Approches Sémantique du Document Electronique

Actes du septième Colloque International sur le Document Électronique : CIDE.7
22 – 25 juin 2004, La Rochelle, France

Dans le cadre de la Semaine du Document Numérique

Coordinateurs :

Patrice ENJALBERT
Université de Caen

Mauro GAIO
Université de Pau

Préface

Pour sa septième édition, CIDE s'intègre dans la semaine du Document numérique (SDN). La SDN est une manifestation tout à fait exceptionnelle consistant à regrouper en un même lieu 20 manifestations scientifiques portant sur l'objet « Document numérique ». Il s'agit là d'une volonté du Réseau Thématique Pluridisciplinaire Document du département STIC du CNRS de fédérer les activités portant sur le document dans le monde francophone autour de ces manifestations (conférences, workshop, journées, ateliers…) sur une semaine. L'objectif est avant tout de provoquer des échanges entre communautés scientifiques travaillant sur un même objet dans des disciplines aussi diverses que les informaticiens, historiens, linguistes, cogniticiens, psychologues, pédagogues… Cet objectif vise à renforcer les liens interdisciplinaires entre les individus afin qu'émergent des méthodologies innovantes s'appuyant sur des outils, des usages et des concepts complémentaires aux disciplines.

Pour chacune des éditions de `CIDE` le comité de conférence choisi de focaliser le débat sur une thématique particulière qui, au-delà des communautés établies, des disciplines scientifiques et des démarches spécifiques, témoigne des problèmes et des enjeux de cette objet de recherche : le document électronique.

Pour cette édition, le thème des `approches sémantiques du document électronique` a été retenu. Nous reprendrons ici les attendus de l'appel à communication :

> *La mise en avant du « sens » a en effet longtemps été regardée avec beaucoup de scepticisme au profit de traitements dits « de surface », s'attachant à « la forme » par opposition au « contenu ». Cette perception est en train de changer. Des progrès significatifs ont été réalisés au cours des dernières années, d'abord sur le document textuel (extraction d'informations, question answering, résumé automatique...), puis relayés de plus en plus dans les autres médias (extraction d'information et indexation de documents sonores et vidéo par le contenu, résumé d'oeuvres...). Par ailleurs, les travaux déployés autour du thème du « web sémantique » visent à décrire le contenu des documents ou ressources de toutes sortes de manière à les rendre accessibles et interopérables.*
>
> *Un autre point de vue, plus radical, serait de considérer que même les traitements dits « de surface » ou « numériques » sont en fait, à y bien regarder, sémantiques. Si le « sens » ne se réduit pas à « l'information », produire de l'information, n'est-ce pas produire du sens ? Un désambiguïseur utilisant une méthode statistique, même si la méthode ne se réclame d'aucune*

théorie linguistique, résout bien une ambiguïté sémantique lexicale. Un segmenteur thématique va repérer des récurrences lexicales que d'autres appelleront isotopies. Un extracteur de descripteurs thématiques produit bien ce sens minimal : « de quoi parle ce document », etc.

On le voit, l'appel était volontairement ouvert, laissant aux auteurs toute latitude pour décliner à leur guise le terme « sémantique », en fonction de leurs propres objectifs et méthodes de recherche.

Il en est résulté 31 propositions de communications, dont 15 ont pu être retenues. Les présentations ont été réparties en 5 sessions dont 1 commune avec la conférence CIFT : *Structuration de documents, Analyse textuelle, Organisation et exploration de corpus documentaires multimédia, Sémantique linguistique et applications documentaires et Structuration de documents et recherche d'information (commune avec CIFT)*.

Trois conférences de synthèse sont également au programme. La première, proposée par *Hugues Vinet (Directeur du département Sciences et technologies du son et de la musique, Ircam-CNRS)* a pour objet de dresser un état de l'art synthétique des recherches en matière de description automatisée de document musicaux et sonores. Des exemples de travaux récents viennent appuyer le propos, en particulier issus des recherches de l'Ircam et du projet européen CUIDADO (Content-based Unified Interfaces and Descriptors for Audio/music Databases available Online), s'appuyant sur un modèle des niveaux de représentation des informations musicales.

