ÉMERGENCE PAR NAVIGATION INTERTEXTUELLE : UNE NOUVELLE APPROCHE POUR LA VEILLE STRATEGIQUE

Éric TRUPIN (*), **Maryvonne HOLZEM** (*), **Jean-Luc BOURDON** (**), **Pierre BEUST** (***),
Eric.Trupin@univ-rouen.fr , Maryvonne.Holzem@univ-rouen.fr , Jean-Luc.Bourdon@u-cergy.fr , Pierre.Beust@info.unicaen.fr
Stéphane FERRARI (***), **Youssouf SAIDALI** (*), **Jacques LABICHE** (*)
Stephane.Ferrari@info.unicaen.fr , Youssouf.Saidali@univ-rouen.fr , Jacques.Labiche@univ-rouen.fr

(*) Laboratoire LITIS, EA 4051,
Avenue de l'Université, BP 8, 76801 Saint-Étienne-du-Rouvray Cedex, France.
(**) UFR Sciences et Techniques, Département des Sciences Informatiques,
2 avenue A. Chauvin, BP 222, 95302 Cergy-Pontoise Cedex, France.
(***) Laboratoire GREYC, UMR 6072,
Campus 2, Université de Caen, 14032 Caen Cedex, France.

Mots clefs :
Veille scientifique et technologique, gestion des connaissances, ingénierie des connaissances, modélisation des connaisssances, innovation, collecte d'informations
Keywords:
Scientific and technical observation, knowledge management, knowledge engineering, innovation, knowledge modeling, information gathering
Palabras clave :
Escudriñar científico y tecnológico, administración del conocimiento, ingeniería del conocimiento, innovación, formalización del conocimiento, reunir de información

Résumé

Une approche originale pour la veille stratégique est présentée ici. Dans le cadre particulier de la veille juridique dans le champ du transport, nous développons un système destiné à provoquer l'émergence de concepts lors de la navigation de l'utilisateur dans une base de documents juridiques.
L'utilisateur est amené à utiliser le système dans un cadre professionnel, mais il peut avoir plus ou moins de compétences dans le domaine juridique. La présentation interactive des documents et des différentes portions de texte font l'objet de traitements informatiques et linguistiques qui placent l'utilisateur au centre du processus. Le socle du système est composé d'outils, génériques ou spécialisés, développés par l'équipe lors de différents projets. Ces outils, comme la méthodologie générale et le champ d'application, sont détaillés dans cette communication.

1 Introduction

Lors du récent congrès « Un demi siècle d'Intelligence Artificielle » [17], a été établi le constat qu'il n'existe pas de système intelligent sans homme dans le système. Ce repositionnement amène d'une part à considérer que l'homme et la machine se complètent dans l'action et d'autre part à adopter une démarche expérimentale pour la conception d'un système, notamment pour la veille stratégique.
L'objectif principal vers lequel nous nous orientons est d'offrir à un utilisateur un éventail d'actions le plus large possible sur des données que le système produit. L'idée est de ne pas considérer que les buts des utilisateurs sont fixés *a priori*, mais de laisser libre cours à leur parcours interprétatif dans une base de documents textuels. C'est typiquement le cas lorsqu'un utilisateur est en veille d'une information stratégique pour son activité professionnelle. L'analyse de ce parcours interprétatif d'un utilisateur, dans une situation de recherche d'information, aidera ensuite à la sélection des fonctionnalités retenues pour le système et qualifiera plus précisément le contenu informationnel de la base de documents.
Notre propos se place donc en rupture avec les trois conceptions historiques de l'Intelligence Artificielle (IA) : forte, faible et technologique [2]. Il amène à définir le cadre expérimental nécessaire pour que soit posé un problème qui puisse être résolu et non à tenter de résoudre un problème posé a priori. Les interactions entre personne et système sont donc centrales pour notre approche ; les compétences cognitives des usagers du système seront mises à contribution dans un environnement informatique susceptible de favoriser leur expression dans le cadre du paradigme enactif [18].
Le genre textuel retenu pour notre expérimentation concerne l'information juridique pour le transport international et les enjeux qui s'y réfèrent sont les suivants :

- La conception d'un dispositif et d'un protocole expérimental pour le couplage structurel d'un utilisateur avec un système informatique, au sens enactif ;
- L'extension des facultés cognitives des utilisateurs dans différentes situations de navigation (imagination, sérendipité, raisonnements abductifs, raisonnement par analogie) ;
- La mise à disposition et/ou la création de nombreux outils d'aide à la navigation dans une base documentaire ;
- La mise en évidence des évolutions temporelles et des émergences terminologiques sur ce corpus.

À la lumière d'une étude que nous avons récemment menée sur le cycle de vie du document au sein d'une organisation [7], nous situerons le cadre théorique dans lequel nous aborderons cette veille. Nous définirons le processus de veille comme un processus d'interprétation en contexte organisationnel et présenterons le corpus juridique au sein duquel s'ancrera cette interprétation. Cette communication se consacrera alors à la problématique de l'interprétation assistée de documents numériques. Nous poserons ensuite les bases d'une navigation intertextuelle et enfin, avant de préciser la méthodologie retenue pour notre application et de conclure, nous décrirons les méthodes employées ainsi que les outils déjà développés ou en cours de développement.

