Canopée : un environnement numérique de sémantique augmentée

Pierre Beust[1], Roger Cozien[2], Serge Mauger[1], Dominique Haglon[2]

[1]GREYC CNRS UMR 6072 & Pôle ModesCoS de la MRSH – Université de Caen Basse-Normandie – 14032 Caen Cedex, France

[2]eXo maKina – www.exomakina.fr – 234 rue Championnet – 75018 Paris

Abstract

This paper deals with a project, still in developpement, which is called Canopée. Canopée is a user-centered tool aiming at assist its user in a texts interpretation task. We look to build an human-machine interaction in which the meaning arise from the user's navigation within the texts and process results as statistics and vizualisations.

Résumé

Cet article présente le projet en cours de développement intitulé Canopée. Il vise à la construction d'un environnement d'assistance à l'interprétation des textes. Le but est de médiatiser l'activité interprétative du sujet et lui proposant des résultats statistiques et des visualisations qui, pour lui, soient porteurs de signification. C'est dans une interaction personne-système que se mettent en place de manière incrémentale les parcours interprétatifs.

Mots-clés : Sens, Sémiotique, Interaction, Interprétation, Lexicométrie.

1. Introduction

Cet article rend compte d'un travail de recherche et de développement au sein de la société eXo maKina[1] en collaboration avec le laboratoire GREYC CNRS UMR 6072 de l'université de Caen Basse-Normandie. L'objectif est de développer et de commercialiser un outil d'assistance à l'interprétation dédié à l'accès personnalisé aux contenus dans des collections documentaires numériques textuelles.

1 eXo maKina est une jeune société spécialisée dans la conception et la production de logiciels spécifiques à destination d'usages professionnels. Elle développe notamment des solutions logicielles innovantes dans les technologies de l'image : marquage implicite d'images résistant aux altérations (logiciel Tantale), analyse d'images contrefaites destiné à la photo-interprétation avancée (logiciel Tungstène). Les solutions consistent à mettre en évidence l'ensemble des traces insolites et autres singularités présentes dans les fichiers photographiques. Au vu de ces traces, il devient possible de montrer par exemple les montages et toutes les tentatives de manipulation par l'image. Au delà de la technique informatique, eXo maKina propose une méthode d'étude et d'analyse des formes de communication par l'image basée sur des principes de sémiotique opérationnelle. Ses principaux clients sont les agences de presses et les services de défense et de renseignement.

eXo maKina s'intéresse à l'extraction des éléments linguistiques qui font sens dans les documents (numériques) textuels (et plus particulièrement ceux qui s'accompagnent d'images, par exemple de dépêches d'agence de presse). L'approche entreprise vise à offrir à chaque utilisateur/analyste la possibilité de définir son propre lexique et ses collections de centres d'intérêt pour explorer des corpus documentaires et en mettre en évidence de manière personnalisée les grandes tendances sémiotiques. C'est le projet Canopée que cet article présente.

Dans une première partie nous dressons une typologie des différentes méthodes d'accès au contenu textuel. Ensuite nous nous en démarquerons en affirmant une démarche avant tout centrée utilisateur. Nous présenterons l'outil Canopée comme un développement industriel mettant en œuvre cette approche centrée utilisateur. Cela nous amènera à préciser ce que nous entendons comme étant un environnement de sémantique augmentée.

2. Accès au contenu et recherche d'information

Les technologies de l'information (notamment sur l'Internet) forment un domaine d'application direct du TAL, de la lexicométrie et de l'accès au contenu des documents. La taille des données textuelles à traiter ainsi que le nombre et la variété des traitements à réaliser rendent incontournable le développement de méthodes d'analyses automatiques les plus fiables possibles et rapides.

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

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

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

On entend par extraction d'information les méthodes qui consistent à rechercher dans un corpus très homogène (par exemple des dépêches d'actualité dans ou encore des articles scientifiques) des informations dont on sait qu'elles s'y trouvent. Ainsi on cherche par exemple dans un corpus d'actualité boursière à extraire les transactions de rachats et de fusions de sociétés ce qui revient à chercher à remplir des sortes de formulaires électroniques indiquant notamment qui a acheté qui, à quel prix et quand. Il est donc souvent visé ici d'alimenter de manière automatique des bases de données préexistantes à partir de corpus soigneusement sélectionnés. Les méthodes dites de Questions/Réponses n'ont pas le même objectif. Elles consistent à chercher un fragment de texte extrait d'un corpus volontairement assez généraliste dans lequel un sujet interprétant a de bonnes chances de trouver la réponse à une question qu'il aura formulée en langue naturelle. Par exemple extraire une séquence du style « (…) la vie de Baudelaire, auteur des Fleurs