La seconde conférence est faite par *Georges Vignaux, (Directeur de Recherche au CNRS, Laboratoire Communication et Politique)* et intitulée « Du Corpus à l'Hypertexte ». S'appuyant particulièrement sur le programme CoLiSciences, qui aboutit aujourd'hui à la mise en ligne d'un grand corpus balisé de textes scientifiques « historiques », G. Vignaux interroge les notions d'hypertextualité, de corpus, de parcours de lecture, l'impact des technologies… dans une réflexion sur la construction du sens par un « lecteur » confronté aux nouvelles formes documentaires, numériques et hypertextuelles.

La troisième est une tentative de notre part pour synthétiser l'ensemble des communications de `CIDE` et les resituer dans les perspectives générales évoquées dans l'appel à communications. Entreprise difficile ! Mais qui révèle, nous semble-t-il, un champ de recherche particulièrement riche, un ensemble de convergences réelles, parfois saisissantes, toujours stimulantes. Bref : un lieu d'émergence d'idées nouvelles et prometteuses pour le document numérique.

Finalement, nous avons tenu à organiser une Table Ronde, avec la participation de représentants d'autres communautés (ATALA, CIFT, Terminologie…) de manière à mener collectivement ce travail de synthèse.

Nous remercions les chercheurs, dont la diversité des origines témoigne depuis la création du colloque de son assise pluridisciplinarité, d'avoir choisi `CIDE.7` et de permettre ainsi de faire vivre un lieu d'expression et de débat au-delà des communautés établies.

Nous saluons le travail du Comité de Programme, pour sa participation à l'élaboration du projet scientifique de CIDE 7 et son travail d'évaluation des communications et de mise au point du programme, ainsi que l'ensemble des collègues ayant contribué à la lecture des articles soumis.

Nos remerciements vont aussi aux organismes qui ont permis CIDE.7 : l'université de Caen, les laboratoires GREYC, LIUPPA et Paragraphe, les sociétés MEMODATA, et JOUVE, le CNRS, le programme ComSciences du Poitou-Charentes, et sans oublier le RTP DOC et les organisateurs de la Semaine du Document Numérique.

Et enfin, nous voulons dire que CIDE.7 DOIT beaucoup à plusieurs personnes de l'université de Caen, en particulier à Antoine Widlöcher et Christophe Turbout pour la gestion du site web de la conférence, à Lydie Sauvé pour son efficacité dans l'édition des actes, et à l'ensemble du comité d'organisation.

Les présidents de CIDE.7
Patrice ENJALBERT et Mauro GAIO

Table des matières

Conférences de synthèse

Session 1. Structuration de documents

Session 2. Analyse textuelle

Session 3. Organisation et exploration de corpus documentaires multimédia

Session 4. Sémantique linguistique et applications documentaires

Session 5. Indexation et recherche d'information

Comité de programme

M.H. Antoni, Université de Poitiers, France
T. Baccino, Université de Nice, France
B. Bachimont, INA et UTC, France
F. Cerbah, Dassault Aviation, France
J.P. Desclés, Lalicc, Université de Paris 4, France
P. Enjalbert, Université de Caen, France (coordinateur)
C. Faure, ENST, France
S. Ferrari, Université de Caen, France
C. Fluhr, CEA, France
M. Gaio, Université de Pau, France (coordinateur)
B. Grau, LIMSI, France
P. Laublet, Lalicc, Université de Paris 4, France
G. Mourad, Lalicc, Université de Paris 4, France
A. Napoli, LORIA, France
M.P. Pery Woodley, Université de Toulouse 2, France
I. Saleh, Université de Paris 8, France
K. Tombre, LORIA, France
B. Victorri, CNRS-ENS, France
G. Vignaux, CNRS-LCP, France
H. Vinet, IRCAM, France
J. Vivier, Université de Caen, France

Comité d'organisation

F. Bilhaut, Université de Caen, France
P. Enjalbert, Université de Caen, France
E. Faurot, Université de Caen, France
S. Ferrari, Université de Caen, France
M. Gaio, Université de Caen, France
V. Perlerin, Université de Caen, France
L. Sauvé, Université de Caen, France
C. Turbout, Université de Caen, France
A. Widlöcher, Université de Caen, France