2 L'analyse du cycle de vie d'un document au sein d'une organisation : un préalable à la veille

La veille a été définie par l'un de ses pionniers en France comme *le moyen pour l'entreprise de faire émerger les éléments stratégiques de la masse d'information disponible aujourd'hui* [5]. Cette définition nous semble contenir en germe les points de vue théoriques que nous développerons ici. Parler d'émergence nous conduit en effet à nous situer dans une optique non représentationnelle, et donc à ne pas considérer les documents comme des dépôts de connaissances exprimées en termes, ou de candidats termes, du domaine. Avant de détailler notre approche de l'émergence, nous voudrions la situer par rapport aux démarches liées à la fouille de textes dans une perspective de veille.

La recherche de mots pertinents (souvent assimilés aux éléments stratégiques de la définition ci-dessus) oriente la plupart des outils logiciels de Traitement Automatique des Langues (TAL) vers la décontextualisation (extraction), la lemmatisation (élimination de la variation suffixale) et jusqu'à la cartographie basée sur des proximités syntaxiques (morphologie dérivationnelle entre autres) ou sémantiques (relations hyper/hyponymie par marqueurs, par exemple « *est un* »). Les spécialistes de la veille, côté fouille de données ou de textes, parlent alors de KDD (Knowledge Discovery in Databases) ou de KDT (Knowledge Discovery in Texts) [9] comme d'une science qui découvre des connaissances contenues dans des textes. Qu'il s'agisse de fouille de données ou bien de textes, l'essentiel est de trouver des modèles permettant de séparer l'information intéressante (signal pour l'action) du bruit.

Partant du postulat qu'un corpus de documents est un construit en fonction d'objectifs pour la pratique en cours, nous avons abordé [7] la transformation d'un mémoire technique en brevet d'invention à la lumière des apports de la linguistique textuelle [13]. La lignée de réécriture, i.e. le genre dialogique, dans laquelle s'inscrivent les textes conditionne alors grandement la diffusion de leur terminologie [6], tout comme leur interprétation au sein d'une pratique à l'image du discours procédural [1]. En ce qui concerne le terrain de la propriété industrielle, nous nous sommes penchés sur l'activité éditoriale qui consistait à transformer un texte scientifique original et à le rendre compatible avec le fonctionnement du secteur de la propriété industrielle. Une analyse sociolinguistique a alors permis de témoigner d'une montée en généricité (termes déjà connus du domaine et ayant une couverture sémantique assez large pour protéger l'invention), perceptible sur le plan lexical comme accompagnement d'un processus de légitimation sociale (faire acte dans le domaine de la propriété industrielle). De ce point de vue, la recherche de mots exceptionnels, de sigles, de co-occurrences de termes scientifiques dans un processus de veille sur ce type de corpus, serait vouée à l'échec : le travail d'interprétation de l'ingénieur brevet consistant justement à *encapsuler* d'une certaine manière ce vocabulaire spécifique tout en structurant le document d'un point de vue argumentatif.

De ce point de vue, concevoir le document comme une simple structure de communication, réceptacle d'informations à partager, autorise certes à le modéliser dans un format rigide, totalement prévisible et déterminé, mais occulte le rôle prédominant de l'usage comme véritable expérience au sens phénoménologique du terme : c'est-à-dire comme processus partagé à l'œuvre dans le travail d'interprétation d'un utilisateur. Cette problématique de la modélisation dynamique de l'expérience, en débat au sein de la communauté de la pensée complexe [10] qui s'appuie, d'après Husserl [8], sur l'expérience du temps et la dynamique des attentes, nous semble particulièrement appropriée pour rendre opératoire ce processus émergeant et commun aux acteurs logiciels et humains dans le cadre d'une veille. C'est le point de vue que nous tenterons de développer après avoir présenté le corpus de notre étude.

3 Le corpus : la documentation juridique du transport

Il s'agit ici de la base documentaire de l'Institut du Droit International du Transport (IDIT) accessible en ligne. Elle est stockée dans une base MySQL, et son interrogation se fait via le système d'information (SI) de l'IDIT codé en PHP. Elle s'adresse à des adhérents spécialistes du droit mais elle est difficilement utilisable par un novice, comme un transporteur, qui y chercherait des informations particulières, bien que stratégiques pour la mise en place de conditions de transport de marchandises idoines à la législation en vigueur.