du mal, fut (…) » à la question « Qui a écrit les Fleurs du mal ? ». La bonne construction linguistique de la réponse n'est pas ici visée car il ne s'agit que de fournir une « fenêtre » 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 TREC9, les systèmes de questions/réponses avaient pour consigne de rendre des réponses de moins de 250 caractères à partir de 980 000 documents et de 700 questions. A la différence des méthodes d'extraction d'information, on s'en remet à l'interprétation d'un sujet humain quant à la qualité des réponses trouvées. Les méthodes de résumé automatique s'appuient aussi au final sur l'interprétation de celui à qui est destiné le résumé. Bien souvent il est plus juste de parler de condensation ou de réduction de textes plutôt que de résumé (dans le sens de ce qu'est un résumé quand il est rédigé par un sujet humain). L'enjeu technique est de rechercher des phrases dont on pense qu'elles ont un statut assez significatif (par exemple une phrase qui commencerait par « en somme, on constate que (…) » a de bonnes chances de synthétiser ce qui est dit avant) et de les juxtaposer dans un « résumé » dont on fait l'hypothèse que celui qui le lira pourra rétablir une certaine cohérence textuelle, par exemple relativement aux rattachements anaphoriques.

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

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

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

Dans leurs interactions avec les utilisateurs, les moteurs de recherche sont souvent assez rudimentaires. Il faut bien souligner que l'utilisateur et son objectif de recherche sont

uniquement considérés sous la forme d'une liste de mots clés (dont la casse et l'accentuation et même l'ordre sont d'ailleurs rarement pris en compte) considérés pour une seule recherche dans la mesure où toutes les requêtes sont traitées indépendamment les unes des autres. Dans la pratique on s'aperçoit que pour mener à bien une recherche sur le web, il convient en fait d'interroger successivement plusieurs fois le (ou les) moteur(s) en ajoutant ou en précisant certains mots clés en fonction des résultats rendus à chaque étape. C'est donc le plus souvent à l'utilisateur seul qu'il convient de développer des stratégies efficaces pour trouver des mots clés adaptés à sa recherche. Certaines tentatives sont mises en place par certains moteurs pour aller un peu plus loin que la simple prise en compte de mots clés. Par exemple, Google[2] permet de rechercher un mots clé ou un de ses synonymes avec l'opérateur tilde ~ (par exemple une recherche sur powerpoint ~help effectuera une recherche sur powerpoint ET help ou tips, faq, tutorial). Cependant, c'est le moteur lui-même qui établit ses listes de synonymes et il serait peut être plus judicieux que celles-ci soit validées par les utilisateurs quand ils les utilisent.

3. Approche centrée-utilisateur

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

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

Notre objectif est de mettre au point des Environnements Numériques de Travail (ENT) où les problématiques sémiotiques et linguistiques sont des points clés de l'interaction homme-machine. Dans ce type d'ENT (par exemple, des environnements pour la veille stratégique et/ou documentaire, pour la GED, pour le e-learning ...), il convient de modéliser et d'expérimenter des ressources (principalement terminologiques) et des modèles d'analyse informatiques des textes et du sens. Viser des formes d'instrumentations dites centrées utilisateurs consiste à construire les applications et les ressources manipulées avant tout autour des spécificités socio-linguistiques des utilisateurs. La priorité en terme de description concerne leurs centres d'intérêt, leurs habitudes terminologiques, leurs parcours interprétatifs dans les textes.

2 A propos de Google, on trouve sur le blog de Jean Véronis (aixtal.blogspot.com) un autre exemple de résultat assez malheureux : celui de l'opérateur define. L'opérateur define (disponible pour les pages en français depuis avril 2005) sert à rechercher à propos d'un mot des pages Web où ce mot ferait visiblement l'objet d'une définition. L'expérience relatée consiste à rechercher ainsi sur Google une définition du mot femme avec la requête define:femme. Les résultats donnés sont pour le moins plus que contestables. On aurait donc bien tort de croire à la fiabilité de l'opérateur define (qui pourtant est présenté par Google comme un outil de recherche de définition sans plus de détails) comme on aurait tort aussi de considérer le Web dans son ensemble comme une encyclopédie dans lequel on puisse rechercher des définitions attestées, notamment d'un point de vue moral.