Comité permanent de la conférence CIDE

M. Bellafkih (Maroc)
J. Caelen (France)
J. Ducloy (France)
M. Gaio (France)
J. Gardes (France)
J.L. Hainaut (Belgique)

P. King (Canada)
J. Labiche (France)
M. Leonard (Suisse)
J.P. Raysz (France)
J.M. Robert (Canada)
Z. Sahnoun (Algérie)
M. Szmurlo (France)
L. Thomazo (France)
E. Trupin (France)
C. Vanoirbeek (Suisse)
J. Virbel (France)
J. Vivier (France)
A. Widlöcher (France)
K. Zreik (France, coordinateur)

Organisateurs de la Semaine du Document Numérique

R. Mullot, Université de La Rochelle, France
J.M. Salaun, ENSSIB, Lyon, France

Conférences

de synthèse

Actualité d'une approche sémantique du document électronique

Patrice Enjalbert[1], Mauro Gaio[2]

[1]*GREYC, Université de Caen,*
Campus II, Bd du Mal Juin, 14032 Caen Cedex - France

Patrice.Enjalbert@info.unicaen.fr

[2]*LIUPPA, Université de Pau et des Pays de l'Adour,*
Avenue de l'Université, BP 1155, 64013 Pau Cedex - France

Mauro.Gaio@univ-pau.fr

Résumé :

Nous tentons dans cet article, à partir d'une analyse des contributions à CIDE 7 au sein d'un ensemble plus vaste de recherches actuelles, de cerner ce que peut être une approche sémantique du document numérique, d'en discerner les caractéristiques et les lignes de force. Nous distinguons ce qui relève des objectifs et des méthodes, pour nous interroger sur ce qui fonde l'unité de ces travaux sous le signe du « sémantique ».

Mots-clés : sémantique, document numérique, méthodologie.

Abstract:

In this paper, we try to synthesise the various contributions to CIDE 7 among a wider range of current research, in order to determine what a semantic approach of numerical document can be, and draw out the main trends and characteristics. We distinguish objectives and methods, questioning on what « semantic » ground the unity of these works can rely.

Key-words: semantics, numerical document, methodology.

1. Introduction

La décision de proposer le thème des « approches sémantiques » pour la septième occurrence de CIDE est liée à l'observation et à une analyse de certaines tendances actuelles de la recherche concernant les documents numériques.

Une première constatation, assez répandue, touche aux limites qui paraissent atteintes par les techniques « standard » de recherche documentaire. Celles-ci sont, on le sait et pour faire bref, basées sur des analyses de type statistiques de formes directement « perceptibles » par un programme : les mots (ou graphies) qui constituent un texte. A ces techniques on souhaiterait opposer des méthodes traitant véritablement du « contenu » des documents, même, et bien sûr, appréhendé de manière très partielle pour des raisons d'efficacité. Le gain attendu est à la fois en termes de rappel (plusieurs mots peuvent tomber sous le même concept objet de la recherche), de richesse de l'interrogation (dépasser la combinaison booléenne d'indicateurs), mais aussi d'appréhension par l'utilisateur des résultats de sa requête.

Une autre version, peut-être moins « radicale », serait de poser le problème en termes d'échelle du corpus documentaire ciblé : au « tout venant » des méthodes à base d'analyse de données, au prix d'une « finesse sémantique » moins grande — des méthodes plus riches, atteignant le « sens » de manière plus profonde, pouvant et devant être développées pour des espaces thématiques plus restreints et des tâches plus spécifiques.

Liée à cette première préoccupation est la volonté de décrire, de structurer des corpus documentaires, selon des espaces de connaissances — souvent baptisées ontologies — de référence. Cette structuration étant réalisée « à la main » au cours de la constitution du corpus, ou bénéficiant de traitements de contenu des documents comme évoqué plus loin. Il s'agit de présenter à l'utilisateur « l'espace documentaire » selon des concepts qui lui sont a priori familiers, qui « font sens » pour lui. On reconnaîtra ici le projet du « Web Sémantique », qui concerne d'ailleurs l'accès à des objets plus généraux que des documents (du moins dans une acception usuelle) tels que des services de toute nature.