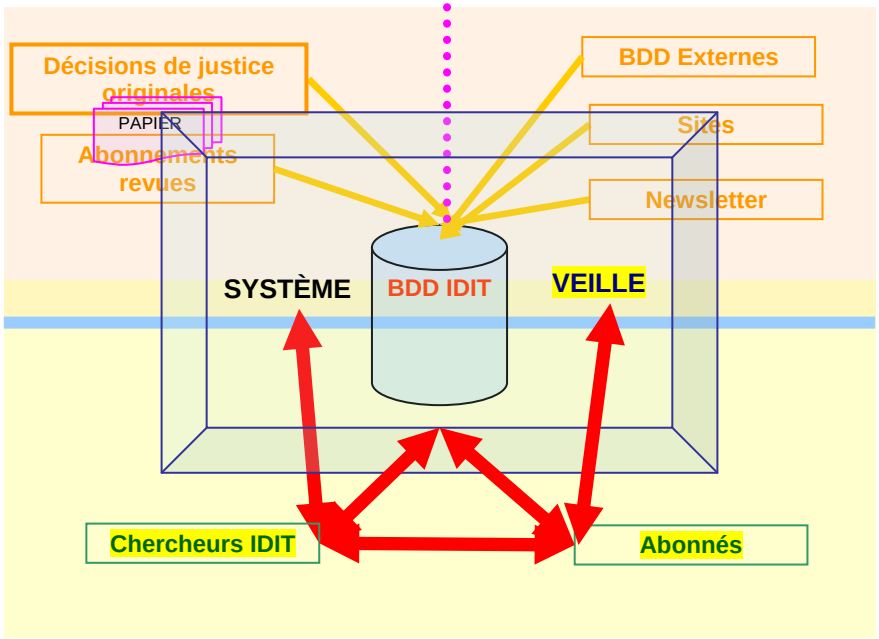

Figure 1 : Organisation de la base documentaire de l'IDIT.

Cette base documentaire est associée à un thésaurus hiérarchisé « maison » pour améliorer son interrogation. Elle est renseignée manuellement à partir de revues ou après interrogation d'autres sources de données en ligne auxquelles l'IDIT a accès. Elle impose aussi la saisie manuelle de compte-rendus (CR) de cours d'appel, de jurisprudence et d'arrêts sous la forme de fiches. Cette captation de l'information et la veille représentent deux difficultés majeures pour renseigner et mettre à jour le système d'information. Notre objectif est de tenter d'automatiser certaines tâches de recherche d'information en fonction d'une analyse fine du corpus : typologie de la structure argumentative en lien avec la présentation matérielle des données et les attentes d'adhérents hétérogènes (juristes, transporteurs, …). Le SI est validé à des fins de diffusion vers les adhérents afin que ces derniers puissent gérer dans les meilleures conditions leurs entreprises et sécuriser leurs activités. Les acteurs du transport et de la logistique se doivent de rechercher et d'analyser des informations de plus en plus nombreuses. Ainsi le suivi et l'anticipation des cadres juridiques communautaire, législatif et réglementaire sont des éléments de gestion incontournables. Or, la fragmentation de l'information relative au droit des transports et de la logistique qui couvre des domaines aussi variés que le droit commercial, le droit des sociétés, le droit de l'environnement, le droit administratif, le droit pénal, le droit social, rend difficile l'accès à l'information (information éparse, réglementation pléthorique, accès difficile et coûteux…), d'où la nécessité d'une mise en relief de celle-ci.

Deux éléments nécessaires se distinguent alors : un outil de veille globale pour renseigner le SI et un outil de diffusion. Le premier doit permettre, par exemple, de faire un état des lieux hebdomadaire selon un domaine précis (avec une requête) par interrogation automatique (à l'aide d'un moteur de recherche). Le second doit permettre une veille personnalisable sur le SI pour un adhérent spécialiste, ou non, du droit. C'est sur ce second aspect que nous nous focaliserons ici. Il s'agit de créer des alertes pour les adhérents hétérogènes de la base dans le domaine de la sureté dans le transport pour les compagnies d'assurance de logistique. Une étude de faisabilité portée par le pôle de compétitivité Logistique Seine Normandie démontre la nécessité de la mise en place d'un tel système, avec un spectre de recherche à

la fois traditionnel (droit des transports) s'élargissant à des matières annexes s'inscrivant dans le développement des activités de prestations logistiques (droit de l'environnement, droit douanier, dématérialisation documentaire).
Si les fiches qui enrichissent cette base émanent de quatre sources : jurisprudence, article, texte et fonds documentaires, c'est à partir des arrêtés de jurisprudence que l'information est la plus dense. C'est donc à partir de ce corpus de fiches que nous testerons notre contribution à la veille. Nous présentons et commenterons ci-dessous les rubriques à partir desquelles s'organisera la navigation intertextuelle :

Thèmes :
Criminalité dans les transports

Date de la décision : 05/04/1993
Mode de transport : transport aérien
Pays : France
Objet :
Embarquement d'un fret interdit (bateau zodiac avec un moteur de hors bord, fusées de détresse, spray de peinture) - Faux étiquetage et fausse déclaration sur la LTA - Infraction d'entrave à la circulation aérienne -
Sommaire :
La sécurité de l'aviation civile exige un respect absolu des formalités d'embarquement du fret à bord d'un avion-cargo; le fait pour des agents d'une compagnie de tromper volontairement le commandant de bord d'un avion sur la nature d'un chargement constitue une entrave à la navigation aérienne.
Référence :
Cour d'appel de Paris 12e ch. 5 avril 1993
Ministère public et Air France c/. M. Pinot, Geromini et autres
Observation :
Décision intégrale disponible à l'IDIT