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

On se place ici dans la suite des travaux de Jacques Coursil qui montre que le sens est au moins autant du côté de l'interprétant que du côté du sujet parlant. Ainsi c'est parce que le sujet parlant est aussi le premier interprétant de ce qu'il dit, au moment où il le dit, qu'il peut donner forme à sa parole en continue. C'est le principe de non préméditation de la chaîne parlée (Coursil, 2000). Il n'est bien sûr pas question ici de comprendre qu'un sujet parlant ne sait pas ce qu'il veut dire au moment où il commence à le dire mais plutôt qu'il ne sait pas exactement comment il va le dire avant de proférer une parole. Ce principe de non préméditation et une mise en évidence d'un couplage intéressant entre le sujet parlant/interprétant et l'environnement dans lequel il produit sa parole. Dans ce rapport à l'interprétation, notre travail trouve naturellement un positionnement scientifique et épistémologique dans les travaux en Sémantique Interprétative (SI) de François Rastier (Rastier, 1987), elle-même en filiation avec les travaux en sémantique structurale dont l'origine remonte aux travaux de Saussure. Dans la SI le sens est vu comme une perception sémantique, perception forcément individuelle, dont toute tentative d'objectivation est une sommation incomplète de points de vue.

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

Pour nous, le sens n'est pas le résultat d'un calcul, c'est une activité au centre d'une interaction (activité qui de plus n'est pas forcément préalablement finalisée dans le temps). Ainsi, nous remettons en cause l'idée que quelque chose ait du sens (ou non) pour défendre plutôt l'idée qu'il y a des choses qui font sens (ou pas) pour quelqu'un. On se rapproche en cela des principes de la sémiotique triadique de Peirce (Peirce, 1978). L'approche adoptée nous amène donc à préférer une instrumentation interactive du sens à une construction compositionnelle (ou à un calcul) du sens (comme c'est le cas dans beaucoup de travaux de sémantique en TAL). Nous ne cherchons pas une représentation formelle extralinguistique du sens (comme c'est le cas par exemple en logique). On rejoint ici l'avis de (Nicolle, 2005) pour qui le sens n'est jamais capturé par ses représentations. Toute représentation formelle du sens est forcément incomplète. De plus, et

c'est aussi une différence importante par rapport à une formalisation qui se veut objective, le sens est subjectif dans la mesure où il n'y a pas forcément de consensus d'un interprétant à un autre sur une explicitation (partielle ou complète) du sens d'un texte, si court soit-il.

Nous préférons donc de loin l'idée d'une instrumentation du sens à celle de la construction du sens. Nous cherchons à assister les compétences interprétatives mais surtout pas à les remplacer. Il s'agit donc de considérer que le sens tel que le produit une interprétation humaine n'est pas à la portée d'un seul traitement informatique. Cela ne veut pas dire pour autant que les machines ne puissent pas avoir une activité d'analyse des textes et être d'une utilité certaine dans une problématique d'accès au contenu. C'est ici que la lexicométrie nous semble tout à fait pertinente. Les machines ont une approche calculatoire de l'accés au contenu là où les sujets humains ont une approche interprétative. Ces deux formes de rapports aux textes ne sont pas en concurrence car l'activité de la machine n'a en aucun cas le but de supplanter celle de l'utilisateur. Au contraire, dans l'environnement numérique de travail elle doivent être complémentaires. L'environnement doit avoir comme objectif de produire dans l'interaction des signes (notamment, par exemple, au moyen de techniques de visualisation adéquates) qui vont participer aux interprétations du ou des utilisateurs et ainsi prolonger le couplage.

4. Le logiciel Canopée

Canopée est un produit qui est encore en phase de développement. Sa sortie commerciale et la communication de la part de la société qui va de paire est prévue dans l'année 2012. Le projet Canopée s'inscrit dans une suite aux projets développés au GREYC : ThemeEditor (Beust, 2002), ProxiDocs (Roy, 2007), Lucia (Perlerin, 2004).