La troisième observation est celle d'un essor, et de succès remarquables, de procédures de traitement du contenu « à grande échelle ». Certes, il s'agit de traitements relativement limités, mais certains résultats nous semblent tout à fait impressionnants. Le cas d'école en la matière est constitué par la technologie dite de « l'extraction d'information ». Nous reviendrons plus loin sur cette technologie, mais rappelons ici qu'il s'agit de remplir, à partir de textes courts et ciblés (tels que des dépêches d'agence) des « fiches » collationnant les informations factuelles principales. Les meilleurs systèmes sont parvenus à une qualité de l'ordre de 80 % par rapport aux performances humaines avec des temps de développement qui deviennent proches de l'industriellement acceptable. On peut montrer que les méthodes développées dans ce cadre ont des retombées et des prolongements très important en termes de recherche d'information « par le contenu » comme évoqué plus haut. Une autre remarque cruciale est que ces techniques d'analyse sémantique limitée de documents textuels ont leur exact pendant dans d'autres modalités, qu'il

s'agisse d'images fixes, de vidéo ou de documents sonores, notamment musicaux. L'idée d'un accès au « sens », à « l'information elle-même » (encore une fois, fut-ce de manière partielle) au delà des formes directement perceptibles, semble ainsi s'imposer comme une direction de recherche crédible « en vraie grandeur ».

Complétons encore par deux autres observations, que nous développerons moins ici, mais d'importance. La première concerne plus les « traitements humains » que les « traitements machine ». Elle concerne les *usages spécifiques* suscités par la forme numérique du document : c'est par exemple l'idée de la navigation dans de vastes ensembles de documents, qui conduit à de nouvelles stratégies d'appropriation, de « construction du sens ». Ce qu'il convient d'étudier à la fois en tant que procédure cognitive nouvelle (donc informative sur la cognition en tant que telle) et pour en tirer des indications sur la bonne manière d'organiser la navigation et structurer les espaces documentaires. L'ultime remarque concerne l'impact de la disponibilité de corpus numériques, et des procédures documentaires qui les accompagnent, pour des études *en* sémantique On touche-là à certains aspects d'une « linguistique de corpus » dont l'actualité n'est plus à démontrer. On peut supposer – mais les auteurs sont moins informés sur ce point – que ce type de démarche se développe ou peut se développer aussi par rapport à d'autres média, en terme d'analyse de « documents », artistiques par exemple.

Si l'on partage peu ou prou ces constats, il devient à coup sûr scientifiquement pertinent parler d'approches sémantiques du document numérique et de chercher à confronter et mieux asseoir des démarches de ce type. Tel a été l'objectif de CIDE 7. L'appel (dont les grandes lignes sont reprises dans la préface) a été volontairement très ouvert, de manière à permettre le plus large « balisage », avec le moins d'*a priori* possible, de ce nouveau champ.

Nous voudrions maintenant, en nous appuyant fortement sur ces différentes contributions[1], tenter de préciser ce qui nous paraît être quelques orientations de recherche pertinentes et prometteuses, et esquisser une cartographie possible du champ de recherche. Classiquement, nous commencerons par la question des *objectifs*, des *tâches*, visés, avant de nous intéresser aux *méthodes* pour les atteindre ou réaliser. Nous pourrons alors poser et discuter la question de l'*unité* de travaux ainsi rassemblés, et somme toutes assez divers, sous le signe de la sémantique.

2. Objectifs

Nous allons donc examiner ici des objectifs de recherche –en relation avec le document numérique – que l'on peut, à notre sens, et à un titre ou un autre, qualifier de sémantiques. Cette qualification est à l'évidence problématique si l'on considère la diversité des objectifs en question. Aussi tenterons-nous, en même temps qu'une description de fait, d'interroger le terme même de « sémantique ». En accord avec

1 Dont l'interprétation dans les lignes qui suivent est évidemment de notre seule responsabilité.

l'esprit d'ouverture de l'appel à communication rappelé plus haut, nous prendrons au sérieux la « revendication » par les auteurs d'une telle qualification de leurs travaux, la considérant comme une bonne heuristique dans notre réflexion. Il nous semble pouvoir mettre en évidence trois champs d'étude (ou trois facettes du même champ) que nous allons examiner succinctement. Il ne s'agit évidemment pas d'une « nomenclature » figée et complète. Mais quelques lignes de force, au moins à titre d'hypothèse, nous semblent se dégager.

2.1 Organisation et description de corpus documentaires

L'optique est ici essentiellement macroscopique. Il s'agit de considérer les *collections documentaires* et leur organisation et description en vue d'un usage donné : on retrouve là la problématique de l'indexation en recherche documentaire (RD) « traditionnelle », mais aussi la structuration hypertextuelle ou toute autre structuration propre à « navigation » dans les bases documentaires, l'organisation spécifique de documentations techniques ou de corpus artistiques (musicaux par exemple), etc.