Fiche n°1

Si les intitulés des rubriques sont très réguliers, leur contenu varie sensiblement d'un point de vue langagier. La rubrique « *Thèmes* » peut en effet comporter un seul syntagme nominal (Fiche n°1), comme rendre compte, à la façon d'un langage documentaire, de toute une hiérarchie d'items comme dans l'exemple ci-après (Fiche n°2) :

Thèmes :
Accident routier

CMR (Transport routier international)
  prescription (art. 32)
    faute lourde - délai 3 ans
Commissionnaire de transport
  qualification du commissionnaire de transport
Expéditeur
  responsabilité
Faute lourde
  fait constitutif
Matières dangereuses
  responsabilité du chargeur
Matières dangereuses
  responsabilité du transporteur

Fiche n°2

De même, les rubriques « *Objet* » et « *Sommaire* » sont-elles variables en contenu, présentation et longueur. La répartition entre ces deux rubriques n'est pas du type thème (le sujet abordé), rhème (l'actualisation du sujet en contexte), mais plutôt du type résumé/développement étant donné que les développements en *sommaire* sont annoncées (et même numérotés) de façon identique dans l'*objet*.
La seule différence remarquable entre ces deux rubriques est d'ordre typographique puisque les tirets entre items de l'*objet* sont remplacés par des points de fin de phrases dans la rubrique *sommaire*. Nous avons ainsi constaté que la structuration initiale des jugements de la cours d'appel sur lesquels s'appuie cette jurisprudence n'était pas respectée lors de la réécriture des fiches puisque les appréciatifs du type « *traitement désinvolte de..* », *« insuffisance du système de…* » qui figurent après l'exposé du litige dans les actes du jugement rendu, se trouvaient dès les premières lignes dans la rubrique *objet*.
Les rédacteurs des fiches ont ainsi voulu répondre au plus vite aux attentes pressenties des adhérents de la base vis-à-vis d'une information capitale en droit : la qualification des faits par le tribunal. C'est en effet à partir de cette catégorisation (la dénomination de l'acte) que peut intervenir le jugement. À de nombreuses reprises les rédacteurs des fiches ont également eu recours à l'opposition de parenthèses contenant un affirmatif ou une négation juste après la dénomination juridique, objet même du procès. C'est le cas, par exemple, de : « *faute inexcusable (non)* » - « *responsabilité du… (non)* » etc. Là encore, les rédacteurs ont eu le même souci d'efficacité et de concision dans l'information donnée. Malheureusement ces règles de réécriture ne sont pas systématiques. Un exemple relevé dans cette même zone : « *exclusion du droit du transporteur à …* » alors que l'on aurait pu s'attendre à « *Droit du transport à … (non)* » ou bien « *faute lourde du transporteur résultant de l'inobservation des indications figurant sur les étiquettes mettant en évidence la fragilité de la marchandise et la position des colis* », périphrase longue qui gagnerait sûrement à être réécrite comme ci-dessus. Qu'elle soit réécrite en langage documentaire, en périphrase, syntagme fleuve sans verbe conjugué, ou phrase complète d'un point de vue grammatical comme : « *Qualité de commissionnaire de transport de la société qui n'a pas émis les connaissements directs mais a organisé le transport terrestre* » ; la zone *objet* nous semble être celle à partir de laquelle devrait s'initier la navigation intertextuelle, étant donné ses reprises et développements vers les autres zones de la fiche.

4 Émergence conceptuelle par la navigation intertextuelle

Nous expliciterons ici notre démarche en prenant comme exemple une navigation à partir des notions « d'étiquette » ou (le seul opérateur booléen actuellement disponible sur le site de l'IDIT) « d'étiquetage »[1] . À partir d'une première requête d'un adhérent, portant par exemple sur le domaine[2] de l'étiquetage, une recherche plein texte sélectionnerait les fiches contenant un des éléments du champ lexical. Les mots *étiquette, étiquetage, étiqueter* (morphologie dérivationnelle) sont alors recherchés, mais nous pourrions poursuivre la sélection des fiches en y ajoutant les sémèmes des supports matériels de l'étiquette dans le domaine du transport : *porte, conteneur, colis*, etc. ainsi que la classe des sémèmes pouvant ou devant figurer sur ces étiquettes : *code postal, code barre, produit dangereux, inflammable* avec les sèmes spécifiques obligatoires (faute) non obligatoire (erreur) permettant de les distinguer. La construction de telles classes sémantiques peut sembler fastidieuse mais elle pourrait, d'une part, être construite au fur et à mesure en lien avec la navigation d'un utilisateur dans la base et, d'autre part, se révèlerait rapidement opérationnelle selon nous.
Une fois les fiches sélectionnées, seul le champ *objet* de chaque fiche apparaîtrait ce qui, en limitant la quantité de textes à lire, permettrait à l'utilisateur de pouvoir naviguer au sein d'un corpus d'*objet* eux-mêmes liés à des modes de transports (les taxèmes maritime/routier/ferroviaire pouvant être à ce niveau discriminants).

