chaque forum, suivi de près par le domaine *Collectif*. Le domaine *Ressources* est également présent dans les différents forums, mais l'est beaucoup plus faiblement dans le forums $E1_{2002-2003}$. Le domaine *Personnel* est présent de façon significative uniquement dans le forum $E2_{2003-2004}$. Les autres domaines *IUFM* et *Information* sont abordés de façon moins importante dans les différents forums.

Les cartes des forums $E1_{2002-2003}$ et $E2_{2003-2004}$ présentent chacune un groupe de messages du domaine majoritaire *Pratique* de taille très importante (groupes A et B de la carte). En consultant les rapports associés à ces grands groupes, nous avons observé que le domaine du collectif est également très présent, comme l'illustre la sous-carte (partie gauche de la figure 4.19) pour le groupe A du forum $E1_{2002-2003}$.

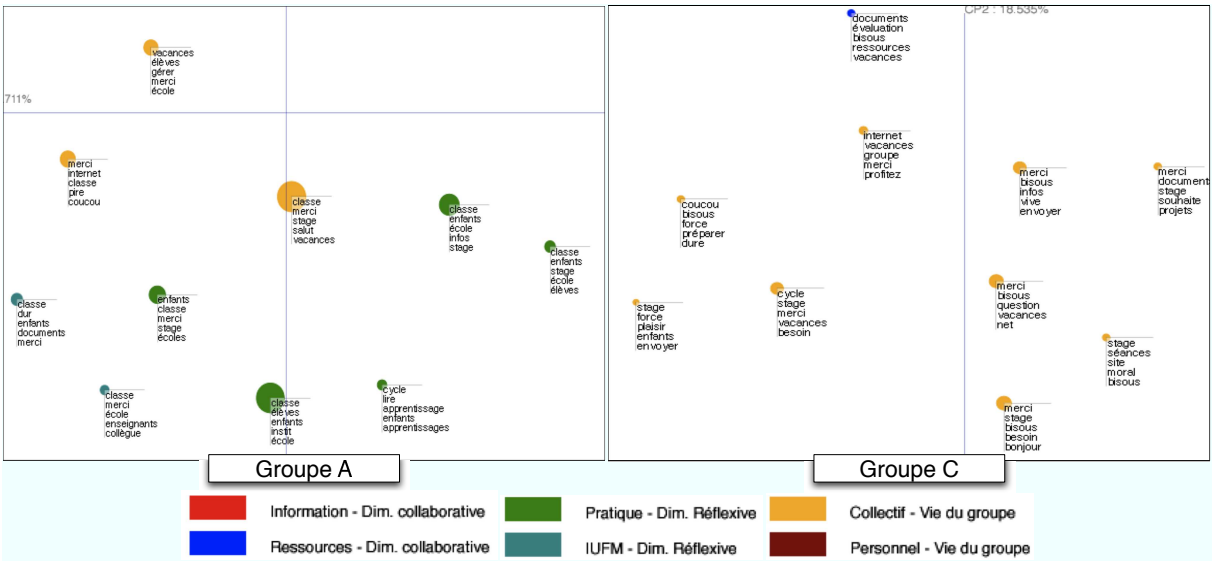


FIG. 4.19 – Sous-cartes des groupes A et C de la carte du forum $E1_{2002-2003}$ présenté en figure 4.18.

Cette sous-carte, accessible dans le rapport du groupe, met en évidence cinq sous-groupes du domaine *Pratique*, trois du domaine *Collectif* et deux du domaine *IUFM*. Le domaine *Pratique* est toujours le plus important mais des domaines secondaires comme *Collectif* et *IUFM* apparaissent de façon significative. La présence de ce dernier domaine dans deux sous-groupes est assez intéressante étant donné qu'il est quasiment absent de la carte globale. En consultant les différents messages contenus dans le groupe principal, les trois domaines énoncés précédemment sont souvent abordés dans de mêmes messages, mais dans des phrases différentes, sans lien étroit entre les domaines, comme l'illustre l'extrait suivant d'un message du groupe :

> *(...) [`Pratique` :] Quant à moi il m'a conseillé de plus laisser des moments d'échanges élèves/élèves avant de faire des mises en commun collective qui sont plutôt difficile à gérer quand elle deviennent longues. Voilà pour ces deux premiers jours. (...) [`Collectif` :] Si quelqu'un a des billes pour améliorer les exos de maths en page 4, ou pour mener quelques exercices de découvertes supplémentaires, de manipulation , quelques situations problèmes qui sont construites différemment...elles sont les bien*

> *venues. (...) [IUFM :] je n'ai aucune info particulière si ce n'est que les dates des commissions de SR3 seront soi envoyées par mail, soit envoyées par courrier, soit pas envoyées et on est mal pour le 21 mai!!!*

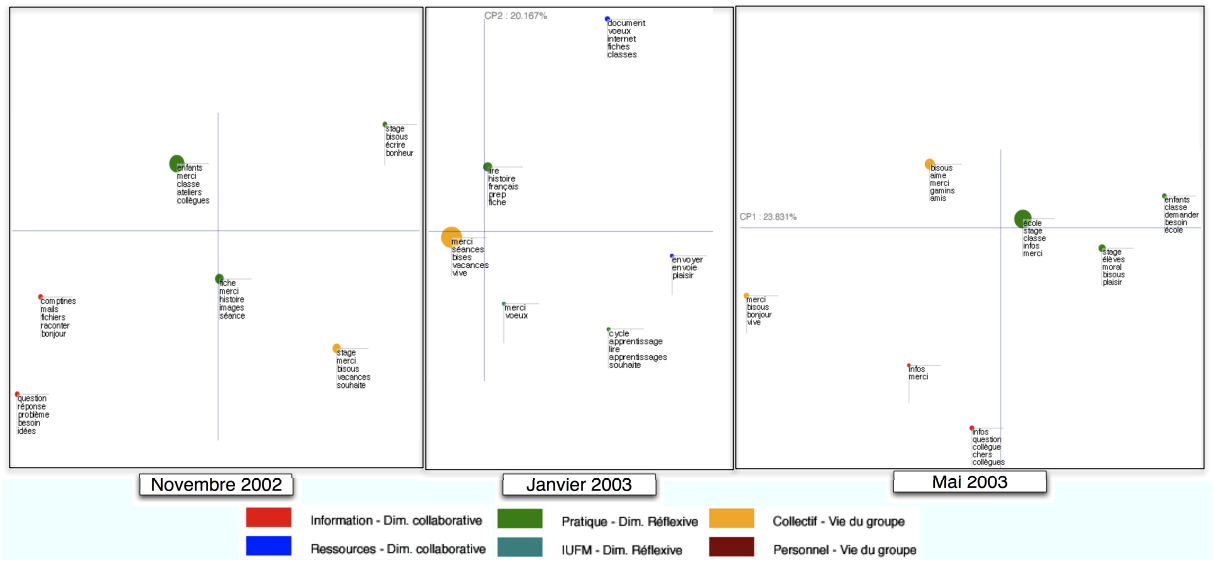
Dans le cas précédent, la sous-carte a mis en évidence la présence conjointe de différents domaines dans des messages. Le groupe C de la carte du forum $E1_{2002-2003}$ est lui de domaine majoritaire *Collectif*. En observant sa sous-carte (figure 4.19, partie droite), nous avons remarqué un usage beaucoup plus restreint des autres domaines avec une certaine homogénéité des messages dans le domaine du collectif, comme l'illustre le message suivant extrait du groupe :

> *Coucou tout le monde, SNIFFF c'est fini, non que je ne suis pas contente de vous revoir mais je me suis comme qui dirait un peu attaché à ces petits bouts. Je travaille encore demain mais je n'ai pas d'enfants le samedi. Alors comment cela se passe pour samedi, chacun raméne sa bouffe ou on fait quelque chose pour tout le monde ou presque (on ne va pas nourrir tout l'IUFM) ? Sinon pour ma part je suis invitée vendredi midi au repas de noël de l'école, sympa!!!! Bisous à vous tous (...)*

Comme nous l'avons constaté précédemment, le forum $E1_{2002-2003}$ aborde très peu le domaine des ressources pédagogiques, contrairement aux autres forums où ce domaine est présent de façon significative. Un autre forum se « démarque », $E2_{2003-2004}$ possède un groupe (le groupe D) assez important du domaine *Personnel*, contrairement aux autres forums où ce domaine n'apparaît pas de façon majoritaire dans les groupes de messages. Le groupe D est assez homogène dans ce domaine, les sujets abordés dans les messages qu'il contient sont très souvent liés à des difficultés personnelles rencontrées pendant les premières heures d'enseignement, comme l'illustre l'extrait suivant provenant d'un message du groupe :

> *Salut les E2... Moi c'est M. Jaquard qui vient, comme pout Caroline, lundi après-midi ...apparemment il restera jusqu'à la récré uniquement, bonne nouvelle. (...) Je me répète peut-être, mais si vous avez une idée d'activité arts plastiques à mener en atelier en grande section, suceptible de plaire à M. Jaquard, ça m'intéresse ! Sinon, mon premier jour a été très difficile, très fatigant, très démoralisant, mes ateliers sont partis en vrille, mes consignes mal comprises, mon emploi du temps à refaire...une horreur. Heureusement, aujourd'hui c'était mieux, même si mes nuits blanches ont commencé à peser brutalement en début d'après-midi (une séance sieste accompagnée n'aurait pas été de trop..) Voilà pour le moment... (...)*

Après cette analyse globale des différents forums, nous avons commencé à interroger la façon dont les domaines sont abordés au fil du temps dans les forums. La figure 4.20 présente trois extraits de la carte temporelle associée aux forums $E1_{2002-2003}$, chaque extrait représente les messages échangés sur une période d'un mois.



Période	Nombre de messages	Domaines les plus présents	Mots les plus fréquents
novembre 2002	49	*Pratique* (114 occ.), *Collectif* (71 occ.), *Ressources* (31 occ.)	*merci*, *enfants*, *classe*, *séances*
janvier 2003	56	*Collectif* (67 occ.), *Pratique* (44 occ.), *Ressources* (27 occ.)	*merci*, *vive*, *cycle*, *séances*
mai 2003	60	*Pratique* (111 occ.), *Collectif* (87 occ.), *Personnel* (48 occ.)	*école*, *stage*, *infos*, *classe*

FIG. 4.20 – Description de trois extraits de la carte temporelle du forum $E1_{2002-2003}$.

Au mois de novembre, les échanges portent principalement sur le domaine *Pratique*. Les messages abordent essentiellement l'organisation des enseignements et la demande de ressources. Une sorte de mise en route du forum est en quelque sorte observée. Au mois de janvier, le domaine *Pratique* est moins présent, le domaine *Collectif* prenant le relais avec un grand nombre de messages de soutien et d'encouragement échangés entre les stagiaires. La demande de ressources pédagogiques est, quant à elle, toujours présente. Enfin au mois de mai, le domaine *Pratique* redevient le plus important juste devant le domaine *Collectif*. Le domaine *Personnel* apparaît également, les stagiaires commencent à faire part de leurs sentiments personnels sur leur expérience d'enseignement déjà vieille de plusieurs mois.

Les différentes cartes présentées ici ont entraîné un premier retour sur l'acquisition de l'identité professionnelle. Par la projection des différents domaines abordés dans les forums, l'entraide des stagiaires, le partage de ressources pédagogiques, les difficultés rencontrées, l'organisation des séances ont pu être mis en évidence. L'organisation du forum au fil de l'année a également pu être interrogée, confirmant ainsi une phase de « mise en route », une phase plutôt liée aux difficultés des stagiaires et une dernière phase de bilan personnel des premières expériences d'enseignement et de prise de conscience de la profession d'enseignant.

La mise en parallèle des cartes des différents forums a permis de comparer ces derniers. Ces comparaisons ont aussi bien permis d'isoler un forum abordant peu le domaine des ressources pédagogiques, et un autre se démarquant en abordant de façon significative le domaine *Personnel*. Ces différentes comparaisons peuvent permettre une intervention ciblée du formateur, soit pour insister sur un point manquant, soit pour aider les stagiaires en éventuelle difficulté.

La collaboration avec Georges Ferone, qu'a impliquée cette expérimentation, a débuté récem-

ment, en janvier 2007 et est encore active actuellement. Notre principale perspective de recherche porte sur l'analyse du contenu d'un forum en cours de réalisation. L'objectif est de pouvoir suivre les domaines abordés par les stagiaires au fil du temps et ainsi de mettre en avant d'éventuels manques ou originalités. Également, la notion de fil de discussion n'est pas prise en considération dans les analyses. En tenir compte pourrait permettre de mieux caractériser le contenu de messages et surtout les liens entre ces derniers.

Dans cette expérimentation, les domaines considérés étaient assez généraux et décrits de façon assez simple, par un petit nombre d'éléments. Ces domaines ont ensuite été projetés sur des forums, sans consigne donnée au préalable et où les formateurs n'intervenaient que très peu. La section suivante présente un autre travail d'analyse de forums de discussion, prenant place également dans l'ERTé CALICO. Cette fois-ci les domaines considérés seront décrits plus finement. Les forums étudiés seront particulièrement guidés par les formateurs afin de faire interagir les stagiaires sur un problème bien précis.

4.4.2 Observation des usages d'une terminologie professionnelle

La tâche visée par ce second travail sur les forums de discussion consiste à observer l'usage sur des forums dédiés d'une terminologie professionnelle enseignée pendant la formation. Cette tâche a été menée dans le cadre d'une collaboration avec Nicole Clouet et Marie-Laure Compant-Lanfontaine, formatrices à l'IUFM de Caen intervenant dans le CAPES de documentaliste.

Projection d'une terminologie professionnelle existante sur des forums pédagogiques

Par « terminologie professionnelle » (appelée également « lexique professionnel » ou encore « lexique de référence »), les formatrices désignent l'ensemble des termes-clés enseignés aux stagiaires durant leur formation d'enseignants-documentalistes. Ces termes-clés appartiennent à différents domaines liés à la discipline tels la pédagogie, la documentation, l'informatique, le droit, etc. Au début de cette collaboration, une première terminologie a été fournie par Nicole Clouet. Cette terminologie a été construite par la formatrice au fil du temps, en rassemblant des termes qu'elle jugeait pertinents. La terminologie ainsi produite regroupe 273 formes graphiques de termes correspondant soit à des formes lemmatisées de lexies (comme *bibliographie*), soit à des formes fléchies de lexies (comme *bibliothèques électroniques*), soit à des sigles (comme *FADBEN*), répartis en douze catégories définies par Nicole Clouet. Chaque catégorie correspond à un domaine de la formation des enseignants documentalistes : *Document*, *Droit*, *Documentaliste*, *Lecteur/Usager*, *Bibliothéconomie*, *Traitement documentaire*, *Espace documentaire*, *Fonds documentaire*, *Politique documentaire*, *Recherche*, *Information*, *Technologies de l'information*. Chaque catégorie contient un nombre très variable de termes allant de moins de 10 pour certaines à plus de 50 pour d'autres[171].

Afin d'observer l'usage de cette terminologie professionnelle dans des forums de discussion où échangent les stagiaires, nous avons réalisé une première projection des différents domaines sur ces forums. Les forums de discussion étudiés sont élaborés dans le cadre de ce que les formatrices appelle des « études de cas » [Clouet, 2005]. Une étude de cas a pour objectif de confronter les stagiaires à des situations pédagogiques problématiques. Une étude de cas est un procédé se déroulant en plusieurs étapes :

1. Tout d'abord, le récit d'un problème professionnel vécu est présenté aux stagiaires ;
2. Une analyse guidée du cas est ensuite réalisée afin de garantir la compréhension de la situation ;

[171] Cette terminologie professionnelle « version initiale » est disponible en annexe C.

3. À la suite de cette analyse, un forum non modéré est créé afin de permettre aux stagiaires d'échanger et de débattre autour du problème, ceci sur une période d'environ quatre semaines ;

4. À la clôture du forum, la trace des échanges, exploitée en tant que texte, devient le matériel de travail de deux journées de formation en présentiel.

Les forums de discussion que nous analysons prennent donc place dans un contexte bien précis avec un fort objectif de formation. Deux forums de discussion ont été tout particulièrement étudiés :

- le premier, que nous intitulons *Utilisation de l'informatique*, traite de la correcte utilisation de l'informatique au CDI. Un problème lié à un élève (Maxime) est traité, cet élève utilisant les ordinateurs du CDI de façon inappropriée en se connectant à Internet et en exécutant des applications non autorisées. Ce forum contient 113 messages échangés du 11/11/2003 au 15/12/2003.
- le second, que nous intitulons *Confidentialité des prêts*, traite de la confidentialité des prêts de livres au CDI. Le cas d'un élève (Florian) est abordé, ce dernier ressentant une forte gêne à emprunter des livres sur les relations sexuelles et à inscrire son nom sur les fiches de prêt associées aux livres. Ce forum contient 62 messages échangés sur la période du 12/11/20004 au 06/12/2004.

Des projections des différents domaines de la terminologie professionnelle sur ces deux forums ont été réalisées. La figure 4.21 présente les cartes des groupes de messages en 2 dimensions. Ces cartes ont été obtenues avec un comptage relatif, associé à une ACP (taux d'inertie autour de 40% pour chacune des cartes) et suivie d'une CHA automatique (7 groupes ont été obtenus pour les deux forums).

Les deux cartes ainsi présentées révèlent les domaines majoritairement représentés dans les forums. Ainsi, la carte du forum *Utilisation de l'informatique* met principalement en évidence des groupes majoritairement du domaine *Technologie de l'information* (5 groupes sur 7, dont un groupe de taille très importante à la gauche de la carte). La carte du forum *Confidentialité des prêts* révèle plusieurs petits groupes, le domaine *Information* est le plus présent (4 groupes sur 7). Ces deux cartes proposent ainsi un accès global au contenu des forums en isolant les domaines principalement abordés.

La navigation sur les cartes a permis d'avoir un premier retour sur les usages de la terminologie professionnelle dans les forums. Il est ressorti un aspect assez générique de l'usage d'éléments de la terminologie professionnelle, comme l'illustrent les dix graphies de la terminologie les plus fréquentes dans le forum *Utilisation de l'informatique* : *charte*, *internet*, *formation*, *site*, *recherche*, *organisation*, *classe*, *consultation*, *portail*, *fichier*. La faible présence d'éléments de la terminologie dans les forums a également été perçue (225 occurrences de graphies de la terminologie dans les 113 messages de *Utilisation de l'informatique* et seulement 73 occurrences pour les 62 messages de *Confidentialité des prêts*). Cette faible présence a entraîné un retour, pas vraiment sur les forums, où les stagiaires ont correctement échangé, mais plutôt sur la terminologie professionnelle utilisée. En effet, cette terminologie, dont la conceptrice (Nicole Clouet) avait déjà repéré certaines limites, s'est révélée un peu trop générique, voire incomplète sur certains domaines (notamment dans les domaines de la formation, de la pédagogie ou encore de la communication).

Une version enrichie de cette terminologie a alors été mise au point et de nouvelles analyses ont eu lieu. C'est ce que nous détaillons dans la section suivante de cette thèse.

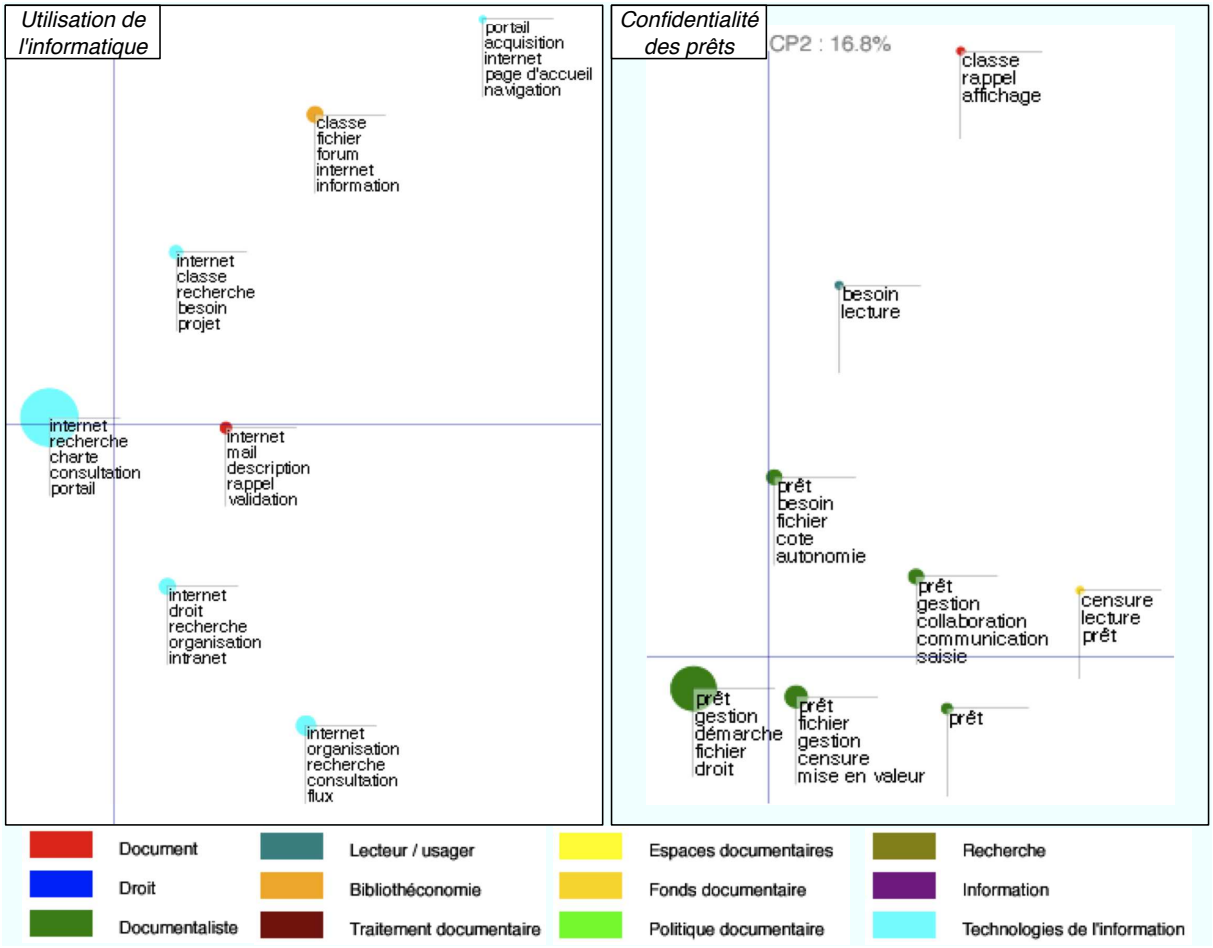


FIG. 4.21 – À gauche, carte des groupes du forum *Utilisation de l'informatique*, à droite, carte des groupes du forum *Confidentialité des prêts*.

Élaboration d'une nouvelle terminologie métier

Pour élaborer une nouvelle version de cette terminologie professionnelle, le besoin de retourner sur des textes faisant référence dans la profession d'enseignant documentaliste a été ressenti. Ces textes (environ vingt) rédigés par des spécialistes issus du corps de l'inspection générale, de la recherche, du monde professionnel, prennent place dans les différents domaines de la discipline[172]. Afin d'extraire de ces textes des lexies candidates à la terminologie professionnelle, nous avons principalement utilisé les outils *Memlabor* et *FlexiSemContext*, respectivement pour extraire les graphies répétées des textes et consulter leurs contextes d'apparition dans les textes.

Après sélection par Nicole Clouet et Marie-Laure Compant-Lanfontaine de graphies proposées par les outils précédents et leur éventuelle composition en lexies (dans leur forme lemmatisée), une terminologie de 800 termes a pu être mise au point[173]. Tout comme pour la première terminologie, les différents termes ont été catégorisés en différents domaines. Afin de compléter cette terminologie, mais également d'affiner sa structuration en domaines, il a ensuite été décidé d'utiliser le thésaurus *MOTBIS*[174]. Ce thésaurus est réalisé par le Centre National de Documentation Pédagogique (CNDP), il propose de recenser les différents termes liés à la profession d'enseignant. Les termes du thésaurus sont organisés hiérarchiquement, avec principalement des « méta-termes », des « termes génériques » et des « termes spécifiques ». Dans l'exemple de la figure 4.22 nous présentons une sortie de *MOTBIS* sur le terme *bibliothéconomie*. Le thésaurus considère ce terme comme un terme générique regroupant les termes spécifiques cités en dessous du terme dans la figure. Le terme *bibliothéconomie* est lui-même caractérisé par le méta-terme *documentation*.

L'utilisation de ce thésaurus a permis de compléter en nombre notre terminologie, qui regroupe maintenant environ 1 200 termes, mais aussi d'amorcer une description plus fine des différents domaines. Ainsi, deux grandes catégories de domaines ont été distinguées : ceux liés à l'*éducation* et ceux liés à la *discipline*. La première de ces grandes catégories, en cours de structuration à l'écriture de ces lignes, contient dix domaines communs à la profession d'enseignant, comme le domaine de l'administration du système éducatif, des élèves, de la psychologie, de la formation, etc. La seconde catégorie, mieux finalisée, regroupe sept domaines relatifs à la discipline des enseignants documentalistes. Ces domaines sont les suivants : *Bibliothéconomie* (141 termes[175]), *Communication* (34 termes), *Politique documentaire* (140 termes), *Recherche d'information* (96 termes), *Document* (68 termes), *Métier du documentaliste* (115 termes) et *Source d'information*. De cette manière, 594 termes se répartissent dans les sept domaines de la catégorie de la discipline.

Une première projection de ces domaines sur les forums *Utilisation de l'informatique* et *Confidentialité des prêts* est présentée en figure 4.23 (des ACP ont été utilisées à l'issue de comptages relatifs, taux d'inertie de l'ordre de 55% dans les deux cartes, suivies de CHA automatiques).

[172] Notamment, des textes issus des sites spécialisés *Savoirs-CDI* - `http://savoirscdi.cndp.fr` (page consultée le 6 juillet 2007) et *Doc pour docs* - `http://docsdocs.free.fr` (page consultée le 6 juillet 2007) ont été consultés.

[173] La construction de cette terminologie ne s'est pas basée sur la terminologie précédente et est repartie d'une liste vide. Nous avons tout de même pu vérifier que la majorité des termes de la première terminologie se retrouvait dans la nouvelle terminologie.

[174] Thésaurus consultable en ligne : `http://www.thesaurus.motbis.cndp.fr/site` (page consultée le 6 juillet 2007).

[175] Ce nombre regroupe les termes et certaines de leurs formes fléchies jugées appropriées. Par exemple, pour le terme *référence bibliographique*, la flexion plurielle *références bibliographiques* a été ajoutée, par contre, pour le terme *désherbage de documents*, la flexion plurielle n'a été jugée pertinente (aucune autre flexion d'ailleurs).

	DF NA EP TA
TG bibliothéconomie	TA
TS aménagement d'un centre documentaire	
TS catalogage	EP TA
TS coopération documentaire	EP
TS gestion documentaire	
TS accès aux documents	EP
TS libre accès	
TS acquisition de documents	EP
TS désherbage de documents	EP
TS prêt de documents	
TS PEB	EP
TS stockage de documents	TA
TS archivage de documents	TA
TS classement de documents	TA
TS conservation de documents	
TS identification des documents	TA
TS dépôt légal	TA
TS ISBN	EP
TS ISSN	EP
TS politique documentaire	

FIG. 4.22 – Extrait du thésaurus *MOTBIS* pour le terme *bibliothéconomie*.

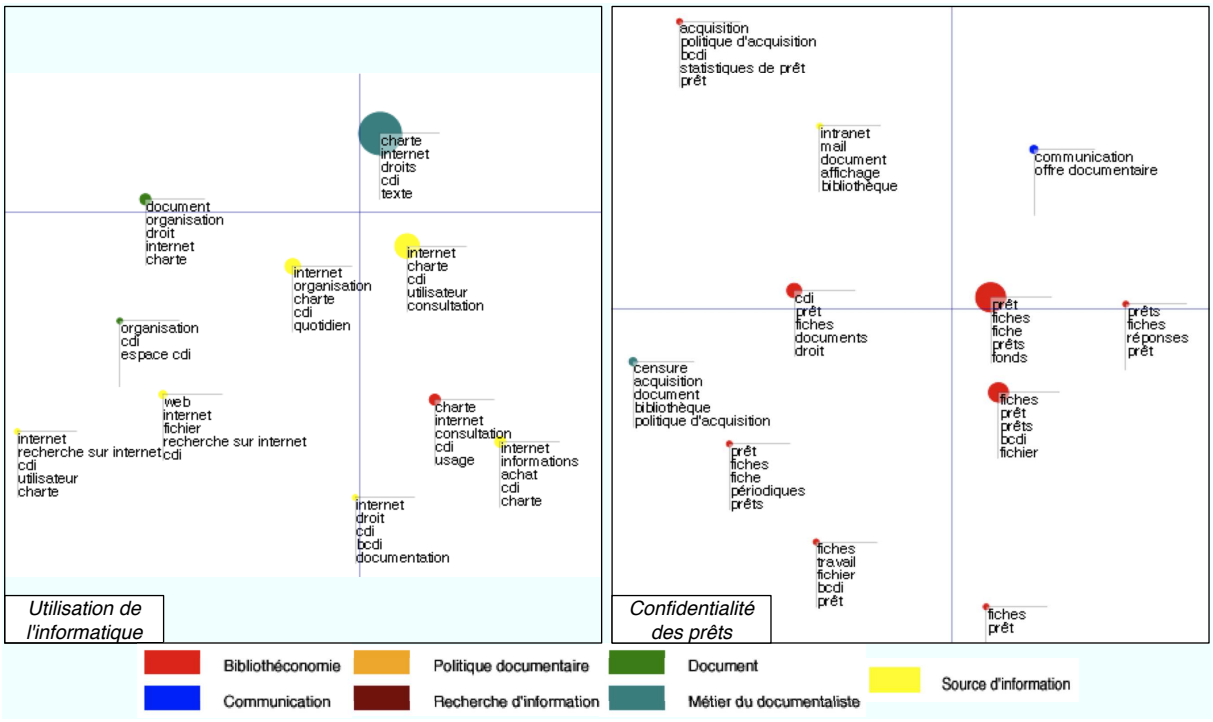


FIG. 4.23 – À gauche, la carte des groupes du forum *Utilisation de l'informatique*. À droite, la carte des groupes du forum *Confidentialité des prêts*, ces cartes ont été construites avec la nouvelle version de la terminologie.

La carte du forum *Utilisation de l'informatique* met ainsi tout particulièrement en évidence le domaine *Métier du documentaliste* (principalement avec les termes *charte*, *droit*, *travail*, *partenaire*) et suivi du domaine *Source d'information* (principalement avec *internet*, *CDI*, *texte*, *web*). Avec la première version de la terminologie, le domaine *Technologies de l'information* était très majoritaire. Ce domaine était lié au contenu du forum mais restait très général. Dans la nouvelle carte, les deux domaines liés à l'utilisation de l'informatique et au rôle de l'enseignant documentaliste ressortent tout particulièrement. Ces domaines caractérisent bien les sujets abordés dans le forum, et surtout l'idée de la création et de l'application d'une charte informatique au CDI.

Dans la carte du forum *Confidentialité des prêts*, le domaine *Bibliothéconomie* est majoritaire, principalement avec les termes *prêt*, *fiche*, *BCDI*, *fonds*, *fichier*. Le domaine *Métier de documentaliste* ressort de façon plus secondaire avec les éléments *censure*, *travail*, *collaboration*, rôle. Dans la carte de ce même forum avec la première terminologie, le domaine *Documentaliste* était majoritaire. Là encore, ce dernier domaine s'était révélé assez général et la nouvelle carte fait mieux ressortir les particularités du forum, à savoir la problématique des différentes étapes du prêt et du rôle du documentaliste confronté à des situations d'embarras des élèves.

Ces toutes premières projections donnent un retour assez positif sur la nouvelle terminologie, en tout cas sur la partie liée aux domaines de la discipline. Une analyse un peu plus précise des différents forums a ainsi pu être réalisée. Une première navigation sur les cartes des deux forums a permis de mettre en évidence l'usage de termes « professionnels » présents de façon au moins équivalente aux termes plus courants[176].

Ce travail de mise au point d'une terminologie métier propre à la profession d'enseignant documentaliste a débuté fin 2006 et des travaux sont toujours en cours à l'heure actuelle, principalement pour structurer la terminologie en domaines, et plus particulièrement dans la catégorie principale de l'éducation. La principale perspective de ce travail est d'affiner encore plus la stucturation des différents domaines. Par exemple, au domaine *Bibliothéconomie*[177], nous pourrons associer les cinq sous-catégories données au tableau 4.7.

Nom de la catégorie	Lexies associées
Accès au document	*accès à distance*, *accès à Internet*, *accès aux documents*, *accès aux ressources*, *modalités d'accès*, etc.
Accueil du public	*accueil*, *cahier de suggestion*, *ouverture*, *prêt*, *relance*, etc.
Fonds documentaire du CDI	*achat*, *acquisition*, *archivage*, *budget*, *conservation de documents*, etc.
Circuit du document	*bulletinage*, *cataglogage*, *catalogue partagé*, *chaîne du document*, *dépouillement*, etc.
Espace du CDI	*aménagement*, *banque de prêt*, *bureau de prêt*, *équipement*, *espace CDI*, etc.

TAB. 4.7 – Sous-catégories liées au domaine de la bibliothéconomie.

L'élaboration d'un dispositif *LUCIA* à partir de chaque domaine tenant compte de telles sous-catégories constitue certainement ce que nous aborderons dans la prochaine phase de notre collaboration avec Nicole Clouet et Marie-Laure Compant-Lafontaine à travers notre double objectif de mise au point d'une terminologie professionnelle et de mise évidence dans des forums de l'usage de cette terminologie.

[176] Par exemple, nous avons pu observer en navigant sur les cartes que le terme *acquisition* était présent sur les deux forums à six reprises, autant que son pendant « courant » *achat*.

[177] Ce terme fait référence à l'ensemble des techniques de gestion et d'organisation des bibliothèques.

4.4.3 Apports de vues globales et interactives dans l'accès au contenu de forums pédagogiques

Nous avons présenté dans cette section deux expérimentations dont l'objet d'étude était un forum de discussion pédagogique. Dans la première expérimentation, nous avons exploité des ressources simples pour étudier une certaine acquisition de l'identité professionnelle par des stagiaires professeurs des écoles lors de leurs premières semaines d'enseignements. Dans la seconde expérimentation, des ressources plus complexes ont été construites afin d'interroger l'usage d'une terminologie professionnelle par des stagiaires enseignants documentalistes. Dans tous les cas, les cartes proposées ont permis d'appréhender, tout d'abord de façon globale, les domaines choisis par les formateurs dans les forums. Ensuite des analyses plus locales au niveau des messages ont pu être menées afin d'apporter de premiers éléments de réponses aux différentes problématiques.

Notre schéma d'évaluation appliqué à ces expérimentations est alors le suivant :

- *Étape 1 : phase de constitution des ressources*
  De grandes différences ont été rencontrées entre les deux expérimentations. La première a nécessité la construction de six domaines très simples. Une telle construction n'a nécessité que quelques heures de travail pour le formateur. Dans la seconde expérimentation, l'objectif visé a entraîné la construction de ressources terminologiques plus complètes. Après avoir exploité une première version de cette terminologie dans de premières analyses, il a été décidé de faire une nouvelle version plus complète. La construction de cette terminologie a débuté il y a plusieurs semaines et n'est pas encore finalisée actuellement. Dans les deux expérimentations, la représentation des domaines d'intérêt des utilisateurs en ensembles de graphies a été utilisée. Dans la première expérimentation cette représentation s'est révélée suffisante, alors que dans la seconde, un besoin de plus de finesse de description a été ressenti. Les outils *Memlabor* et *FlexiSemContext* ont été tous les deux utilisés pour cette phase de construction des ressources.
- *Étape 2 : phase d'exécution des outils*
  ProxiDocs a été utilisée pour projeter les ressources sur les forums. Plusieurs exécutions ont été nécessaires afin de tenir compte des différentes versions des ressources. La phase d'exécution a été assez courte, de l'ordre de quelques dizaines de minutes.
- *Étape 3 : phase d'évaluation des sorties produites*
  Les cartes de textes et de groupes de textes en 2 dimensions ont principalement été analysées. Une carte temporelle a également été présentée pour la première expérimentation. Les premières analyses permises par les cartes ont révélé aussi bien un retour sur les forums et leur contenu que sur les ressources utilisées.