Quelle peut être la caractérisation d'une approche sémantique de la question ? Nous proposons l'idée suivante : la mise en évidence d'un certain *espace de « valeurs », « notions », « concepts »* (selon les points de vue ou les *a priori* théoriques) *stabilisé* et doté d'une *organisation propre*, auquel les documents sont *rapportés*, et qui peut « faire sens » (pour parler intuitivement) pour l'usager.

Plusieurs articles de la conférence peuvent se discuter sous cet angle. Dans (Crestan *et al.*)[2] deux espaces sont envisagés (dans une finalité de RD classique) : le premier est structuré en « environ 800 dimensions » correspondant à des « concepts » représentés par des « sacs de mots », et censés permettre de repérer tout mot de la langue française ; le second est constitué d'*entités nommés* (personnes, lieux, dates…) dont il faut bien voir que ce sont en effet des *entités concrètes* (par nature typées), différentes de leurs réalisations langagières, qui peuvent être multiples. (Bénel) s'intéresse spécifiquement aux documents archéologiques, et à l'annotation de segments documentaires, permettant une indexation et un « arpentage » des collections, dans l'optique « d'offrir (…) des assistants à la construction du sens dans les bibliothèques numériques ». La référence à un corps de connaissances archéologiques est extrêmement nette, l'un des points traités étant la gestion de points vue divers selon les experts. (Bonardi et Rousseau) étudient l'indexation d'œuvres musicales : ils montrent comment les collections de CD sont actuellement rangés selon des critères fixes de genre, auteur etc., critiquent cette pratique et prônent une approche centrée sur des notions de prototype et de similarité : on voit donc là un débat entre deux modes d'indexation, un traditionnel qui serait sans doute celui des « ontologies » et un mode original, d'inspiration cognitive (prototype). Il faudrait encore mentionner le travail de (Besson *et al.*) sur les bases d'image mais nous y reviendrons plus loin.

2 Les références sous cette forme portent sur des articles du présent volume.

Les contributions de (Charlet *et al.*) et (Cerbah) proposent un regard un peu différent. La première présente une méthode de constitution semi-automatique d'un index d'ouvrage scientifique : pour être bref, disons qu'il s'agit d'une application de procédures d'acquisition de terminologie. Elle nous paraît bien relever du cadre proposé plus haut, dans la mesure où les auteurs la positionnent (entre autres) en terme d'ingénierie des connaissances : il s'agirait en quelque sorte de *faire émerger et d'acquérir le corps de connaissances terminologiques*, auquel l'ouvrage pourra alors être « rapporté ». (Cerbah) s'inscrit dans « une approche structurée de la documentation technique », conduisant à « fragmenter tout fonds documentaire en unités autonomes au contenu clairement spécifié » mises en relation par des hyperliens. Son insistance sur la nécessité d'une « interprétation plus ou moins profonde du contenu textuel balisé » (pour être mis en hypertexte) nous paraît de nouveau adéquate avec l'exigence mise en avant dans notre proposition d'une identification claire, en termes de domaine de connaissance (ici : un domaine technique), de l'espace de repérage.

Bien évidemment, la problématique que nous décrivons est aussi au cœur du dit « web sémantique » et de ses fameuses « ontologies ». Quelques remarques pour conclure ce premier aspect. D'abord pour souligner que la question de cette « indexation sémantique » peut être abordée de manière indépendante de celle des traitements : quels sont les bons principes et modes de description/structuration ? L'indexation elle-même pouvant être réalisée « à la main » comme dans (Bénel) ou dans beaucoup de travaux du web sémantique. Par contre on peut aussi s'interroger sur la manière d'indexer automatiquement (ou semi-automatiquement) les documents une fois le mode d'organisation choisi, ce qui nous conduit inévitablement au problème des traitements « sémantiques » (ou « du contenu » etc.) des documents. Enfin, notons que la question se pose quel que soit le média.