[1] Sur les 28 fiches sélectionnées, deux seulement comportent les deux items étiquette et étiquetage.
[2] En sémantique textuelle les domaines sont constitués d'un groupe de taxèmes lié à une pratique sociale. Le domaine est commun aux divers genres (textuels) propres au discours qui correspond à cette pratique. Le taxème est quant à lui la classe sémèmes (ou plus petite unité de signification) reflet de pratiques concrètes comme « métro-train-autobus-autocar » relèvent du domaine du transport moyens collectifs articulé en deux taxèmes : ferré et routier eux-mêmes différenciés par les sèmes spécifiques intra/extra-urbain.

C’est à partir de la lecture de ces différentes zones « objet » issues de fiches échelonnées dans le temps que l’utilisateur serait alors invité à se construire sa propre représentation du sens et pourra alors, par l’extraction qu’il opérera entre les différents items, construire sa propre hyphologie[3] du domaine pour les besoins de sa pratique en cours.

5 Méthodes et outils

Nous nous situons dans le contexte très large de la présentation et de la recherche d'information par navigation ainsi que dans celui du traitement automatique de la langue naturelle pour lequel il existe de multiples outils opérationnels. Notre plate forme mettant en oeuvre ces outils doit être munie d’une interface permettant de choisir différents modes de présentation des données textuelles et autorisant une navigation dans une base de documents. Notre positionnement entend s'appuyer sur leur opérationnalité pour inventer de nouvelles pratiques de navigation dans un corpus textuel. Il s’appuie également sur nos résultats de recherche présentés dans ce qui suit.

Nous avons déjà développé une plate forme nommée ACTI_VA [16] permettant l’acquisition et la valorisation de connaissances dans le domaine du traitement d’images de documents. Cette plate forme repose sur l’exploitation d’une bibliothèque d’outils à chaîner lors du processus complexe d’interprétation d’image. L’originalité du modèle d’ACTI_VA en termes d’interaction réside dans la présentation de l’enchaînement des outils avec une gestion des historiques. Cette présentation permet à des utilisateurs, dont les univers de référence sont différents, d’effectuer des modifications dans leur parcours interprétatif d'image à plusieurs niveaux. L’utilisateur a donc la possibilité de tester plusieurs modes de présentation et n’a pas à reformuler l’ensemble de sa requête en cas de résultat non pertinent, tout en gardant les traces des scénarios déjà joués. Cette plate-forme permet l’accès à une grande masse d’information capitalisée pour présenter un ensemble pertinent de documents en réponse à une requête utilisateur. Cette plate-forme est dédiée à l'interprétation d'image mais servira de base au développement de notre dispositif expérimental tout en soulignant les problèmes non résolus que son expérimentation a montrés :

- L’évaluation du résultat dépend essentiellement de l’utilisateur (pour un utilisateur ayant un besoin d’information, quels que soient ses critères, sa requête est souvent imprécise, il en résulte que l’objet de sa recherche est a priori inconnu) ;
- L’appropriation de champs scientifiques (traitement d’images initialement pour ACTI_VA) pour plusieurs types d’utilisateurs dont les référentiels sont différents.

Certaines pistes à explorer sont d'ores et déjà envisagées comme :

- La présentation par regroupement de grands ensembles documentaires ;
- Des outils graphiques d’aide à la formulation et création des requêtes complexes ;
- Les techniques algorithmiques de reformulation au vu de restituer les documents pertinents ;
- La présentation graphique des données textuelles, résultats de la recherche ;
- Une présentation des documents textuels réellement interactive au sens de la perception active.

Divers outils de TAL peuvent être utilisés et/ou éventuellement intégrés dans cette nouvelle plateforme inspirée de ACTI_VA. Nombre de ces outils permettent en effet des analyses fines de corpus pour l’extraction de termes ou de relations, l’étiquetage de données, l’acquisition de classes sémantiques, l’analyse en dépendance

[3] En référence à la théorie du texte de Roland Barthe. Hyphologie renvoie à la toile, le tissu, la trame (origine étymologique de texte textualité texture) au sein de laquelle se dissout l’auteur en attendant une réinterprétation par un utilisateur en contexte.

fonctionnelle, etc. Cependant la conception de ces outils est sous-tendue par des choix théoriques que l'utilisateur adopte d'une façon plus ou moins consciente et qui oriente son activité.