La problématique de l'accès au contenu de forums de discussion constitue une thématique de recherche très active à l'heure actuelle. La nature complexe de cet objet d'étude[178] a entraîné différents travaux sur sa représentation et sa visualisation. Pour une présentation de tels travaux, nous renvoyons à [Huynh-Kim-Bang et Bruillard, 2005]. Dans cet article, les auteurs proposent également l'outil *Bobinette* présenté en figure 4.24. Pour améliorer notre contribution à l'accès visuel et personnalisé à des forums de discussion, une perspective intéressante de recherche et de développement serait d'associer nos représentations cartographiques de forums à des visualisations comme celles proposées par l'outil *Bobinette* où la visualisation est particulièrement centrée sur les fils de discussion du forum, dimension qui est ignorée dans nos analyses. Par ce « couplage », les visualisations retournées pourraient mieux prendre en considération les particularités des forums, et en particulier l'enchaînement des messages.

[178] Selon [Mangenot, 2002], les forums de discussion possèdent quatre dimensions (écrite, asynchrone, publique et structurée), rendant la représentation de forums particulièrement difficile.

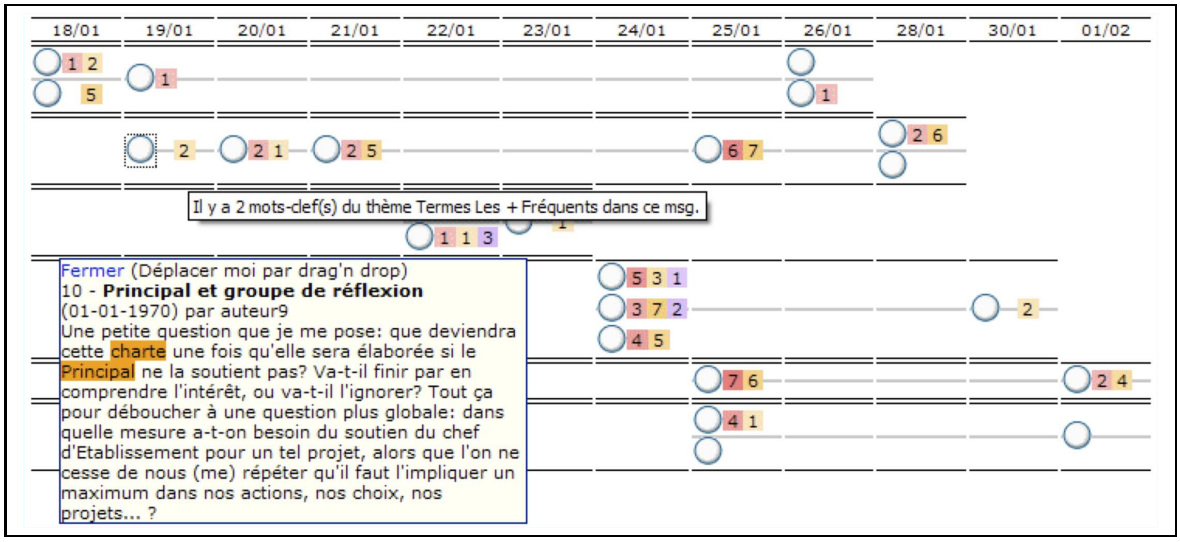


FIG. 4.24 – Un écran de l'outil *Bobinette* de [Huynh-Kim-Bang et Bruillard, 2005], outil disponible en ligne : `https ://wims.crashdump.net/www/forum/bobinette.php` (page consultée le 7 juillet 2007).

L'aspect visuel des supports d'analyses que nous proposons s'est ainsi révélé particulièrement important pour chacune des différentes expérimentations présentées dans ce chapitre. C'est souvent en quelques regards posés sur les cartes que les utilisateurs isolent les principales informations sur les domaines présents dans l'ensemble documentaire, informations qui orientent fortement toute l'analyse de l'utilisateur. Afin d'interroger la lisibilité des visualisations que nous proposons aux utilisateurs, et en particulier la façon dont leurs premiers regards se posent sur les vues, nous proposons en partie suivante une expérimentation réalisant une telle interrogation auprès d'un large panel d'utilisateurs.

4.5 Mesurer les premiers regards portés sur des cartes d'ensembles documentaires

4.5.1 Motivations et contexte

Les expérimentations précédentes ont permis de mettre tout particulièrement en évidence la pertinence des supports cartographiques que nous proposons pour l'accès au contenu d'ensembles documentaires. Par différentes interactions sur les cartes et les RTO, il a été possible d'isoler des informations pertinentes décrivant aussi bien le contenu de textes ou de sous-ensembles de textes de l'ensemble documentaire que de l'ensemble dans sa globalité. Au cours de ces expérimentations, les différents utilisateurs nous ont fait ressentir la grande importance du premier accès visuel aux cartes. Ce premier regard sur les cartes, sans manipulation de la part des utilisateurs, semble tout particulièrement orienter l'analyse globale de l'ensemble documentaire.

Pour interroger la variabilité de ce regard, nous nous sommes intéressés à des protocoles expérimentaux en psychologie portant sur le suivi du regard, appelé encore *eyetracking*[179]. À

[179] Nous renvoyons à [Salvucci, 2001] pour une présentation détaillée des principes d'études portant le suivi du regard.

l'aide d'un dispositif approprié[180], il est proposé de suivre le regard d'un sujet sur une image, le plus souvent fixe. Un tel suivi permet, entre autres, d'isoler des zones auxquelles le sujet porte un intérêt particulier. À l'opposé, des zones peu visitées sont également mises en évidence. L'ordre de parcours des différents éléments peut également être représenté.

De nombreux travaux en psychologie portent sur le suivi de regard sur des documents textuels [Rayner, 1998] (la figure 4.25 présente un tel parcours du regard sur du texte) et des images [Breux *et al.*, 2007]. Plus récemment, ces travaux se sont également intéressés aux interfaces de logiciels [Goldberg et Kotval, 1999] et aux pages de sites Internet [Outing et Ruel, 2004].

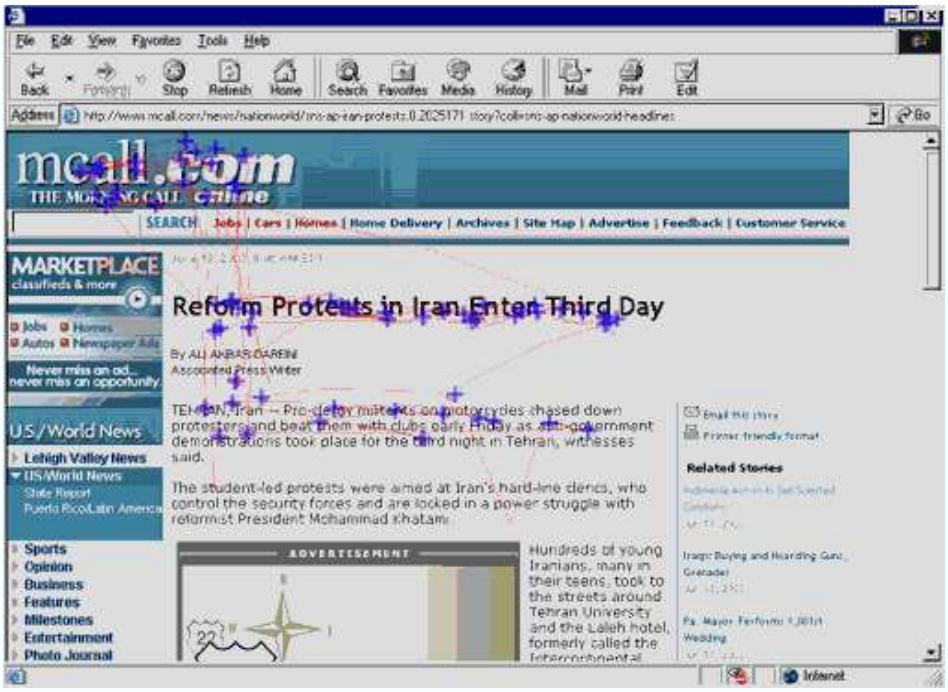


FIG. 4.25 – Suivi du regard sur un texte, `http://www.alexpoole.info/academic/lecturenotes_fr.html` (page consultée le 9 juillet 2007). Le suivi du regard mis en évidence est caractéristique de la lecture d'un texte avec un parcours linéaire des différents mots.

Aux intérêts scientifiques portant sur l'étude des processus cognitifs liés à la visualisation d'information, des intérêts commerciaux sont venus s'ajouter afin, par exemple, d'isoler sur des pages de sites Internet des zones candidates à la présence de publicités[181]. Par exemple, une telle étude a été réalisée sur la première page de résultats proposée par le moteur de recherche *Google*. Il est ainsi ressorti l'importance du coin supérieur gauche (appelé « triange d'or »), lieu où les regards des sujets viennent principalement se poser (figure 4.26 ; partie gauche). Dans [Outing et Ruel, 2004], cité précédemment, les auteurs sont même allés plus loin en proposant un parcours du regard « typique » des pages d'accueil de sites Internet, symbolisé dans la figure 4.26 (partie droite) par le chemin tracé en rouge.

[180] Un tel dispositif est une caméra filmant les mouvements de l'œil, reliée à un ordinateur retranscrivant ces mouvements en fixations et déplacements sur l'image visualisée.

[181] Nous avons également présenté au début de ce chapitre, l'utilisation de telles techniques pour améliorer la disposition des éléments sur une page d'un site Internet (figure 4.1, page 119).

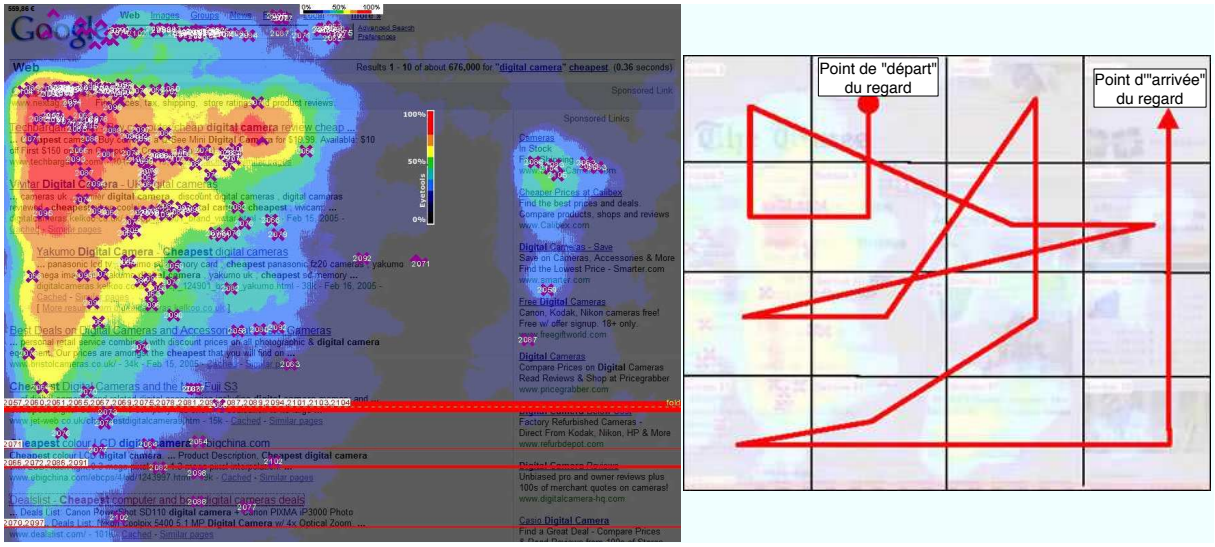


FIG. 4.26 – À gauche, suivi du regard sur une page de résultats du moteur de recherche *Google*, `http://www.eyetools.com/inpage/research_google_eyetracking_heatmap.htm` (page consultée le 9 juillet 2007). À droite, parcours du regard « typique » sur une page d'accueil d'un site Internet selon [Outing et Ruel, 2004].

Par une telle expérimentation appliquée à notre problématique, notre principal objectif est de mettre en évidence la façon dont les utilisateurs parcourent les visualisations cartographiques que nous leur proposons et d'observer si des zones d'intérêt se distinguent (non pas pour y mettre des publicités, mais pour éventuellement ajuster la disposition des éléments sur les cartes). Il est également intéressant d'étudier si des catégories de sujets émergent selon leurs parcours oculaires, afin d'adapter le plus possible nos supports de visualisation d'ensembles documentaires aux utilisateurs.

Ce travail a été réalisé dans le cadre du pôle pluridisciplinaire *ModeSCo* (Modélisation en Sciences Cognitives) regroupant des informaticiens, des linguistes, des neuropsychologues, des psychologues et des chercheurs des Sciences et Techniques des Activités Physiques et Sportives de l'Université de Caen. En collaboration avec Henri Roussel et Stéphane Breux, psychologues, Pierre Beust et moi-même avons débuté en 2006 une expérimentation visant à enregistrer les parcours du regard de différents sujets sur les visualisations cartographiques que nous proposons pour l'accès au contenu d'ensembles documentaires.

4.5.2 Cadre expérimental

Pour réaliser cette expérience, nous avons dû élaborer un cadre expérimental approprié. Tout d'abord, nous avons utilisé un appareil adapté pour mesurer les mouvements du regard d'un sujet sur une image. Cet appareil, appelé également oculomètre[182], prend la forme d'une caméra fixe positionnée devant le sujet[183] et filmant l'un de ses yeux durant sa phase d'exploration de l'image étudiée. La caméra est reliée à un ordinateur afin d'enregistrer les mouvements du regard

[182] Le modèle utilisé est de la marque *ASL Eyetracking System*, plus de détails sont accessibles sur le site de la société : `http://www.a-s-l.com` (page consultée le 8 juillet 2007).

[183] Des modèles mobiles prenant la forme d'un casque existent également, l'intérêt de ces derniers est de permettre aux sujets de bouger contrairement au dispositif que nous avons utilisé où les sujets doivent rester immobiles.

du sujet sur l'image. Les résultats de cet enregistrement sont ensuite projetés sur les images parcourues à l'aide de logiciels dédiés[184]. La figure 4.27 présente un tel dispositif.

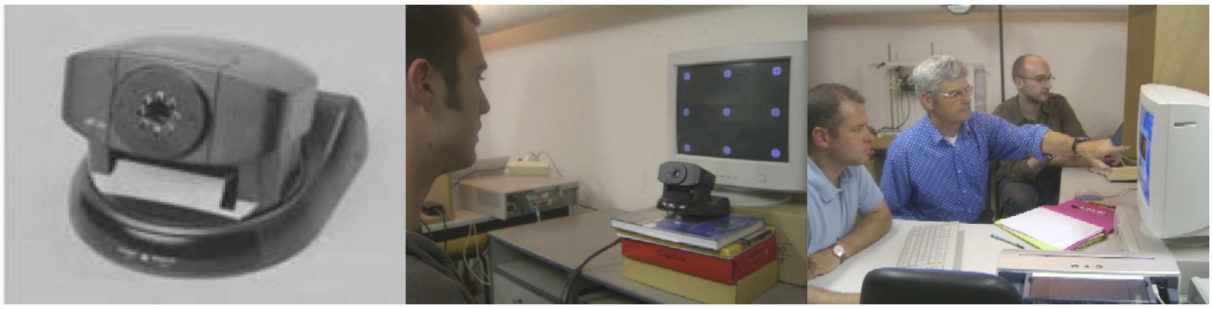


FIG. 4.27 – En partie gauche de la figure, la caméra utilisée ; au centre, le sujet positionné face à l'écran diffusant les images et à la caméra filmant l'un de ses yeux ; à droite, les expérimentateurs devant leurs écrans de contrôle pour la diffusion des images et l'enregistrement des mouvements oculaires du sujet.

La première tâche a consisté à sélectionner les différentes cartes que nous souhaitions présenter aux sujets. L'idée principale qui a guidé cette expérimentation est la « mise en condition » des sujets. Nous avons donc choisi de les mettre face à une tâche d'accès au contenu d'un ensemble documentaire (339 articles du journal *Le Monde* extraits aléatoirement de l'ensemble des articles du journal de l'année 1989) tenant compte de domaines particuliers (six domaines construits selon le modèle de représentation en ensemble de graphies : *justice*, *religion*, *violence*, *éducation*, *guerre*, *travail*). Des projections cartographiques selon le modèle *AIdED* de ces domaines sur l'ensemble documentaire vont être les différentes vues que nous allons proposer aux sujets. Quatre cartes statiques en 2 dimensions et une carte animée en 3 dimensions constituent ces vues. Parmi les cartes statiques en 2 dimensions, nous avons sélectionné une carte des documents et trois cartes des groupes de documents, les deux dernières faisant chacune figurer une lexie particulière mise en évidence par le survol d'une de ses occurrences sur la carte (figure 4.28). La carte en 3 dimensions est une carte des groupes de documents, à la différence des cartes précédentes représentées par des images fixes, cette carte est représentée par une vidéo.

Ces différentes cartes sont présentées aux sujets pendant 10 secondes chacune (durée déterminée après plusieurs essais afin d'avoir suffisamment de données sur le trajet du regard des sujets). Avant chaque carte, le sujet est mis en scène avec un texte lui disant quelle manipulation lui a permis d'obtenir la carte présentée. Techniquement, ces différentes vues et scénarios associés prennent place dans un diaporama où des informations liées au calibrage de l'appareil sont intégrées. C'est lors du passage de ce diaporama incluant les vues que nous mesurons les regards portés par les sujets sur ces dernières. Un tel diaporama est présenté en annexe D de cette thèse. Avant le passage du diaporama, nous avons choisi de lire un texte à chaque sujet expliquant succinctement ce qu'il allait voir et l'invitant également à ne pas bouger la tête. Après le passage du sujet, un questionnaire lui est posé afin de noter son profil, son degré d'utilisation de l'informatique et surtout ce qu'il a retenu du diaporama et des visualisations proposées, par exemple avec des questions portant sur le nombre de groupes visualisés, le nombre de fois où une lexie mise en relief a été visualisée, etc.

[184]Les logiciels *Eyenal* et *Fixplot* de la société *ASL Eyetracking System* ont été utilisés dans nos différentes analyses.

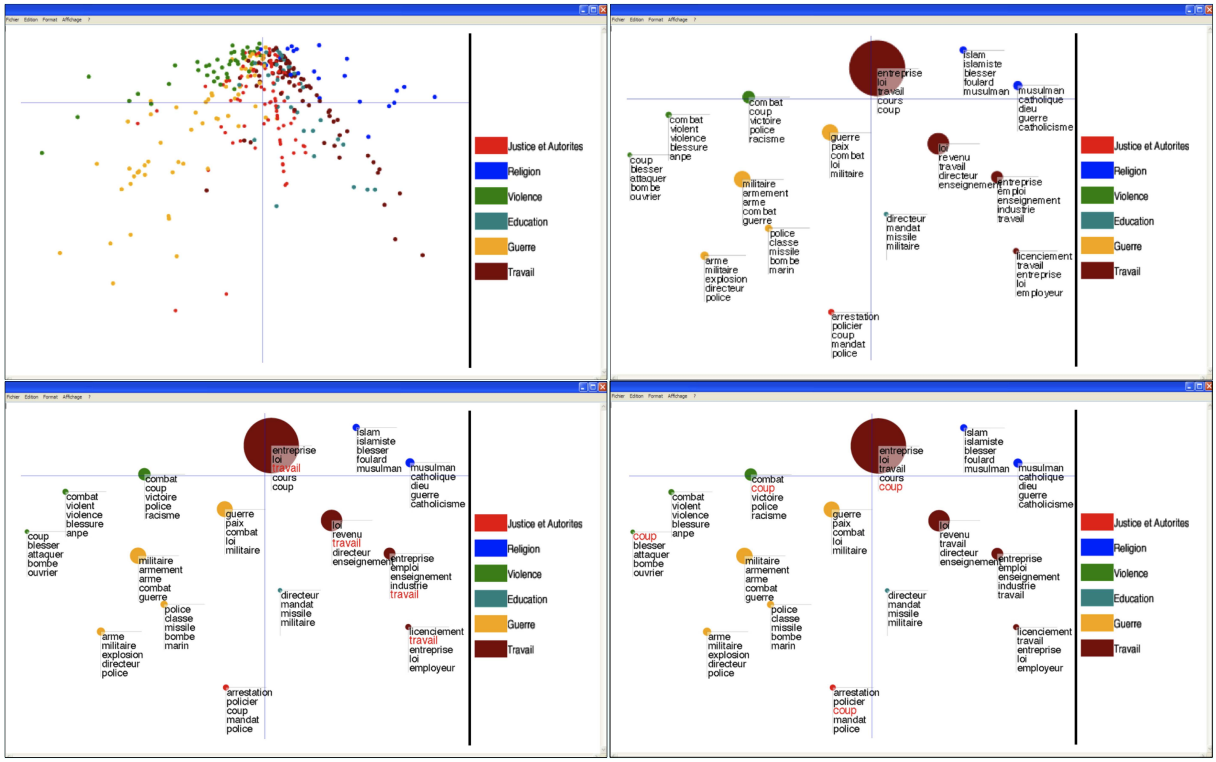


FIG. 4.28 – Les cartes en 2 dimensions proposées aux sujets. En partie supérieure gauche, la carte des documents, en partie inférieure droite, la carte des groupes de documents. Les cartes en partie inférieure reprennent la carte des groupes de documents avec les lexies *travail* (carte de gauche) et *coup* (carte de droite) mises en évidence.

De cette manière, nous avons proposé notre diaporama à 23 sujets. Les premiers sujets ont servis de « calibrage » et nous ont permis d'isoler des manques et des limites dans notre façon de procéder, et ainsi affiner notre protocole expérimental. Par exemple, nous avons pu rencontrer un certain nombre de difficultés d'ordre technique, avec des sujets portant des verres correcteurs rendant difficile l'enregistrement de leurs mouvements oculaires.

Des difficultés d'ordre plus « théorique » ont également été rencontrées. Ainsi, les premières versions de notre diaporama incluaient des questions à l'issue de chaque carte. Nous notions la réponse à cette question, mais de telles interactions directement avec les sujets au cours de diaporama entraînaient des mouvements oculaires particuliers liés à des tâches de mémorisation. Nous avons ainsi pu remarquer que pour « bien » répondre aux questions que nous leur posions, les sujets semblaient être particulièrement attentifs aux différentes cartes et cherchaient à mémoriser le plus d'éléments possibles. Lors d'une telle tâche de mémorisation, le sujet quitte souvent du regard l'objet mémorisé (en l'occurrence, la carte), cette perte du regard créant alors de grandes difficultés dans les enregistrements (se reporter à [Kinsbourne, 1972] pour une présentation et une explication de ce phénomène neuropsychologique).

Sur les 23 sujets que nous avons étudiés, 5 n'ont pas eu des mouvements oculaires exploitables. Les 18 sujets restants ont fait l'objet d'une analyse plus poussée, détaillée en section suivante.

4.5.3 Analyse des résultats et retour sur les cartes d'ensembles documentaires

Pour chacun de ces 18 sujets, nos avons obtenu cinq enregistrements de leurs parcours oculaires, un par carte proposée. En projetant ces parcours sur les cartes, nous avons pu mettre en évidence les différentes façons dont les sujets appréhendaient les cartes. Ces projections représentent le déplacement du regard par des lignes bleues reliant deux fixations. Une fixation est une immobilité du regard sur un point pendant une durée supérieure ou égale à 0,1 seconde. Plus la fixation est longue, plus elle est représentée par un disque de grande taille sur la projection. Les fixations sont numérotées selon leur ordre de parcours par le sujet. Juste avant de présenter chaque carte aux sujets, nous avons choisi de les contraindre à fixer un point au centre de l'écran. De cette manière, une fois la carte présentée, nous connaissons le point d'entrée du regard du sujet sur la carte. La figure 4.29 décrit les parcours du regard de deux sujets différents sur la première carte présentée, la carte des documents.

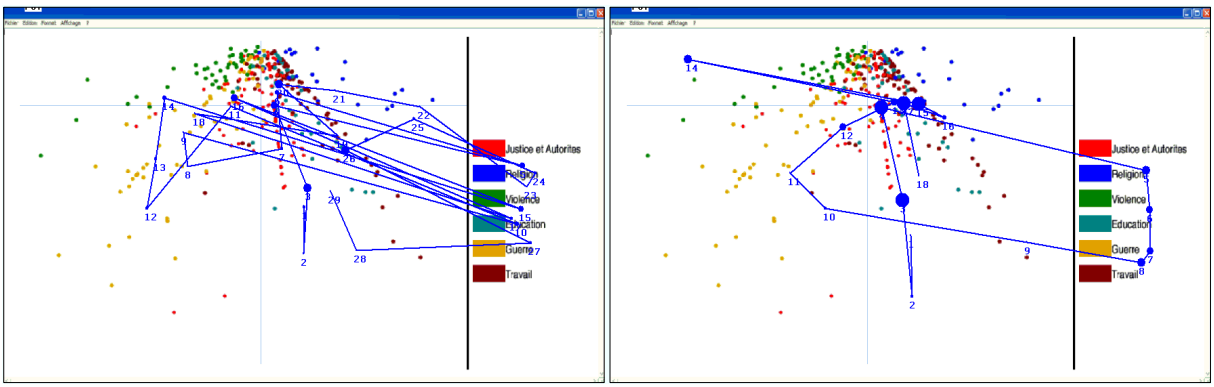


FIG. 4.29 – Parcours du regard de deux sujets différents sur la carte des documents.

Sur la carte de gauche, nous pouvons observer de grands et nombreux déplacements du regard du sujet avec beaucoup d'allers-retours entre la carte et la légende. 29 fixations très courtes (0,27 seconde en moyenne) ont été observées sur la carte. La carte de droite met en évidence un parcours du regard assez différent : les déplacements sont moins nombreux, il y a beaucoup moins d'allers-retours avec la légende (un seul parcours de toute la légende est réalisé). 18 fixations plus longues (0,53 seconde en moyenne) sont réalisées. De telles différences de comportements se retrouvent dans la carte des groupes de documents qui est la deuxième carte présentée aux sujets. Deux exemples de parcours de regard sur cette carte sont présentées en figure 4.30 (les sujets dont les parcours sont présentés sont différents de ceux dont les parcours sont présentés en figure 4.29).

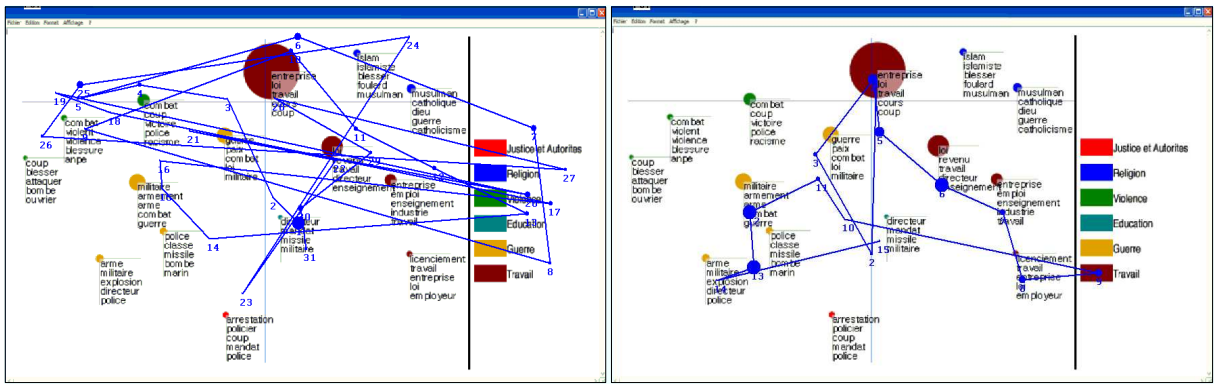


FIG. 4.30 – Parcours du regard de deux sujets différents sur la carte des groupes de documents.

Là encore, deux types de parcours sont observés. À gauche, le sujet a parcouru rapidement toute la carte avec 31 fixations très courtes (0,31 seconde de moyenne). Beaucoup d'allers-retours avec la légende sont également constatés. À droite, le parcours de la carte est plus partiel, les fixations étant moins nombreuses (15 fixations) et ne couvrant qu'une partie de la carte. Les fixations sont par contre plus longues (0,57 seconde de moyenne) et un seul allers-retour avec la légende a eu lieu. Les deux cartes présentées ensuite aux sujets illustrent le résultat d'interactions basiques sur la carte des groupes de documents (figure 4.31). Sur la première de ces deux cartes, le « scénario » mis en place dans le diaporama explique à l'utilisateur qu'il a visualisé la lexie *travail* étiquetant le groupe le plus important en partie haut/centre de la carte des groupes de documents. Pour la seconde de ces deux cartes, c'est la lexie *coup* qui est mise en évidence.

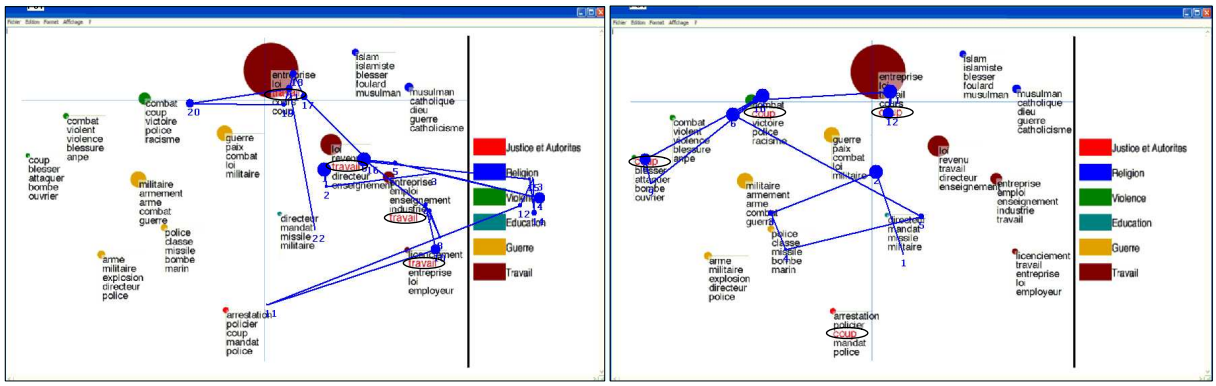


FIG. 4.31 – Parcours de deux sujets différents sur les cartes des groupes de documents mettant en évidence le résultat d'interactions.

Sur la carte de gauche, la lexie *travail* est donc mise en avant (pour plus de lisibilité, nous avons entouré ses différentes occurrences sur la carte). Pour celle de droite, c'est la lexie *coup* qui est mise en avant (dont les occurrences sont également entourées sur la carte). Sur la première carte, nous pouvons observer que toutes les occurrences de la lexie sont survolées par le regard du sujet. Au contraire, sur la seconde carte, l'occurrence de *coup* située en partie inférieure semble ignorée par le sujet. Cette occurrence, assez éloignée des autres, est d'ailleurs souvent ignorée, la mise en relief utilisée (couleur rouge et agrandissement de la taille des caractères) ne semble pas suffisante pour attirer le regard des différents sujets sur toutes les occurrences de la lexie, et surtout sur celles qui sont particulièrement éloignées des autres.

L'analyse de la carte en 3 dimensions a été plus délicate, les logiciels fournis avec l'oculomètre permettant de projeter le trajet d'un regard uniquement sur des images fixes et non sur des animations. La solution adoptée a consisté à projeter ce trajet sur une image vierge, puis à observer les différents tracés pour chaque sujet en parallèle avec l'animation de la carte en 3 dimensions, comme l'illustre la figure 4.32.

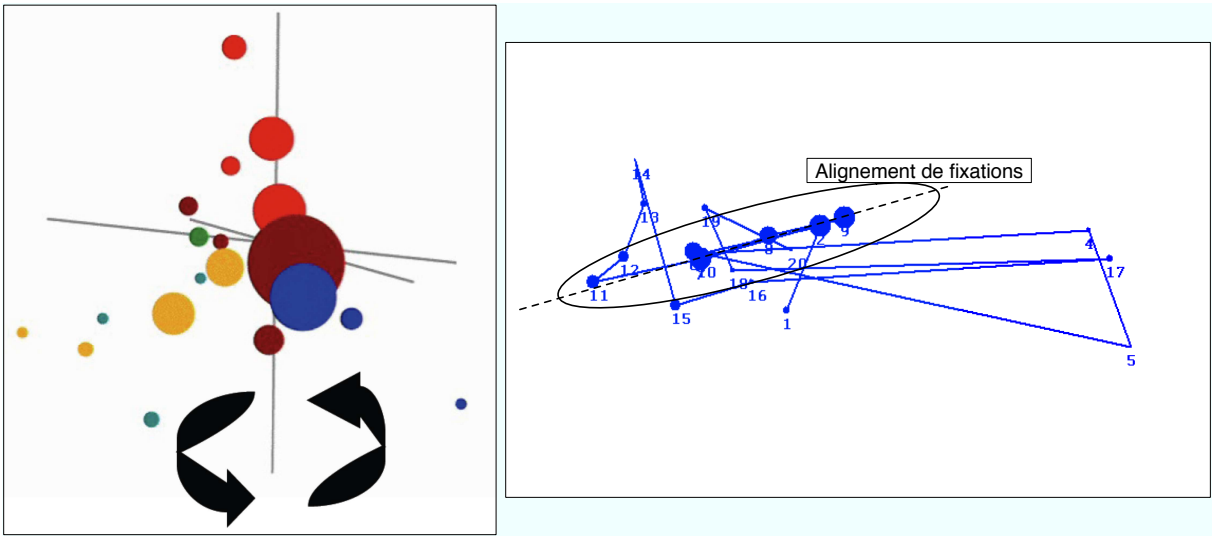


FIG. 4.32 – Mouvement du regard d'un sujet sur l'animation de la carte des groupes de documents en 3 dimensions.

La principale observation que nous avons pu réaliser sur cette carte en 3 dimensions est une poursuite du regard des sujets sur la carte. Cette poursuite, représentée sur la carte précédente par un alignement des fixations, semble révéler une certaine attraction du regard vers les plus grandes sphères représentant d'importants groupes de documents. Les sujets semblent alors suivre tout particulièrement l'une de ces sphères (souvent la plus grande) et ignorent les autres.

Une fois l'ensemble des tracés oculaires des sujets observés, différentes zones d'intérêt ont pu être mises en évidence. Ainsi, la légende s'est révélée particulièrement visitée par certains sujets surtout lors des premières cartes. La partie supérieure de la carte contenant le disque le plus important a également fait l'objet de fixations assez soutenues. La partie inférieure de la carte a par contre été survolée de façon plus irrégulière par les différents sujets, certains parcourant rapidement toute la carte, d'autres restant plus localisés près de la légende. La figure 4.33 illustre l'ensemble des fixations des sujets sur les première et deuxième cartes leur étant présentées (chaque fixation est représentée par un triangle plus ou moins grand selon la durée la fixation).

Nous avons pu observer précédemment des comportements assez différents entre sujets : ceux fréquentant beaucoup la légende, ceux survolant toute la carte, etc. De cette manière, nous

avons isolé cinq zones d'intérêt (appelées encore « aires d'intérêt ») communes aux cartes en 2 dimensions. Ces zones sont présentées en figure 4.34. Ces zones d'intérêt sont marquées en rouge sur la carte. La zone 5 est associée à la légende de la carte. La zone 1 fait référence au coin supérieur gauche, alors que la zone 2 fait référence à la partie supérieure droite de la carte, incluant le groupe le plus important. Les zones 3 et 4 séparent en deux la partie inférieure de la carte.