Dans les tâches numériques quotidiennes de la majeure partie des utilisateurs l'appréhension de collections documentaires revêt une importance croissante. Ainsi, l'accès à un dossier ou à un répertoire est déjà une confrontation à une collection potentiellement assez grande. La gestion régulière d'une boite de courrier électronique est aussi une façon d'appréhender une collection qui par nature est de taille croissante. La réponse même d'une requête à un moteur de recherche est aussi une collection assez vaste (en témoigne les estimations de nombres de réponses fournies) dont l'utilisateur ne regarde dans presque tous les cas que les premiers documents. Enfin, les flux d'informations (dépêches d'actualités, flux RSS, etc.) sont aussi des collections à part entière avec, en plus, la spécificité chronologique. Le projet Canopée cherche à produire une assistance logicielle dans le rapport entre un utilisateur particulier (voir un groupe d'utilisateurs) et une collection, rapport sous-tendu par les capacités interprétatives de l'utilisateur.

Les fonctionnalités de Canopée qui sont déjà opérationnelles concernent :

- la saisie de ressources termino-ontologiques personnalisées (cf. Figure 1),
- l'extraction des thèmes (cf. Figure 2),
- la cartographie de corpus et le coloriage thématique (cf. Figure 3),
- les regroupements thématiques de documents par technique de clustering (cf. Figure 4),
- la recherche documentaire (cf. Figure 5).

Les ressources termino-ontologiques de Canopée sont principalement des descriptions thématiques et componentielles. Dans chaque thème on décrit des lexies. Pour chaque lexie on liste des flexions et on étiquette librement les lexies par des sèmes représentés par une simple description textuelle. Le tout se fait dans une partie de l'application qui gère l'encodage XML des ressources, le partage des sèmes utilisés, les flexions des lexies (module appelé Kotoba, cf. Figure 1).