L'approche enactive dans la conception et l'intégration d'outils de TAL marque une différence de point de vue avec les méthodes compositionnelles classiques en favorisant une démarche scientifique expérimentale. Ainsi plutôt que de considérer une conduite de projets de recherche en TAL sur le schéma classique « modélisation → implémentation → tests → évaluations comparatives », nous préférons mettre en avant l'expérimentation comme une boucle de conception où la modélisation n'est pas une étape initiale, pas plus que les évaluations (non nécessairement comparatives) ne sont des étapes finales. L'objectif ici n'est pas de chercher à faire mieux certaines tâches déjà réalisées avec des méthodes éprouvées, mais il est plutôt question de chercher à inventer de nouveaux usages du TAL dans les interfaces Homme-machine (IHM) ainsi que de nouvelles façons d'utiliser des ordinateurs dans des recherches sur le langage.

Il ne s'agit donc pas comme dans la plupart des systèmes de Traitement Automatique des Langues de proposer une fonctionnalité complexe (extraction de termes, de relation, classification automatique, annotation automatique) orientant le parcours interprétatif du lecteur sur le texte, mais plutôt d'utiliser des fonctionnalités élémentaires qui permettent à l'utilisateur de faire émerger des fonctionnalités de plus haut niveau par combinaison. Des fonctions atomiques proposées émergent alors de nouvelles fonctionnalités actualisées par l'interaction entre les utilisateurs et le corpus. L'idée est de proposer à l'utilisateur des rapprochements de contextes syntagmatiques (alignement) et de le laisser en inférer des classes sémantico-lexicales (apparaissant selon des patrons, des contextes particuliers).

Les fonctionnalités atomiques sont de plusieurs types et correspondent aux niveaux de traitements linguistiques possibles, c'est-à-dire les niveaux phonétique, morphologique, syntaxique, sémantique et pragmatique pour le parcours d'un texte. Sur chacun de ces niveaux il est possible d'effectuer des traitements atomiques basés sur des analyses manuelles ou des analyses statistiques. Les analyses manuelles étant basées le plus souvent sur des règles linguistiques définies sur analyse d'un corpus, il paraît plus cohérent d'opter pour des traitements statistiques non supervisés, bruités certes mais moins dépendants de choix théoriques a priori.

- opérations au niveau phonétique/phonologique (proximité phonétique) :
    - syllabation, longueur de mots ;
    - mise en correspondance phonème / graphème ;
    - anagramme, allitération ou assonance ;
- opérations au niveau morphologique (proximité morphologique) :
    - racinisation (stemming), étymologie ;
    - flexion (suffixation), dérivation (affixation) ;
    - combinaison, lexie simple ou complexe ;
- opérations au niveau syntaxique :
    - collocation, co-occurrence ;
    - dépendance fonctionnelle.

Les opérations plus complexes donneront lieu à un étiquetage de la part de l'utilisateur. Par exemple un étiquetage des relations paradigmatiques. Cet étiquetage permet de générer des patrons liés à différents paliers d'analyse linguistique. Notre perspective, liée à la sémantique interprétative, nous invitant à relier ces différents paliers :

- au niveau syntaxique :
    - patron ;
- au niveau sémantique :
    - étiquetage de / classification selon des relations syntagmatiques :
        - hyponymie, méronymie, holonymie, synonymie, antonymie, isonymie ;
    - étiquetage de / classification selon des relations paradigmatiques :

        - relations prédicatives (action objet, fonctions, agent, application…), relations script (ou scénario), relations type analogique, relations type lexical, relations liens personnels ;
    - étiquetage de / classification selon les Fonctions Lexicales (FLs) de la TST
- au niveau pragmatique :
    - présentation selon un étiquetage (format de la notice, position dans le texte) ;
    - mise en relation entre les utilisateurs authentifiés ;
    - mise en relation des productions des utilisateurs, des utilisateurs et des notices ou d'autres paramètres.

On pourra utiliser ici des outils comme l'analyseur morpho-syntaxique de Jacques Vergne [19], la plateforme de prototypage de chaînes de traitement LinguaStream [20].

Dans le but de faire émerger de nouveaux usages dans le domaine de la recherche d'information et de la consultation de corpus, nous nous sommes intéressés à la navigation intertextuelle dans un environnement de cartographie interactive de corpus. Ces logiciels (LUCIA [12], ProxiDocs [15], ThemeEditor [3]).sont centrés utilisateurs dans la mesure où ils tiennent compte avant tout des spécificités socio-linguistiques de leurs utilisateurs (par exemple leurs centres d'intérêt, leurs habitudes terminologiques).