2.2 Analyse du « contenu » des documents

Nous pourrons partir ici de la technologie dite de l'*Extraction d'Information (EI)* [PIA 97] [POI 03], évoquée plus haut comme emblématique de progrès récents en « analyse de contenu ». Il ne sera sans doute pas évident pour tout le monde de l'appréhender dans son caractère sémantique. Beaucoup d'auteurs, peut-être la majorité, y voient une question « technologique » relativement neutre, concernant par exemple l'utilisation d'automates (ou transducteurs) pour reconnaître les « motifs » ou « patrons » textuels porteurs de l'information à extraire, et de méthodes d'apprentissage pour acquérir ces motifs. Et lorsque des considérations linguistiques sont invoquées, c'est bien souvent sous l'angle de l'analyse syntaxique, dite en l'occurrence « légère » (« shallow parsing »).

Pourtant il est facile de montrer la filiation avec les projets de « compréhension automatique » développés en Intelligence Artificielle dans les années 1980, dans une mutuelle fécondation avec d'autres traditions d'ingénierie linguistique et documentaire [DUP 02] [POI 03]. Or qui dit « compréhension », automatique ou non, dit évidemment « sémantique ». Le fait qu'elle soit en EI partielle, limitée

orientée, ciblée … n'y change rien[3]. Nous n'insisterons pas d'avantage ici sur cette technologie, non représentée en tant que telle dans CIDE 7. Mais nous pensons qu'elle a constitué une véritable rupture dans l'histoire des traitements automatiques du « contenu » textuel et constitue à ce titre un repère méthodologique majeur. Par ailleurs, nous esquissons dans [DUP 02] quelques pistes pour des travaux en sémantique linguistique, susceptibles de repousser certaines « limites» des systèmes actuels.

A priori, l'EI constitue une tâche disjointe de la Recherche Documentaire ou Recherche d'Information (RI) traditionnelle : dans le modèle « de base », des techniques de RD sont mises en œuvre *en amont* d'un système d'EI, pour extraire d'un flux textuel les documents qui, de par leur thématique, sont susceptibles d'être traités. En fait la situation est aujourd'hui beaucoup plus riche et complexe, et la recherche a largement évolué vers un croisement entre méthodes d'EI d'une part, et tâches et techniques (éventuellement requalifiées) de RD/RI de l'autre. De nombreux exemples en témoignent, caractérisés par : 1) des requêtes « structurées », dépassent le traditionnel assemblage de mots ; et 2) une recherche de « motifs » textuels (comme en EI) susceptibles de les « matcher » en un sens ou un autre. Exemple : « Trouvez les documents (ou segments documentaires) concernant les transactions financières en Europe d'un montant supérieur à 1 MEuro » [CIR 99] ou « le retard scolaire dans l 'Ouest dans les années 1980 / la sécurité maritime dans la Manche » [BIL 03].

Le représentant le plus avancé de cette tendance est la technologie dite des systèmes de Question/Réponse (« Question Answering »), dans lequel il s'agit de répondre automatiquement à des questions telles que : « Qui est l'auteur du "Dernier tango à Paris", Quels autres films a-t-il réalisé récemment ? », ou (dans un manuel Unix) : « Comment fait-on pour changer les droits d'accès? Que fait la commande "tar" ». On trouvera sur les sites de TREC de nombreuses références sur cette nouvelle problématique de recherche [VOO 01].

Avec des objectifs plus classiques, deux articles de CIDE 7 proposent des analyses de contenu au service de tâches de type RD, (Crestan) déjà mentionné et (Hurault-Plantet). (Valette) pose une question un peu différente : le problème est d'identifier les sites web racistes. L'auteur y montre qu'une analyse de contenu assez fine et débarrassée de beaucoup d'*a priori* communs est nécessaire : prise en compte de la mise en forme visible au delà du lexique lui-même (le rouge et les majuscules sont de bons indicateurs de sites racistes), importance du contexte (les termes a priori racistes/non racistes s'échangent de manière étonnante), importance du « genre » (ici : pages web), etc.

Nous abordons maintenant une question très intéressante, et représentée significativement dans CIDE 7, à savoir un parallèle frappant entre ces problématiques textuelles et des travaux qui se développent pour des documents relevant d'autres média : image, vidéo, son. La question de l'accès à des bases d'images à travers non des légendes ou autres descriptifs (ou pas seulement), mais

3 A la vérité, en est-il jamais autrement ?

grâce à une analyse des images elles-mêmes, apparaît aujourd'hui comme une nécessité, posant d'ailleurs des problèmes scientifiques et techniques difficiles. Il en est de même pour la vidéo et le son. Dans cette problématique de l'accès au contenu, si la tendance majeure est sans doute aujourd'hui d'identifier ce « contenu » à des traits, des descripteurs, disons « physiques » (dominances de couleur, de textures, spectres de fréquences…) une tendance se développe qui se positionnera par exemple en termes d'*interprétation* d'image. Autrement dit, une tendance que l'on peut effectivement qualifier de sémantique. Nous nous référerons ici à deux contributions de CIDE 7.