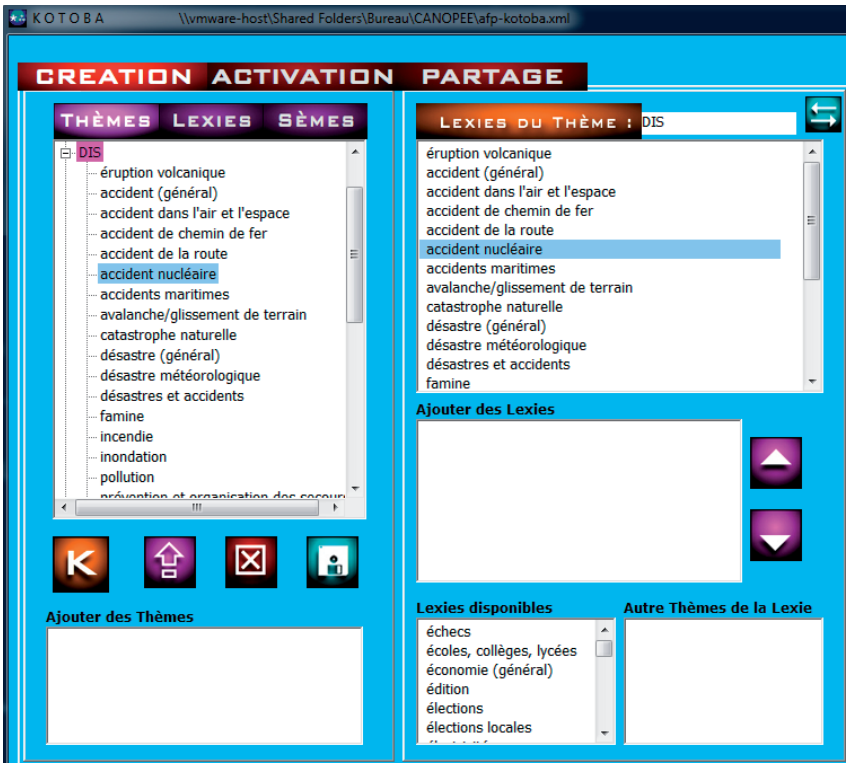

Figure 1 : Interface de description du contenu lexical

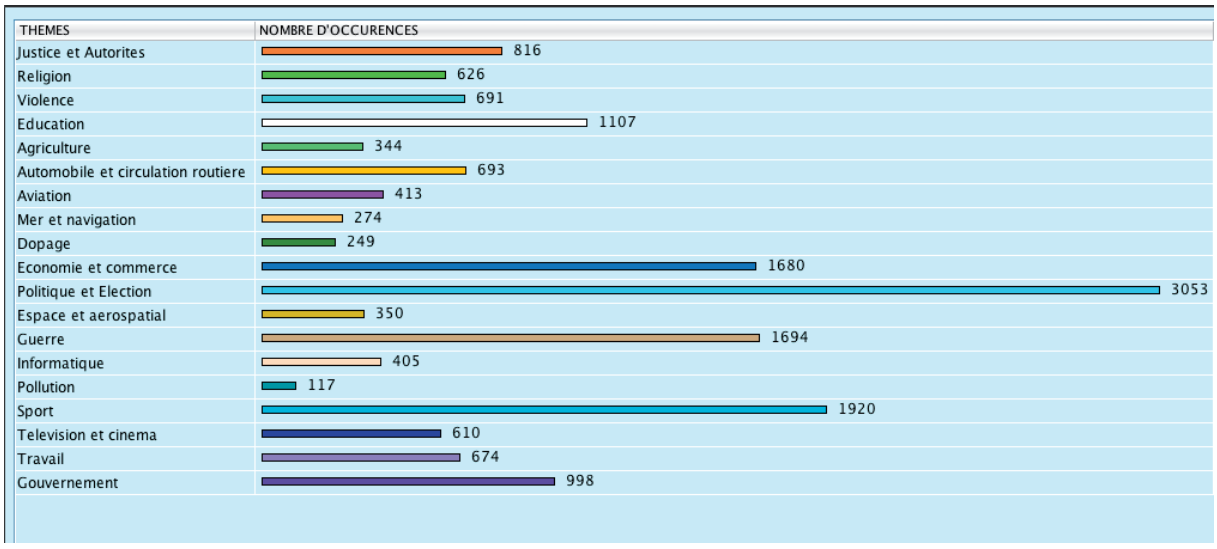

Figure 2 : Importances relatives des thèmes dans un flux documentaire

A partir des ressources lexicales qu'il a définies, l'utilisateur demande au logiciel de calculer une cartographie de la collection documentaire. Cette cartographie est une représentation 3D de l'espace vectoriel des documents de la collection relativement aux thèmes de l'utilisateur, ce qui s'apparente à une matrice de dimension n (n étant le nombre de thèmes) contenant p vecteurs (p étant le nombre de documents).

	Th1	Th2	…	…	…	…	Thn-1	Thn
Doc1	x11[3]	x12	…	…	…	…	x1n-1	x1n
…								
Docp	xp1	xp2	…	…	…	…	xpn-1	xpn

Les cartographies produites forment un environnement interactif où l'on peut à tout moment revenir sur les ressources et reformer en conséquences les visualisations, obtenir des résultats statistiques sur les visualisations, produire des coloriages thématiques de textes, rechercher des documents via des requêtes, comparer différentes projections. C'est un environnement numérique de travail pour appréhender une collection documentaire de manière personnalisée en fonction de ses centres d'intérêts.