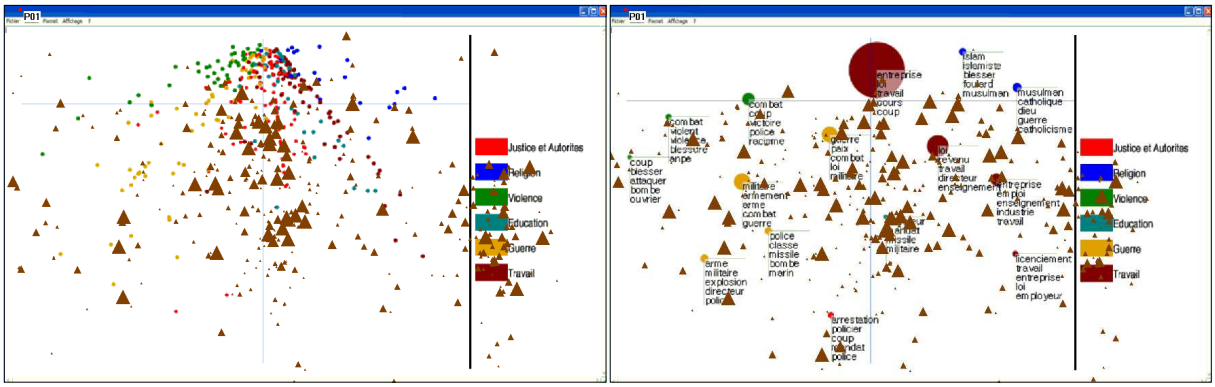


FIG. 4.33 – Fixations de l'ensemble des sujets sur les cartes des documents (à gauche) et des groupes de documents (à droite).

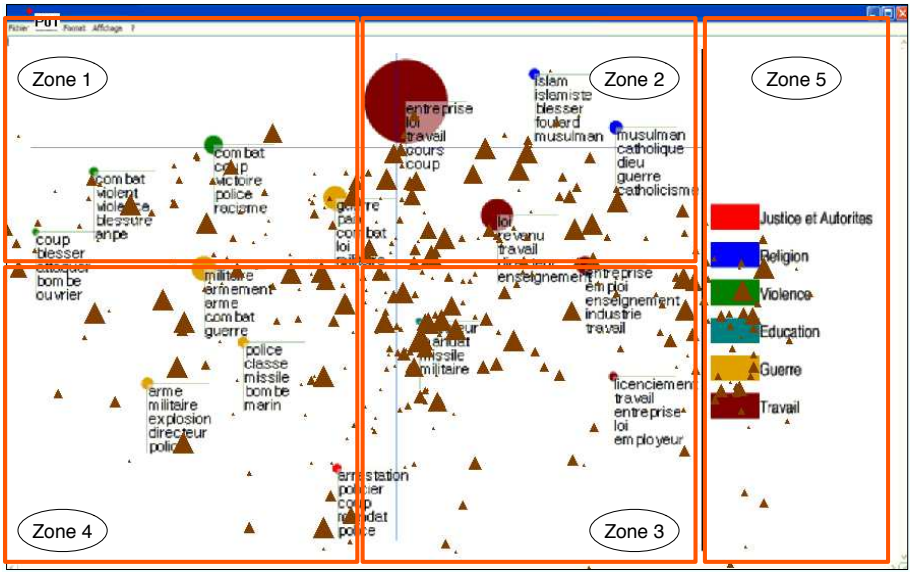


FIG. 4.34 – Les zones d'intérêt isolées sur les cartes en 2 dimensions.

Traditionnellement, le découpage en zones d'intérêt se fait selon un quadrillage régulier ne prenant pas en considération les particularités de l'image visualisée [Outing et Ruel, 2004]. Le découpage que nous proposons prend en considération les particularités des cartes proposées en isolant des zones suscitant un intérêt particulier pour les sujets.

Ces zones ont été définies afin d'intervenir dans la constitution de catégories de sujets[185]. Une catégorisation des sujets peut, selon nous, prendre en considération les trois critères suivants :

1. pour chaque carte, le nombre moyen de fixations et la durée moyenne de ces fixations ;
2. pour chaque zone abordée précédemment, le nombre moyen et la durée moyenne de fixations dans la zone ;
3. les réponses au questionnaire du sujet, dans le sens où elles peuvent révéler une bonne ou une mauvaise compréhension des cartes.

Pour l'instant, nous avons travaillé essentiellement sur le premier critère, qui nous a déjà permis de mettre en évidence des premières catégories de sujets. Pour chaque image, nous avons étudié le nombre moyen de fixations et leur durée moyenne pour chaque sujet, ainsi que leurs positionnements par rapport aux moyennes de tous les sujets sur la même image. Ainsi, nous avons pu mettre en évidence les cinq catégories suivantes :

1. les sujets « mobiles longs » - 3 sujets : ce sont des sujets qui réalisent beaucoup de fixations (20 fixations par image), en moyenne assez longues (0,5 seconde) ;
2. les sujets « 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) ;
3. les sujets « sédentaires » - 5 sujets : cette catégorie caractérise des sujets peu mobiles (assez faible nombre de fixations, environ 15) et dont la durée moyenne de fixation est assez longue (0,6 seconde) ;
4. les sujets « 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) ;
5. les sujets « 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).

Cette première catégorisation nous a permis d'isoler les différentes stratégies des sujets pour parcourir les cartes que nous leur proposions. Nous travaillons actuellement sur une façon de tenir compte dans cette catégorisation des différentes zones d'intérêt que nous avons isolées précédemment. Nous cherchons également à intégrer les réponses des différents sujets aux questionnaires. Sur ce dernier point, de premières analyses laissent paraître que les sujets répondant correctement aux questions se situent très majoritairement dans les catégories 1 à 3 énoncées précédemment, ayant comme point commun des durées moyennes de fixations assez longues. Une telle analyse préliminaire, laissant penser à un lien entre compréhension des cartes et temps de fixation, doit cependant être approfondie dans des travaux futurs.

Cette expérimentation avait pour objectif d'interroger le premier regard porté par des individus sur des visualisations cartographiques d'ensembles documentaires. Malgré le biais initial de l'expérimentation, à savoir l'absence d'interaction sur les cartes (absence « comblée » par une mise en situation du sujet avant et durant le diaporama) et des temps de visualisation réduits (10 secondes par carte), nous avons pu faire ressortir une très grande variété de stratégies de parcours visuels chez les sujets. De grandes différences entre eux ont ainsi été observées, que cela soit au

[185] Pour un travail similaire, nous renvoyons à [Zangemeister *et al.*, 1995] où les auteurs proposent une catégorisation de sujets à partir de leurs parcours oculaires sur des photographies.

niveau des nombres et temps de fixations, ou au niveau des zones parcourues ou des réponses aux formulaires. Ces différences nous ont amené à établir différentes catégories de sujets et surtout à nous interroger sur de nouveaux types de cartes à leur proposer. Ainsi, pour certains, la légende s'est révélée particulièrement importante et de nombreux allers-retours entre la carte et cette dernière ont été réalisés. Les éléments de grande taille ont également paru assez « attractifs » pour les sujets, peut-être au détriment d'éléments plus petits non visités. Les parcours des cartes de certains sujets se sont révélés très partiels, où seules une ou deux zones étaient survolées. Dans tous les cas, il ressort que les cartes des ensembles documentaires ne sont pas parcourues comme du texte, bien qu'elles contiennent, pour la plupart, des étiquettes textuelles. Elles constituent une modalité visuelle à part entière qu'il convient de mieux caractériser. Des réflexions sur la façon de faire évoluer nos vues cartographiques par rapport à ces différents éléments doivent donc être menées afin de s'adapter le plus possible aux particularités de chaque utilisateur dans sa tâche d'accès au contenu d'ensembles documentaires.

Conclusion : valeur ajoutée et flexibilité du modèle *AIdED*

Les différentes expérimentations présentées dans ce chapitre nous ont permis d'avoir plusieurs retours d'utilisateurs dans des tâches variées d'accès au contenu d'ensembles documentaires.

Par exemple, pour des tâches de recherche documentaire, nous avons pu mettre en évidence la valeur ajoutée par nos supports de visualisation pour accéder au contenu aussi bien de pages de sites Internet généralistes que de documents scientifiques médicaux. Il a ainsi été possible de dégager les différents thèmes abordés dans les ensembles et également de décrire assez finement le contenu de différents sous-groupes de textes.

La seconde expérimentation présentée concernait l'étude de métaphores conceptuelles. Par cette expérimentation, nous avons pu mettre en évidence ce que nous avons appelé le degré de métaphoricité des textes d'un corpus. Les vues proposées sur le corpus ont ainsi permis de distinguer des emplois littéraux et des emplois métaphoriques des domaines sources des métaphores conceptuelles étudiées. Par des vues dynamiques représentant la dimension temporelle du corpus, un lien étroit entre la présence de certaines métaphores conceptuelles et l'actualité a également été mis en évidence.

Un travail portant sur l'analyse de forums de discussion a ensuite été détaillé. Ce travail a tout d'abord permis d'isoler les thématiques abordées dans les forums de discussion. Un accès plus local aux groupes de messages et aux messages a permis d'observer assez finement la façon dont les participants des forums échangeaient.

Toutes ces expérimentations ont permis d'illustrer nos propositions en abordant les différentes « étapes » instaurées par *AIdED*, à savoir la construction de RTO personnelles, leur projection en ensembles documentaires et l'interaction avec les supports de visualisation cartographique. C'est cette interaction avec les supports qui a permis à nos différents utilisateurs d'atteindre le contenu des ensembles documentaires selon leurs propres domaines d'intérêt. En plus de cet accès au contenu de l'ensemble documentaire, un retour pertinent sur les RTO utilisées a souvent été obtenu à travers les cartes. Un tel retour sur les ressources permet leur évolution et leur maintenance au fil du temps, comme nous l'avons résumé dans [Roy et Beust, 2007].

Une expérimentation de nature différente a également été menée au cours de cette thèse. Son objectif était d'interroger la lisibilité des cartes que nous proposons, en observant la façon dont des utilisateurs parcouraient visuellement de telles cartes. Une très grande variété de parcours a alors été révélée avec des sujets appréhendant de manières très différentes les cartes que nous leur proposions. Ce retour évaluatif, différent des précédents, nous a également donné des éléments

d'évaluation de nos propositions et nous a surtout amené à nous interroger sur la nature même des cartes et sur la disposition des différents éléments sur ces dernières.

Un schéma générique d'évaluation en trois étapes *constitution des ressources / utilisation des logiciels / analyse des sorties produites* a été proposé et appliqué aux différentes expérimentations d'accès au contenu. Ce schéma a permis de réaliser, en plus une évaluation intrinsèque des logiciels, une évaluation extrinsèque complète, laissant les utilisateurs totalement libres d'exprimer leur point de vue. Ce schéma d'évaluation nous semble particulièrement adapté pour déterminer la pertinence de systèmes informatiques laissant une place importante aux utilisateurs.

Une importante valeur ajoutée a ainsi pu être évaluée dans chacune de ces expérimentations, pourtant très variées. Le modèle *AIdED* a ainsi pu être appliqué aux différentes tâches visées, parfois avec quelques adaptations logicielles. Mais dans tous les cas, notre modèle s'est montré particulièrement flexible et en adéquation avec chaque tâche visée. Nous avons illustré une telle flexibilité d'*AIdED* dans deux publications [Roy et Ferrari, 2006, Roy et Ferrari, 2007], reprenant des éléments énoncés dans ce chapitre.

Pour chacune de ces expérimentations, nous avons soulevé au cours de ce chapitre un certain nombre de perspectives de recherche. Une perspective commune à l'ensemble de ces expérimentations est selon nous de conduire une utilisation à plus grande échelle et à plus long terme de nos logiciels. Une telle utilisation, prenant toujours place dans des contextes expérimentaux variés, comme nous l'avons fait ici, nous permettrait d'interroger la robustesse de notre modèle par rapport à des tâches quotidiennes d'accès au contenu d'ensembles documentaires. Nous revenons sur une telle perspective en conclusion de cette thèse.

Conclusion

Le chemin parcouru durant le travail de thèse présenté dans ce manuscrit nous a permis de collaborer avec un grand nombre de chercheurs issus de différentes disciplines, l'informatique principalement, mais également la linguistique, et dans une moindre mesure les sciences de l'éducation et la psychologie. Cette pluri-disciplinarité a été au cœur de cette thèse, avec une problématique faisant cohabiter l'informatique et la linguistique pour apporter un éclairage nouveau à une tâche d'accès personnalisé au contenu d'ensembles documentaires. Au terme de cette thèse, nous faisons le bilan de ses différents apports et des perspectives de recherche ouvertes.

Apports de ce travail de thèse

Dans un premier temps, nous avons dressé un panorama des systèmes informatiques existants pour l'accès au contenu d'ensembles documentaires. Des systèmes interactifs ont été particulièrement mis en évidence, de même que les systèmes exploitant des techniques de visualisation et de cartographie. Une telle revue de l'existant nous a permis d'isoler les points qui, selon nous, empêchaient un réel accès au contenu d'ensembles documentaires : le manque de prise en considération du point de vue de l'utilisateur et le manque de visualisation globale et interactive sur son ensemble documentaire.

Pour combler de tels manques dans l'existant, nous avons proposé le modèle *AIdED*. Ce modèle, extension du modèle *LUCIA* de Vincent Perlerin, lui-même dans la filiation du modèle *Anadia* de Pierre Beust, poursuit une certaine problématique de recherche en cours depuis plusieurs années à l'Université de Caen / Basse-Normandie. Empruntant de nombreux éléments théoriques à la Sémantique Interprétative de François Rastier, *AIdED*, tout comme ses prédécesseurs, a pour rôle d'apporter une aide aux utilisateurs dans l'accès au contenu de textes. Contrairement à ses prédécesseurs dédiés plus particulièrement à l'accès au niveau du texte, *AIdED* se positionne à un niveau textuel supérieur, celui de l'ensemble documentaire.

Les principales propositions que nous faisons avec *AIdED* consistent, tout d'abord, en la prise en compte de la globalité de l'ensemble documentaire et de son influence sur ses éléments plus locaux. La visualisation cartographique des principales récurrences d'éléments de signification, choisis et décrits par l'utilisateur, à l'intérieur des ensembles documentaires est ensuite proposée. De telles propositions sont réalisées en projetant des représentations sémiques différentielles des domaines d'intérêt de l'utilisateur sur des cartes interactives et multi-échelles de l'ensemble documentaire. Ces cartes mettent en évidence entre autres des isotopies intertextuelles parcourant l'ensemble documentaire et délimitant des sous-ensembles de textes partageant de mêmes caractéristiques. Ces visualisations orientent l'utilisateur dans sa tâche d'accès au contenu de l'ensemble documentaire en médiatisant ses parcours interprétatifs à l'intérieur de l'ensemble.

Nos différentes propositions, réunies dans le modèle *AIdED*, ont ensuite été mises en outils et en instruments logiciels. Nous avons repris des outils existants et en avons développé de

nouveaux, afin d'assister le plus possible l'utilisateur dans sa phase de construction de RTO *LUCIA*. Le développement de la plate-forme *ProxiDocs* a ensuite été réalisé afin de projeter les RTO représentant les domaines d'intérêt de l'utilisateur dans son ensemble documentaire, et ainsi, de construire les supports de visualisation interactive de l'ensemble. Ces différents éléments logiciels sont tous interactifs et laissent, le plus possible, l'utilisateur manipuler le matériau textuel, que cela soit dans la phase de construction des domaines ou dans celle de projection de ces domaines en ensemble documentaire. Comme nous avons pu le voir au cours de cette thèse, de nombreux échanges sont naturellement présents entre ces deux phases : l'élaboration de domaines se fait selon le ou les ensembles documentaires à analyser, et la projection des domaines dans un ensemble documentaire entraîne, en plus d'un accès au contenu de ce dernier, un retour important sur les domaines, pouvant impliquer leur évolution.

L'évaluation du modèle *AIdED* et de son instrumentation logicielle a ensuite été proposée à travers différentes expérimentations, prenant chacune place dans le cadre de collaborations avec des objectifs bien précis en accès au contenu d'ensembles documentaires. Les premières expérimentations présentées prennaient place dans le cadre de tâches de la recherche documentaire. Pour de telles tâches, nous avons pu mettre en évidence la façon dont les visualisations interactives proposées sur les ensembles documentaires permettaient d'isoler assez finement les différents sujets abordés dans les ensembles. Un tel accès, d'ordre thématique, a été observé aussi bien dans le cadre d'une recherche généraliste sur Internet, que dans le contexte particulier d'une recherche de documents médicaux.

La deuxième expérimentation présentée était dédiée à l'étude de certaines métaphores conceptuelles dans un corpus d'articles boursiers. Une telle étude a alors permis de mettre en évidence ce que nous avons appelé le degré de métaphoricité des textes du corpus, sorte de baromètre des types de métaphores observées, partant des emplois non métaphoriques pour aller jusqu'aux métaphores variées, en passant par des emplois littéraux de métaphores. Ce degré reste, cependant, à mieux caractériser dans des travaux futurs. Dans le cadre de cette expérimentation, des liens étroits entre la présence de certaines métaphores conceptuelles et des faits d'actualité ont également pu être observés.

La troisième expérimentation concernait l'analyse de forums de discussions pédagogiques. Là encore, un accès thématique a pu être fourni aux utilisateurs par l'interaction sur les cartes, aussi bien sur la globalité du forum, que dans son évolution dans le temps. Les visualisations cartographiques ont également permis d'observer et d'illustrer les usages de certaines terminologies dédiées à la formation.

Ces différentes expérimentations ont pris place dans le cadre de collaborations entre chercheurs, et elles ont visé des tâches bien précises, fixées durant les collaborations. De tels échanges avec des spécialistes de différentes disciplines ont été particulièrement enrichissants, tout d'abord, par le travail collaboratif et la communication impliqués, mais surtout, par les retours évaluatifs très pertinents et objectifs sur nos propositions. Ces expérimentations, interrogeant plus particulièrement la pertinence de nos supports de visualisation pour l'accès au contenu d'ensembles documentaires, ont été suivies par une expérimentation, proche d'une problématique d'ergonomie, portant sur l'aspect visuel et la disposition des informations sur nos cartes. Des analyses des premiers résultats expérimentaux ont permis de mettre en évidence une catégorisation de sujets selon leurs parcours des cartes. Des zones particulièrement visitées par les sujets sur les cartes sont également ressorties, nous invitant même à repenser la disposition de certains éléments sur les cartes que nous proposons.

La mise en place de ces dispositifs expérimentaux, liés à des objectifs très différents, a permis de mettre en évidence la valeur ajoutée et la flexibilité de notre modèle pour chacune des tâches visées. L'évaluation de nos propositions n'aurait pas été possible sans la réalisation de telles

expérimentations. Ces expérimentations ont pris place dans des contextes bien délimités avec des utilisateurs visant des tâches précises d'accès au contenu d'ensembles documentaires donnés selon certains domaines. En proposant un schéma d'évaluation en trois étapes : *constitution des RTO personnelles / exécution des logiciels / analyse des sorties produites*, nous avons essayé de centrer le plus possible cette phase d'évaluation sur les utilisateurs et, ainsi, de prendre réellement en considération leurs points de vue. Une remise en question des évaluations traditionnelles en informatique, et plus particulièrement en TAL, peut alors être impliquée, afin d'intégrer au mieux, comme nous avons essayé de le faire, la part de l'évaluation extrinsèque, au sens de [Spark Jones et Galliers, 1995], dans les évaluations de systèmes de TAL.

Perspectives de recherche soulevées

L'ensemble de nos propositions nous invite à penser à un grand nombre de perspectives de recherche qui devront être, selon nous, abordées en continuité de cette thèse.

La première d'entre elle est liée à l'approfondissement de la notion de parcours interprétatif réalisé au niveau de l'ensemble documentaire. De tels parcours interprétatifs intertextuels restent, selon [Rastier, 2001a, pages 110-111], entièrement à élaborer. Dans ce travail de thèse, nous avons cependant proposé une première médiatisation de tels parcours *via* les cartes d'ensembles documentaires. Cette médiatisation a permis d'assister la création de parcours interprétatifs de l'utilisateur au sein de son ensemble documentaire, l'aidant, ainsi, à accéder au mieux au contenu de l'ensemble. Malgré cela, de tels parcours restent influencés par ce que François Rastier appelle des *contrats interprétatifs*, éléments généralement implicites liés aux discours et aux genres. De tels contrats font, par exemple, que nous ne lisons pas une liste de références bibliographiques comme un roman. La caractérisation et la prise en considération des contrats interprétatifs est, selon nous, une voie de recherche très pertinente à explorer, afin de représenter au mieux la dimension intertextuelle de l'ensemble documentaire et de son interprétation.

La notion d'anagnose proposée par Théodore Thlivitis [Thlivitis, 1998] oriente également les parcours interprétatifs intertextuels. Comme nous l'avons énoncé au chapitre 2 de cette thèse, l'anagnose regroupe, en quelque sorte, tout le « passé » textuel d'un individu, et donc un grand nombre d'intertextes. Dans nos travaux, nous avons représenté l'intertexte par l'ensemble documentaire analysé. Pour aller vers une véritable prise en considération de l'anagnose et en tenir compte au sein des parcours interprétatifs, il serait possible de passer à un palier textuel encore supérieur à celui de l'ensemble documentaire : le palier regroupant différents ensembles documentaires faisant partie du passé textuel de l'utilisateur. En considérant des isotopies partagées entre ces différents ensembles documentaires[186], il pourrait alors être possible de mieux guider les parcours interprétatifs de l'utilisateur et de lui permettre un accès plus efficace au contenu.

Dans cette perspective de recherche, l'analyse de très grands ensembles documentaires doit également être envisagée, un nombre de plus en plus grand de documents électroniques étant accessibles aux utilisateurs. Dans les expérimentations présentées dans cette thèse, les ensembles documentaires contenaient d'une petite centaine à un demi-millier d'éléments. Un véritable « passage à l'échelle » de nos traitements doit être envisagé afin de permettre l'analyse d'ensembles documentaires de plusieurs milliers de textes. Ce passage à l'échelle mettra à l'épreuve la robustesse et la pertinence des méthodes statistiques utilisées jusqu'à présent. L'intégration de nouvelles méthodes de projection et de catégorisation pour traiter de grands ensembles documentaires pourra alors être envisagée, notamment des méthodes issues de la fouille de données, des probabilités ou de l'apprentissage, telles celles abordées dans [Crochemore *et al.*, 2005]. Au-

[186] Théodore Thlivitis nomme de telles isotopies des *inter-interisotopies* [Thlivitis, 1998].

delà d'une mise à l'épreuve des méthodes numériques utilisées, la conception de nouveaux modes de visualisation et d'interaction avec l'ensemble documentaire devra certainement être envisagée afin de tirer partie au mieux de cette nouvelle échelle de travail.

Hormis ce passage à l'échelle en taille des ensembles documentaires, la multiplication des utilisateurs et des contextes d'utilisations des cartes que nous proposons devra être envisagée dans des tâches répétées, voire quotidiennes, comme des tâches de recherche documentaire sur Internet, d'accès au contenu de boîtes aux lettres électroniques, de suivi de forums de discussion, etc. De multiples utilisations des visualisations cartographiques pourraient permettre de pousser encore plus loin notre évaluation, de faire ressortir des profils d'utilisateurs, différents degrés d'utilisations de la cartographie, et de ainsi montrer également la valeur ajoutée de nos propositions sur des périodes plus longues et dans des tâches quotidiennes.

Les différentes analyses réalisées sur les ensembles documentaires ont pu nous donner un retour pertinent sur les RTO utilisées. De telles RTO, représentant les domaines d'intérêt de l'utilisateur, sont construites par ce dernier avec des aides logicielles que nous avons présentées dans le chapitre 3 de cette thèse. Cette tâche peut malheureusement se révéler assez longue pour certains utilisateurs, surtout lors d'une première construction de domaines. Pour amorcer une telle construction, il serait intéressant d'orienter des recherches vers la construction automatique de « pré-ensembles » de graphies employées dans des contextes proches et liées à de mêmes domaines, à la manière des travaux détaillés dans [Rossignol et Sébillot, 2003]. L'utilisateur serait alors libre de reprendre des éléments de tels ensembles pour construire et structurer ses propres domaines.

Dans cette thèse, nous avons légèrement étendu le modèle *LUCIA* en permettant de placer en valeurs d'attributs, des formes non textuelles, comme des *smileys* (cf. chapitre 3). De telles formes sont de plus en plus fréquentes dans les documents électroniques, comme le montre [Véronis et Guimier de Neef, 2006]. Une importante réflexion, portant sur la prise en considération de telles formes dans les RTO *LUCIA*, se doit alors d'être poursuivie, afin de permettre de les positionner, non seulement en valeur d'attribut comme nous le faisons déjà, mais aussi en attribut et, même, en entrée de table, au même titre que les lexies le sont actuellement. Une aide à l'interprétation dans un ensemble documentaire de telles formes, et des informations qu'elles portent, pourrait ainsi être fournie à l'utilisateur afin de compléter l'accès au contenu de son ensemble documentaire.

Depuis 2006, une participation à un projet de recherche du Pôle Universitaire Normand a été initiée. Ce projet porte sur la théorie de l'énaction et son application à l'interprétation de documents électroniques. En se fondant, entre autres, sur le constat réalisé lors du congrès *Un demi siècle d'intelligence artificielle*[187] établissant qu'il n'existe pas de système intelligent sans homme dans le système, les compétences cognitives des utilisateurs de tels systèmes se doivent, alors, d'être mises à contribution dans des environnements informatiques susceptibles de favoriser leur expression. C'est ce que nous proposons d'analyser dans ce projet *via* le paradigme scientifique en émergence de l'énaction.

L'énaction est une théorie issue des sciences cognitives qui a été mise au point par Francisco Varela [Varela, 1996]. L'un de ses principes fondamentaux est qu'un organisme donne forme à son environnement en même temps qu'il est façonné par ce dernier. Le travail pluridisciplinaire initié dans cette thèse s'inscrit alors dans une démarche énactive pour l'aide à l'interprétation de documents numériques. Il convient ainsi d'analyser et d'exploiter, le plus finement possible, les différents aspects sémiotiques des documents numériques tels que leur dimension intertextuelle,

[187] Congrés réalisé dans le cadre de la plate-forme AFIA en novembre 2006 : `http://afia.lri.fr/node.php?node=1175` (page consultée le 16 juillet 2007).

leur ancrage temporel ou encore leur nature multimodale. Nous envisageons de poursuivre un tel travail de recherche à l'issue de cette thèse en mettant à profit la plate-forme *ProxiDocs* pour produire de véritables interfaces de navigation énactive dans des ensembles documentaires, en développant, par exemple, de nouvelles visualisations et de nouvelles interactions permettant à l'utilisateur d'être le plus possible « plongé » dans l'environnement formé par son ensemble documentaire.

Ces différentes perspectives de recherche entraînent aussi bien des réflexions et des études plus poussées en linguistique que des modélisations et des développements informatiques. L'ensemble de ces perspectives, ainsi que les différentes propositions que nous avons pu faire et évaluer au cours de cette thèse, partagent toutes le même objectif : celui de mieux caractériser la dimension intertextuelle et subjective que nous attribuons à la sémantique des langues. Une telle dimension nous semble être une voie de recherche en informatique et en linguistique particulièrement ouverte et pertinente afin de proposer aux utilisateurs les meilleurs accès possibles à leurs ensembles documentaires.

Au-delà de cet accès documentaire, ce sont les différentes évolutions d'un document électronique textuel qu'il faut considérer : de sa création à son partage, en passant par les différentes lectures, modifications et ajouts de chacun, toutes ces étapes étant facilement réalisables avec le développement des nouveaux moyens de communication et de présentation de l'information. C'est une telle culture numérique qu'il convient de prendre en considération afin de concevoir des médias informatiques assistant toujours plus la communication et l'échange entre les hommes *via* les machines.

Annexe A

Extraits de fichiers XML décrivant les RTO utilisées

A.1 Extrait d'un fichier de RTO *LUCIA* simple

```
<ThemeList>
  <Theme>
    <Theme_Name>Education</Theme_Name>
    <Lexie>éducation</Lexie>
    <Lexie>école</Lexie>
    <Lexie>scolaire</Lexie>
    <Lexie>scolariser</Lexie>
    <!-- etc. -->
  </Theme>
  <Theme>
    <Theme_Name>Economie et commerce</Theme_Name>
    <Lexie>bourse</Lexie>
    <Lexie>commerce intérieur</Lexie>
    <Lexie>commerce extérieur</Lexie>
    <Lexie>exporter</Lexie>
    <!-- etc. -->
  </Theme>
  <!-- etc. -->
</ThemeList>
```

FIG. A.1 – Extrait d'un fichier XML contenant des ensembles de lexies construits avec l'outil *ThemeEditor*.

A.2 Extraits de fichiers décrivant un dispositif *LUCIA*

```
<!-- le dictionnaire d'attributs -->
<dictionnaire_attributs>
  <attribut id="attr1">
    <attribut_nom>Evaluation</attribut_nom>
    <valeur id="attr1val0">bien</valeur>
    <valeur id="attr1val1">mal</valeur>
  </attribut>
  <!-- etc. -->
</dictionnaire_attributs>

<!-- le dictionnaire de lexies -->
<dictionnaire_lexie>
  <lexie id="sante_1">
    <lemme>douleur</lemme>
    <flexion>douleurs</flexion>
  </lexie>
  <lexie id="sante_2">
    <lemme>guérison</lemme>
    <flexion>guérison</flexion>
  </lexie>
  <!-- etc. -->
</dictionnaire_lexie>

<!-- le dispositif -->
<dispositif id="disp_La_santé">
  <dispositif_nom>La santé</dispnom>
  <table id="dispositif_santé_table_1" attributs="attr1" color="red" ligne_héritée="">
    <table_nom>Phénomènes</table_nom>
    <ligne id="dispositif_santé_table_1_ligne_1" vals="attr1val0">
      <lexie lem="douleur" ref="sante_1"/>
    </ligne>
    <ligne id="dispositif_santé_table_1_ligne_2" vals="attr1val1">
      <lexie lem="guérison" ref="sante_2"/>
    </ligne>
  </table>
  <!-- etc. -->
</dispositif>
```

FIG. A.2 – Extraits des fichiers XML décrivant un dispositif construit avec *VisualLuciaBuilder*

Annexe B

Retour sur les méthodes de projection et de classification utilisées

Sommaire

B.1 Attribution d'un espace numérique à l'ensemble documentaire . . . 188
B.2 Méthodes de projection de l'espace obtenu à l'issue du comptage . 188
B.2.1 Méthodes de projection simples développées durant cette thèse 188
B.2.2 La méthode de l'analyse en composantes principales 190
B.2.3 La méthode de projection de Sammon 194
B.2.4 L'analyse factorielle des correspondances 195
B.3 Méthodes de classification de l'espace visualisé 197
B.3.1 La classification hiérarchique ascendante 197
B.3.2 La méthode des K-Means 198
B.3.3 Le choix du nombre de groupes d'une classification 199
B.4 Comparaisons des méthodes de projection et de classification 201
B.4.1 Cartes des textes 201
B.4.2 Cartes des groupes de textes 204

Dans cette annexe, nous proposons de présenter plus en détail les méthodes numériques exploitées dans la plate-forme *ProxiDocs*. Nous présentons les différentes algorithmes associés aux méthodes qui ont ainsi été implémentées, afin de permettre aussi bien leur compréhension que leur reprise par des travaux futurs.

B.1 Attribution d'un espace numérique à l'ensemble documentaire

Les méthodes de comptage, présentées au chapitre 3 de cette thèse, ont nécessité de développer des algorithmes permettant de les implémenter. L'algorithme, présenté ci-dessous, correspond au comptage relatif des lexies des domaines de l'utilisateur dans les textes d'un ensemble documentaire.

Entrée : Un ensemble documentaire E et un ensemble de domaines D
Sortie : L'espace relatif des textes construit par rapport aux domaines

```
2.1  Domaines_expressions_regulieres = ∅;
2.2  Espace_ensemble_documentaire = ∅;
2.3  pour chaque domaine de D faire
2.4  |  Domaines_expressions_regulieres [domaine] ← ∅;
2.5  |  pour chaque lexie de domaine faire
2.6  |  |  Domaines_expressions_regulieres [domaine] += lexie_courante + '|';
2.7  |  fin
2.8  fin
2.9  pour chaque texte de S faire
2.10 |  chaine_texte ← contenu du texte;
2.11 |  Espace_ensemble_documentaire [texte] = ∅;
2.12 |  pour chaque domaine de Domaines_expressions_regulieres faire
2.13 |  |  nombre_de_matches = application de l'expression régulière
     |  |  Domaines_expressions_regulieres [domaine] à chaine_texte;
2.14 |  |  Espace_ensemble_documentaire [texte][domaine] = nombre_de_matches;
2.15 |  fin
2.16 fin
```

Algorithme 2 : Algorithme de la méthode de comptage relative des domaines en ensemble documentaire.

B.2 Méthodes de projection de l'espace obtenu à l'issue du comptage

B.2.1 Méthodes de projection simples développées durant cette thèse

Principes généraux

Pour visualiser l'espace numérique obtenu à l'issue de l'étape de comptage, de nombreuses méthodes existent. Elles permettent de réduire un espace à grande dimension en un espace à deux ou trois dimensions qui soit visualisable. Nous verrons dans la suite de cette annexe de telles méthodes, faisant référence en statistiques et analyses des données, et que nous avons implémentées dans la plate-forme *ProxiDocs*, telle la méthode de Sammon, ou encore, la méthode de l'Analyse en Composantes Principales. Avant cela, nous présentons des méthodes que nous avons développées. Ces méthodes sont plus simples, donnent des résultats moins précis que les méthodes de référence, mais elles se révèlent plus rapides à l'exécution. Nous les avons développées

afin de laisser aux utilisateurs le plus large choix possible de méthodes de projection, selon la finesse de la tâche d'accès au contenu visée.

Ces différentes méthodes de projection que nous avons développées se basent sur le même principe général que la méthode de l'Analyse en Composantes Principales, qui détermine les deux axes (ou composantes) les plus « caractéristiques » de l'espace de départ, puis réalise une projection des points de l'espace de départ sur ces deux axes. Par exemple, si nous cherchons à représenter le plus fidèlement possible un objet en trois dimensions sur une feuille, autrement dit, un espace en deux dimensions, il semble judicieux de le dessiner suivant une vue montrant le plus de surface possible de l'objet. Si nous voulons dessiner un poisson sur cette feuille, il faudra éviter d'en faire une représentation avec une vue de dessus, cette représentation ne rendrait pas compte de la forme du poisson. La représentation « idéale » serait celle le montrant avec une vue de côté, révélant ainsi sa véritable forme. Les deux axes principaux de cette représentation sont alors ceux dans la figure B.1.

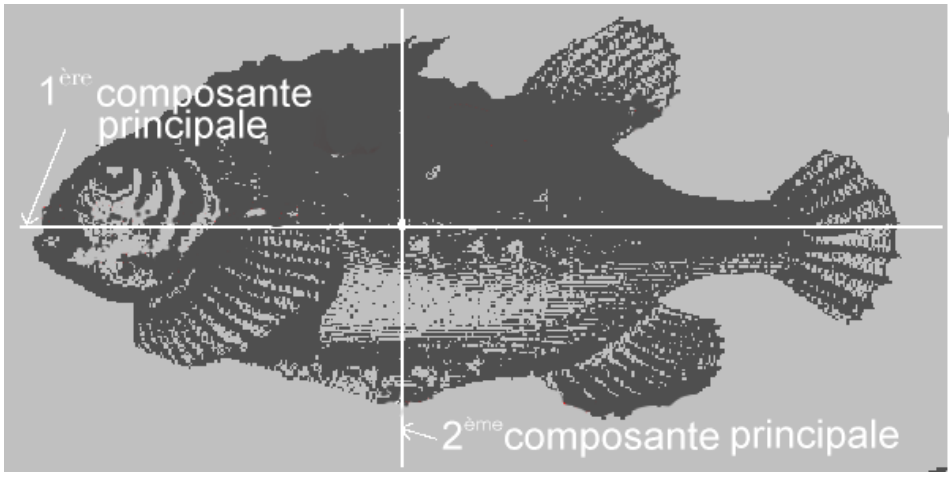


FIG. B.1 – Les deux composantes optimisant une représentation d'un poisson en 2 dimensions.