Le but du logiciel d'étude ProxiDocs est de plonger son utilisateur (ou un petit groupe d'utilisateurs) dans des interactions qui offrent la possibilité de notamment mieux discerner l'homogénéité thématique d'un corpus, de mettre en évidence sa densité, d'en extraire les principales tendances thématiques et de permettre un accès rapide à tel ou tel document ou passage de document. Au sein de ces interactions l'outil permet de produire et de naviguer dans des représentations graphiques personnalisées que l'on appelle des cartes. En préalable, l'utilisateur décrit un ensemble de termes qui sont ceux qui l'intéressent et les donne en entrée au logiciel ainsi que son (ou ses) corpus. Ces termes peuvent être représentés selon deux modèles, soit sous forme d'une liste de graphies, soit sous la forme d'un dispositif de représentation sémique différentielle des significations des termes. À ces deux formes de ressources lexicales correspondent des outils interactifs qui permettent à des utilisateurs de les constituer de manière incrémentale : l'outil ThemeEditor pour l'extraction interactive de classes thématiques de graphies à partir de textes et l'outil VisualLuciaBuilder [12] pour la construction en interaction de ressources terminologiques componentielles et différentielles. Avec ces ressources, ProxiDocs construit des cartes dynamiques et interactives (en 2 ou 3 dimensions statiques ou bien animées) ainsi que des visualisations des textes agrémentées d'une visualisation par coloriage des isotopies.

D'un point de vue expérimental, il s'agira de chercher à savoir comment ce type d'interface et le couplage qu'il induit permettent l'émergence par enaction d'une perception sémantique du corpus. Si on considère que le sens d'un texte provient de l'expérience de l'interprétant face à ce texte et à son intertexte, alors de toute évidence des sujets différents sont amenés à déceler différents sens étant donné que leurs expériences sont différentes. Suivant cette remarque, on peut tout à fait considérer que « l'expérience » qu'une machine fait d'un document (par exemple dans une tâche d'un traitement statistique dans le but d'une cartographie ou encore une indexation automatique) est une certaine forme de sens.

Dans le couplage personne-système les interprétations des utilisateurs et des machines ne sont pas en concurrence car l'une n'a en aucun cas le but de supplanter l'autre. Au contraire, nous les pensons comme complémentaires dans le sens où l'interprétation d'une machine a pour objectif de produire dans l'interaction des traces qui vont participer aux interprétations du ou des utilisateurs. Nous suivons bien ici l'idée de Dionisi et Labiche [4] qui consiste à caractériser des « processus logiciels » impliqués dans des « processus expérientiels », eux-mêmes impliquant des « processus cognitifs ».

Enfin, nous avons également développé une aide en ligne basée sur des fiches de description terminologique. Nous nous proposons de reprendre le modèle de ces fiches qui permettent de fixer les termes dans un environnement sémantique. Leur originalité tient d'une part, à la finesse des relations décrites (par exemple, l'isonymie qui procède par contraste minimal pour saisir sous un même hyperonyme, les traits distinctifs de sens entre unités), mais surtout, aux relations prédicatives :

- {obj typ} : l'objet typique est l'objet sur lequel porte l'action ;

- {fonct typ} : le rôle joué par une entité au sein d'un ensemble ;
- {action typ} : l'action typique réfère à l'action produite ;
- {ag typ} : l'agent typique est celui qui permet ou effectue l'action sur l'objet ;
- {appl typ} : l'application typique est le but dans lequel l'action est produite.

Celles-ci, au nombre de cinq dans l'aide en ligne d'ACTI_VA, peuvent être modulées pour les besoins des utilisateurs et du type de Base de connaissances que nous voulons construire.

6 Conclusion et perspectives

L'interprétation de documents numériques dans le cadre de systèmes informatiques ne doit pas être envisagée comme un traitement autonome, indépendant du cadre spatio-temporel des documents. Quelles que soient les visées applicatives d'un système d'aide à l'interprétation de documents numériques (interface de lecture rapide, accès au contenu, indexation, navigation, archivage, veille …), il convient d'analyser et d'exploiter le plus finement possible les différents aspects sémiotiques des documents numériques tels que leur dimension intertextuelle (mettant en évidence la complexité des rapports local/global), leur caractère acté car liés à une pratique, leur inscription et leurs évolutions dans le temps ou encore leur nature multimodale. Dans cette perspective, la sémantique interprétative de [14] et la sémiotique de Peirce [11] sont des ancrages théoriques intéressants. Dans la conception de systèmes d'aide à l'interprétation ou de systèmes d'information documentaire, il convient également et surtout de donner une place centrale à l'utilisateur (voire à un groupe d'utilisateurs) en prenant en compte des spécificités socio-linguistiques telles que son domaine d'activité, ses centres d'intérêt, son lexique et ses terminologies. Ces spécificités ainsi que les rapports entre sujets interprétants et documents manipulés rendent nécessaire plus d'interaction pour un meilleur couplage système/utilisateur. L'interprétation, et l'aide à l'interprétation, sont vues ici comme des activités sémiotiques complexes et nécessairement interactives car perpétuellement en aller-retour entre le développement de ressources termino-ontologiques, l'analyse de corpus et la visualisation de données dans le but de faire sens pour l'utilisateur.