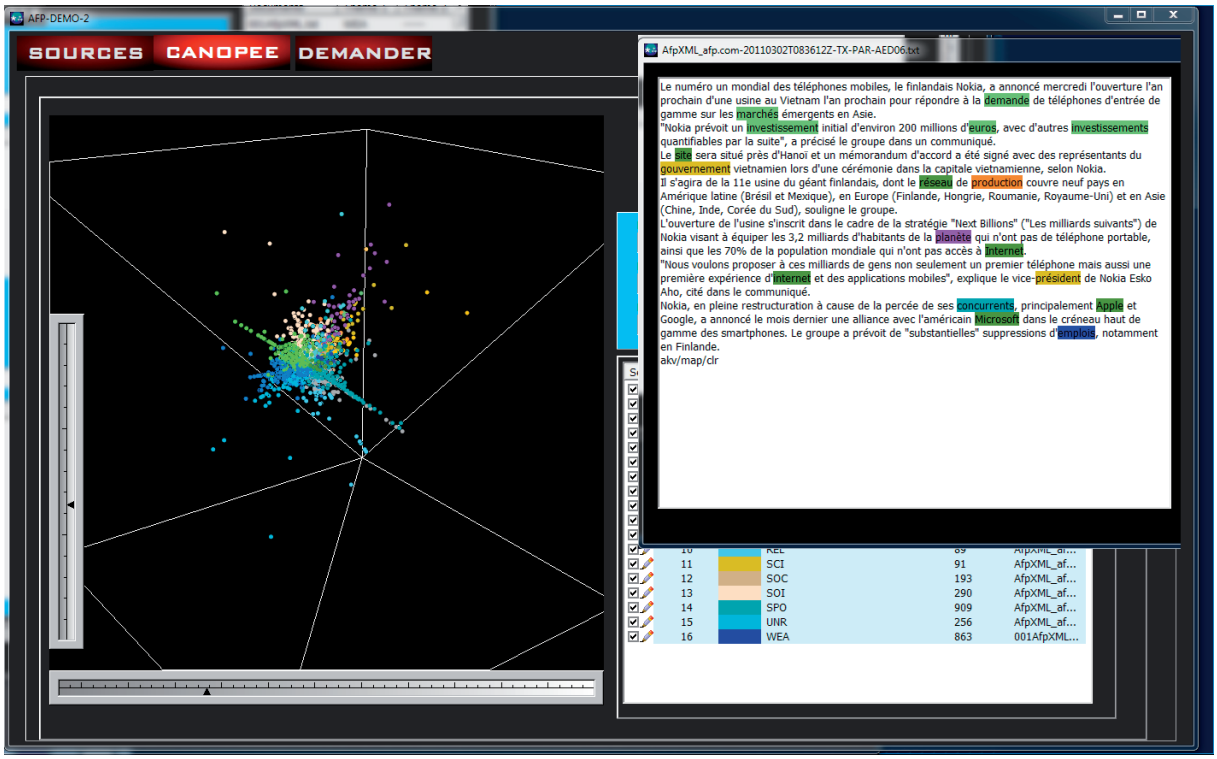

Figure 3 : Visualisation d'un flux documentaire avec Canopée et coloriage thématique de document.

3 x_{yz} représente le nombre de relatifs d'occurrences des lexies du thème z par rapport au nombre total de mots du document y.

A partir de la projection 3D de l'espace vectoriel, on peut en déduire par une méthode de clustering (i.e. k-means ou classification hiérarchique ascendante) des grandes classes thématiques de documents représentant les grandes tendances de la collection relativement aux centres d'intérêt de l'utilisateur (cf. Figure 4).

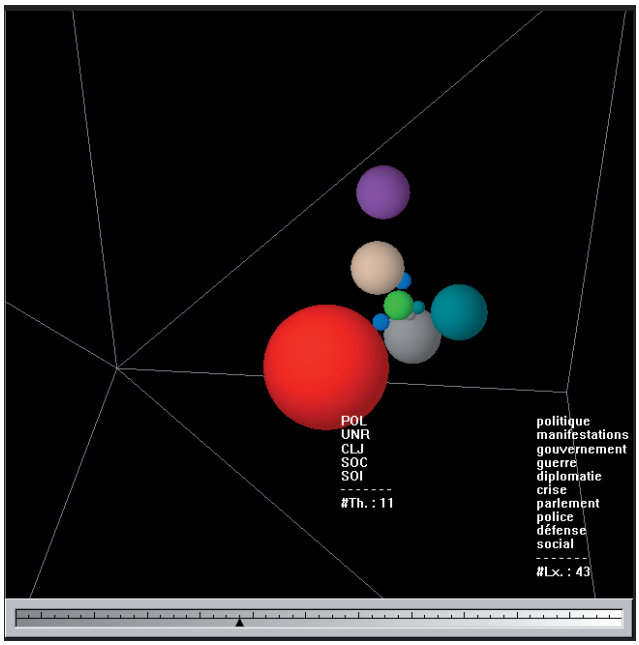

Figure 4 : Regroupement de documents par clustering

En tant qu'outil pour la veille documentaire, Canopée se doit d'intégrer une fonctionnalité de recherche de documents dans les collections parce que c'est une attente forte des utilisateurs/clients. La majeure partie des moteurs de recherche dans des collections sont construits sur le même modèle opérationnel de la base de données d'index de documents. Dans l'approche centrée utilisateur et cartographique qui est la nôtre nous avons cherché à proposer une fonctionnalité de recherche différente basée sur la topologie de l'espace documentaire personnel plus que sur des entrées d'indexation indépendantes. C'est également le point de vue considéré en recherche d'information par le modèle vectoriel de Salton (Salton et McGill, 1983) dont Canopée ou encore la sémantique latente (Landauer *et al.*, 1998) sont des déclinaisons.