La méthode des deux plus grandes distances

Dans l'idée de sélectionner les deux composantes les plus « importantes » dans l'espace de départ pour servir d'axes de projection dans l'espace d'arrivée, nous avons tout d'abord choisi les deux axes de plus grandes longueurs dans l'espace d'origine. Pour cela, nous avons considéré les différents couples de textes dans l'espace de départ, puis nous avons calculé la distance euclidienne entre chaque couple, à l'aide de la formule suivante :

$$\mathrm{dist}(texte_1, texte_2) = \frac{\sum_{i=1}^{n} (\mathrm{coord}_i(texte_1) - \mathrm{coord}_i(texte_2))^n}{n}$$

Dans la formule ci-dessus, $\mathrm{coord_i}(texte_1)$ désigne la i^{ieme} coordonnée de $texte_1$. L'idée de cette méthode, que nous avons appelée « méthode des plus grandes distances », revient à prendre, comme axes principaux, les deux couples de points les plus éloignés les uns des autres (dont la distance euclidienne est la plus grande). Une projection sur ces deux axes de l'espace de départ permet alors de le visualiser dans un plan (projection que nous verrons dans une section suivante de cette annexe).

La méthode du produit scalaire

La méthode précédente propose de sélectionner, comme axes de projection, les axes formés par les couples de textes les plus éloignés dans l'espace de départ. Dans cette méthode, intitulée « méthode du produit scalaire », l'idée est de conserver comme premier axe principal, l'axe formé par les deux textes les plus éloignés dans l'espace de départ. Par contre, le second axe principal sera celui lui étant le plus orthogonal dans l'espace de départ.

Dans cette méthode, chaque couple de textes dans l'espace de départ est représenté par un vecteur V décrit par la formule suivante :

$$\text{coord}_i(V) = \text{coord}_i(texte_1) - \text{coord}_i(texte_2), \text{pour } 1 \leq i \leq n$$

Nous considérons ensuite la formule du produit scalaire suivante pour deux vecteurs V_1 et V_2 désignant deux couples de textes :

$$\text{ProduitScalaire}(V_1, V_2) = \sum_{i=1}^{n} (\text{coord}_i(V_1) * \text{coord}_i(V_2))$$

Le vecteur considéré comme le plus orthogonal avec le premier axe principal, sera celui ayant le produit scalaire le plus proche de 0 avec ce dernier. Un tel vecteur sera alors retenu comme second axe de projection.

La projection sur les axes principaux

Une fois les deux axes principaux obtenus, il faut y projeter les textes de l'espace de départ. Pour effectuer une telle opération, nous réalisons les étapes ci-dessous, pour chacun des deux axes principaux sélectionnés :

1. Soit (C_1C_2), la droite caractérisant l'un des deux axes principaux et, O, le centre du segment $[C_1C_2]$. On calcule ensuite les coordonnées du vecteur $\overrightarrow{OC_2}$.
2. Soit T, un texte à projeter sur notre axe principal. Il faut alors calculer les coordonnées du vecteur $\overrightarrow{OT}$.
3. Une fois ce vecteur déterminé, on peut calculer le produit scalaire P entre $\overrightarrow{OC_2}$ et $\overrightarrow{OT}$.
4. Soit α, l'angle séparant les vecteurs $\overrightarrow{OC_2}$ et $\overrightarrow{OT}$, on obtient alors l'égalité suivante : $P = |\overrightarrow{OC_2}| * |\overrightarrow{OT}| * cos(\alpha)$, l'angle α peut donc être calculé.
5. Soit x, le projeté de T sur la droite (C_1 C_2). Il est alors possible de calculer la distance entre le point O et le projeté x, par l'égalité suivante : $|\overrightarrow{Ox}| = |\overrightarrow{OT}| * \cos(\alpha)$.

Ces opérations sont alors à effectuer sur chaque texte de l'espace de départ et pour les deux axes principaux. Le premier axe est considéré comme l'axe des abscisses, le second, comme l'axe des ordonnées. On obtient ainsi pour chaque texte, deux coordonnées, chacune obtenue par projection sur l'une des deux composantes principales. Une telle projection n'est réalisée que pour les deux méthodes simples présentées dans les deux sections précédentes.

B.2.2 La méthode de l'analyse en composantes principales

Les méthodes de projection, présentées précédemment, ont été réalisées afin de proposer à l'utilisateur des méthodes simples, assez peu coûteuses en mémoire et en temps de calcul. Des méthodes, beaucoup plus attestées, ont également été implémentées dans la plate-forme *Proxi-Docs*. Parmi ces méthodes de référence, nous pouvons citer l'Analyse en Composante Principale

(ACP) [Bouroche et Saporta, 1980], qui est une méthode de calcul très utilisée dans le domaine de l'analyse des données. Son utilisation est très répandue dans des domaines très variés tels la génétique [Rouvier, 1966], la météorologie [Pottier, 1991] et, plus proche de nous, le TAL avec, par exemple, [Illouz et Jardino, 2001] ou encore [Buisine et Martin, 2006].

Dans les paragraphes suivants, nous allons tout d'abord donner une idée de l'intérêt de cette méthode, mais aussi ses limites et ses inconvénients. Il est ensuite abordé la réalisation informatique de cette méthode.

Une approche statistique

L'ACP est une méthode de structuration et de synthèse de données numériques. Elle permet de présenter un résumé descriptif, accompagné de représentation graphique, d'un ensemble d'observations mesurées sur un ensemble de variables numériques. On utilise cette méthode lorsqu'il s'agit de décrire et de « visualiser » au mieux l'information contenue dans un tableau de données quantitatives où n individus ont été évalués en fonction de p variables.

Le graphique des individus permet de visualiser les proximités entre individus qui s'interprètent en terme de similitude de comportement vis-à-vis des variables. L'étude de la forme du nuage des individus permet de distinguer d'éventuels regroupements et de différencier des individus ou des groupes d'individus selon leurs réponses à l'ensemble des variables.

La méthode ACP cherche à déterminer un nombre restreint de variables, représentant les mêmes données et non corrélées entre elles. C'est-à-dire synthétiser les variables, ou encore, tenter de résumer l'information contenue dans un tableau de données, en un ensemble réduit de combinaisons linéaires des variables initiales, en veillant à minimiser la perte d'information du fait de cette réduction. Ces nouvelles variables synthétiques, appelées « composantes principales », possèdent les propriétés suivantes :

- elles sont non corrélées (les coefficients de corrélation linéaire des composantes prises deux à deux sont nuls) ce qui évite la redondance de l'information déjà résumée ;
- les composantes principales, notées $(C^1, C^2, .., C^q)$, sont des combinaisons linéaires des variables initiales (pour nous, les domaines) $(X^1, X^2, ..., X^p)$: $C^j = a_1X^1 + ... + a_pX^p$ pour tout $j = 1$ à q, avec $q \leq p$ (où q est le nombre de domaines et p le nombre de textes) ;
- la première composante porte plus d'informations que la seconde, qui porte plus d'informations que la troisième, et ainsi de suite, de sorte qu'en se limitant aux deux premières composantes on dispose, en général, d'un bon résumé de l'information contenue dans les données.

Il est ainsi cherché des axes orthogonaux (les plus indépendants possibles) qui ont la propriété d'extraire conjointement le maximum d'informations sur les individus.

Une approche informatique

Afin de mettre en œuvre une ACP dans un programme informatique, nous nous sommes basés sur la description de la méthode donnée dans [Bouroche et Saporta, 1980]. Cette description fait intervenir différentes notions mathématiques, et particulièrement, des notions issues de l'algèbre linéaire et du calcul matriciel. Nous ne détaillons pas de telles notions dans cette thèse, nous renvoyons à [Bouroche et Saporta, 1980] pour plus de détails.

Nous proposons le schéma d'algorithme ci-contre, reprenant ces différents calculs et proposant une mise œuvre logicielle de la méthode. C'est un tel schéma d'algorithme qui a été implémenté dans la plate-forme *ProxiDocs*.

Entrées : Une matrice numérique $M_{initiale}$ issue de l'application de l'une des méthodes de comptage définies précédemment et le nombre k de dimensions souhaitée pour l'espace d'arrivée

Sortie : Une matrice numérique $M_{resultat}$ contenant les positions de chaque texte dans l'espace d'arrivée

3.1 *Calcul de la matrice X centrée réduite de $M_{initiale}$*

3.2 **pour chaque** *colonne i de $M_{initiale}$* **faire**

3.3 Calculer $Moyenne_i = \frac{\sum M_{initiale}\_[i][j]}{Nombre\ de\ lignes\ de\ M_{initiale}}$, avec j variant de 1 jusqu'au nombre de lignes de $M_{initiale}$

3.4 **fin**

3.5 **pour chaque** *colonne i de $M_{initiale}$* **faire**

3.6 Calculer $Ecart\_Type_i = \sqrt{\frac{\sum M_{initiale}\_[i][j] - Moyenne_i{}^2}{Nombre\ de\ lignes\ de\ M_{initiale}}}$, avec j variant de 1 jusqu'au nombre de lignes $M_{initiale}$

3.7 **fin**

3.8 **pour** *i variant de 1 jusqu'au nombre de colonnes de $M_{initiale}$* **faire**

3.9 **pour** *j variant de 1 jusqu'au nombre de lignes de $M_{initiale}$* **faire**

3.10 Calculer $X_{ij} = \frac{M_{initiale}\_[i][j] - Moyenne_i}{Ecart\_Type_i}$, les valeurs X_{ij} forment la matrice X centrée et réduite de $M_{initiale}$

3.11 **fin**

3.12 **fin**

3.13 *Calcul de la matrice des corrélations*

3.14 Calculer X^t, la matrice transposée de X (les colonnes de X deviennent les lignes de X^t)

3.15 Calculer $R = X^t.X$, R est la matrice des corrélations associée à $M_{initiale}$, c'est une matrice carrée dont la taille correspond au nombre de colonnes de $M_{initiale}$

3.16 *Calcul et tri des valeurs propres et des vecteurs propres associés à la matrice des corrélations R*

3.17 Calculer D, la matrice résultant de la diagonalisation de R

3.18 Calculer les valeurs propres val_i associées à D, i allant de 1 à jusqu'au nombre de colonnes de R

3.19 Calculer les vecteurs propres vec_i associées aux valeurs propres val_i, i allant de 1 à jusqu'au nombre de colonnes de R

3.20 Trier les couples (val_i, vec_i) par ordre croissant des valeurs propres, avec i allant de 1 à jusqu'au nombre de colonnes de R, placer la liste des couples triées dans la liste $Composantes$

3.21 *Extraction des composantes principales et calcul du taux d'inertie*

3.22 Extraire les k premiers couples de la liste composantes

3.23 Calculer l'inertie associée aux k premiers composantes, $inertie = \frac{\sum_{n=1}^{k} val_n}{\sum_{m=1}^{|composantes|} val_m}$

3.24 *Calcul des coordonnées des éléments de départ dans l'espace formée par les k composantes principales*

3.25 Construire la matrice V, où chaque colonne n correspond aux vecteurs propres vec_n, n allant de 1 à n

3.26 Calculer C, la matrice résultant du produit $X.V$, chaque ligne i de la matrice C représente les coordonnées de élément de la ligne i de $M_{initiale}$ dans un espace à k dimensions.

Algorithme 3 : Schéma d'algorithme de la méthode de l'ACP.

Évaluer et interpréter une analyse en composantes principales

Une fois l'ACP réalisée, un nuage de points est retourné à l'utilisateur. Un tel nuage est cependant complexe à analyser. Pour réaliser une telle analyse, nous proposons la grille suivante, passant en revue les différents points à prendre en considération dans une ACP.

1. **Examiner les valeurs et les vecteurs propres des composantes principales :**
   Chaque composante principale est représentée par une valeur propre et un vecteur propre. La valeur propre d'une composante principale représente la proportion de caractères résumés par la composante (la somme des valeurs propres des composantes principales est donc égale au nombre de caractères). Les valeurs propres ne sont pas élevées, si :
   – le nombre de caractères est important ;
   – la matrice de données à analyser est « creuse ».
   Plus une valeur propre est élevée, plus elle résume des caractères. L'examen du vecteur propre associé à une composante principale permet de mettre en évidence quels caractères sont résumés par cette composante (c'est-à-dire les caractères dont les coordonnées sont les plus éloignés de 0 dans le vecteur, aussi bien en positif, qu'en négatif). Le plan choisi pour la projection est généralement constitué par les deux premières composantes principales. Les composantes suivantes peuvent également intervenir si leurs valeurs propres sont élevées et proches des valeurs propres des deux premières composantes.
2. **Examiner la qualité de la représentation des individus dans le plan :**
   Afin de mesurer la représentativité d'un point dans le plan de projection, il faut mesurer l'angle θ entre le vecteur v représentant le point dans l'espace initial et sa projection p dans le plan. Pour cela, il suffit de calculer la valeur suivante :
   $$cos^2\ (\theta) = (\frac{produit\ scalaire(v,p)}{(||v||.||p||)})^2$$
   Plus cette valeur est proche de zéro, plus la représentativité du point dans le plan est douteuse.
3. **Construire et examiner le cercle des corrélations :**
   Le cercle des corrélations permet d'observer les « positions » des caractères dans la projection (la plate-forme *ProxiDocs* retourne le cercle des corrélations en même temps que les cartes, comme nous avons pu le voir au chapitre 4). La position d'un caractère sur la composante principale se calcule en déterminant le coefficient de corrélation entre le caractère et la composante :
   $$coefficient\ de\ relation(composante, caractere) = composante\ du\ vecteur\ propre\ correspondant\ au\ caractere * \sqrt{valeur\ propre}$$
   NB : il faut appliquer cette formule pour chaque composante principale (par exemple, les deux composantes pour un projection sur un plan) afin d'obtenir un nombre de coordonnées suffisant à positionner le caractère.
   L'examen du cercle de corrélations permet de repérer les groupes de caractères liés ou opposés entre eux (à condition que les points représentant les caractères soient éloignés de l'origine du repère). Il est également possible de caractériser les axes du plan avec des groupes d'attributs. Des « superpositions » relatives entre les individus sur le plan de projection, et les caractères sur le cercle des corrélations, peuvent indiquer des expressions fortes de ces caractères chez ces individus.
4. **Examiner la place des individus sur le plan de projection :**
   À partir du plan de projection, il est possible de réaliser des déductions sur les individus

d'origine, à condition que la majorité de leurs projetés dans le plan soient représentatifs (cf. deuxième point). Les projetés situés aux extrémités positives et négatives des composantes du plan peuvent servir à caractériser ces composantes. Il est alors possible de déduire des similarités entre des points proches les uns des autres sur le plan de projection. Il arrive que des lignes ou des courbes formées par des points soient visibles sur le plan de projection, cela signifie qu'il existe des corrélations fortes entre les individus représentés.

L'ACP est une méthode de projection particulièrement utilisée et attestée pour des tâches très différentes prenant place dans différents domaines. Dans cette thèse, nous avons choisi de la présenter tout particulièrement, en abordant ses principes de fonctionnement ainsi qu'en détaillant les bases de son implémentation. D'autres méthodes de projection sont également intégrées à la plate-forme *ProxiDocs*, ces méthodes sont présentées dans les sections suivantes avec des niveaux de détails légèrement inférieurs à celui utiliser pour présenter la méthode l'ACP.

B.2.3 La méthode de projection de Sammon

Afin de proposer à l'utilisateur différentes méthodes statistiques pour la projection de ses domaines en ensembles documentaires, nous avons implémenté d'autres méthodes dont la méthode de Sammon. Cette méthode, présentée dans [Sammon, 1969], permet, comme l'ACP, de projeter des données prenant place dans des espaces numériques de grandes dimensions vers des espaces de plus faibles dimensions. L'utilisation de cette méthode peut permettre d'avoir une nouvelle vue sur un ensemble documentaire, de faire ressortir des informations passées sous silence lors d'une autre projection, etc. La méthode de Sammon a été utilisée dans différentes applications.

En TAL, dans [Illouz *et al.*, 1999], les auteurs proposent de projeter des mots (extraits d'un grand corpus d'articles de presse) représentés par des vecteurs de 108 dimensions vers un espace à 2 dimensions afin de les visualiser et de mettre en évidence une grande hétérogénéité des articles desquels ils ont été extraits. Dans [Renaux, 2003], l'auteur présente un logiciel dédié à l'analyse de contrefaçons de noms de marques. Ce logiciel utilise la méthode de Sammon afin de projeter, sur un plan, des vecteurs d'indicateurs linguistiques représentant des marques, et ainsi, visualiser d'éventuelles similarités entre noms de marques, déposées ou non. Une autre utilisation, très différente, détaillée dans [Dybowski *et al.*, 1996], consiste à utiliser des projections de Sammon afin de visualiser des données génétiques.

La tâche réalisée par la méthode de Sammon est identique à celle effectuée par les méthodes abordées précédemment, c'est-à-dire la projection d'un espace à n dimensions vers un espace à k dimensions (avec $n > k$). Son fonctionnement général sera le suivant :

1. Placer chaque texte aléatoirement dans l'espace d'arrivée.
2. Pour chacun de ces textes, tester si les distances (dans l'espace de départ à n dimensions) entre ce dernier et les autres textes sont respectées.
3. Si ce n'est pas le cas, les autres textes peuvent effectuer un léger déplacement afin de tendre vers une situation où les distances entre chacun des textes sont respectées.
4. Reprendre à l'étape 2. jusqu'à ce que les distances entre chaque texte soient respectées dans l'espace d'arrivée à une faible approximation près.

Afin d'implémenter cette méthode, nous avons utilisé l'algorithme décrit ci-dessous. Voici les notations qui seront utilisées dans cet algorithme :

N : le nombre de vecteurs à projeter (le nombre de textes de l'ensemble documentaire).

A, B, C, ... : les points (ou textes) de l'espace de départ à i dimensions (à projeter).

A',B',C',... : les points (ou textes) de l'espace de projection à k dimensions (projeté).

A_i : la i^{eme} dimension du vecteur A.

AB : le vecteur de l'espace de départ.

A'B' : le vecteur de l'espace de projection.

E : l'erreur de projection par cycle sur l'ensemble des textes.

d_{AB} : la distance euclidienne entre les points A et B (dans l'espace de départ).

$d_{A'B'}$: la distance euclidienne entre les points A' et B' (dans l'espace de projection).

Del_1 *et* Del_2 : représentent respectivement les dérivées première et seconde des points.

c : la somme des d_{AB} sur l'ensemble des paires AB de l'espace de départ.

MF : le *magic factor* de régression linéaire (on prendra entre 0,3 et 0,4, valeurs déterminées empiriquement), ce facteur permet de pondérer les coordonnées des projetés à chaque nouvelle coordonnée calculée.

Seuil : représente le seuil à partir duquel on considère l'erreur de projection tolérable pour obtenir une cartographie correcte de l'ensemble documentaire.

L'algorithme ci-contre décrit alors la méthode de projection de Sammon. L'inconvénient majeur de cette méthode est lié au placement aléatoire des points au début de l'algorithme. Ainsi, cet algorithme exécuté plusieurs fois sur un même ensemble documentaire, donnera des projections différentes, pouvant révéler différemment les informations de l'ensemble de départ. Il est à noter que la complexité de cet algorithme peut être assez élevée, étant donné qu'une boucle sur le calcul des projetés est effectuée jusqu'à avoir atteint un certain seuil d'erreur acceptable. Il faudra donc choisir le seuil d'erreur permettant d'obtenir des résultats en un temps raisonnable.

B.2.4 L'analyse factorielle des correspondances

Comme pour les méthodes présentées précédemment, nous avons choisi d'intégrer la méthode de l'analyse factorielle des correspondances (AFC) afin de donner la possibilité aux utilisateurs de visualiser et de comparer plusieurs projections de leur espace de départ par plusieurs méthodes. L'AFC a été développée par Jean-Paul Benzécri dans le début des années 1980 [Benzécri, 1980]. Les motivations qui ont entraînées la création de cette méthode sont d'ordre linguistique : son auteur cherchait à mettre en évidence une éventuelle structure mathématique du langage par des traitements statistiques (l'auteur se place dans les prolongements des travaux présentés dans [Harris, 1971]).

La méthode de l'AFC est particulièrement populaire dans la communauté de l'analyse de données, et plus récemment, dans la communauté de l'analyse des données textuelles. Des outils, tel Lexico3[188], implantent une telle méthode afin, par exemple, de révéler une certaine proximité entre mots à l'intérieur d'un corpus de documents. Devant le relatif « consensus » existant autour de cette méthode, mais aussi devant son intérêt dans une problématique très proche de la nôtre, nous avons donc choisi de l'intégrer à l'application *ProxiDocs*.

Afin d'implanter la méthode de l'AFC dans nos programmes, nous avons utilisé le lien très étroit entre cette méthode et la méthode de l'ACP. Ainsi dans [Bouroche et Saporta, 1980][page 87], les auteurs expliquent qu'effectuer une AFC revient à réaliser une double ACP, à la fois sur les lignes, et sur les colonnes du tableau de données, en utilisant la distance du χ^2 (ou Khi-deux),

[188] Lexico3 est un logiciel conçu pour le traitement lexicométrique de textes comportant plusieurs centaines de milliers d'occurrences. Il a d'abord été développé par André Salem (ILPGA - Paris 3) au sein du laboratoire « Lexicométrie et textes politiques » de l'E.N.S. de Fontenay-Saint-Cloud. Il est désormais maintenu par l'équipe CLA^{2}T de l'UPRES SYLED.

```
// Calcul initial des distances AB ;
c = 0 ;
pour A allant de 1 à N-1 faire
    pour B allant de 2 à N faire
        d_AB = 0 ;
        pour x allant de 1 à k faire
            d_AB = d_AB + (A_i − B_i)^2 ;
        fin
        d_AB = sqrt(d_AB) ;
        c = c + d_AB ;
    fin
fin
// Calcul initial des points projetés ;
pour A allant de 1 à N-1 faire
    pour x allant de 1 à k faire
        A_x = random(0, 1) ;
    fin
fin
// Réajustements des positions des points dans l'espace d'arrivée ;
tant que E > Seuil faire
    E = 0 ;
    pour A allant de 1 à N-1 faire
        pour B allant de 2 à N faire
            Calculer d_A'B' et stocker le résultat dans un tableau ;
            E = E + ([d_AB − d_A'B']^2 / d_AB)/c ;
        fin
    fin
    Del_1x = 0, Del_2x = 0, Del_1y = 0, Del_2y = 0 ;
    pour A allant de 1 à N faire
        pour J allant de 1 à N faire
            si A ≠ J alors
                Del_1x = Del_1x + [(d_AJ − d_A'J')/(d_AJ * d_A'J')] * (A_x − J_x) ;
                Del_1y = Del_1y + [(d_AJ − d_A'J')/(d_AJ * d_A'J')] * (A_y − J_y) ;
                Del_2x = Del_2x + [1/(d_AJ * d_A'J')] * ((d_AJ − d_A'J') − [(A_x − J_x)^2/d_A'J'] *
                [1 + (d_AJ − d_A'J')/d_A'J']) ;
                Del_2y = Del_2y + [1/(d_AJ * d_A'J')] * ((d_AJ − d_A'J') − [(A_y − J_y)^2/d_A'J'] *
                [1 + (d_AJ − d_A'J')/d_A'J']) ;
            fin
        fin
        Del_1x = −2 * Del_1x/c, Del_2x = −2 * Del_2x/c ;
        Del_1y = −2 * Del_1y/c, Del_2y = −2 * Del_2y/c ;
        Del_x = Del_1x/|Del_2x|, Del_y = Del_1y/|Del_2y| ;
        A_x = A_x − MF * Del_x, A_y = A_y − MF * Del_y ;
    fin
fin
```

Algorithme 4 : Algorithme de la méthode de projection de Sammon.

plutôt que la distance euclidienne utilisée traditionnellement dans l'ACP. Les lignes du tableau de données représentent les textes de l'ensemble étudié, alors que les colonnes représentent les domaines de l'utilisateur. Les projections que nous construisons mettent en évidence les textes de l'espace documentaire étudié, nous ne souhaitons donc faire figurer que les textes dans les représentations graphiques.

B.3 Méthodes de classification de l'espace visualisé

Afin d'aider les utilisateurs dans leur analyse des visualisations d'ensembles documentaires, nous leur proposons différents regroupements automatiques entre textes sur les cartes. Pour cela, les deux méthodes, présentées ci-dessous, sont utilisées.

B.3.1 La classification hiérarchique ascendante

La méthode de la classification hiérarchique ascendante (CHA) est issue de l'analyse des données [Bouroche et Saporta, 1980] et permet de regrouper entre eux des éléments proches, décrits par des valeurs numériques. Son utilisation est fréquente, nous pouvons, par exemple, citer son usage pour l'alignement de textes bilingues [Zimina, 2000] ou encore pour le document structurés [Despeyroux *et al.*, 2005].

Cette méthode a un schéma de fonctionnement global assez simple, pouvant se résumer par les deux étapes suivantes :

- Parmi les n entités à classer, chercher les deux entités les plus proches. Ces deux entités sont ensuite agrégées en un nouveau groupe.
- Calculer les distances entre le nouveau groupe et les entités restantes. La configuration est alors identique à celle de l'étape 1., sauf qu'il reste seulement $n - 1$ entités à classer.

Et ainsi de suite, on cherche de nouveau les deux entités ou groupes les plus proches, que l'on agrége et ceci jusqu'à ce qu'à obtenir le nombre de groupes choisi par l'utilisateur.

L'application de cette méthode à notre problématique est alors assez simple, puisque nous assimilons les textes aux points d'un espace. Calculer la proximité de deux textes, revient donc à calculer la distance euclidienne entre les points les représentant. Ainsi, la distance entre deux groupes de textes s'obtiendra en déterminant le centre de gravité de chacun de ces groupes, puis en calculant la distance euclidienne entre ces deux points. Le texte supposé le plus représentatif d'un groupe est alors celui étant le plus proche du centre de gravité de ce groupe. La figure B.2 illustre le principe de la CHA dans la classification de 5 documents, chacun de ces textes est représenté par un couple de réel (X, Y).

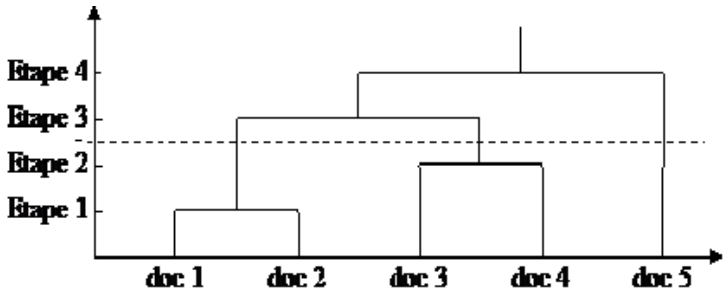


FIG. B.2 – Application de la CHA dans la classification de 5 documents. La ligne en pointillés indique met en évidence les groupes obtenus à l'issue de l'étape 2, soient le groupe formé par doc 1 et doc 2, le groupe formé par doc 3 et doc 4 et le groupe formé par doc 5.

Afin d'implanter cette méthode de classification, nous avons utilisé l'algorithme suivant[189].

```
    Entrées : Un ensemble documentaire et un nombre de groupes
    Sortie  : Un ensemble de groupes de documents
5.1 // étape d'initialisation ;
5.2 liste_groupes = ∅ ;
5.3 pour chaque texte de l'ensemble documentaire faire
5.4 |   Créer un groupe avec le texte ;
5.5 |   Placer ce groupe dans liste_groupes ;
5.6 fin
5.7 // étape de classification ;
5.8 tant que la taille de liste_groupe ne correspond pas au nombre de groupes demandé
    par l'utilisateur faire
5.9 |   distance_mini = +∞ ;
5.10|   groupes_à_fusionner = (∅, ∅) ;
5.11|   pour chaque groupe G_i de liste_groupe faire
5.12|   |   pour chaque groupe G_j de liste_groupe tel que G_j ≠ G_i faire
5.13|   |   |   Soit distance_courante, la distance euclidienne entre G_i et G_j ;
5.14|   |   |   si distance_courante < distance_mini alors
5.15|   |   |   |   distance_mini = distance_courante ;
5.16|   |   |   |   groupes_à_fusionner = (G_i, G_j) ;
5.17|   |   |   fin
5.18|   |   fin
5.19|   fin
5.20|   Fusionner les deux groupes de groupes_à_fusionner ;
5.21 fin
```

Algorithme 5 : Algorithme de la Catégorisation Hiérarchique Ascendante.

B.3.2 La méthode des K-Means

D'autres méthodes de classification sont également très populaires. Tout particulièrement, une méthode, appelée *K-Means* [MacQueen, 1967], est fréquemment utilisée. Son usage principal est souvent lié à la segmentation des images [D'Hondt et Khayati, 2005], mais son usage se retrouve également en TAL, par exemple, dans [Bellot et El-Bèze, 2000], où les auteurs proposent d'utiliser cette méthode dans une tâche de recherche documentaire. Intégrer cette méthode des K-Means à l'application *ProxiDocs* semble être particulièrement pertinent et pourrait nous permettre d'obtenir des regroupements intéressants et différents de ceux obtenus avec la méthode de la CHA.

Le fonctionnement de cette méthode peut se résumer par les étapes suivantes :

1. Placer aléatoirement[190] les k centres dans l'espace d'arrivée. Ces points représentent les centres initiaux des groupes.

[189] La formule de la distance euclidienne, utilisée dans cet algorithme, a été définie précédemment dans cette annexe

[190] Le placement initial peut se faire de façon non aléatoire en respectant la topologie de l'ensemble en essayant de distribuer « intelligemment » les centres. Par exemple dans [Déjean, 2005], l'auteur propose de réaliser tout d'abord une CHA afin d'obtenir une classification, puis d'utiliser les centres de gravité des groupes obtenus comme position initiale des centres de la méthode des K-Means.

2. Assigner chaque élément de l'espace au groupe duquel il est le plus proche du centre.
3. Quand tous les éléments sont affectés à un groupe, recalculer les centres de gravité des k groupes, ces nouveaux centres remplacent les précédents.
4. Répéter les étapes 2 et 3 jusqu'à ce que les centres ne changent (presque) plus.

L'algorithme de la méthode des K-Means que nous avons implémenté dans la plate-forme *ProxiDocs* est le suivant.

```
    Entrées : Un ensemble documentaire D et un nombre de groupes k
    Sortie  : Un ensemble de groupes de documents
6.1 // étape d'initialisation ;
6.2 pour i allant de 1 à k faire
6.3 |   Positionner aléatoirement le point k_i dans l'espace, ce point représente la position
    |   initiale du i-ème centre ;
6.4 fin
6.5 // étape de classification ;
6.6 tant que les centres se déplacent faire
6.7 |   pour chaque D[i] de l'espace documentaire D faire
6.8 |   |   Calculer les distances entre D[i] et chaque centre ;
6.9 |   |   Affecter D[i] au groupe duquel il est le plus proche du centre ;
6.10|   fin
6.11|   pour i allant de 1 à k faire
6.12|   |   Replacer le centre k_i sur le centre de gravité du groupe i ;
6.13|   fin
6.14 fin
```

Algorithme 6 : Algorithme de la méthode des K-Means.

B.3.3 Le choix du nombre de groupes d'une classification

Choisir un nombre de classes à construire à partir d'un ensemble d'éléments est un problème particulièrement difficile. Dans [Bock, 1996], l'auteur dresse un panorama très complet de ces différentes méthodes numériques cherchant à déterminer le nombre de classes contenues dans un ensemble et il en conclut que :

> *The estimation of the* ***true*** *number of classes has been recognized as one of the most difficult problems in cluster analysis.*

De notre point de vue, les principales raisons de cette difficulté semblent liées au fait que les concepteurs de ces méthodes cherchent à trouver **le** bon nombre de classes ainsi que **la** bonne classification. En effet, pour évaluer des méthodes de classification, les experts comparent les résultats obtenues par ces méthodes, avec les résultats obtenus avec des méthodes de classification, où les classes à déterminer sont décrites et explicitées en machine (par exemple, se reporter respectivement à [Fraley et Raftery, 1998] et à [Still et Bialek, 2004] pour des présentations de deux méthodes différentes de classification ainsi que de leur évaluation). Il est alors, en effet, très difficile de retrouver et de remplir automatiquement des classes pré-définies par des experts humains.

L'objectif de notre travail est de fournir aux utilisateurs des supports de visualisations et d'interactions les aidant à appréhender le contenu d'ensembles documentaires par rapport à leurs

domaines d'intérêts. Les méthodes que nous proposons se doivent d'être les plus pertinentes possible, sans pour autant chercher à répondre à des critères d'évaluation très stricts et pré-établis, comme, par exemple, chercher à retrouver des classes que l'on aurait définies auparavant.

Cette définition préalable des classes de textes à obtenir ne pourrait alors être pertinente qu'en demandant à l'utilisateur quels types de regroupements il envisage d'observer dans son ensemble documentaire, par rapport à sa tâche. Une telle opération pourrait être envisagée dans la suite de nos travaux pour laisser à l'utilisateur un moyen supplémentaire d'exprimer ses connaissances et son point de vue sur sa tâche, même si, la plupart du temps, les regroupements de textes mis en évidence sont assez éloignés d'éventuels attendus.

L'intérêt majeur des méthodes de classification, par rapport à des méthodes de classification, est lié au fait qu'il n'est pas besoin d'avoir des connaissances préalables sur les groupes à déterminer. Les méthodes utilisées réalisent, chacune, une telle opération de classification, à partir d'un nombre de groupes choisi par l'utilisateur. Ce choix peut alors amener à des analyses très différentes des classifications obtenues :

- si le nombre choisi de groupes est petit, alors la classification sera générale et risque de forcer le regroupement d'éléments assez éloignés uniquement pour atteindre le nombre de groupes fixés ;
- d'une manière opposée, si le nombre choisi de groupes est grand, alors la classification est plus fine et risque de retourner plusieurs groupes distincts d'éléments pourtant très proches et qui pourraient être regroupés au sein d'un seul groupe.

Bien évidemment, l'utilisateur peut utiliser l'application de manière cyclique et s'il se rend compte qu'une première classification à n groupes propose des regroupements trop généraux, il peut alors relancer une nouvelle classification avec un nombre de groupes supérieur au nombre n initial. Dans le cas opposé, où les regroupements sont trop fins, l'utilisateur peut alors réaliser une nouvelle classification avec un nombre de groupe inférieur à n.

Malgré cela, et toujours pour assister un peu plus l'utilisateur, nous avons développé une nouvelle version de la méthode CHA, ne prenant pas de nombre de groupes en entrée. Cette méthode consiste à s'inspirer du critère de Ward [Ward, 1963] expliquant qu'une « bonne » classification est obtenue lorsque l'on minimise la perte d'inertie inter-classe résultant de l'agrégation de deux entités. Cette opération est effectuée dans le but d'obtenir les groupes les plus homogènes et les plus distants les uns des autres. Afin de reprendre cette idée, nous avons alors intégré de nouveaux traitements à la méthode de la CHA. Ces traitements consistent toujours à calculer, à chaque pas de classification, la distance entre chaque couple de groupes. Les plus petite et plus grande distances, respectivement p et D, sont mémorisées et le rapport p/D est calculé. Si ce rapport est inférieur à un certain seuil S[191], fixé empiriquement, alors la classification continue, nous considérons que les groupes ne sont pas assez distants les uns des autres, et dans le cas contraire, la classification s'arrête.

Ceci revient à remplacer la condition de la boucle principale de l'algorithme de la CHA :

`Tant que le nombre de groupes demandés n'est pas atteint, faire :`

Par :

`Tant que le rapport` p/D `de la distance` p `entre les deux groupes les plus proches avec la distance` D `entre les deux groupes les plus éloignés est inférieur à un certain seuil` S`, faire :`

[191] Le problème du nombre est en quelque sorte déplacé : on ne choisit plus de nombre de groupes mais on fait intervenir un seuil qui, selon sa valeur va faire évoluer le nombre de groupes ainsi que la finesse des classifications obtenues (plus S est grand, plus le nombre de groupes sera grand, et inversement).

Durant nos expériences, le seuil S a été fixé empiriquement à une valeur de 1/8. Cette méthode, assez simple, permet alors à l'utilisateur de réaliser des classifications avec la CHA sans préciser de nombre de groupes au préalable.