On est ici dans le cadre d'une interprétation temporellement différée entre producteur et utilisateur de document. L'intention du producteur et l'interprétation de son discours écrit cessent de coïncider (car liées en particulier à des contextes temporels différents) et il ne peut y avoir réglage de sens co-référentiel comme dans un dialogue Homme-Homme. Pour ces raisons, le cadre émergent de l'enaction [18] semble constituer un paradigme tout à fait adapté pour analyser l'interprétation dans un couplage système/utilisateur. Dans une approche enactive de l'interprétation des documents nous cherchons à penser différemment les systèmes à concevoir. Plutôt que de viser un usage particulier (qui en général est contourné par les utilisateurs qui en inventent de nouveaux), nous pensons préférable ne pas contraindre au préalable les besoins et les finalités des utilisateurs car justement, de manière enactive, ils se définissent dans l'interaction et c'est en partie cela qu'il est particulièrement intéressant d'étudier.

7 Remerciements

Nous remercions tout d'abord M. Ludovic Couturier (Directeur administratif de l'IDIT) pour sa disponibilité et son investissement dans la mise en place de la collaboration avec le LITIS.

Nous tenons à remercier également le réseau STIC-SHS du Pôle Universitaire Normand (PUN) qui a permis la mise en relation des chercheurs du GREYC et du LITIS pour élaborer des projets communs, et notamment celui dont cet article se fait l'écho.

8 Bibliographie

[1] ADAM J-M., *Types de textes ou genres de discours ? Comment classer les textes qui disent de et comment faire ?*, Langages, n° 141, 2001, p. 10-27

[2] BACHIMONT B., *L'artéfacture entre herméneutique de l'objectivité et de l'intersubjectivité : un projet pour l'intelligence artificielle*, In Herméneutique : textes, sciences, PUF Philosophie d'aujourd'hui, Paris, 1997

[3] BEUST P., *Un outil de coloriage de corpus pour la représentation de thèmes*, 6èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT 2002), Saint Malo (France), Mars 2002, p. 161-172

[4] DIONISI D. et LABICHE J., *Enaction et informatique : les enjeux de l'opérationnalisation technologique d'une théorie de la cognition*, Colloque ARCO, Paris, 2006, 13 pages

[5] DOU H., *La veille technologique*, La Recherche, numéro spécial, 1994

[6] GAMBIER Y., *Problèmes terminologiques des pluies acides : pour une socioterminologie*, Meta, 32(3), 1987, p. 314-320

[7] HOLZEM M., DIONISI D., LABICHE J., TRUPIN E., *Le Document dans son agir organisationnel : le modèle de l'organisation dans l'interaction usager système*, dans ZREIK. K., ed, Document Electronique Dynamique : Le multilinguisme, Actes du huitième Colloque International sur le Document Electronique : CIDE.8, 25-28 mai 2005 Beyrouth, p. 133-154

[8] HUSSERL E., *Leçons pour une phénoménologie de la conscience intime du temps,* PUF, 1930, édition de 1991

[9] KODRATOFF Y., *Knowledge discovery in texts : A definition and applications, in Foundation of Intelligent systems*, Ras & Skowron (eds.), Lecture Notes in Artificial Intelligence, n° 1609, Springer-Verlag, 1999, p. 16-29

[10] LERBET-SERENI F. et all, *Expérience de la modélisation et modélisation de l'expérience,* L'Harmattan, 2004

[11] MARTY R., *L'algèbre des signes, Essai de sémiotique scientifique d'après C.S. Peirce*, Collection "Foundations of Semiotics", John Benjamins, Amsterdam/Philadelphie, 1990, 406 pages

[12] PERLERIN V., *Sémantique légère pour le document - Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Doctorat en informatique de l'université de Caen / Basse-Normandie, 7 décembre 2004, 272 pages

[13] RASTIER F., *L'action et le sens pour une sémiotique des cultures*, Journal des anthropologues, n°85-86, 2001, p. 183-219

[14] RASTIER F., *Pour une sémantique des textes théoriques*, Revue de Sémantique et Pragmatique, 17, 2005, p. 151-180

[15] ROY T. et BEUST P., *ProxiDocs, un outil de cartographie et de catégorisation thématique de corpus*, 7èmes Journées internationales de l'Analyse statistiques des Données Textuelles (JADT 2004), Louvain-la-Neuve (Belgique), Mars 2004, p. 978-987

[16] SAIDALI Y., TRUPIN E., HOLZEM M., BAUDOUIN N., *Pour une aide à l'interprétation de connaissances traiteurs d'images : une approche terminologique*, EGC 07 : 7èmes journées francophones : Extraction et gestion des connaissances : atelier ECOI, Institut d'informatique FUNDP 23-26 janvier 2007, Namur, Belgique, p. 25-37

[17] SEBAG M. (sous la direction de), *Un demi-siècle d'Intelligence Artificielle*, http://afia.lri.fr/node.php?node=1175, Ministère de la Recherche, Paris, Novembre 2006

[18] VARELA F., *Invitation aux sciences cognitives,* Point Seuil, 1996

[19] VERGNE J. et GIGUET E., *Regards Théoriques sur le "Tagging"*, in Actes de la cinquième conférence Le Traitement Automatique des Langues Naturelles (TALN 1998), Paris, France, 10-12 juin 1998

[20] WIDLOCHER A. et BILHAUT F., *La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus*, Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN), Dourdan, 2005