Du point de vue de l'utilisateur, le changement par rapport aux moteurs de recherche réside dans la nature de la requête. Dans les moteurs de recherche classiques à base d'indexation full-text, une requête s'exprime comme une combinatoire (plus ou moins structurée) de mots clés. Ici la requête et les documents recherchés sont de même nature, c'est-à-dire textuels. L'utilisateur peut écrire en quelques lignes de texte la thématique et les spécificités du document qu'il cherche (il peut également en contournant l'usage faire un copier/coller d'un texte pour chercher ceux de son corpus qui s'en rapprochent le plus). La formulation de la requête est considérée comme un texte à part entière au même titre que les documents du corpus (elle est même, comme tout autre texte, rattachée automatiquement à un thème principal). On peut donc associer à la requête un vecteur dans l'espace vectoriel et rendre comme réponses à la requête les vecteurs de l'espace qui sont les plus proches en terme de distance du vecteur de la requête. Cela s'apparente à une

sorte de recherche documentaire par l'exemple. De plus, en comparaison avec les moteurs de recherche où chaque requête est indépendante des autres, on peut simultanément soumettre plusieurs requêtes et en évaluer les proximités réciproques.

Dans l'interface de Canopée la réponse à une recherche est un point mis en évidence dans l'espace 3D (et plusieurs dans le cas de requêtes multiples comme c'est le cas en Figure 5). Ce point est mis en exergue sous la forme d'un petit cube rouge pour le différencier des autres points qui représentent les documents. L'ensemble des documents formant les réponses trouvées dans le corpus se visualise comme une sphère (dont on peut paramétrer le rayon) autour du cube rouge et incluant les documents recherchés.

En complément on pourrait également penser que connaître les documents de la collection les plus éloignés de la requête pourrait être aussi une information utile. C'est aussi une forme de réponse à la requête, forme à laquelle les utilisateurs ne sont pas du tout habitués étant donné qu'un moteur de recherche classique ne peut pas, de manière similaire, répondre en creux à une requête en donnant les documents qui a priori n'auraient rien à voir. C'est une fonctionnalité à expérimenter qui nous semble intéressante et potentiellement génératrice de couplage avec l'utilisateur.