B.4 Comparaisons des méthodes de projection et de classification

Nous proposons dans cette section une comparaison des différentes méthodes de projection et de classification présentées précédemment. L'ensemble documentaire documentaire analysé est constitué de 789 articles du journal *Le Monde* de 1989 choisis aléatoirement parmi l'ensemble des articles de l'année. Neuf domaines ont été utilisés : la religion, l'éducation, l'agriculture, l'automobile, l'économie, la politique, la guerre, le sport et la télévision. La représentation utilisée est l'ensemble de lexies, la taille moyenne des ensembles est de 40 lexies. La méthode de comptage relative a été utilisée avec une prise en compte des flexions des lexies avec *BDLex*. Une restriction sur les textes contenant moins de 2 occurrences des domaines a été réalisée. Les cartes présentées ci-dessous mettent ainsi en évidence 679 documents.

B.4.1 Cartes des textes

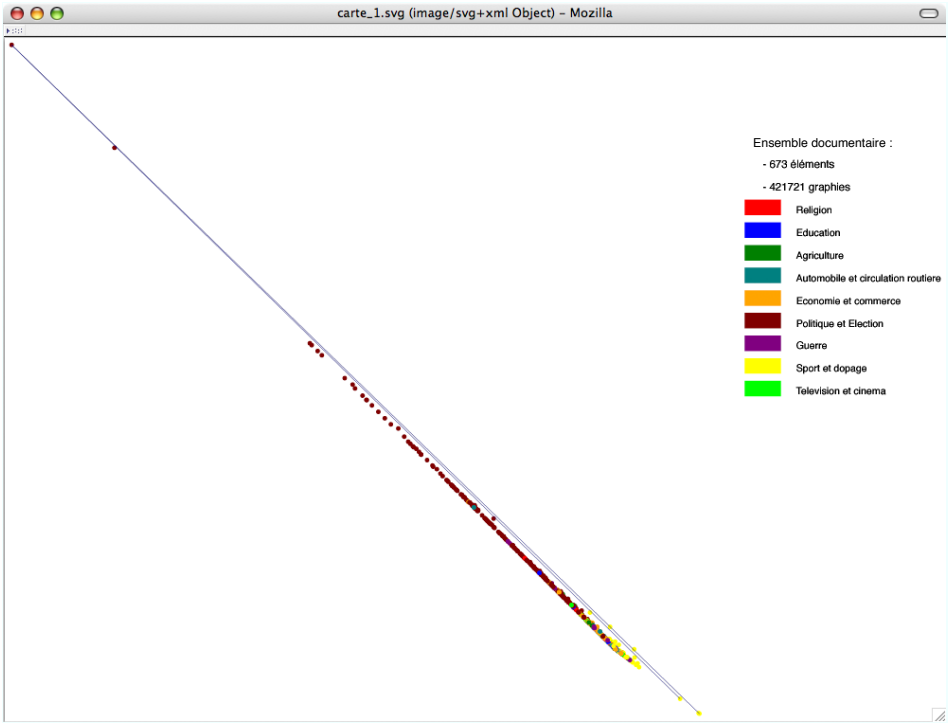


FIG. B.3 – Carte de l'ensemble documentaire réalisée avec la méthode des plus grandes distances.

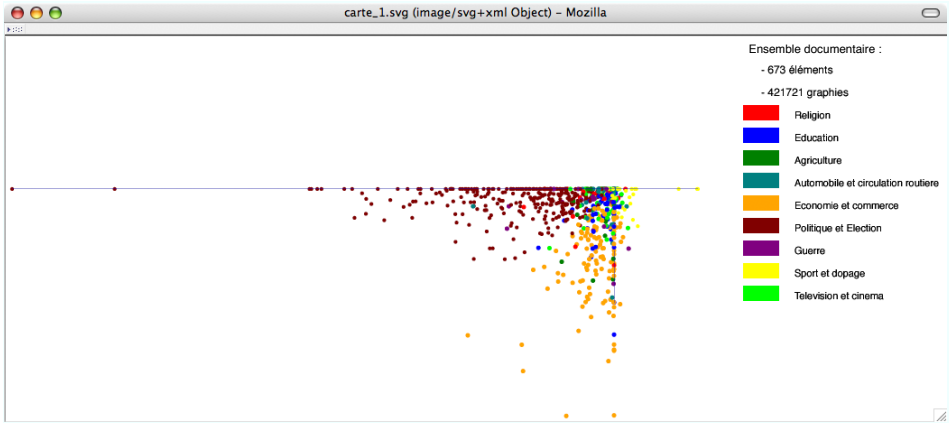


FIG. B.4 – Carte de l'ensemble documentaire réalisée avec la méthode du produit scalaire.

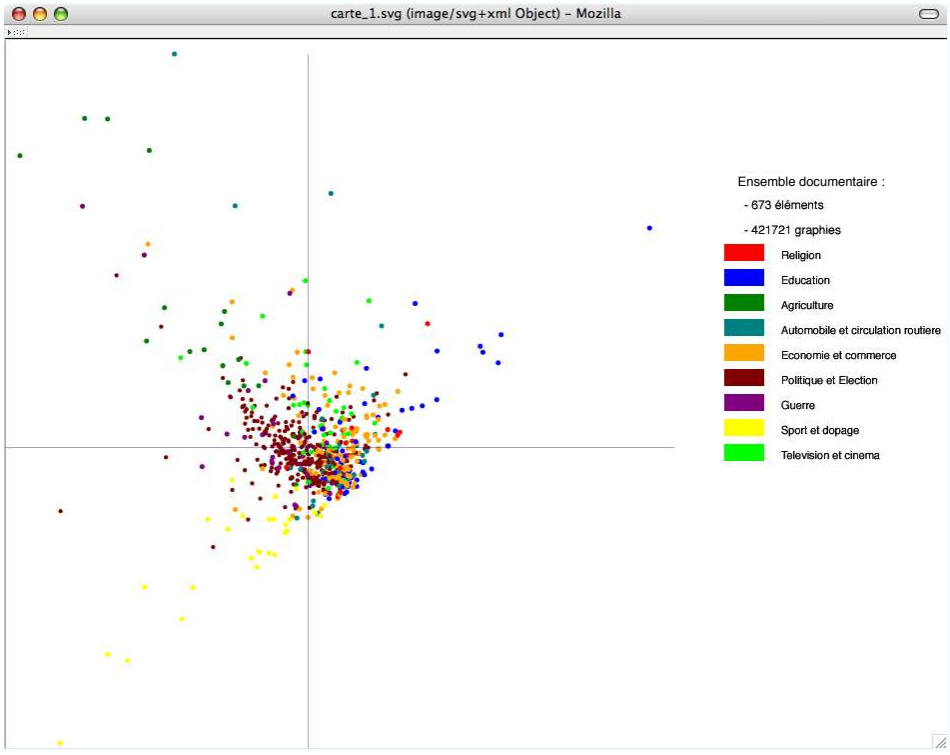


FIG. B.5 – Carte de l'ensemble documentaire réalisée avec la méthode de l'ACP.

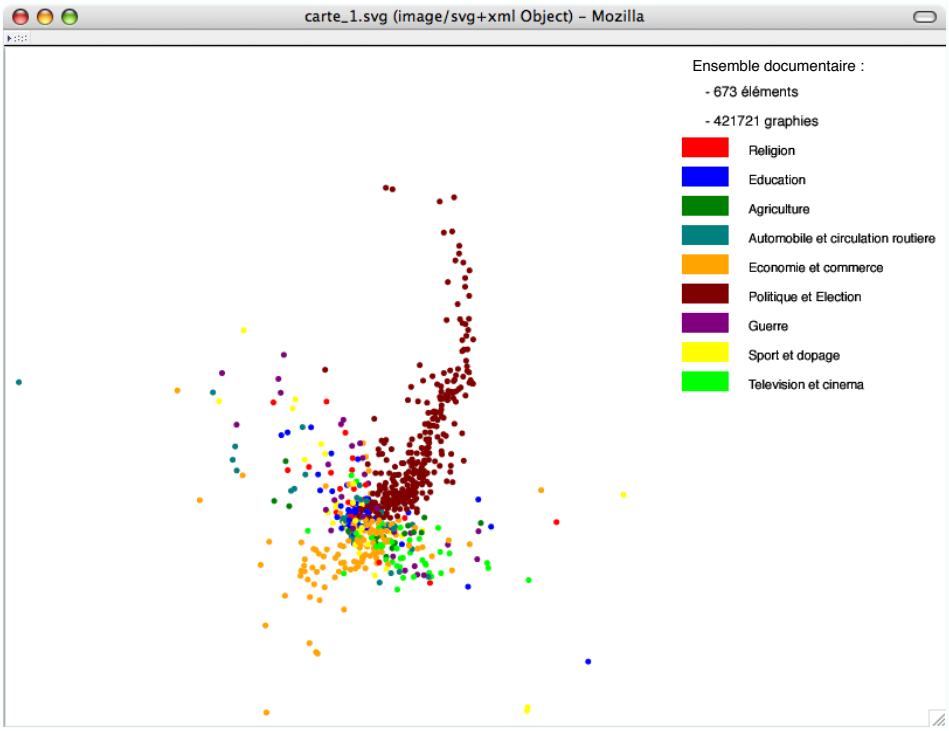


FIG. B.6 – Carte de l'ensemble documentaire réalisée avec la méthode de Sammon.

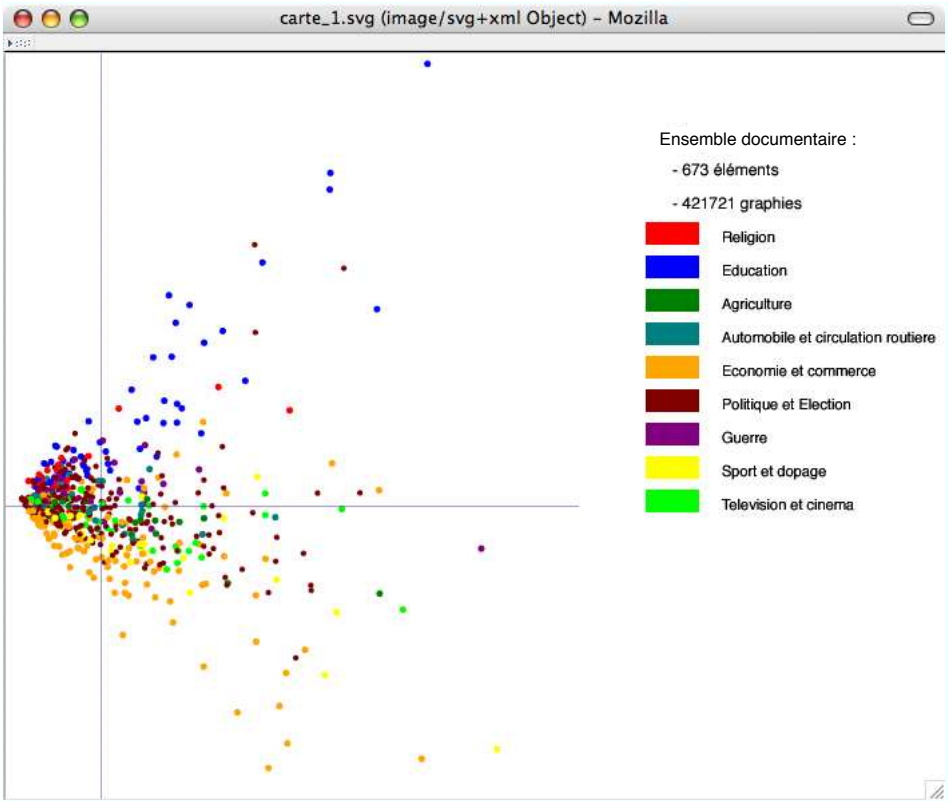


FIG. B.7 – Carte de l'ensemble documentaire réalisée avec la méthode de l'AFC.

La méthode des plus grandes distances et la méthode du produit scalaire donnent des cartes où les points sont particulièrement rapprochés. En particulier, la première de ces deux méthodes a retenu deux axes séparés par un angle très aigu, entraînant une projection des points dans une zone très restreinte formée par ces deux axes. Les méthodes de l'ACP, de Sammon et de l'AFC, présentent des ensembles de points beaucoup plus espacés, laissant mieux entrevoir des ensembles de textes de mêmes domaines majoritaires.

B.4.2 Cartes des groupes de textes

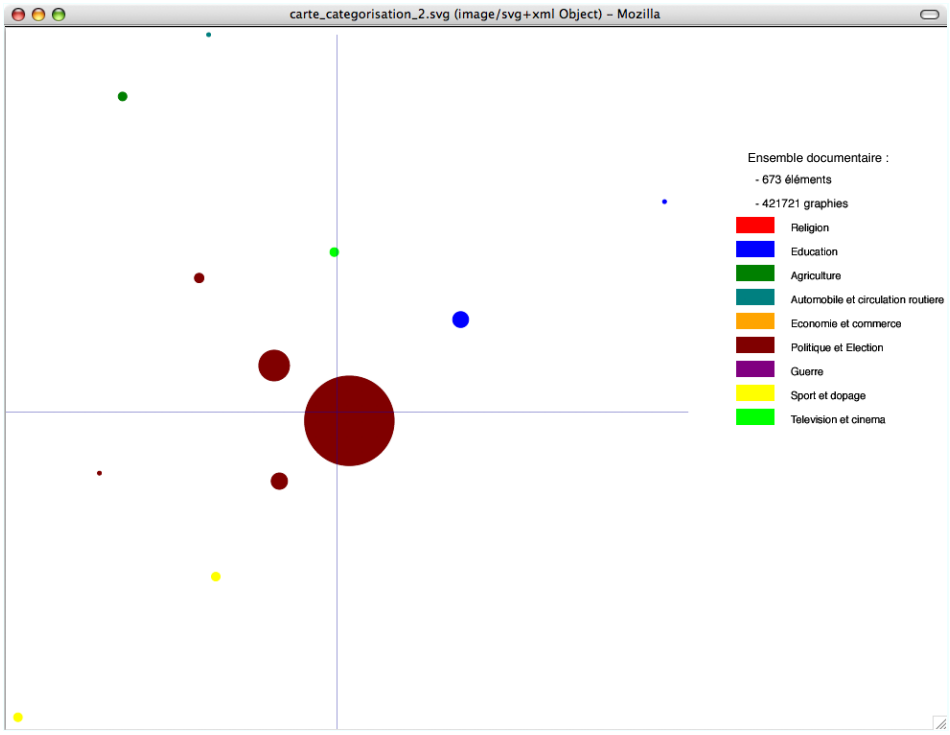


FIG. B.8 – Carte des groupes de textes de l'ensemble documentaire catégorisé avec une CHA à partir d'une ACP.

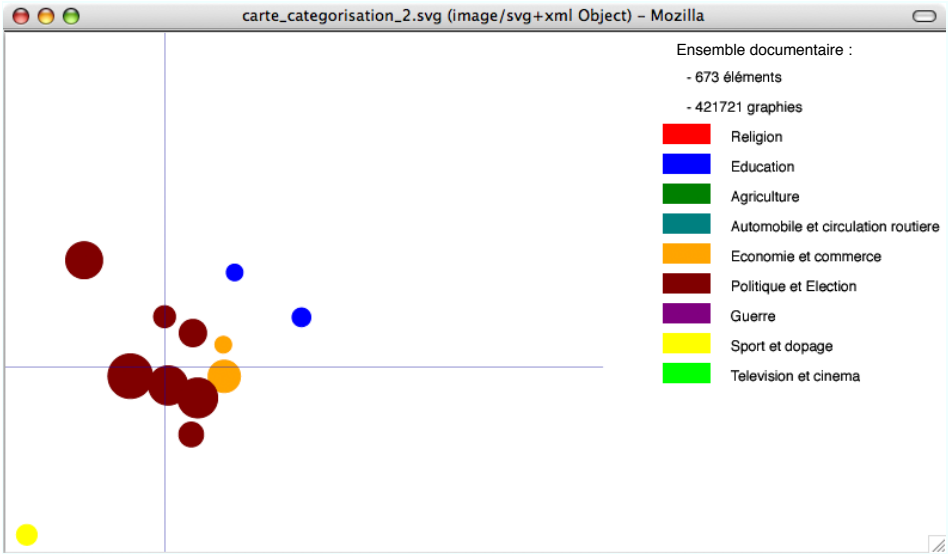


FIG. B.9 – Carte des groupes textes de l'ensemble documentaire catégorisé avec la méthode des K-Means à partir d'une ACP.

Les deux cartes précédentes mettent en évidence des classifications en 12 groupes (nombre choisi empiriquement) à partir de la carte de l'ACP. La méthode de la CHA laisse paraître une classification avec un grand groupe principal et des groupes de plus petites tailles « gravitant » autour de ce grand groupe. Au contraire, la méthode des K-Means propose des groupes de taille équivalente mais beaucoup plus rapprochés.

Dans les différentes expériences que nous avons présentées au chapitre 4 de cette thèse, nous avons principalement utilisé la méthode de projection de l'ACP et la méthode de classification de la CHA. De telles utilisations étaient principalement liées aux meilleures observations que nous permettaient de faire ces méthodes par rapport aux autres. Selon les objectifs de l'utilisateur, des méthodes pourront être plus appropriées que d'autres. Dans tous les cas, ce dernier est libre de choisir parmi les différentes méthodes que nous lui proposons. Il pourra comparer les cartes obtenues par différentes méthodes de projection et de classification, et retenir celles répondant le mieux à ses attentes.

Annexe C

Dispositifs LUCIA utilisés au cours des différentes expérimentations du modèle *AIdED*

Sommaire

C.1 Dispositifs utilisés pendant l'expérimentation de recherche documentaire sur Internet 208

C.2 Dispositifs utilisés dans l'étude des trois métaphores conceptuelle durant le projet *IsoMeta* 214

C.3 Ressources lexicales utilisées pendant les expérimentations sur les forums de discussions 217

C.3.1 Expérimentation sur la détection de l'identité professionnelle 217

C.3.2 Expérimentation sur l'usage d'une terminologie professionnelle 218

Dans cette annexe, nous présentons les différentes RTO utilisées dans les différentes expérimentations détaillées au chapitre 4 de cette thèse. Les dispositifs LUCIA présentés sont accessibles à l'adresse suivante : `http://www.info.unicaen.fr/~troy/dispositifs` (page consultée le 15 juillet 2007).

C.1 Dispositifs utilisés pendant l'expérimentation de recherche documentaire sur Internet

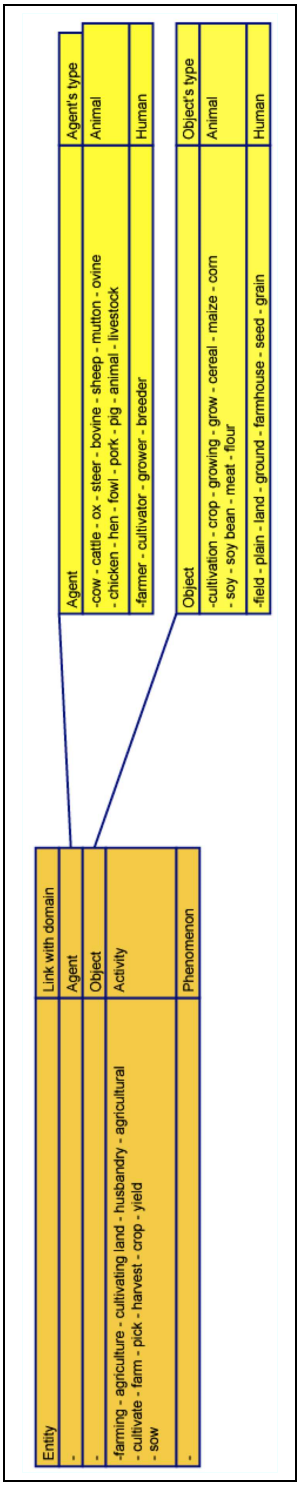


FIG. C.1 – Le dispositif de l'agriculture construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet.

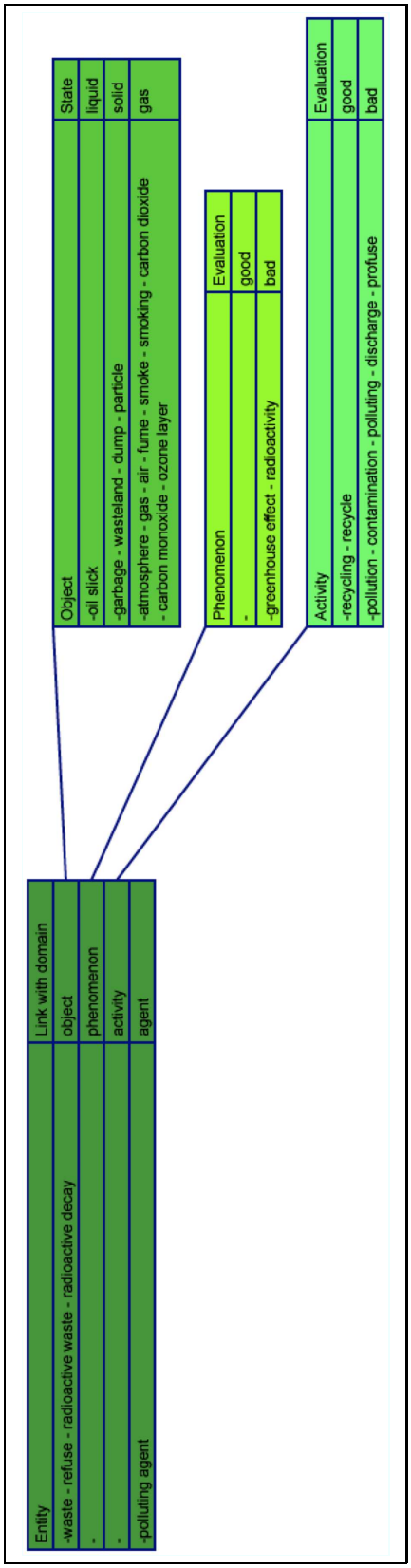


FIG. C.2 – Le dispositif de la pollution construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet.

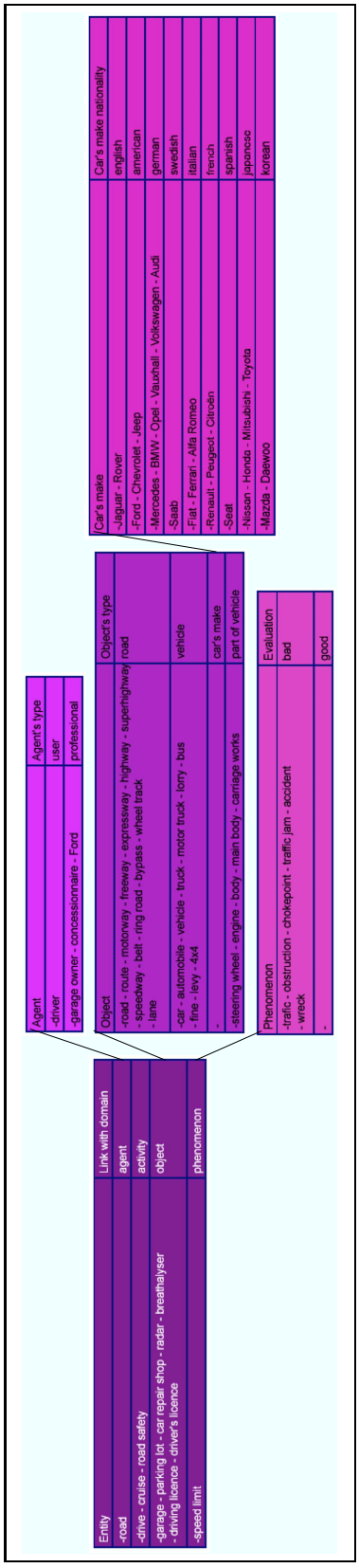


FIG. C.3 – Le dispositif de la sécurité routière construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet.

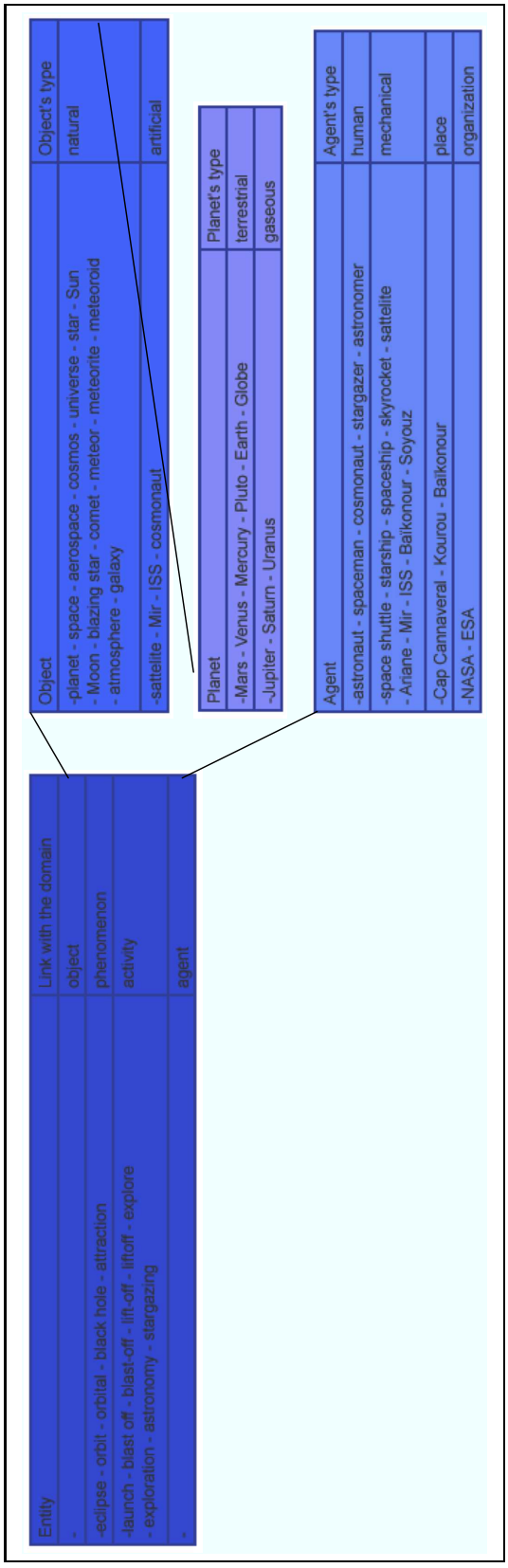


FIG. C.4 – Le dispositif de l'espace construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet.

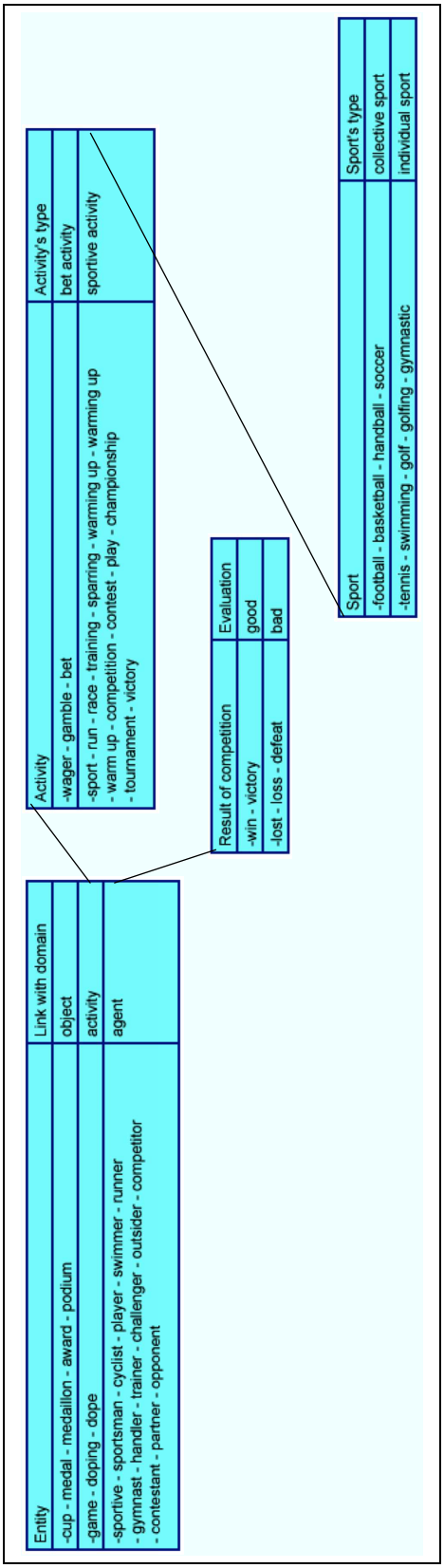


FIG. C.5 – Le dispositif du sport construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet.

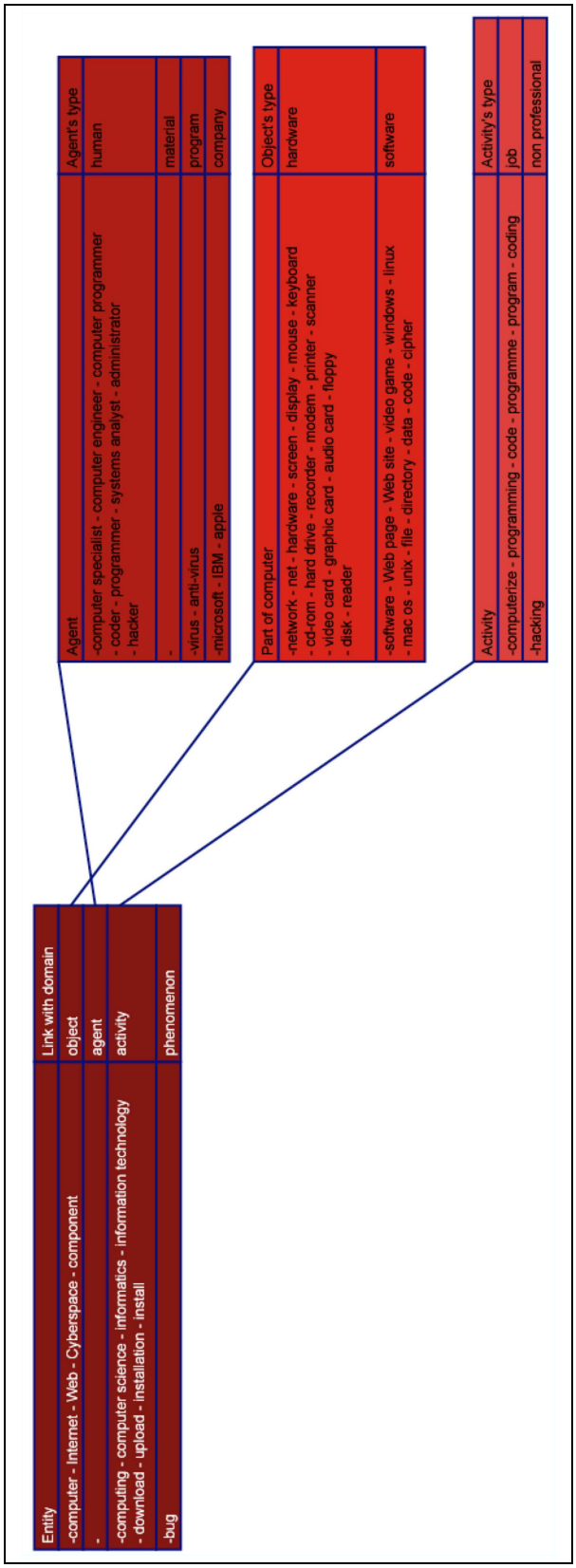


FIG. C.6 – Le dispositif de l'informatique construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet.

C.2 Dispositifs utilisés dans l'étude des trois métaphores conceptuelle durant le projet *IsoMeta*

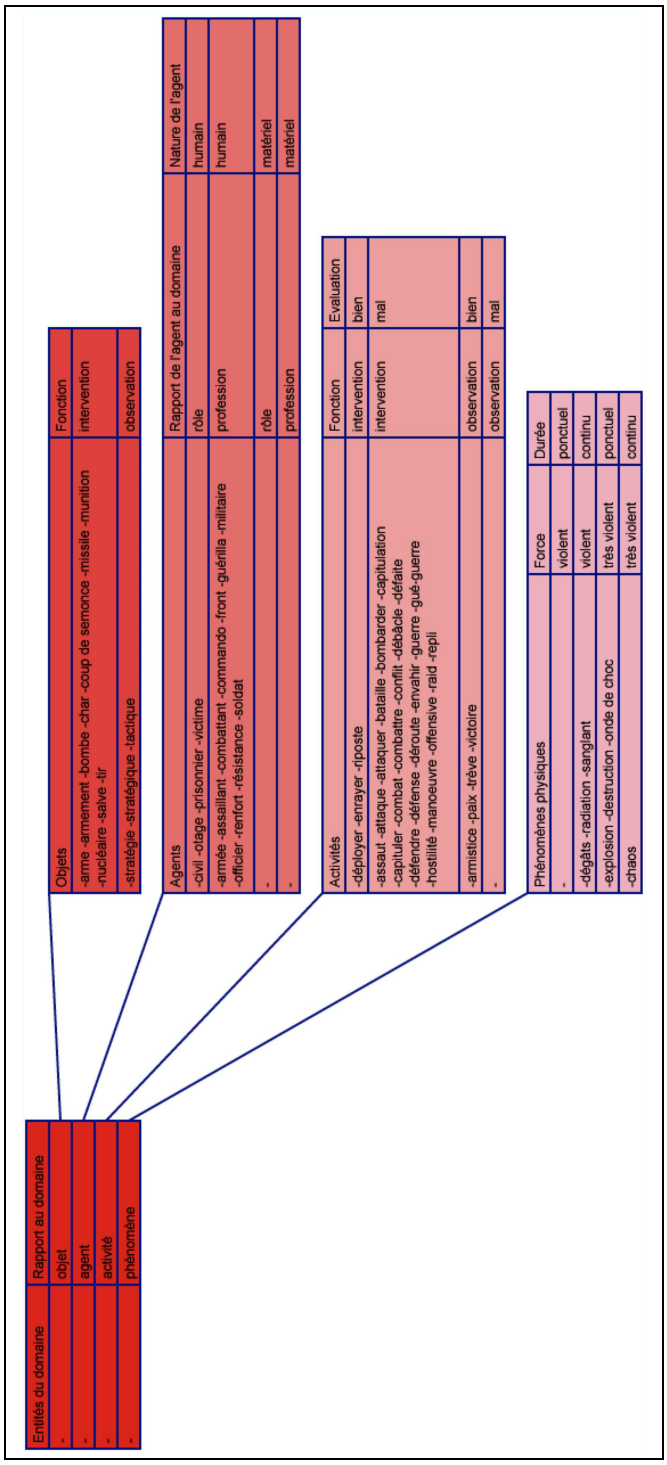


FIG. C.7 – Le dispositif de la guerre mis au point et utilisé au cours de l'expérimentation associée au projet *IsoMeta*.

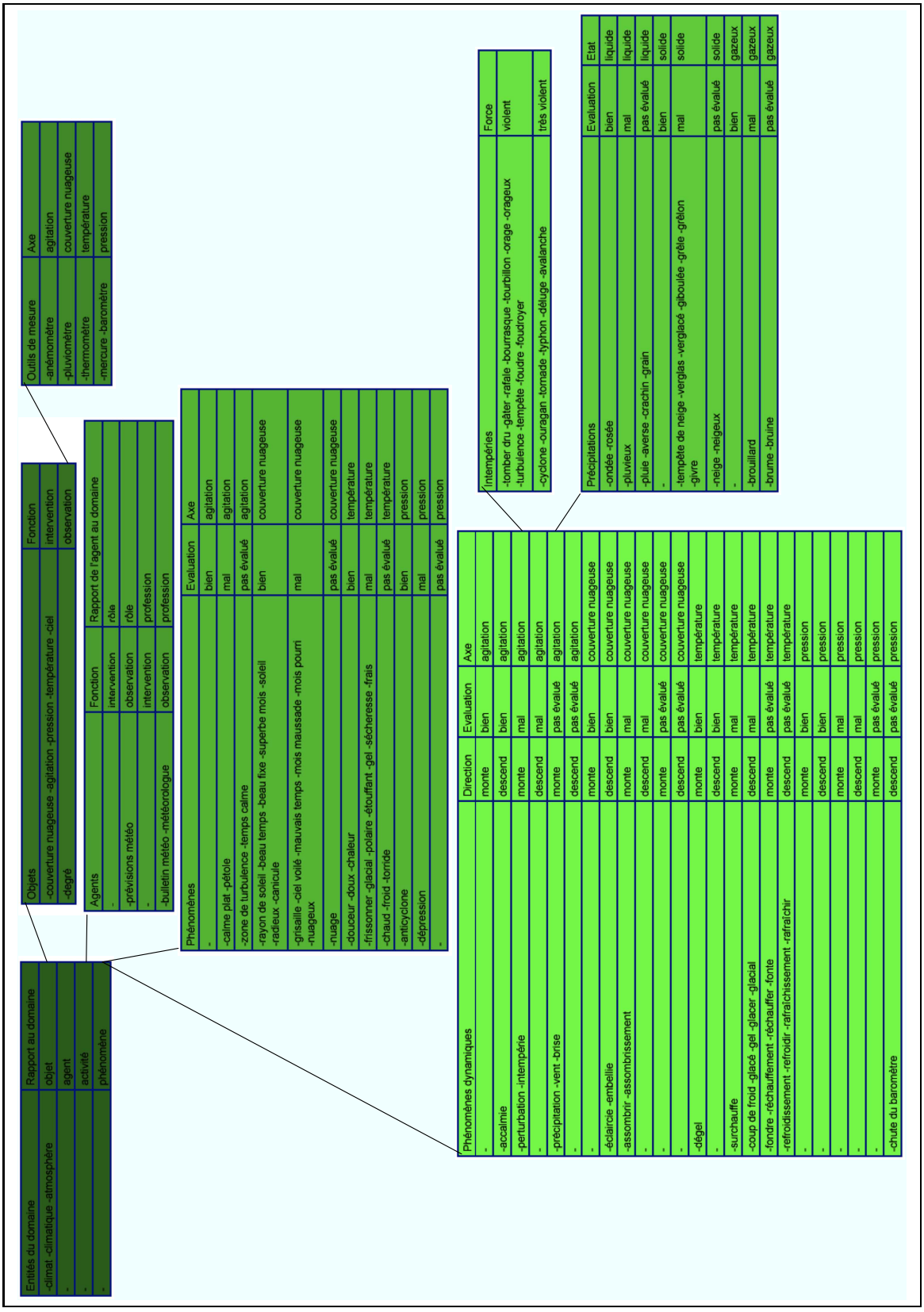


FIG. C.8 – Le dispositif de la météorologie repris et utilisé au cours de l'expérimentation associée au projet *IsoMeta*.

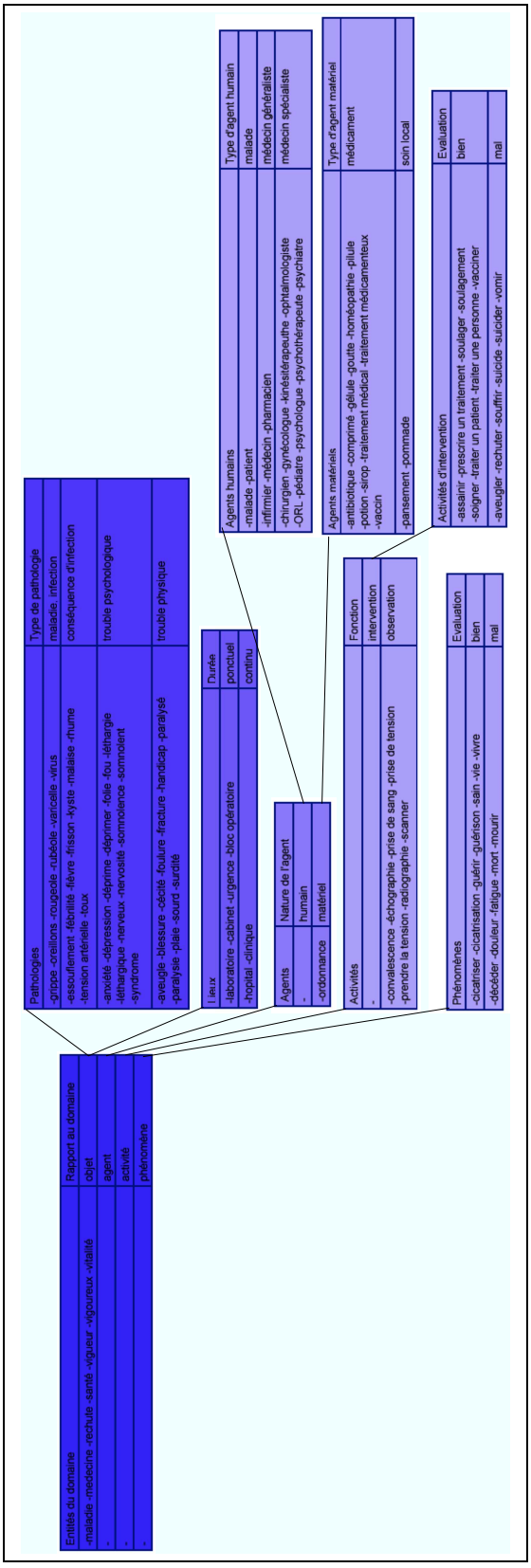


FIG. C.9 – Le dispositif de la santé mis au point et utilisé au cours de l'expérimentation associée au projet *IsoMeta*.

C.3 Ressources lexicales utilisées pendant les expérimentations sur les forums de discussions

C.3.1 Expérimentation sur la détection de l'identité professionnelle

Information (dimension collaborative) *demander, envoyer, info, information, informations, infos, mails, nouvelles, question, répondre, réponse, yahoo, communiquer*
Ressources (dimension collaborative) *besoin, demande, doc, docs, document, documents, échanges, envoie, envoyer, fiche, fiches, fichiers, http, idée, idées, images, intéresse, Internet, liste, mayetic, net, photocopies, prep, prép, préparation, préparations, preps, publication, quickplace, ressources, site, sites, trucs, www*
Pratique (dimension réflexive) *activité, activités, album, apprendre, apprentissage, apprentissages, arts, atelier, ateliers, bio, chansons, classe, classes, compétences, compréhension, comprendre, comptines, connaissance, conseils, danse, discipline, école, élève, élève, élèves, enfant, enfants, eps, évaluation, français, gamins, gestion, histoire, lecture, lire, math, maths, matière, motricité, musique, niveau, objectifs, oral, pédagogique, pratique, préparer, progressions, projet, projet, projets, punir, punition, sciences, séance, séances, situation, sport, stage, stages, texte, géo, gérer, maître, maîtresse*
IUFM (dimension réflexive) *ecole, affectation, affectations, ais, atsem, ce2, circonscription, cm1, cm2, collègue, collègues, contact, cp, cycle, directeur, directrice, écoles, enseignant, enseignants, équipe, famille, gs, inspection, instit, instits, mater , maternelle, métier, mouvement, ms, parents, placement, poste, postes, rentrée, scolaire, section, titulaire, voeux*
Collectif (vie du groupe) *aime, ambiance, amis, anniversaire, bises, bisou, bisous, bonjour, bonsoir, bouffe, bravo, chers, coucou, courage, e1, e2, e3, embrasse, félicitations, fête, fêtes, groupe, manger, merci, moral, news, ouf, pique-nique, profitez, raconter, remerciements, repas, resto, salut, souhaite, soutien, stagiaires, stress, sympa, vacances, vive*
Personnel (vie du groupe) *aider, bonheur, choix, contente, difficile, difficiles, difficultés, dormir, dur, dure, écrire, envie, fatiguée, force, groupes, help, impression, malade, manque, mort, penser, peur, pire, plaisir, points, pression, problème, problèmes, santé, service, soleil*

C.3.2 Expérimentation sur l'usage d'une terminologie professionnelle

Document *support, contenu, structure , hypertexte, paratexte, intertexte, auteur, métadonnées, lisibilité, pertinence, validité, fiabilité, prix, cote , indice, ISBN, ISSN, identification, indexation, authentification, validation, accès, recherche , bruit, silence, rappel, précision, classement, rangement, production, modification, description, condensation, reproduction, mode de diffusion*
Droit *propriété intellectuelle, copyright, dépôt légal, oeuvre de l'esprit, droit de reproduction, droit de prêt, droit*
Documentaliste *gestionnaire, administrateur, formateur, médiateur, fournisseur de prestations , fournisseur de produits d'information, animateur, concepteur d'un système d'information, concepteur , organisateur, politique documentaire, coordonnateur, accueil , prêt, service, déontologie, veille, IFLA, ADBS, FADBEN, projet, évaluation, réajustements, négociation, concertation, gestion de moyens , gestion du budget, bilan, collecte , gestion, diffusion, communication, promotion, catalogage partagé, travail en réseau, produits documentaires, revue de presse, dossiers documentaires, travail d'équipe, dialogue, partenariat, collaboration*
Lecteur / usager *compétences, connaissances, besoin d'information, besoin, attente, individu, groupe, lecture, formation , apprentissage, production, recherche, sélection, questionnement, usage, autonomie*
Bibliothéconomie *circuit du document, chaîne du document, traitement documentaire, techniques documentaires, fonds documentaires, catalogue, fichier, collection, langage documentaire, langage libre, langage contrôlé, thésaurus , liste d'autorité, cotation, norme, référence, classe, classification, description bibliographique, acquisition, collecte, commande, dépouillement, bulletinage , publication en série, récolement, enregistrement, saisie, désherbage*
Traitement documentaire *unité documentaire, document hôte, notice bibliographique, Cote, résumé, description bibliographique, mention de responsabilité, collation, descripteur générique, descripteur spécifique, descripteur associé, microthésaurus, champ sémantique , analyse documentaire, description bibliographique, condensation, catalogage, indexation, catalogage partagé*
Espaces documentaires *zone, accès libre, accès contrôlé, accès indirect, affichage, classement, classification, signalétique, espace physique, espace virtuel, aménagement, consultation, orientation*
Fonds documentaire *exhaustivité, collection , archive, politique d'acquisition, inventaire, ressources électroniques, ouvrages de référence, livres et monographies, périodiques et revues, bibliographies, textes officiels, guides et atlas, brochure, produits documentaires, censure, organisation, mise en valeur, projet intellectuel, stockage, flux*
Politique documentaire *système d'information documentaire, ingénierie documentaire, cahier des charges, tableau de bord, indicateur, rapport d'activité, charte , acquisition, accès, management, marketing documentaire, service, usager, collectivité, étude de besoins, analyse de situation, traitement des données, conduite de projet, programmation, planification, formalisation, mise en oeuvre, évaluation et contrôle, analyse de la valeur, démarche-qualité*
Recherche *recherche documentaire, recherche bibliographique, recherche d'information, requête, mot-clé, descripteur, équation de recherche, opérateurs booléens, logiciel documentaire, plein texte, texte intégral, banque de données, bibliographie, bruit, silence, rappel, précision, autopostage, troncature, langage pré coordonné, langage post coordonné, langage naturel, résultat, consultation, démarche, stratégie, zapping, étapes, compétences, critères de recherche, recherche en ligne, navigation, recherche multicritères*
Information *économie de, droit de, science de, information scientifique et technique, information brute , information spécialisée, information stratégique, surinformation , sous information*
Technologies de l'information *internet, réseau de l'établissement, intranet, bibliothèques électroniques, banques de données bibliographiques, forum, mail, chat, liste de diffusion, liste de discussion, outils de recherche, indexation automatique, navigateur, portail, hypermédia, hypertexte, site, URL, page d'accueil, métadonnées, consultation, recherche, diffusion, publication, communication*

Annexe D

Suivi du regard : les différentes diapositives proposées aux sujets

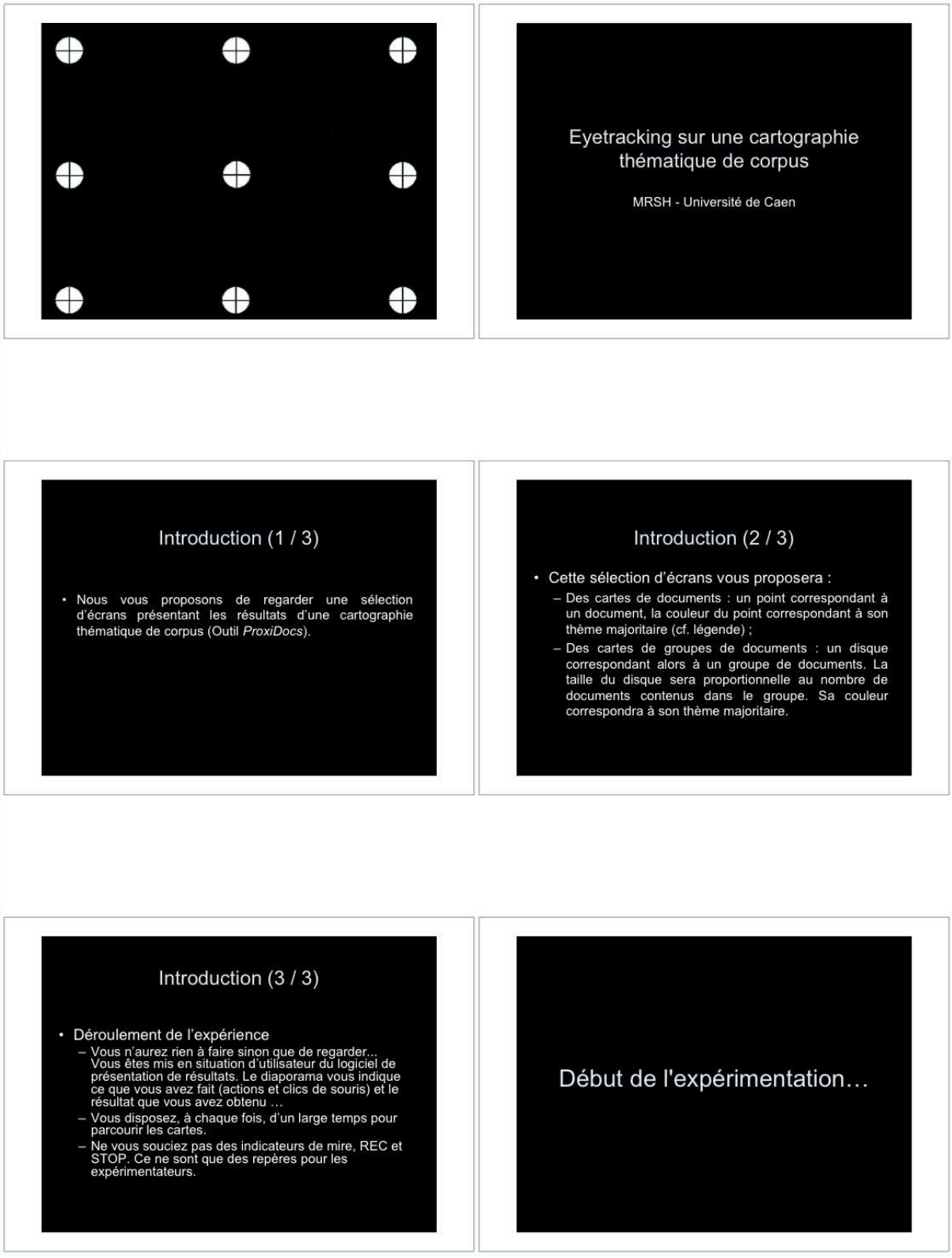


Fig. D.1 – Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 1 à 6.

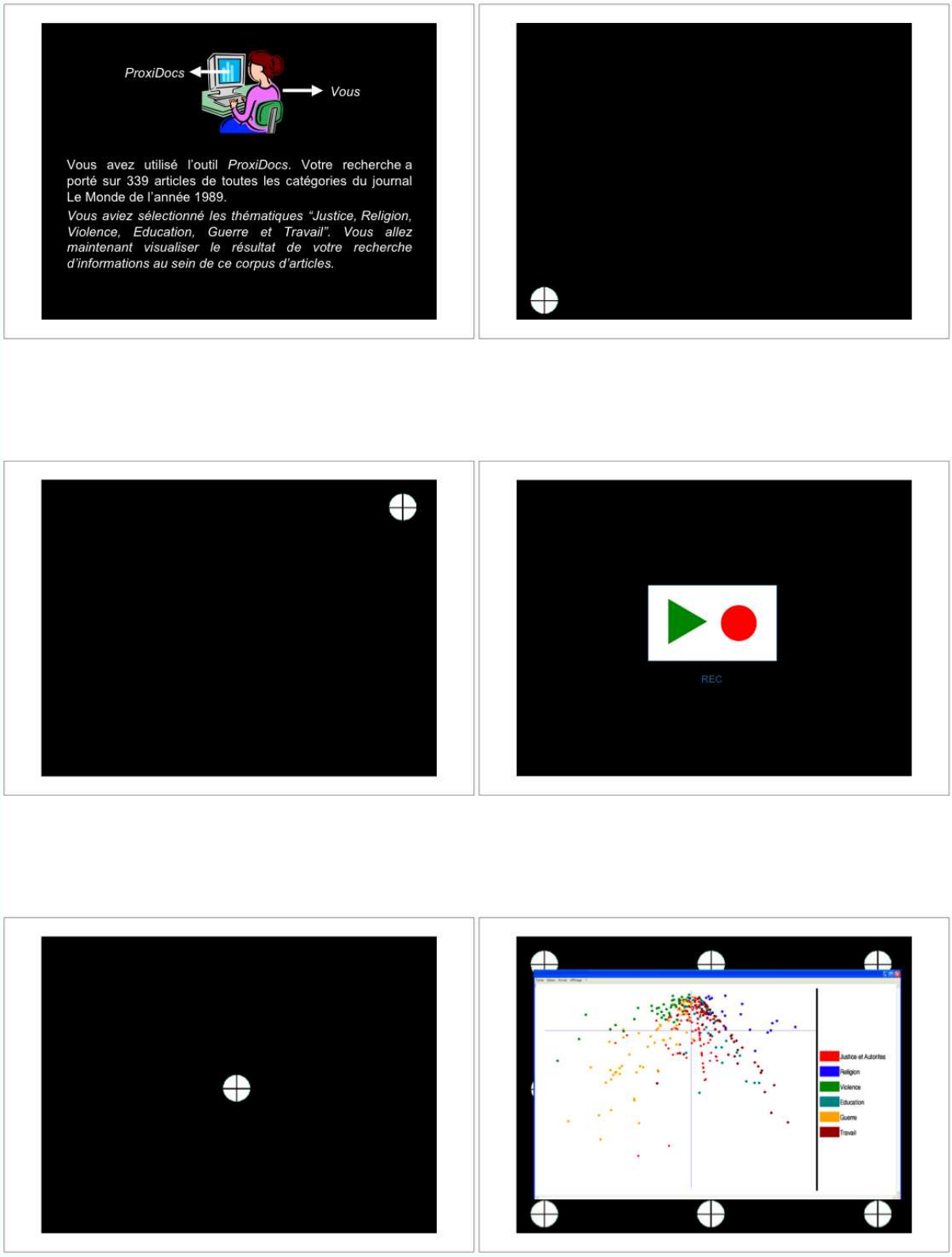


FIG. D.2 – Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 7 à 12.

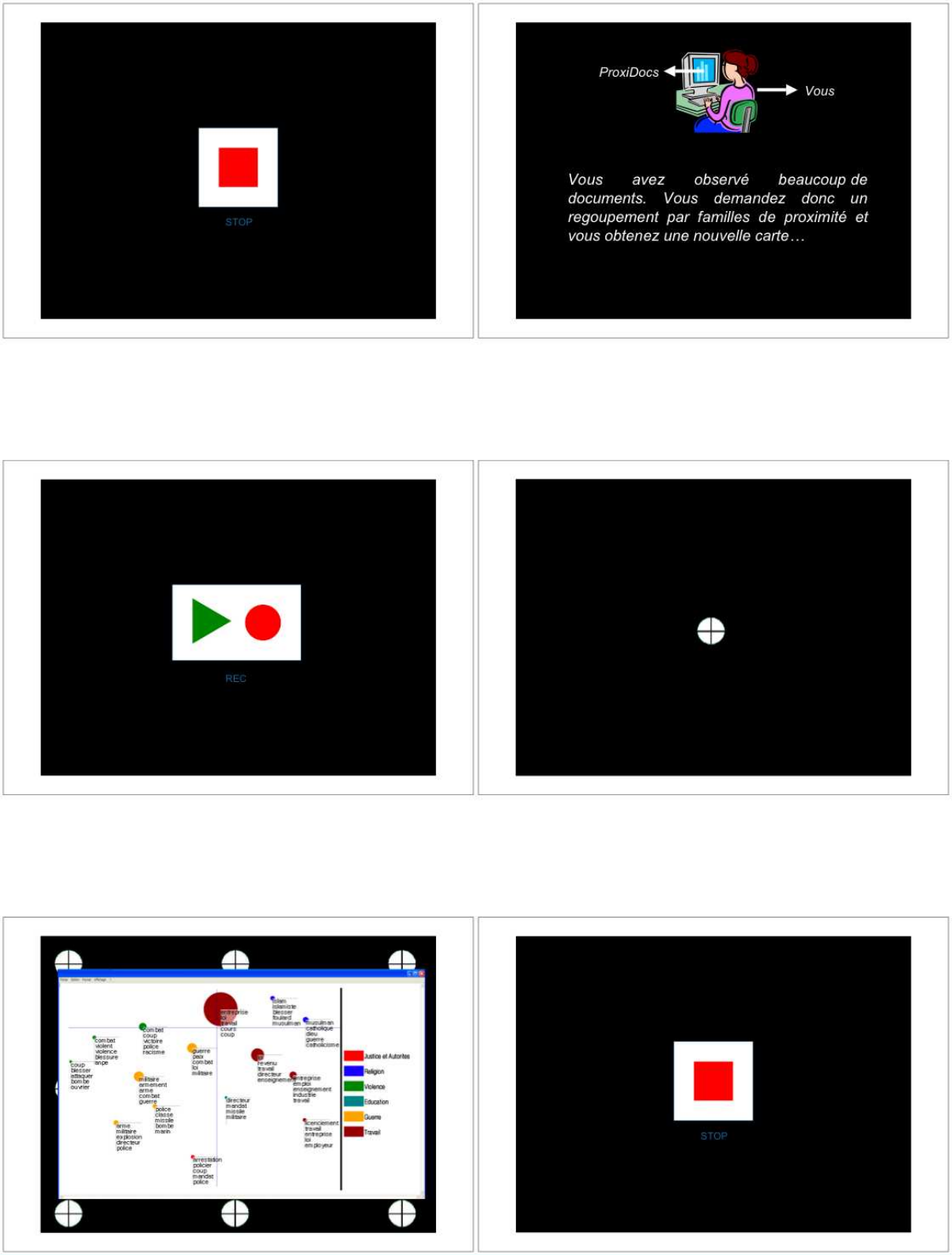


FIG. D.3 – Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 13 à 18.

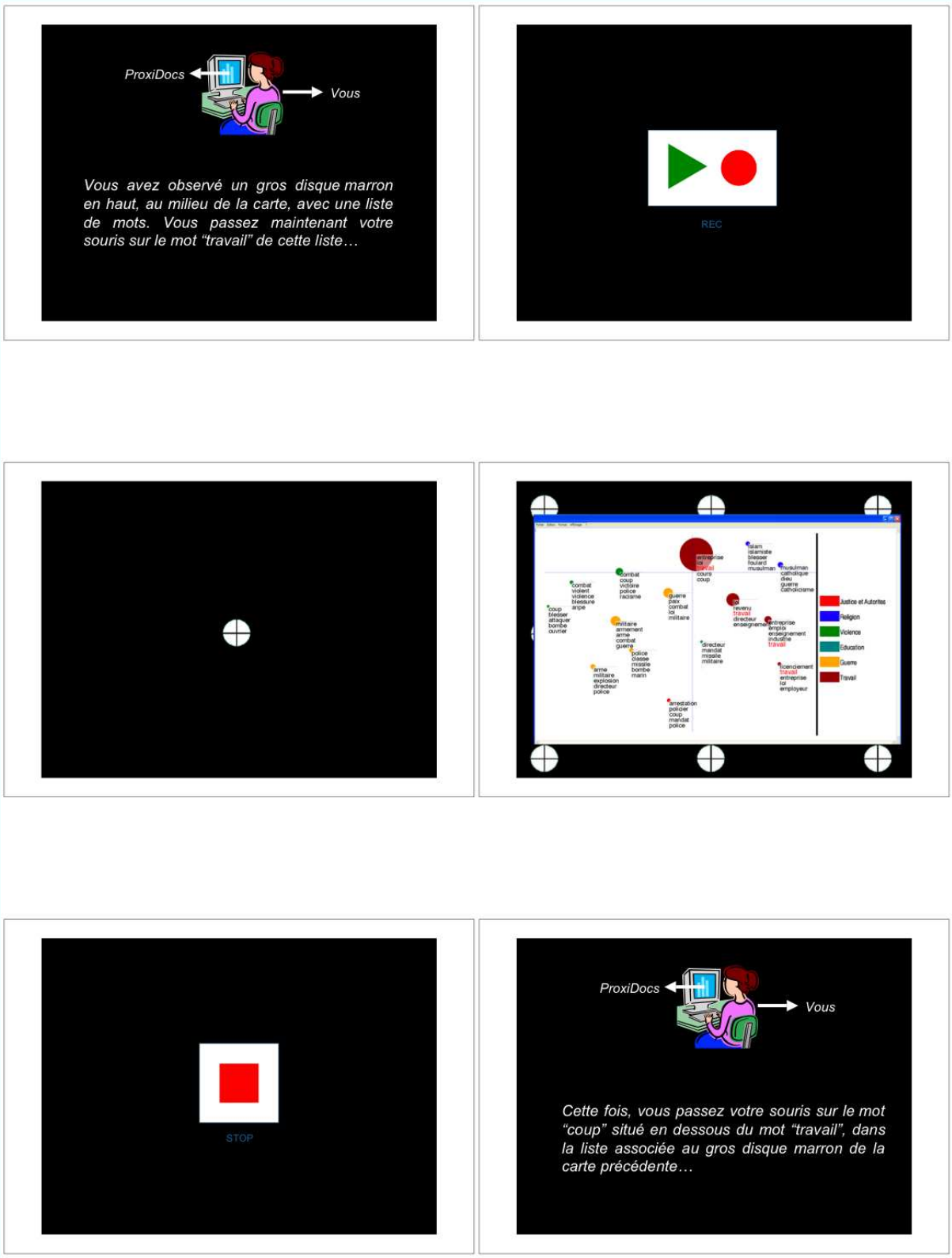


FIG. D.4 – Diaporama pour l’étude du suivi du regard sur les cartes : diapositives 19 à 24.

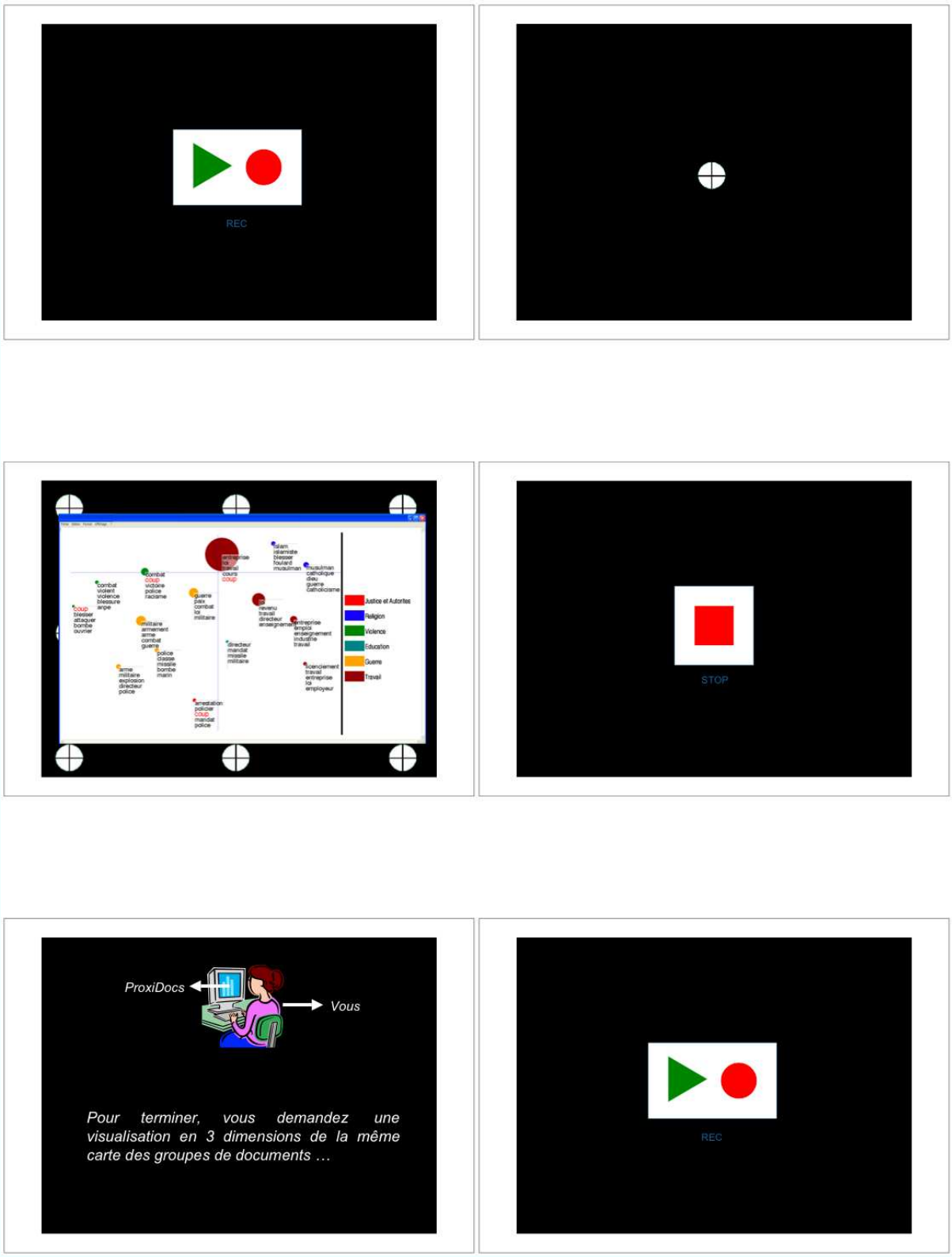


FIG. D.5 – Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 25 à 30.

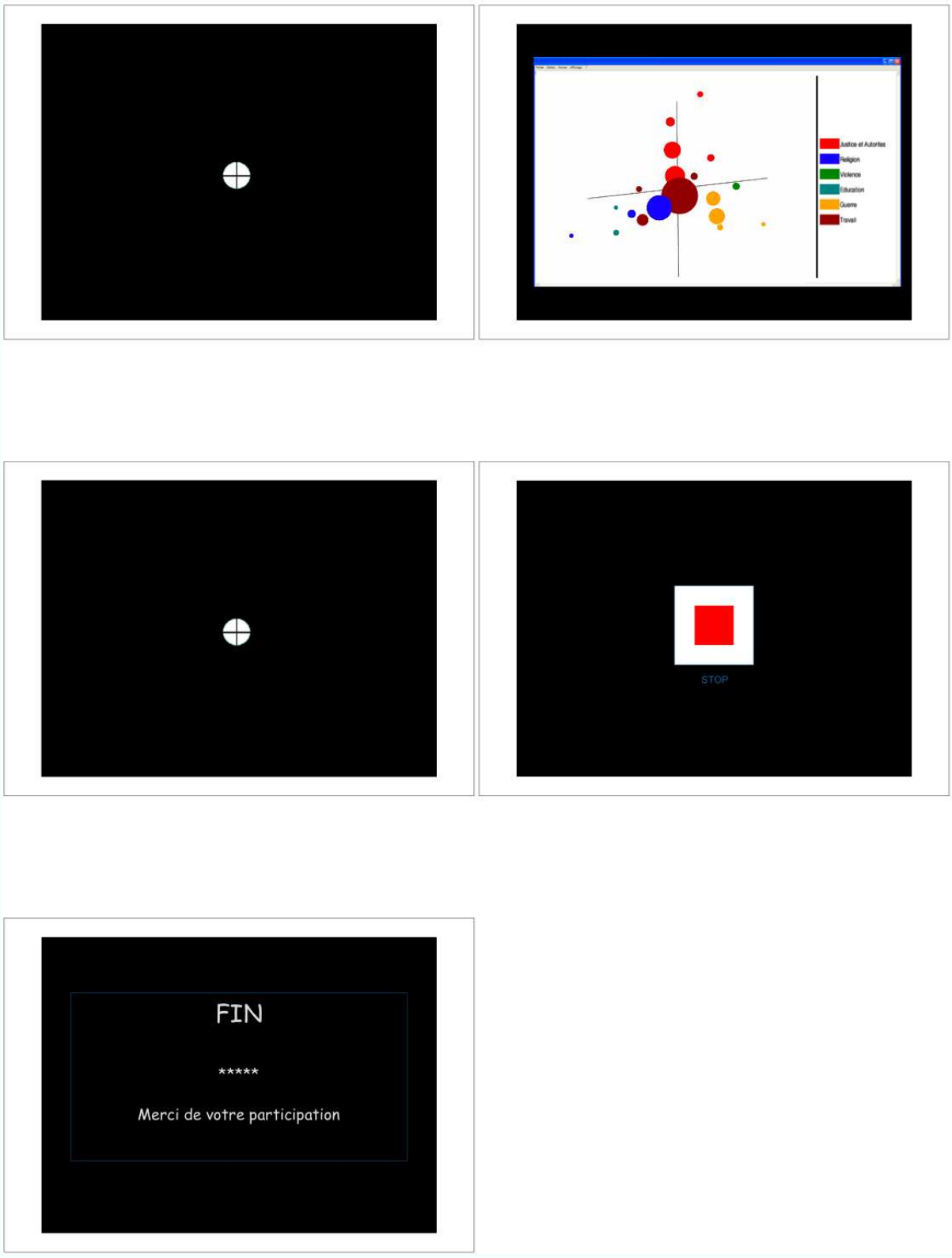


FIG. D.6 – Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 31 à 35.

Bibliographie

[Achour *et al.*, 2007] Achour, M., Betz, F., Dovgal, A., Lopes, N., Olson, P., Richter, G., Seguy, D., Vrana, J. et several_others (2007). *PHP Manual*. `http://php.net/manual/en/` (page consultée le 8 février 2007).

[Alonge et Castelli, 2003] Alonge, A. et Castelli, M. (2003). Encoding information on metaphoric expressions in Word-Net-like resources. *In Proceedings of the ACL Workshop on the Lexicon and Figurative Language*, pages 10–17.

[Alpac, 1966] Alpac (1966). A report by the automatic language processing advisory committee (ALPAC). *Language and Machines. Computers in Translation and Linguistics.*

[Alphonse *et al.*, 2004] Alphonse, E., Aubin, S., Bessiéres, P., Bisson, G., Hamon, T., Laguarigue, S., Nazarenko, A., Manine, A.-P., Nédellec, C., Vetah, M. O. A., Poibeau, T. et Weissenbacher, D. (2004). Event-based information extraction for the biomedical domain : the Caderige project. *In Proceedings of the International Workshop on Natural language Processing in Biomedicine and its Applications (JNLPBA)*, pages 43–49.

[Andries, 2002] Andries, P. (2002). Introduction à Unicode et à l'ISO 10646. *Revue Document numérique, Unicode, écriture du monde ?*, 6(3-4):51–88.

[Asher, 1993] Asher, N. (1993). *Reference to Abstract Objects in Discourse*. Kluwer, Dordrecht.

[Aubin *et al.*, 2006] Aubin, S., Derivière, J., Hamon, T., Nazarenko, A., Poibeau, T. et Weissenbacher, D. (2006). A robust linguistic infrastructure for efficient web content analysis : the ALVIS project. *In Proceedings of the Symposium on Digital Semantic Content across Cultures*, Paris.

[Auffret, 2000] Auffret, G. (2000). *Structuration de documents audiovisuels et publication électronique : constitution d'une chaîne éditoriale numérique pour la mise en ligne de collections audiovisuelles*. Thèse de Doctorat en Informatique, université de Technologie de Compiègne.

[Aussenac-Gilles *et al.*, 2007] Aussenac-Gilles, N., Condamines, A. et Sèdes, F. (2007). Evolution et maintenance des ressources termino-ontologiques : une question à approfondir. *Hors-série 2006 de la revue Information - Interaction - Intelligence : Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance*, pages 7–14.