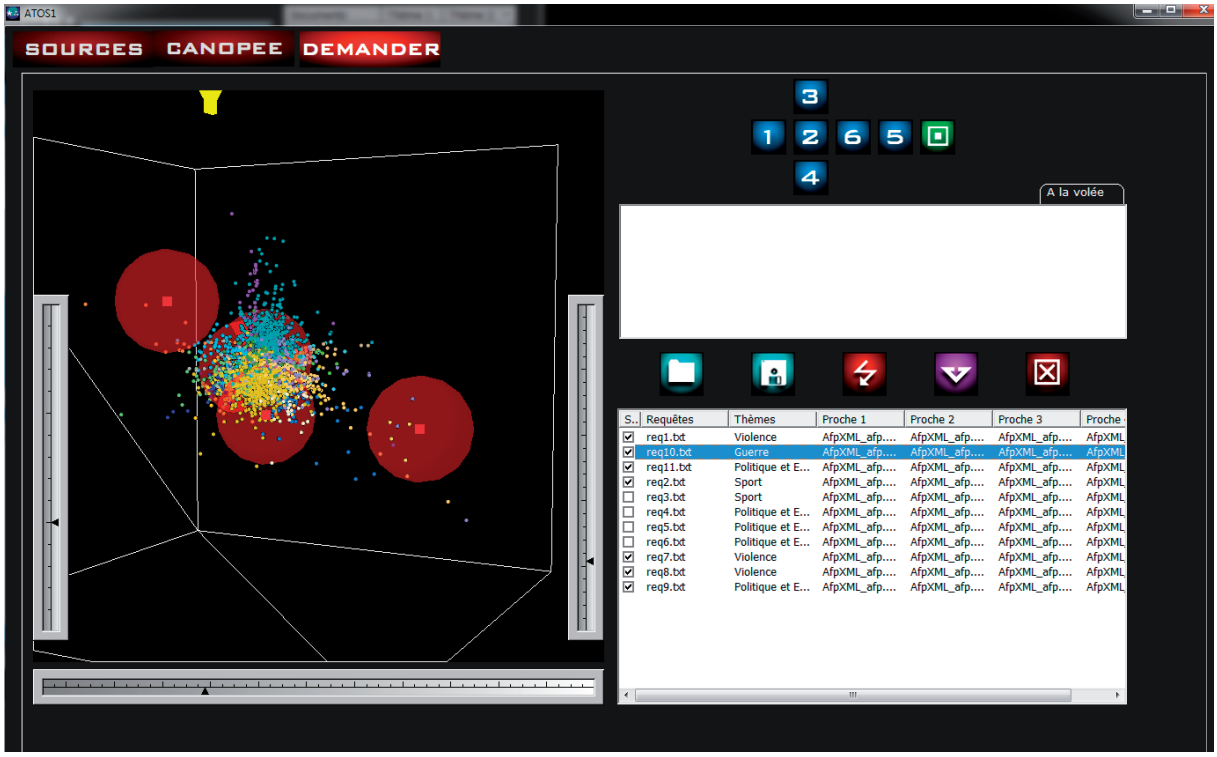

Figure 5 : Interface graphique pour le requêtage multiple en recherche documentaire

D'autres fonctionnalités de Canopée sont encore en cours de développement comme celle de cartographie dynamique de la collection documentaire d'une certaine date à une certaine date en précisant une taille de fenêtre de déplacement.

C'est le cas de la cartographie des ressources. La boucle interactive de l'outil permet des aller-retours entre le corpus et les ressources. Cette fonctionnalité supplémentaire vise à accroitre encore ces aller-retours. Quand on fait une cartographie du corpus en fonction des ressources, on projette les ressources sur le corpus. L'idée serait de faire une sorte de cartographie duale en proposant un « mappage » du corpus sur la ressource. C'est à dire projeter le corpus sur les ressources en créant un espace vectoriel de lexies où chaque vecteur de dimension P (avec P le nombre de documents du corpus) indique la façon dont chaque document porte un nombre d'occurrence de la lexie en question. Il suffit d'appliquer exactement les mêmes traitements ACP que dans la carte initiale. On le fait juste sur la matrice transposée de la matrice initiale : $M_T[i;j] = M[j;i]$

Cet espace vectoriel transposé de l'espace initial est une matrice d'indexation au sens de (Salton et Yang, 1975). Cette cartographie issue de la matrice d'indexation présenterait l'intérêt d'amener l'utilisateur à une réflexion supplémentaire sur ses ressources. Certains thèmes ou certaines lexies sont peut-être peu occurrentes là ou d'autres le sont beaucoup plus et peut-être de manière corrélée avec d'autres thèmes ou lexies. La cartographie des documents visualise des rapprochements entre textes. Ici on vise plutôt des rapprochements entre lexies et entre thèmes. Ce serait complémentaire de l'histogramme d'importance des thèmes (cf. Figure 2). Une cartographie de la ressource permettrait de mettre les thèmes de l'utilisateur dans un espace topologique d'où on pourrait tirer des informations utiles sur l'adéquation globale corpus-ressource en répondant par exemple aux questions suivantes :

- Quelle sont les thèmes les plus centraux dans le corpus ?
- Quels sont les thèmes corrélés par le corpus (proches dans la cartographie) ? ce qu'on pourrait interpréter par « de quoi parle-t-on en général de manière conjointe ? ». De manière duale, quels sont les thèmes qui semblent s'exclure ?

Canopée est clairement un produit ouvert en termes de fonctionnalités additionnelles. La version en cours de finalisation devrait s'enrichir de ces fonctionnalités prochainement.

5. Conclusions

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

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

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

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

Références

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.

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

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

Condamines A. (sous la direction de) (2005). Sémantique et corpus, Hermès, Paris, ISBN : 2-7462-1055-X.

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

Landauer T.K., Foltz P. W. et Laham D., Introduction to Latent Semantic Analysis, dans Discourse Processes, vol. 25, 1998, p. 259-28

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

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

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

Perlerin, V. (2004) Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu, Thèse de doctorat en Informatique – Université de Caen Basse Normandie.

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

Roy T. (2007) Visualisations interactives pour l'aide personnalisée à l'interprétation d'ensembles documentaires, Thèse de doctorat en Informatique – Université de Caen Basse Normandie.

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

Salton G., Yang C.S. (1975) *A vector space model for automatic indexing*, Communication of the ACM, vol. 18 (11), nov., p. 613-620.