[Authesserre et Courvalet, 2005] Authesserre, S. et Courvalet, J. (2005). Elaboration d'un métamoteur de recherche personnalisé à cartographie thématique. Rapport de projet de Master d'informatique première année à l'Université de Caen (universitaire 2004-2005) disponible à l'adresse suivante : `http://users.info.unicaen.fr/~troy/projet/2004-2005` (page consultée le 30 mai 2007).

[Ayache *et al.*, 2005] Ayache, C., Grau, B. et Vilnat, A. (2005). Campagne d'évaluation EQueR-EVALDA - évaluation en question-réponse. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, volume 2, pages 63–72, Dourdan, France. Hermès - Lavoisier.

[Ballabriga, 2005] Ballabriga, M. (2005). Sémantique textuelle 2. *Revue en ligne* Texto !, article disponible en ligne : `http://www.revue-texto.net/Reperes/Cours/Ballabriga2/index.html` (page consultée le 21 juin 2007).

[Balvet *et al.*, 2005] Balvet, A., Embarek, M. et Ferret, O. (2005). Minimaliste en question-réponse : le système Oedipe. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, volume 2, pages 77–80, Dourdan, France. Hermès - Lavoisier.

[Bar-Hillel, 1960] Bar-Hillel, Y. (1960). The present status of automatic translation of languages. *Advances in Computers*, 1:91–141.

[Bellies, 2002] Bellies, L. (2002). *La conception : processus d'élaboration et d'évaluation de représentations pour l'action.* Thèse de Doctorat en Ergonomie, École Pratique des Hautes Études, Paris.

[Bellot et El-Bèze, 2000] Bellot, P. et El-Bèze, M. (2000). Classification locale non supervisée pour la recherche documentaire. *Revue Traitement Automatique des Langues (T.A.L.)*, 41-2:335–366.

[Benzécri, 1980] Benzécri, J.-P. (1980). *L'analyse des données - tome 2 : l'analyse des correspondances.* Bordas.

[Berners-Lee, 1998] Berners-Lee, T. (1998). What the Semantic Web can represent ? *W3C*, article disponible en ligne : `http://www.w3.org/designissues/rdfnot.html` (page consultée le 29-06-2006).

[Bernhard, 1998] Bernhard, P. (1998). Apprendre à maîtriser l'information : des habiletés indispensables dans une société du savoir. *Les bibliothèques à l'ère électronique dans le monde de l'éducation*, article disponible en ligne : `http://www.w3.org/designissues/rdfnot.html` (page consultée le 29-06-2006), XXVI(1).

[Bertin, 1983] Bertin, J. (1983). *Semiology of Graphics.* University of Wisconsin Press, Madison, WI.

[Beust, 1998] Beust, P. (1998). *Contribution à un modèle interactionniste du sens - Amorce d'une compétence interprétative.* Thèse de Doctorat en Informatique de l'Université de Caen, Caen.

[Beust, 2002] 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, 2005] Beust, P. (2005). La réflexivité dans l'évaluation en traitement automatique des langues. *In Actes des 12e journées de Rochebrune "Réflexivité et auto-référence dans les systèmes complexes"*, pages 11–23.

[Beust et Roy, 2006a] Beust, P. et Roy, T. (2006a). Prendre en compte la dimension globale d'un corpus dans la contextualisation du sens : expérimentations en informatique linguistique. *revue Glottopol - Traitements automatisés des corpus spécialisés : contextes et sens*, 8:53–72.

[Beust et Roy, 2006b] Beust, P. et Roy, T. (2006b). Utiliser des traces de la dimension globale d'un corpus pour l'accès au contenu des documents. *In Actes des Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels : Traces, Enigmes, Problèmes : Emergence et Construction du Sens (Rochebrune 2006)*, pages 47–60, Megève (France).

[Bihanic, 2003] Bihanic, D. (2003). Les hypermédias graphiques explorateurs. *In Actes des Journées Francophones de la Toile (JFT'03)*, pages 237–246.

[Bilhaut, 2006] Bilhaut, F. (2006). *Analyse automatique de structures thématiques discursives - Application à la recherche d'information.* Thèse de Doctorat en Informatique de l'Université de Caen, Caen.

[Bilhaut et Widlöcher, 2006] Bilhaut, F. et Widlöcher, A. (2006). Linguastream : An integrated environment for computational linguistics experimentation. *Proceedings of the 11th Conference of the European Chapter of the Association of Computational Linguistics (Companion Volume).*

[Biébow et Szulman, 2000] Biébow, B. et Szulman, S. (2000). Terminae : une approche terminologique pour la construction d'ontologies du domaine à partir de textes. *Actes de Reconnaissance de formes et Intelligence Artificielle*, 2:81–90.

[Blaudez *et al.*, 2005] Blaudez, E., Crestan, E. et de Loupy, C. (2005). SQuAr : Prototype de moteur de questions réponse. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, volume 2, pages 73–76, Dourdan, France. Hermès - Lavoisier.

[Bock, 1996] Bock, H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. *In* Arabie, P., Hubert, L. et Soete, G. D., éditeurs : *Clustering and Classification*, pages 378–453. World Scientific.

[Bourigault, 1994] Bourigault, D. (1994). *LEXTER un Logiciel d'EXtraction de TERminologie. Application à l'extraction des connaissances à partir de textes.* Thèse de Doctorat en Mathématiques, Informatique appliquée aux Sciences de l'Homme, École des Hautes Études en Sciences Sociales de Paris, Paris.

[Bourigault et Aussenac-Gilles, 2003] Bourigault, D. et Aussenac-Gilles, N. (2003). Construction d'ontologies à partir de textes. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'03)*, volume 2, pages 27–47.

[Bourigault et Slodzian, 1999] Bourigault, D. et Slodzian, M. (1999). Pour une terminologie textuelle. *Terminologies Nouvelles*, 19:29–31.

[Bouroche et Saporta, 1980] Bouroche, J.-M. et Saporta, G. (1980). *L'analyse des données.* Presse Universitaires de France, Paris.

[Breux *et al.*, 2007] Breux, S., Caillaud, B., Giacalone, J. et Roussel, H. (2007). Comportement visuel et complexité, exploration de structures aléatoires à 1 et 2 dimensions. *Art et Complexité* article disponible en ligne sur le site de la Maison de la Recherche en Sciences Humaines de l'Université de Caen / Basse-Normandie : `http://www.unicaen.fr/mrsh/publications/online.php` (page consultée le 11 juillet 2007).

[Bruillard, 2007] Bruillard, E. (2007). Le forum de discussion : un cas d'école pour les recherches en EIAH. *Revue STICEF, Numéro spécial Forum de discussion en éducation*, en ligne : `http://sticef.univ-lemans.fr/classement/encours.htm`, 13.

[Brusilovsky, 1996] Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3):87–129.

[Buckland, 1997] Buckland, M. K. (1997). What is a "document" ? *Journal of the American Society of Information Science*, 48(9):804–809.

[Buckland, 1998] Buckland, M. K. (1998). What is a "digital document" ? *Document Numérique*, 2(2):221–230.

[Buisine et Martin, 2006] Buisine, S. et Martin, J.-C. (2006). L'étude de corpus par ACP. *In Actes du 2ème Workshop sur les Agents Conversationnels Animés (WACA'06).*

[Card *et al.*, 1999] Card, S. K., Mackinlay, J. D. et Shneiderman, B. (1999). *Readings in Information Visualization : Using Vision to Think*. Morgan Kaufmann Publishers, San Francisco, États-Unis.

[Caron *et al.*, 2003] Caron, Y., Vincent, N. et Makris, P. (2003). Détection de régions d'intérêt par l'utilisation de lois de puissance. *Actes de Compression et Représentation de Signaux Audiovisuels (CORESA'03)*, pages 239–242.

[Charlet *et al.*, 2003] Charlet, J., Laublet, P. et Reynaud, C. (2003). *Web Sémantique. Rapport de l'Action Spécifique 32 CNRS / STIC. V3*. `http://rtp-doc.enssib.fr/IMG/pdf/ASWebSemantique2003.pdf` (consultée le 29-06-2006).

[Charolles, 1988] Charolles, M. (1988). Les plans d'organisation textuelle : périodes, chaîne, portées et séquences. *Pratiques*, 57:3–12.

[Chaudiron, 2004] Chaudiron, S. (2004). *Évaluation des systèmes de traitement de l'information*. Hermès, Paris.

[Chibout *et al.*, 2001] Chibout, K., Vilnat, A. et Briffault, X. (2001). Sémantique du lexique verbal : un modèle en arborescence avec les graphes conceptuels. *Revue TAL - Lexiques sémantiques*, 42-3:691–727.

[Chomsky, 1969] Chomsky, N. (1969). *Structures syntaxiques*. Seuil, Paris.

[Chung *et al.*, 2002] Chung, W., Chen, H. et Numaker, J. (2002). Business intelligence explorer : A knowledge map framework for discovering business intelligence on the web. *Proceedings of the 36th Hawaii International Conference on System Sciences*.

[Châar *et al.*, 2004] Châar, S. L., Ferret, O. et Fluhr, C. (2004). Filtrage pour la construction de résumés multidocuments guidés par un profil. *Revue Traitement Automatique des Langues (T.A.L.) - Résumé automatique de textes*, 45-1:65–93.

[Ciravegna, 2003] Ciravegna, F. (2003). Designing adaptive information extraction for the Semantic Web in Amilcare. *Annotation for the Semantic Web, Frontiers in Artificial Intelligence and Applications*.

[Claveau et Sébillot, 2004] Claveau, V. et Sébillot, P. (2004). Apprentissage semi-supervisé de patrons d'extraction de couples nom-verbes. *Revue Traitement Automatique des Langues (T.A.L.)*, 45(1):153–182.

[Cleveland, 1993] Cleveland, W. (1993). *Visualizing Data*. Hobart Press, Summit, NJ.

[Clouet, 2005] Clouet, N. (2005). L'étude de cas, un outil pour la professionnalisation des enseignants-documentalistes. `http://labo.eda.free.fr/symposium/resumes/clouet_resum.html` (page consultée le 6 juillet 2007).

[Collot, 1988] Collot, M. (1988). Le thème selon la critique thématique. *Revue Communications*.

[Constantinovici, 2006] Constantinovici, S. (2006). L'œuvre poétique de Tudor Arghezi. La diversité du lexique et le problème de style. *In* Rastier, F. et Ballabriga, M., éditeurs : *Actes du colloque international d'Albi 2006 - Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation*, pages 273–276, Paris : Texto. Carine Duteil and Baptiste Foulquié.

[Coursil, 2000] Coursil, J. (2000). *La fonction muette du langage*. Ibis Rouge Editions, Petit-Bourg (Guadeloupe).

[Crochemore *et al.*, 2005] Crochemore, M., Dias, G. H. et de Sousa, S. M. (2005). Passage à l'échelle - complexité, algorithmique et architecture. *Revue Traitement Automatique des Langues (T.A.L.)*, 46(2).

[Cubaud et Bénel, 2006] Cubaud, P. et Bénel, A. (2006). Au-delà du web : les interfaces de visualisation et d'annotation pour les bibliothèques numériques. *In* Pédauque, R. T., éditeur : *La redocumentarisation du monde*. Cépaduès, Toulouse.

[Darmoni *et al.*, 2005] Darmoni, S., Névéol, A., Renard, J., Gehanno, J., Soualmia, L., Dahamna, B. et Thirion, B. (2005). A MEDLINE categorization algorithm. *BMC Medical Informatics and Decision Making*, 6(7).

[de Calmès et Pérennou, 1998] de Calmès, M. et Pérennou, G. (1998). BDLEX : a lexicon for spoken and written french. *Proceedings of the 1st International Conference on Langage Resources & Evaluation (LREC1998), Grenade*, pages 1129–1136.

[de Saussure, 1972] de Saussure, F. (1972). *Cours de linguistique générale*. Payot, Paris.

[Delépine, 2003] Delépine, L. (2003). *L'assistance à la navigation hyperdocumentaire. Réalisation d'un outil aide à la recherche de documents visités par un utilisateur dans le contexte du Web. Approche sémio-technologique.* Thèse de Doctorat en Informatique, Université de Caen / Basse-Normandie, Caen.

[Despeyroux *et al.*, 2005] Despeyroux, T., Lechevallier, Y., Trousse, B. et Vercoustre, A.-M. (2005). Expériences de classification d'une collection de documents XML de structure homogène. *In* et Suzanne Pinson, N. V., éditeur : *Actes des 5èmes Journées d' Extraction et de Gestion des Connaissances (EGC 2005)*. Cépaduès-Editions.

[D'Hondt et Khayati, 2005] D'Hondt, F. et Khayati, B. E. (2005). étude de méthodes de clustering pour la segmentation d'images en couleurs. *Certificat Applicatifs Multimédia - Projet Traitement de l'Information*.

[Doddington, 2002] Doddington, G. (2002). Automatic evaluation of MT quality using n-gram co-occurrence statistics. *Human Language Technology*, pages 128–132.

[Ducrot et Schaeffer, 1995] Ducrot, O. et Schaeffer, J.-M. (1995). *Nouveau dictionnaire encyclopédique des sciences du langage*. éditions du Seuil.

[Durand, 2004] Durand, N. (2004). *Extraction de clusters à partir du treillis de concepts : Application à la découverte de communautés d'intérêt pour améliorer l'accès à l'information.* Thèse de Doctorat en Informatique, Université de Caen / Basse-Normandie, Caen.

[Durand et Crémilleux, 2002] Durand, N. et Crémilleux, B. (2002). ECCLAT : a new approach of clusters discovery. *In* Bramer, M., Preece, A. et Coenen, F., éditeurs : *Categorical Data - Proceedings of the 22nd SGAI International Conference on Knowledge Based Systems and Applied Artificial Intelligence (ES 2002)*, pages 177–190, Cambridge, UK. BCS Conference Series, Springer-Verlag.

[Dybowski *et al.*, 1996] Dybowski, R., Collins, T. D. et Weller, P. R. (1996). Visualization of binary string convergence by Sammon mapping. *In* Fogel, L. J., Angeline, P. J. et Bäck, T., éditeurs : *Proceedings of the Fifth Annual Conference on Evolutionary Programming (EP'96)*, pages 377–383, Cambridge, MA. MIT Press.

[Déjean, 2005] Déjean, S. (2005). Formation à l'analyse statistique de données d'expression. *In Actes des Journées RNG Transcriptome et Bioinformatique*.

[Ebert *et al.*, 1996] Ebert, D. S., Shaw, C. D., Zwa, A. et Starr, C. (1996). Two-handed interactive stereoscopic visualization. *In Proceedings of the conference on Visualization*, pages 205–210. IEEE Computer Society Press.

[ECMA-International, 1999] ECMA-International (1999). *Standard ECMA-262 - ECMAScript Language Specification.* `http://www.ecma-international.org/publications/files/ecma-st/ECMA-262.pdf` (page consultée le 25 janvier 2007).

[Eco, 1980] Eco, U. (1980). *Le Nom de la Rose (*Il nome della rosa*)*. LGF.

[Ellis, 1938] Ellis, W. D. (1938). *A Source Book of Gestalt Psychology*. The Gestalt Journal Press (réédition de 1997), New-York.

[Enjalbert et Victorri, 1994] Enjalbert, P. et Victorri, B. (1994). Du langage au modèle. *Revue Traitement Automatique des Langues (T.A.L.)*, 35:37–64.

[Farzindar et Lapalme, 2005] Farzindar, A. et Lapalme, G. (2005). Production automatique du résumé de textes juridiques : évaluation de qualité et d'acceptabilité. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, volume 1, pages 183–192, Dourdan, France. Hermès - Lavoisier.

[Farzindar *et al.*, 2004] Farzindar, A., Lapalme, G. et Desclés, J.-P. (2004). Résumé de textes juridiques par identification de leur structure thématique. *Revue Traitement Automatique des Langues (T.A.L.) - Résumé automatique de textes*, 45-1:39–64.

[Fayet-Scribe, 1997] Fayet-Scribe, S. (1997). Chronologie des supports, des dispositifs spatiaux, des outils de repérage de l'information. *Solaris*, 4:article en ligne : `http://biblio--fr.info.unicaen.fr/bnum/jelec/Solaris/d04/4fayet_0intro.html`, page consultée le 29 mars 2007.

[Ferone, 2006] Ferone, G. (2006). Liste de discussion et identité professionnelle des enseignants en formation. *In Actes des premières Journées Communication et Apprentissage Instrumentés en Réseaux (JOCAIR'06)*.

[Ferrari, 2006] Ferrari, S. (2006). Rhétorique et compréhension. *In* Sabah, G., éditeur : *Compréhension des langues et interaction*, chapitre 7, pages 195–224. Lavoisier, Paris.

[Fillol, 1999] Fillol, V. (1999). *Vers une sémiotique de l'énonciation. Du Lieu Commun comme stratégie et des formes et/ou formations discursives comme Lieux Communs de l'énonciation (dans la presse féminine)*. Thèse de Doctorat en Sciences du Language, Université de Toulouse - Le Mirail, Toulouse.

[Foenix-Riou, 2005] Foenix-Riou, B. (2005). *Guide de recherche sur Internet*. Armand Colin.

[Fortier et Kassel, 2004] Fortier, J.-Y. et Kassel, G. (2004). Présentation "sur mesure" d'informations : Une approche appliquée aux mémoires organisationnelles. *Revue d'intelligence artificielle*, 18(4):515–547.

[Foskett et Maniez, 1995] Foskett, D. et Maniez, J. (1995). Indexation. *Encyclopedia Universalis*.

[Fox et Kipp, 1997] Fox, E. et Kipp, N. (1997). Networked digital library of theses and dissertations : An international effort unlocking university resources. *D-lib magazine*.

[Fraley et Raftery, 1998] Fraley, C. et Raftery, A. E. (1998). How many clusters ? which clustering method answers via model-based cluster analysis. *The Computer Journal*, 41(8):578–588.

[Frassincar et Houben, 2002] Frassincar, F. et Houben, G.-J. (2002). Hypermédia presentation adaptation on the Semantic Web. *In Adaptive Hypermedia and Adaptive Web-Based Systems*, pages 133–142. Springer Verlag, Lecture Notes in Computer Science, Malaga.

[Garlatti et Prié, 2003] Garlatti, S. et Prié, Y. (2003). Adaptation et personnalisation dans le Web Sémantique. *In* Charlet, J., Laublet, P. et Reynaud, C., éditeurs : *Rapport final de l'Action spécifique 32 CNRS / STIC : Web sémantique*, pages 71–91. Rapport disponible à l'adresse suivante : `http://rtp-doc.enssib.fr/IMG/pdf/ASWebSemantique2003.pdf` (page consultée le 17 mai 2007).

[Gentner, 1988] Gentner, D. (1988). Analogical inference and analogical access. *In* Prieditis, A., éditeur : *Analogica*, chapitre 3, pages 63–88. Pitman Publishing, Morgan Kaufmann Publishers, London.

[Giguet, 1998] Giguet, E. (1998). *Méthode pour l'analyse automatique de structures formelles sur documents multiligues*. Thèse de Doctorat en Informatique de l'Université de Caen, Caen.

[Giguet et Lucas, 2002] Giguet, E. et Lucas, N. (2002). Intégration d'Unicode - conception d'un agent de recherche d'information sur Internet. *Revue Document numérique, Unicode, écriture du monde ?*, 6(3-4):225–236.

[Goldberg et Kotval, 1999] Goldberg, H. J. et Kotval, X. (1999). Computer interface evaluation using eye movements : methods and constructs. *International Journal of Industrial Ergonomics*, 24:631–645.

[Gosling *et al.*, 2005] Gosling, J., Joy, B., Steele, G. et Bracha, G. (2005). *The Java Language Specification - Third Edition*. `http://java.sun.com/docs/books/jls/` (page consultée le 8 février 2007).

[Grau *et al.*, 2005] Grau, B., Illouz, G., Monceaux, L., Paroubek, P., Pons, O. et Robba, I. (2005). FRASQUES, le système du groupe LIR, LIMSI. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'05)*, volume 2, pages 85–88, Dourdan, France. Hermès - Lavoisier.

[Grau et Magnini, 2005] Grau, B. et Magnini, B. (2005). *Revue TAL - Réponses à des questions*, volume 46-3. Hermès, Lavoisier, Paris.

[Greimas, 1966] Greimas, A.-J. (1966). *Sémantique structurale*. Larousse.

[Gruber et Vemuri, 1996] Gruber, T. et Vemuri, S. (1996). *Model-based Virtual Document Generation*. Knowledge Systems Laboratory, KSL-96-16.

[Guazzini *et al.*, 2004] Guazzini, M. U. E., Bertagna, F. et Calzolari, N. (2004). Senseval-3 : The Italian All-words Task. *In Proceedings of SENSEVAL-3 : Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Association for Computational Linguistics.

[Habert, 2004] Habert, B. (2004). Des corpus, pourquoi faire ? Intervention à la journée d'étude Méthodes en sciences humaines organisée par Catherine Schnedecker, `http://www.limsi.fr/Individu/habert/Paroles/BHabertStrasbourg04.pdf` (page consultée le 12 février 2007).

[Habert, 2005] Habert, B. (2005). Portrait de linguiste(s) à l'instrument. *Revue* Texto !, `http://www.revue-texto.net/Corpus/Publications/Habert/Habert_Portrait.html` (page consultée le 23 janvier 2007).

[Hamadache et Vernier, 2006] Hamadache, K. et Vernier, M. (2006). Réalisation d'une interface d'aide à l'acquisition de lexiques sémantiques : l'outil *VisualLuciaBuilder*. Rapport de projet de Master d'informatique première année à l'Université de Caen (universitaire 2005-2006) disponible à l'adresse suivante : `http://users.info.unicaen.fr/~troy/projet/2005-2006/rapportVLB.zip` (page consultée le 7 juin 2007).

[Harris, 1971] Harris, Z. (1971). *Structures mathématiques du langage*. Dunod.

[Hearst, 1995] Hearst, M. (1995). Tilebars : Visualization of term distribution information in full text information access. *In Proceedings of the ACM's Special Interest Group on Computer-Human Interaction Conference*, pages 59–66. ACM Press.

[Hearst, 1999] Hearst, M. (1999). User interfaces and visualization. *Modern Information Retrieval*.

[Henri et Lundgren-Cayrol, 2001] Henri, F. et Lundgren-Cayrol, K. (2001). *Apprentissage collaboratif à distance. Pour comprendre et concevoir les environnements d'apprentissage virtuels.* Presses Universitaires du Québec, Sainte-Foy, Québec.

[Hill, 1970] Hill, B. (1970). Zipf's law and prior distributions for the composition of a population. *Journal of American Statistical Association*, 65:1220–1232.

[Hjelmslev, 1971] Hjelmslev, L. (1971). *Prolégomènes à une théorie du langage (traduction française 1968).* Editions de Minuit, Paris.

[Holten, 2006] Holten, D. (2006). Hierarchical edge bundles : Visualization of adjacency relations in hierarchical data. *IEEE Transactions on Visualization and Computer Graphics*, 5(12).

[Houben et Rioult, 2006] Houben, F. et Rioult, F. (2006). Étiquetage morpho-syntaxique par classification supervisée : vers une alternative aux dictionnaires ? *In Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, volume 1, pages 507–515.

[Huynh-Kim-Bang et Bruillard, 2005] Huynh-Kim-Bang, B. et Bruillard, E. (2005). Vers une nouvelle interface de lecture pour des forums de discussion dédiés à des élaborations collectives. *In* Saleh, I. et Clément, J., éditeurs : *Créer, jouer, échanger, actes de H2PTM'05*, pages 43–56, Paris. Lavoisier.

[Illouz *et al.*, 1999] Illouz, G., Habert, B., Fleury, S., Folch, H., Heiden, S. et Lafon, P. (1999). Maîtriser les déluges des données hétérogènes. *In* Condamines, A., Fabre, C. et Péry-Woodley, M., éditeurs : *Corpus et traitement automatique des langues : pour une réflexion méthodologique*, pages 37–46.

[Illouz et Jardino, 2001] Illouz, G. et Jardino, M. (2001). Analyse statistique et géométrique de corpus textuel. *In* Daille, B. et Romary, L., éditeurs : *Revue TAL - Linguistique de corpus*, volume 42-2, pages 501–516. Lavoisier, Paris.

[Jacquemin, 2000] Jacquemin, C. (2000). Traitement automatique des langues pour la recherche d'information. *Revue TAL*, 41-2.

[Jacquemin et Jardino, 2002] Jacquemin, C. et Jardino, M. (2002). Une interface 3D multiéchelle pour la visualisation et la navigation dans de grands documents XML. *In IHM '02 : Proceedings of the 14th French-speaking conference on Human-computer interaction (Conférence Francophone sur l'Interaction Homme-Machine)*, pages 263–266, New York, NY, USA. ACM Press.

[Jacquemin et Zweigenbaum, 2000] Jacquemin, C. et Zweigenbaum, P. (2000). Traitement automatique des langues pour l'accès au contenu des documents. *In* LeMaitre, J., Charlet, J. et Garbay, C., éditeurs : *Le document Multimédia pour l'accès au contenu des documents*, pages 71–109. Cépaduès.

[Kamp, 1981] Kamp, H. (1981). A theory of truth and semantics representation. *Formal Methods in the Study of Language.*

[Kennedy, 1998] Kennedy, G. (1998). *An introduction to corpus linguistics.* Longman.

[Kinsbourne, 1972] Kinsbourne, M. (1972). Eye and head turning indicates cerebral lateralization. *Science*, 176(4034):539–541.

[Lakoff et Johnson, 1980] Lakoff, G. et Johnson, M. (1980). *Metaphors we live by.* University of Chicago Press, Chicago.

[Lamping, 1995] Lamping, J. (1995). A focus+context technique based on hyperbolic geometry for viewing large hierarchies. *Proceedings of ACM's Special Interest Group on Computer-Human Interaction*, pages 401–408.

[Lavenus et Lapalme, 2002] Lavenus, K. et Lapalme, G. (2002). Évaluation des systèmes de question réponse. *Revue Traitement Automatique des Langues*, 43(3):181–208.

[Lebart *et al.*, 1998] Lebart, L., Salem, A. et Barry, L. (1998). *Exploring textual data*. Kluwer Academic.

[Lelu et Aubin, 2001] Lelu, A. et Aubin, S. (2001). Vers un environnement complet de synthèse statistique de contenus textuels. Présentation au séminaire Association pour la mesure des sciences et des techniques du 13 janvier 2001.

[Lin et Hovy, 2003] Lin, C.-Y. et Hovy, E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003)*, pages 150–157, Edmonton, Canada.

[Lucas, 2005] Lucas, N. (2005). étude linguistique des procédés d'exposition dans un forum de discussion. *In Actes en ligne du Symposium Formation et Nouveaux Instruments de Communications (Symfonic 2005)*, article disponible en ligne : `http://www.dep.u-picardie.fr/sidir/articles/Nadine_Lucas.htm` (page consultée le 6 juillet 2007), Amiens, France.

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

[Mackinlay et Robertson, 1993] Mackinlay, J. et Robertson, G. (1993). The document lens. *In Proceedings of the ACM User Interface and Software Technology Conference (UIST'93)*, pages 101–108.

[MacQueen, 1967] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 1:281–297.

[Malherbe et Moulec, 2007] Malherbe, F. et Moulec, J. L. (2007). Rapport à l'actualité d'une ressource terminologique sur Internet. Rapport de projet de Master d'informatique première année à l'Université de Caen (universitaire 2006-2007) disponible à l'adresse suivante : `http://users.info.unicaen.fr/projet/2006-2007/LeMoulec-Malherbe.pdf` (page consultée le 11 juin 2007).

[Mangenot, 2002] Mangenot, F. (2002). Ecriture collective par forum sur le Web : un nouveau genre d'écrit universitaire ? Présentation au séminaire Association pour la mesure des sciences et des techniques du 13 janvier 2001, article disponible en ligne : `http://archivesic.ccsd.cnrs.fr/docs/00/06/21/15/HTML/index.html` (page consultée le 7 juillet 2007).

[Martin, 1988] Martin, H.-J. (1988). *Histoire et pouvoir de l'écrit*. Perrin, Paris.

[McTear, 1993] McTear, M. F. (1993). User modelling for adaptive computer systems : a survey of recent developments. *Artificial Intelligence Review*, 7:157–184.

[Mezaille, 2005] Mezaille, T. (2005). Quels mécanismes pour (r)établir la cohésion sémantique textuelle ? sur la prééminence des processus d'assimilation et de dissimilation dans l'interprétation des énoncés contradictoires et métaphoriques. Site Internet de la revue *Texto !* : `http://www.revue-texto.net/Dialogues/Mezaille_Cohesion.html` (consultée le 12 juillet 2007).

[Minel, 2001] Minel, J.-L. (2001). *Filtrage sémantique (du résumé automatique à la fouille de textes)*. Hermès, Paris.

[Minel, 2004] Minel, J.-L. (2004). *Résumés automatiques de textes*, volume 45-1. Revue Traitement Automatique des Langues (T.A.L.).

[Mokrane *et al.*, 2004] Mokrane, A., Arezki, R., Dray, G. et Poncelet, P. (2004). Cartographie automatique du contenu d'un corpus de documents textuels. *In Proceedings of the 7th International Conference on the Statistical Analysis of Textual Data*, volume 2, pages 816–823.

[M.U.C., 1998] M.U.C. (1998). *Proceedings of the seventh Message Understanding Conference*. `http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html` (page consultée le 31 janvier 2007).

[Nazarenko, 2005] Nazarenko, A. (2005). Sur quelle sémantique reposent les méthodes automatiques d'accès au contenu textuel ? *In* Condamines, A., éditeur : *Sémantique et corpus*, chapitre 6, pages 211–244. Lavoisier, Paris.

[Nazarenko et Hamon, 2002] Nazarenko, A. et Hamon, T. (2002). *Revue Traitement Automatique des Langues (T.A.L.) - Structuration de terminologie*, volume 43-1. Hermès, Paris.

[Nicolle, 1996] Nicolle, A. (1996). L'expérimentation et l'intelligence artificielle. *Intellectica*, 22:9–19.

[Nicolle, 2002] Nicolle, A. (2002). Logiciels d'étude des phénomènes complexes - première partie : conception, modélisation. *Cours de Diplôme d'Étude Approfondies Intelligence Artificielle et Algorithmique de l'Université de Caen*, en ligne : `http://users.info.unicaen.fr/~anne/HTML/LogicielsEtude.pdf` (page consultée le 23 janvier 2007).

[Nicolle, 2005] Nicolle, A. (2005). Comparaison entre les comportements réflexifs du langage humain et la réflexivité des langages informatiques. *In Actes des 12e journées de Rochebrune "Réflexivité et auto-référence dans les systèmes complexes"*, pages 137–148.

[Nicolle *et al.*, 2002] Nicolle, A., Beust, P. et Perlerin, V. (2002). Un analogue de la mémoire pour un agent logiciel interactif. *In Cognito*, 21:37–66.

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

[Névéol *et al.*, 2004] Névéol, A., Soualmia, L., Douyère, M., Rogozan, A., Thirion, B. et Darmoni, S. (2004). Using CISMeF MeSH "encapsulated" terminology and a categorization algorithm for health resources. *International Journal of Medical Informatics*, 73(1):57–64.

[Ogden et Richard, 1923] Ogden, C. et Richard, I. (1923). *The Meaning of Meaning*. Routledge, Londres.

[Otlet, 1903] Otlet, P. (1903). Les sciences bibliographiques et la documentation. *Bulletin de l'Institut International de Bibliographie*, pages 125–147.

[Outing et Ruel, 2004] Outing, S. et Ruel, L. (2004). The best of Eyetrack III : What we saw when we looked through their eyes by. Site l'école de journalisme *Poynter Institute* : `http://www.poynterextra.org/eyetrack2004/main.htm` (page consultée le 11 juillet 2007).

[Page *et al.*, 2001] Page, L., Brin, S., Motwani, R. et Winograd, T. (2001). The pagerank citation ranking : Bringing order to the Web. Site Internet de l'Université de Stanford : `http://dbpubs.stanford.edu:8090/pub/1999-66` (page consultée le 13 juin 2007).

[Panaget, 2004] Panaget, F. (2004). Génération d'énoncés au sein de l'agent conversationnel animé Nestor. *In Actes de la journée de Association pour le Traitement Automatique des Langues (ATALA) : Génération en langue naturelle*.

[Papineni *et al.*, 2002] Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. *In Proceedings of the 40th Annual Meeting of the ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

[Papy et Chauvin, 2005] Papy, F. et Chauvin, S. (2005). Pour une approche visuelle et ergonomique dans la recherche et l'exploration d'informations au sein d'un OPAC de SCD. L'exemple du Visual'Catalog. *In* Papy, F., éditeur : *Les bibliothèques numériques*. Hermès, Paris.

[Paquin, 1990] Paquin, L.-C. (1990). Le passage des termes aux concepts. Site Internet de l'Université de Montréal : `http://www.ling.uqam.ca/sato/publications/bibliographie/Termes.htm` (page consultée le 18 juin 2007), première publication du texte lors du Colloque international "Les industries de la langue. Perspectives des années 1990", Thème I : Aspects technologiques : Représentation des connaissances dans le traitement des langues naturelles, Montréal le 22 novembre 1990.

[Partee *et al.*, 1990] Partee, B., ter Meulen, A. et Wall, R. (1990). *Mathematical Methods in Linguistics*, volume 30. Kluwer, Dordrecht.

[Pednault, 2000] Pednault, E. (2000). Representation is everything. *Communications of the ACM*, 43:80–83.

[Peirce, 1879] Peirce, C.-S. (1879). La logique de la science. *Revue philosophique de la France et de l'étranger*, article disponible en ligne : `http://www.psychanalyse-paris.com/La-logique-de-la-science-Comment,664.html` (page consultée le 11 avril 2007), 4(7):39–57.

[Perlerin, 2001] Perlerin, V. (2001). La recherche documentaire, une activité langagière. *In Actes des Rencontres des Etudiants Chercheurs en Informatique et en Traitement Automatique des Langues (RECITAL)*, volume 1, pages 469–479.

[Perlerin, 2002] Perlerin, V. (2002). Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes. *In Actes des Rencontres des Etudiants Chercheurs en Informatique et en Traitement Automatique des Langues (RECITAL)*, volume 1, pages 507–516.

[Perlerin, 2004] Perlerin, V. (2004). *Sémantique légère pour le document*. Thèse de Doctorat en Informatique, Université de Caen / Basse-Normandie, Caen.

[Perlerin et Beust, 2003] Perlerin, V. et Beust, P. (2003). Pour une instrumentation informatique du sens. *In Variation, construction et instrumentation du sens*, pages 197–229. M. Siksou.

[Perlerin *et al.*, 2005] Perlerin, V., Ferrari, S. et Beust, P. (2005). Métaphore et dynamique sémique. *In* Williams, G., éditeur : *La linguistique de corpus*, pages 323–336. Presses Universitaires de Rennes, Rennes.

[Pichon et Sébillot, 1999] Pichon, R. et 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.

[Pincemin, 1999] Pincemin, B. (1999). *Diffusion ciblée automatique d'informations : conception et mise en œuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*. Thèse de Doctorat en Linguistique, Université Paris IV Sorbonne.

[Pincemin *et al.*, 2006] Pincemin, B., Issac, F., Chanove, M. et Mathieu-Colas, M. (2006). Concordanciers : Thème et variations. *Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data*, 2:769–780.

[Ploux et Victorri, 1998] Ploux, S. et Victorri, B. (1998). Construction d'espaces sémantiques à l'aide de dictionnaires informatisés des synonymes. *Revue Traitement Automatique des Langues*, 39-1.

[Poibeau, 2004] Poibeau, T. (2004). Pré-analyse de corpus. *In* Purnelle, G., Fairon, C. et Dister, A., éditeurs : *Le poids des mots, actes des 7èmes Journées internationales d'Analyse statistiques des Données Textuelles*, volume 2, pages 897–903. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.

[Popescu *et al.*, 2006] Popescu, A., Grefenstette, G. et llic, P.-A. M. (2006). Using semantic commonsense ressources in image retrieval. *In* Mylonas, P., Wallace, M. et Angelelides, M., éditeurs : *Proceedings of the 1st International Workshop on Semantic Media Adaptation and Personalization*, pages 31–36. 4-5 December 2006, Athens, Greece, IEEE Computer Science Society.

[Poslad et Zuo, 2006] Poslad, S. et Zuo, L. (2006). A dynamic semantic framework to support multiple user viewpoints during information retrieval. *In Proceedings of the 1st International Workshop on Semantic Media Adaptation and Personalization, IEEE Computer Society*, pages 103–108.

[Pottier, 1991] Pottier, P. (1991). Utilisations de l'analyse en composantes principales pour la prévision statistique en météorologie. *Revue de Statistique Appliquée*, 39-1:37–49.

[Raffy, 2006] Raffy, C. (2006). Visualisation interactive d'ensembles de documents électroniques. Rapport de projet de Master Informatique et Ingénierie de l'Internet seconde année à l'Université de Caen (universitaire 2005-2006) disponible à l'adresse suivante : `http://users.etu.info.unicaen.fr/~craffy/m2/projet/soutenance/` (page consultée le 30 mai 2007).

[Rastier, 1987] Rastier, F. (1987). *Sémantique Interprétative*. Presses Universitaires de France, Paris.

[Rastier, 1998] Rastier, F. (1998). Le problème épistémologique du contexte et le problème de l'interprétation dans les sciences du langage. *Langages*, 129:97–111.

[Rastier, 2001a] Rastier, F. (2001a). *Arts et sciences du texte*. Presses Universitaires de France, Paris.

[Rastier, 2001b] Rastier, F. (2001b). Éléments de théorie des genres. Site Internet de la revue *Texto !* : `http://www.revue-texto.net/Inedits/Rastier/Rastier_Elements.html` (page consultée le 13 juin 2007).

[Rastier, 2005] Rastier, F. (2005). Enjeux épistémologiques de la linguistique de corpus. *In* Williams, G., éditeur : *La Linguistique de Corpus*, pages 31–45. Presses Universitaires de Rennes.

[Rastier, 2007] Rastier, F. (2007). Saussure et la science des textes. *In Actes du Colloque Révolutions saussuriennes*. Genève.

[Rastier *et al.*, 1995] Rastier, F., Béhar, H., Bernard, M., Bourion, E., Bouverot, D., Brunet, E., Erlich, D., Gorcy, G., Mézaille, T. et Surdel, F. (1995). *L'analyse thématique des données textuelles ? 'exemple des sentiments*. éditions Didier Erudition, collection Études de sémantique lexicale, Paris.

[Rastier *et al.*, 1994] Rastier, F., Cavazza, M. et Abeillé, A. (1994). *Sémantique pour l'analyse*. Masson, Paris.

[Rauber et Bina, 2000] Rauber, A. et Bina, H. (2000). Visualizing electronic document repositories : drawing books and papers in a digital library. *In Proceedings of the 5th IFIP 2.6 work. Conference on Visual Databases Systems (VDB5),*, Fukuoka, Japan.

[Rayner, 1998] Rayner, K. (1998). Eye movements in reading and information processing : 20 years of research. *Psycho Bull*, 124:372–422.

[Razmerita, 2003] Razmerita, L. (2003). *Modèle Utilisateur et Modélisation Utilisateur dans les Systèmes de Gestion des connaissances : une approche fondée sur les ontologies*. Thèse de Doctorat en Informatique, Université de Toulouse III.

[Renaux, 2003] Renaux, P. (2003). CATMInE : Computer assisted trade-mark infringement evaluation. *In Proceedings of the 16th Annual Conference on Legal Knowledge and Information Systems (Jurix'03)*, pages 61–70.

[Richard, 1961] Richard, J.-P. (1961). *Proust et le monde sensible*. Seuil, Paris.

[Rijsbergen, 1979] Rijsbergen, C. V. (1979). *Information Retrieval*. University of Glasgow.

[Robertson *et al.*, 1991] Robertson, G. G., Jock, D. M. et Stuart, K. C. (1991). Cone Trees : Animated 3D visualizations of hierarchical information. *In Proccedings of the ACM's Special Interest Group on Computer-Human Interaction*, pages 189–194. ACM Press.

[Rossignol et Sébillot, 2003] Rossignol, M. et Sébillot, P. (2003). Extraction statistique sur corpus de classes de mots-clés thématiques. *Revue Traitement Automatique des Langues (T.A.L.)*, 44(3):217–246.

[Rouse et Rouse, 1982] Rouse, M. A. et Rouse, R. H. (1982). La naissance des index. *In* Martin, H.-J. et Chartier, R., éditeurs : *Histoire de l'édition française, vol. 1 : Le livre conquérant. Du Moyen-Âge au milieu du XVIIe siècle*, pages 77–85. Promodis, Paris.

[Rouvier, 1966] Rouvier, R. (1966). L'analyse en composantes principales : Son utilisation en genetique et ses rapports avec l'analyse discriminatoire. *Revue Biometrics*, 22-2:343–357.

[Roy, 2005] Roy, T. (2005). Une plate-forme logicielle dédiée à la cartographie thématique de corpus. *In Actes de la conférence TALN/RECITAL 2005*, volume 1, pages 545–554, Dourdan, France. Hermès - Lavoisier.

[Roy et Beust, 2004] 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 et Beust, 2005] 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://users.info.unicaen.fr/~troy/rech/ROY_BEUST_ISKO_2005.pdf`.

[Roy et Beust, 2006] Roy, T. et Beust, P. (2006). Construction et exploration de corpus à partir du Web à l'aide d'une plate-forme de cartographie documentaire. *In Actes de la journée de l'Association pour le Traitement Automatique de la Langue : le Web comme ressources pour le TAL*, page Actes en ligne sur le site de la journée : `http://www.atala.org/article.php3?id_article=292`.

[Roy et Beust, 2007] Roy, T. et Beust, P. (2007). Ressources termino-ontologiques différentielles personnelles : construction et projection en corpus. *Hors-série 2006 de la revue Information - Interaction - Intelligence : Textes et ressources terminologiques et/ou ontologiques : évolution et maintenance*, pages 35–60.

[Roy *et al.*, 2007] 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.

[Roy et Ferrari, 2006] Roy, T. et Ferrari, S. (2006). User preferences for access to textual information. *In Proceedings of the 1st International Workshop on Semantic Media Adaptation and Personalization, IEEE Computer Society*, pages 171–176. IEEE Computer Society.

[Roy et Ferrari, 2007] Roy, T. et Ferrari, S. (2007). User preferences for access to textual information : Model, tools and experiments. *In Advances in Semantic Media Adaptation and Personalization*. Studies in Computational Intelligence, Springer Verlag, to appear.

[Roy *et al.*, 2005] Roy, T., Ferrari, S. et Beust, P. (2005). Cartographie de corpus pour l'étude de métaphores conceptuelles. *In Actes des Journées de Linguistique de Corpus (JLC'05)*, à paraître.

[Roy *et al.*, 2006] Roy, T., Ferrari, S. et Beust, P. (2006). étude de métaphores conceptuelles à l'aide de vues globales et temporelles sur corpus. *In* Mertens, P., Fairon, C., Dister, A. et Watrin, P., éditeurs : *Verbum ex machina - Proccedings of TALN'06, the 13th conference Natural Languages Processing*, volume 1, pages 580–589. Presses universitaires de Louvain, Louvain-la-Neuve, Belgium.

[Roy et Névéol, 2006] 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 et Sagit, 2003] Roy, T. et Sagit, O. (2003). Cartographie thématique de corpus. Rapport de projet de Maîtrise d'informatique à l'Université de Caen (universitaire 2002-2003) disponible à l'adresse suivante : `http://users.info.unicaen.fr/~troy/projet/2002-2003` (page consultée le 30 mai 2007).

[Sabah, 1997] Sabah, G. (1997). Le sens dans les traitements automatiques des langues - le point après 40 ans de recherche. *TA-information*, 38(2):91–133.

[Salton, 1989] Salton, G. (1989). *Automatic Text Processing : The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Pennsylvania.

[Salton, 1991] Salton, G. (1991). Developments in automatic text retrieval. *Science*, 253.

[Salton *et al.*, 1995] Salton, G., Allan, J. et Buckley, C. (1995). Automatic analysis, theme generation and summarization of machine readable text. *Science*, 264.

[Salvucci, 2001] Salvucci, D. D. (2001). An integrated model of eye movements and visual encoding. *Cognitive Systems Research*, 1(4).

[Sammon, 1969] Sammon, J. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on computers*, C-18-5:401–409.

[Savoy et Berger, 2005] Savoy, J. et Berger, P.-Y. (2005). Report on CLEF-2005 evaluation campaign : Monolingual, bilingual, and GIRT information retrieval. *Proceedings of the CLEF-2005 conference*.

[Schmidt, 1994] Schmidt, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*.

[Schwab *et al.*, 2005] Schwab, D., Lafourcade, M. et Prince, V. (2005). Extraction semi-supervisée de couples d'antonymes grâce à leur morphologie. *Actes de Traitement Automatique des Langues Naturelles (TALN'05)*, 1:73–82.

[Shneiderman, 1996] Shneiderman, B. (1996). The eyes have it : a task by data type taxonomy for information visualization. *Proceedings of Visual Languages*, pages 336–343.

[Sidir *et al.*, 2006] Sidir, M., Lucas, N. et Giguet, E. (2006). De l'analyse des discours à l'analyse structurale des réseaux sociaux : une étude diachronique d'un forum éducatif. *Revue Sticef (Sciences et Technologies de l'Information et de la Communication pour l'Éducation et la Formation)*, 13.

[Smadja et McKeown, 1990] Smadja, F. et McKeown, K. (1990). Automatically extracting and representing collocations for language generation. *28èmes Rencontres annuelles de Association for Computational Linguistics (ACL90).*

[Spark Jones et Galliers, 1995] Spark Jones, K. et Galliers, J. R. (1995). *Evaluating Natural Language Processing Systems : An Analysis and Review.* Springer, Lecture Notes in Artificial Intelligence.

[Spinat, 2002] Spinat, E. (2002). Pourquoi intégrer des outils de cartographie au sein des systèmes d'information de l'entreprise ? *Colloque Cartographie de l'Information.*

[Still et Bialek, 2004] Still, S. et Bialek, W. (2004). How many clusters ? an information-theoretic perspective. *Neural Computation*, 16(12):2483–2506.

[Tanguy, 1997] Tanguy, L. (1997). *Traitement automatique de la langue naturelle et Interprétation : Contribution à l'élaboration d'un modèle informatique de la Sémantique Interprétative.* Thèse de Doctorat en Informatique, Université de Rennes I, Rennes.

[The-Unicode-Consortium, 2006] The-Unicode-Consortium (2006). *The Unicode Standard 5.0.* Addison-Wesley Professional.

[Thlivitis, 1998] Thlivitis, T. (1998). *Sémantique interprétative Intertextuelle : Assistance anthropocentrée à la compréhension des textes.* Thèse de Doctorat en Informatique, Université de Rennes I, Rennes.

[Thüring *et al.*, 1995] Thüring, M., Hannemann, J. et Haake, J. (1995). Designing for comprehension : A cognitive approach to hypermedia development. *Communications of the ACM*, 38(8):57–66.

[Tricot, 2006] Tricot, C. (2006). *Cartographie sémantique. Des connaissances à la carte.* Thèse de Doctorat en Informatique, Université de Savoie.

[Tufte, 2004] Tufte, E. R. (2004). *The Visual Display of Quantitative Information - Second Edition.* Graphics Press, Cheshire, Connecticut.

[Valette, 2004] Valette, M. (2004). Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet. *In* et Mauro Gaio, P. E., éditeur : *Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique*, pages 215–230. 22-25 juin 2004, La Rochelle.

[Vallet *et al.*, 2006] Vallet, D., Cantador, I., Fernandez, M. et Castells, P. (2006). A Multi-Purpose Ontology-Based Approach for Personalized Content Filtering and Retrieval. *In* Mylonas, P., Wallace, M. et Angelelides, M., éditeurs : *Proceedings of the 1st International Workshop on Semantic Media Adaptation and Personalization*, pages 19–24. 4-5 December 2006, Athens, Greece, IEEE Computer Science Society.

[Varela, 1996] Varela, F. (1996). *Invitation aux sciences cognitives.* Point, Seuil.

[Veale, 2003] Veale, T. (2003). Systematicity and the lexicon in creative metaphor. *In Proceedings of ACL Workshop on the Lexicon and Figurative Language*, pages 28–35.

[Vergne, 2004] Vergne, J. (2004). Un exemple de traitement 'alingue' endogène : extraction de candidats termes dans des corpus bruts de langues non identifiées par étiquetage mot vide - mot plein. *Conférences invitées à l'Université Stendhal Grenoble 3.*

[Véronis, 2006] Véronis, J. (2006). A comparative study of six search engines. *Author's blog : http ://aixtal.blogspot.com/2006/03/search-and-winner-is.html.*

[Victorri, 1998] Victorri, B. (1998). Le projet TACIT : Traitements automatiques pour la compréhension d'informations textuelles. Rapport pour le GIS Sciences de la Cognition.

[Viégas, 2005] Viégas, F. B. (2005). *Revealing individual and collective pasts : Visualizations of online social archives*. Doctorate of Philosophy in Media Arts and Sciences at the Massachusetts Institute of Technology, Massachusetts Institute of Technology.

[Véronis, 2003] Véronis, J. (2003). Cartographie lexicale pour la recherche d'information. *In Actes de la conférence Traitement Automatique des Langues Naturelles (TALN'03)*, pages 265–275.

[Véronis et Guimier de Neef, 2006] Véronis, J. et Guimier de Neef, E. (2006). Le traitement des nouvelles formes de communication écrite. *In* Sabah, G., éditeur : *Compréhension des langues et interaction*, chapitre 8, pages 227–248. Lavoisier, Paris.

[W3C, 2001] W3C (2001). *Synchronized Multimedia Integration Language (SMIL 2.0) - W3C Recommendation*. `http://www.w3.org/TR/2001/REC-smil20-20010807/` (page consultée le 25 janvier 2007).

[W3C, 2003] W3C (2003). *Scalable Vector Graphics (SVG) 1.1 Specification - W3C Recommendation*. `http://www.w3.org/TR/2003/REC-SVG11-20030114/` (page consultée le 25 janvier 2007).

[W3C, 2006a] W3C (2006a). *eXtensible Markup Language (XML) 1.0 (Fourth Edition) - W3C Recommendation*. `http://www.w3.org/TR/2006/REC-xml-20060816` (page consultée le 25 janvier 2007).

[W3C, 2006b] W3C (2006b). *Extensible Stylesheet Language (XSL) Version 1.1 - W3C Recommendation*. `http://www.w3.org/TR/2006/REC-xsl11-20061205/` (page consultée le 25 janvier 2007).

[Ward, 1963] Ward, J. (1963). Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association*, 58:234–244.

[Ware, 2000] Ware, C. (2000). *Information visualization : perception for design*. Morgan Kaufmann Publishers Inc.

[Westerman *et al.*, 2005] Westerman, S., Collins, J. et Cribbin, T. (2005). Browsing a document collection represented in two- and three-dimensional virtual information space. *International Journal of Human-Computer Studies*, 62:713–736.

[Widlöcher, 2006] Widlöcher, A. (2006). Analyse par contraintes de l'organisation du discours. *In Verbum ex machina - Proccedings of TALN'06, the 13th conference Natural Languages Processing*, pages 367–376.

[Williams, 2005] Williams, G. (2005). *La linguistique de corpus*. Presses Universitaires de Rennes, Rennes.

[Witty, 1973] Witty, F. J. (1973). The beginnings of indexing and abstracting : somes notes towards a history of indexing and abstracting in antiquity and the Middle Age. *The indexer, journal of the Society of indexers and of the American Society of indexers*, 8(4):193–198.

[Zangemeister *et al.*, 1995] Zangemeister, W., Sherman, K. et Stark, L. (1995). Evidence for a global scanpath strategy in viewing abstract compared with realistic images. *International Journal of Human-Computer Studies*, 33(8):1009–1025.

[Zimina, 2000] Zimina, M. (2000). Alignement de textes bilingues par classification ascendante hiérarchique. *In Actes de 5èmes Journées Internationales d'Analyse Statistique des Données Textuelles*.

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

[Zouari, 2007] Zouari, L. (2007). *Vers le temps réel en transcription automatique de la parole grand vocabulaire*. Thèse de Doctorat de l'École Nationale Supérieure des Télécommunications, discipline Signal et Images.

Table des figures

1.1 Un exemple de segmentation hiérarchique réalisée avec le logiciel *Thema* sur un texte chinois. . . . 14
1.2 Un écran du système de navigation intra-documentaire *3D-XV*. . . . 15
1.3 Résultats d'une évaluation sur six moteurs de recherche sur Internet. . . . 21
1.4 Carte des pertes humaines de l'Armée Française lors de la campagne de Russie. . 26
1.5 *Stereoscopic Field Analyser* de [Ebert *et al.*, 1996]. . . . 28
1.6 À gauche proposé par *LibViewer* [Rauber et Bina, 2000], à droite, l'atelier de lecture de [Cubaud et Bénel, 2006]. . . . 29
1.7 Visualisations au niveau du texte proposées dans [Perlerin, 2004, pages 199 à 205]. À gauche, schématisation d'un document, à droite, vue en 3 dimensions d'un texte colorié. . . . 29
1.8 À gauche, le *Cone Trees* selon [Robertson *et al.*, 1991], à droite, les *Hyperbolic Trees* selon [Lamping, 1995]. . . . 30
1.9 À gauche, l'interface proposée par le dictionnaire des synonymes pour la visualisation des espaces sémantiques, à droite, la visualisation des champs lexicaux proposées par [Véronis, 2003]. . . . 31
1.10 Différentes vues sur des données textuelles proposées par le logiciel *Alceste* de la société *Image*. . . . 31
1.11 À gauche, projection d'après [Salton, 1989] sur un cercle des documents D répondant à une recherche d'information R, à droite, projection et lien sur un cercle entre différents passages d'un texte d'après [Salton *et al.*, 1995]. . . . 32
1.12 À gauche, visualisation des actualités proposée par *The Big Picture*, à droite, visualisation en nuages de mots du contenu de *blogs* proposée par *TagCloud*. . . . 33
1.13 À gauche, *Mountain* de [Viégas, 2005], à droite, le *Chronologue* de Jean Véronis. 33
1.14 En partie supérieure, un exemple de carte géographique mettant en évidence les différents continents ; en partie inférieure gauche, une carte conceptuelle représentant la répartition des noms de domaines de sites internet ; en partie inférieure droite, une carte de domaines mettant en évidence les distances et les collaborations entre différents auteurs. . . . 35
1.15 Cartographie des outils de cartographie proposée par Laurent Baleydier. . . . 36
1.16 Résultat d'une recherche sur *KartOO* avec le mot-clé *cartographie*. . . . 37
1.17 Résultat d'une recherche sur *MapStan* avec le mot-clé *intelligence économique*. . . 38
1.18 À gauche, une cartographie de dépêches de presse proposée dans [Mokrane *et al.*, 2004], à droite, l'interface de *NewsMap*. . . . 39
1.19 Cartographie thématique produite par *Neuronav* [Lelu et Aubin, 2001]. . . . 39

2.1 Principes généraux du modèle *AIdED* : interactions entre l'utilisateur et l'ensemble documentaire *via* la machine et des interfaces de visualisation interactive globale. 52
2.2 Création de tables de lexies. 55
2.3 Différenciation des lexies au sein de chaque table. 56
2.4 Représentation simple d'un domaine en ensemble de lexies. 56
2.5 Isotopies intra et inter-textuelles parcourant textes et sous-ensembles de textes d'un intertexte. L'isotopie inter-textuelle représentée par la couleur rouge parcourt les textes 1 et 2, délimitant ainsi le sous-ensemble de textes A. 58
2.6 Héritage des traits d'appartenance aux domaines dans les dispositifs d'une même session. 60
2.7 Sèmes génériques et spécifiques au sein d'un dispositif. 60

3.1 Les différents outils logiciels proposés pour l'extraction et la caractérisation de lexies ; leurs interactions sont représentées par des flèches. 77
3.2 Les différents échanges entre les logiciels *Memlabor*, *ThemeEditor* et *VisualLucia-Builder*. 78
3.3 Un exemple de chaîne de traitement *LinguaStream*. 79
3.4 *Memlabor* : Écran présentant les graphies répétées dans l'ensemble documentaire triées par ordre décroissant de fréquence. L'interface présente également deux graphiques mettant respectivement en évidence les graphies en fonction de leur fréquence et les rangs des graphies en fonction de la fréquence. 80
3.5 *Memlabor* : Écran mettant en évidence la constitution d'une liste de lexies à partir de la liste des graphies répétées dans l'ensemble documentaire. 81
3.6 *ThemeEditor* : l'écran de gauche présente un texte de l'ensemble documentaire colorié à l'aide des domaines. L'écran de droite propose des statistiques sur les domaines présents dans le texte (nombre de lexies coloriées, taux de recouvrement du domaine, etc.). 83
3.7 *LUCIABuilder* : Interface montrant la phase de remplissage d'une table d'un dispositif. 84
3.8 *VisualLuciaBuilder* : Interface illustrant les principales zones numérotées manuellement de 1 à 3. Le dispositif sélectionné représente le domaine de la guerre. Il décrit et met en relation les lexies d'une liste constituée à l'aide des outils *MemLabor* et *ThemeEditor* à partir de l'ensemble documentaire abordé précédemment. 85
3.9 *VisualLuciaBuilder* : Recherche de la lexie « rougeole » dans les listes. Une fois la lexie trouvée, elle est automatiquement sélectionnée. Un clic droit sur la lexie permet d'obtenir différentes informations par l'intermédiaire de *Wikipédia*, d'un dictionnaire des synonymes ou encore d'un outil de mise en contexte dans l'ensemble documentaire. 86
3.10 *VisualLuciaBuilder* : Des images (*smileys*) ont été utilisés pour définir l'attribut « évaluation ». De telles images se retrouvent ensuite dans les tables concernées dans le dispositif (dans l'exemple, il s'agit d'un dispositif sur le domaine de la guerre). 87
3.11 *VisualLuciaBuilder* : Résultat de la recherche de la lexie *radiation* dans le dispositif représentant le domaine de la guerre. La lexie a été entourée et reliée à ses attributs manuellement sur la figure. 88
3.12 *VisualLuciaBuilder* : Suppression de l'attribut *fonction* de la table des activités du domaine de la guerre. 88
3.13 *FlexiSemContext* : Interface d'interrogation de l'application. 90

3.14 *FlexiSemContext* : Mise en contexte de la lexie « ministère » dans un ensemble documentaire. 91
3.15 *FlexiSemContext* : Mise en contexte d'un couple *attribut : valeur*. 92
3.16 Illustration des principales sorties de la plate-forme *ProxiDocs*. 94
3.17 Cycle d'interactions entre les outils logiciels d'aide à la construction de RTO personnelles et *ProxiDocs*. Ces différents logiciels constituent l'instrumentation du modèle *AIdED*. 94
3.18 Étape de comptage des domaines de l'utilisateur dans l'ensemble documentaire. . 95
3.19 Étape de projection de la matrice numérique de grande dimension représentant l'ensemble documentaire vers un espace visualisable 97
3.20 Étape de classification de l'ensemble documentaire. 98
3.21 Étape de construction des cartes de l'ensemble documentaire. 99
3.22 Carte des textes en 2 dimensions et les possibilités d'interactions offertes. 100
3.23 Cartes des groupes de textes en 2 dimensions et les possibilités d'interactions offertes. 101
3.24 Mise en évidence des sous-cartographies de deux groupes de textes. 103
3.25 Exemples de cartes des textes (partie supérieure) et de cartes de groupes de textes (partie inférieure) en 3 dimensions. 105
3.26 Illustration de l'enchaînement dynamique proposé par une carte temporelle. Trois extraits capturés à des moments différents sont présentés. 106
3.27 Exemples d'un nuage (à gauche) et d'un anti-nuage (à droite) de lexies d'un domaine (la guerre) dans un ensemble documentaire. 107
3.28 Enchaînement des traitements réalisés par *ProxiDocs*. 108
3.29 Diagramme de classes mettant en évidence les principaux éléments logiciels composant la plate-forme *ProxiDocs*. 110
3.30 Interface de la plate-forme *ProxiDocs*. 111

4.1 Illustration des regards portés sur deux versions d'une page d'accueil d'un site Internet : l'originale (à gauche) et la version « optimisée » (à droite). 119
4.2 Le dispositif de l'informatique utilisé durant cette expérimentation. 123
4.3 Nuage de graphies de l'ensemble documentaire. 124
4.4 Carte des textes de l'ensemble documentaire. 125
4.5 Carte des groupes de textes de l'ensemble documentaire. 125
4.6 En partie gauche de la figure, un exemple de document constituant le corpus, en partie droite, son indexation MeSH au format XML PubMed. 130
4.7 Les 20 métatermes les plus présents dans l'ensemble de notre corpus 131
4.8 Carte des groupes de documents obtenus à partir des 78 métatermes présents dans le corpus. Les groupes de 1 à 6 ont été marqués manuellement sur la carte afin d'en faciliter l'analyse. 132
4.9 Carte des groupes de ressources obtenues à partir des métatermes non transversaux présents dans le corpus. Les groupes de 1 à 3 ont été numérotés manuellement sur la carte afin d'en faciliter l'analyse. 134
4.10 Le graphique de gauche présente le nombre d'articles de `Corpus_Travail` selon le mois de l'année, celui de droite, le nombre moyen par article d'occurrences de graphies des domaines étudiés selon le mois de l'année. 140
4.11 Représentation schématique des ensembles d'articles de `Corpus_Travail` étiquetés `Guerre`, `Météo` et `Santé` et de leurs intersections, les valeurs numériques indiquées sont les cardinalités des différents ensembles résultant des différentes intersections et exclusions des anneaux. 140

4.12 En partie gauche de la figure, extrait du nuage des lexies des trois dispositifs en corpus, en partie droite, extrait de l'anti-nuage des lexies des trois dispositifs en corpus. 141
4.13 Carte en 2 dimensions des textes du corpus. Des zones ont été marquées manuellement sur la carte afin de faciliter son analyse. 142
4.14 Extraits de la carte en 3 dimensions des textes du corpus. 143
4.15 Cartes en 2 dimensions des groupes de textes du corpus. Des groupes et des zones ont été marqués manuellement sur la carte afin de faciliter son analyse. 144
4.16 Mise en évidence des contextes d'actualisation du couple *évaluation* : *mal* dans des textes du groupe 2. 147
4.17 Trois extraits de la carte temporelle du corpus. Le tableau de la partie inférieure détaille chaque extrait. 151
4.18 Cartes des groupes des différents forums étudiés. 156
4.19 Sous-cartes des groupes A et C de la carte du forum $E1_{2002-2003}$ présenté en figure 4.18. 157
4.20 Description de trois extraits de la carte temporelle du forum $E1_{2002-2003}$. 159
4.21 À gauche, carte des groupes du forum *Utilisation de l'informatique*, à droite, carte des groupes du forum *Confidentialité des prêts*. 162
4.22 Extrait du thésaurus *MOTBIS* pour le terme *bibliothéconomie*. 164
4.23 À gauche, la carte des groupes du forum *Utilisation de l'informatique*. À droite, la carte des groupes du forum *Confidentialité des prêts*, ces cartes ont été construites avec la nouvelle version de la terminologie. 164
4.24 Un écran de l'outil *Bobinette* de [Huynh-Kim-Bang et Bruillard, 2005], outil disponible en ligne : `https ://wims.crashdump.net/www/forum/bobinette.php` (page consultée le 7 juillet 2007). 167
4.25 Suivi du regard sur un texte, `http://www.alexpoole.info/academic/lecturenotes_fr.html` (page consultée le 9 juillet 2007). Le suivi du regard mis en évidence est caractéristique de la lecture d'un texte avec un parcours linéaire des différents mots. 168
4.26 À gauche, suivi du regard sur une page de résultats du moteur de recherche *Google*, `http://www.eyetools.com/inpage/research_google_eyetracking_heatmap.htm` (page consultée le 9 juillet 2007). À droite, parcours du regard « typique » sur une page d'accueil d'un site Internet selon [Outing et Ruel, 2004]. 169
4.27 En partie gauche de la figure, la caméra utilisée ; au centre, le sujet positionné face à l'écran diffusant les images et à la caméra filmant l'un de ses yeux ; à droite, les expérimentateurs devant leurs écrans de contrôle pour la diffusion des images et l'enregistrement des mouvements oculaires du sujet. 170
4.28 Les cartes en 2 dimensions proposées aux sujets. En partie supérieure gauche, la carte des documents, en partie inférieure droite, la carte des groupes de documents. Les cartes en partie inférieure reprennent la carte des groupes de documents avec les lexies *travail* (carte de gauche) et *coup* (carte de droite) mises en évidence. . . 171
4.29 Parcours du regard de deux sujets différents sur la carte des documents. 172
4.30 Parcours du regard de deux sujets différents sur la carte des groupes de documents. 173
4.31 Parcours de deux sujets différents sur les cartes des groupes de documents mettant en évidence le résultat d'interactions. 173
4.32 Mouvement du regard d'un sujet sur l'animation de la carte des groupes de documents en 3 dimensions. 174
4.33 Fixations de l'ensemble des sujets sur les cartes des documents (à gauche) et des groupes de documents (à droite). 175

4.34 Les zones d'intérêt isolées sur les cartes en 2 dimensions. 175

A.1 Extrait d'un fichier XML contenant des ensembles de lexies construits avec l'outil *ThemeEditor*. 185
A.2 Extraits des fichiers XML décrivant un dispositif construit avec *VisualLuciaBuilder* 186

B.1 Les deux composantes optimisant une représentation d'un poisson en 2 dimensions. 189
B.2 Application de la CHA dans la classification de 5 documents. La ligne en pointillés indique met en évidence les groupes obtenus à l'issue de l'étape 2, soient le groupe formé par doc 1 et doc 2, le groupe formé par doc 3 et doc 4 et le groupe formé par doc 5. 197
B.3 Carte de l'ensemble documentaire réalisée avec la méthode des plus grandes distances. 201
B.4 Carte de l'ensemble documentaire réalisée avec la méthode du produit scalaire. . 202
B.5 Carte de l'ensemble documentaire réalisée avec la méthode de l'ACP. 202
B.6 Carte de l'ensemble documentaire réalisée avec la méthode de Sammon. 203
B.7 Carte de l'ensemble documentaire réalisée avec la méthode de l'AFC. 203
B.8 Carte des groupes de textes de l'ensemble documentaire catégorisé avec une CHA à partir d'une ACP. 204
B.9 Carte des groupes textes de l'ensemble documentaire catégorisé avec la méthode des K-Means à partir d'une ACP. 205

C.1 Le dispositif de l'agriculture construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet. 208
C.2 Le dispositif de la pollution construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet. 209
C.3 Le dispositif de la sécurité routière construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet. 210
C.4 Le dispositif de l'espace construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet. 211
C.5 Le dispositif du sport construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet. 212
C.6 Le dispositif de l'informatique construit et utilisé au cours de l'expérimentation liée à une recherche documentaire sur Internet. 213
C.7 Le dispositif de la guerre mis au point et utilisé au cours de l'expérimentation associée au projet *IsoMeta*. 214
C.8 Le dispositif de la météorologie repris et utilisé au cours de l'expérimentation associée au projet *IsoMeta*. 215
C.9 Le dispositif de la santé mis au point et utilisé au cours de l'expérimentation associée au projet *IsoMeta*. 216

D.1 Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 1 à 6. . . 220
D.2 Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 7 à 12. . . 221
D.3 Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 13 à 18. . 222
D.4 Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 19 à 24. . 223
D.5 Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 25 à 30. . 224
D.6 Diaporama pour l'étude du suivi du regard sur les cartes : diapositives 31 à 35. . 225

Liste des tableaux

2.1 Exemple de classements avec et sans pondération des isotopies inter-textuelles présentes dans un ensemble de textes. 63

3.1 *Packages* de la plate-forme *ProxiDocs*. 109

4.1 Détails des trois groupes marquées de 1 à 3 sur la carte des groupes de documents de l'ensemble documentaire présentée en figure 4.5. 126
4.2 Description des groupes 1 à 3 de la carte présentée en figure 4.8. 133
4.3 Tableau présentant une première répartition des domaines dans `Corpus_Total` et `Corpus_Travail`. 139
4.4 Détails des différents groupes et zones de groupes marquées sur la carte de la figure 4.15. 145
4.5 Détails des isotopies inter-textuelles parcourant le groupe 5, informations accessibles dans le rapport d'analyse du groupe. Pour rappel, le nombre de répétitions correspond d'un attribut ou d'une valeur d'attribut correspond au nombre d'occurrences de mots les portant des domaines issus des domaines sources. 148
4.6 Détails des différents groupes et zones de groupes marquées sur la carte de la figure 4.15. 150
4.7 Sous-catégories liées au domaine de la bibliothéconomie. 165

Table des algorithmes

1 Schéma d'algorithme de tout traitement multilingue. 72
2 Algorithme de la méthode de comptage relative des domaines en ensemble documentaire. 188
3 Schéma d'algorithme de la méthode de l'ACP. 192
4 Algorithme de la méthode de projection de Sammon. 196
5 Algorithme de la Catégorisation Hiérarchique Ascendante. 198
6 Algorithme de la méthode des K-Means. 199

Résumé : Avec la multiplication des documents électroniques, les utilisateurs se retrouvent face à une véritable montagne de textes difficile à gravir. Cette thèse, prenant place en Traitement Automatique des Langues, a pour objectif d'aider les utilisateurs dans de telles situations. Les systèmes traditionnellement proposés (tels les moteurs de recherche) ne donnent pas toujours satisfaction aux utilisateurs pour des tâches répétées, prenant peu en considération leur point de vue et leurs interactions avec le matériau textuel. Nous proposons dans cette thèse que la personnalisation et l'interaction soient au centre de nouveaux outils d'aide pour l'accès au contenu d'ensembles de textes. Ainsi, nous représentons le point de vue de l'utilisateur sur ses domaines d'intérêt par des ensembles de termes décrits et organisés selon un modèle de sémantique lexicale différentielle. Nous exploitons de telles représentations pour construire des supports cartographiques d'interactions entre l'utilisateur et l'ensemble de textes, supports lui permettant de visualiser des regroupements, des liens et des différences entre textes de l'ensemble, et ainsi d'appréhender son contenu. Afin d'opérationnaliser de telles propositions, nous avons mis au point la plate-forme *ProxiDocs*. Différentes validations de la plate-forme, prenant place dans des contextes pluridisciplinaires variés allant notamment de la recherche d'information sur Internet à l'étude d'expressions métaphoriques, ont ainsi permis de dégager la valeur ajoutée de nos propositions.

Title: *Interactive visualizations for personal help in interpretation of sets of documents.*

Abstract: Since the number of electronic textual documents keeps increasing, it has become more and more difficult for users to access information out of this amount of data. That is why this thesis dealing with Natural Langage Processing aims to help users in such situations. Existing systems (such as search engines on Internet) do not fully satisfy their users in repeated tasks, barely taking into consideration the users' points of view and their interactions with the textual material. In this thesis, we propose to consider personalization and interaction as the core of new tools to access the content of sets of texts. Thus, we represent users' points of view on domains they are interested in with sets of lexical units described and structured according to a differential lexical semantic model. We, then, use such representations to build cartographic supports allowing interactions between users and their sets of texts in order to visualize gatherings, links and differences between texts in a set and, doing so, to reach their content. To computerize such propositions, we have developed the *ProxiDocs* plat-form. Many validations of the plat-form have been realized through multidisciplinary experiments, such as information retrieval on Internet and a study of metaphorical expressions. All of them brought to the fore the efficiency of our propositions.

Mots-clés : traitement automatique des langues naturelles, sémantique, interfaces utilisateurs, cartographie – informatique, gestion électronique de documents, représentations de connaissances, logiciels interactifs et individu-centrés, accès au contenu d'ensembles de textes.

Discipline : Informatique

Laboratoire : Groupe de Recherche en Informatique, Image, Automatique et Instrumentation de Caen - UMR 6072, Université de Caen / Basse-Normandie, Campus Côte de Nacre Boulevard Maréchal Juin, BP 5186 F-14032 Caen Cedex