Les travaux de (Besson *et al.*) concernent la constitution de bases d'images, et se positionnent d'emblée dans une démarche intégrant cette dimension sémantique. Les auteurs discutent explicitement ce qui peut opposer des approches « syntaxiques » et « sémantiques ». Les premières correspondent précisément pour eux à l'extraction de paramètres physiques, tandis que les secondes « consistent à réduire l'image à un ensemble d'objets sémantiques (identifiés par leur signification dans le monde réel) reliés éventuellement par des relations sémantiques (qui correspondent à une interprétation de l'image comme reproduction du monde réel) ». On peut discuter la référence au « monde réel » comme critère de « sémanticité »[4], mais le contraste avec les « paramètres physiques » n'en est que plus patent.

Notre seconde référence sera pour les travaux de Hugues Vinet (Vinet). Nous nous référerons ici particulièrement au projet CUIDADO [VIN 02] visant à développer un système de « Navigation dans des bases de données musicales ». Il s'agit en fait d'un projet très riche, mais dont un des aspects sera de pouvoir répondre (par une sélection d'œuvres) à des requêtes du type : « Je veux des morceaux à tempo rapide, de type Rock / des morceaux proches de ceux-ci / constituer un programme de tempo de plus en plus rapide, avec plus de 60 % de vocal féminin ». Cela est réalisé par une combinaison / enchaînement de modèles et de traitements : traitement du signal, modèles perceptifs, symboliques, cognitifs, modèles de description de contenu… Les dernières « couches » intègrent donc, comme on le voit sur l'exemple, un véritable niveau symbolique, relevant de connaissances musicales élaborées allant jusqu'à une notion de genre musical (ici, le Rock). L'examen détaillé du modèle en couches proposé fait apparaître un parallèle saisissant avec les tâches et niveaux d'analyse en langue.

Finalement, revenant aux documents textuels, nous voudrions relever dans cette rubrique une série d'autres travaux non réductibles à des tâches d'EI/RI, mais impliquant une authentique analyse de contenu. Il s'agit de travaux qui se situent plus dans une perspective théorique (linguistique en l'occurrence) qu'applicative, que l'on peut rapidement situer comme « analyse textuelle ». Ici, c'est plutôt les technologies du document numérique qui sont mises à contribution pour un travail de sémantique linguistique plutôt que l'inverse. Cette tendance est représentée dans

4 Cf. *infra* section 4. Nous dirions que la référence à des « objets et relations » issus d'un *domaine d'expérience et de connaissances* nous situent bien dans le cadre proposé en 2.1.

CIDE 7 par (Lessard *et al.*) et (Leblanc). Le premier s'intéresse à la détection d'une forme particulière, mais récurrente, de structuration du discours narratif : les « topoï narratifs » qui évoquent en quelque sorte, dans des contextes très divers, des scénarios plus ou moins stéréotypés. Le second étudie les occurrences du « je présidentiel » dans les discours de vœux de bonne année, et cette étude est précisément l'occasion de réflexions méthodologiques sur l'utilisation des techniques d'analyses de cooccurrences dans une perspective herméneutique.

2.3 Segmentation et structuration de documents

Ce troisième aspect est plus délicat à cerner, mais clairement présent dans CIDE 7. Deux articles sont particulièrement typiques : (Cerbah) déjà cité, et (Smolczewska, Lallich-Boidin). Ce dernier se présente comme cherchant « à définir un modèle de structuration et d'enrichissement de l'information technique qui constituera la base de la construction d'une représentation du contenu du document technique à partir de son texte intégral (…) résultat de plusieurs étapes intermédiaires telles que :

1. La segmentation du document en unités d'information autonomes ;
2. La caractérisation du contenu de chaque unité ;
3. Le filtrage des unités pertinentes par rapport aux unités non pertinentes pour l'utilisateur ;
4. La construction de liens entre les unités exprimant des parcours de lecture possibles. »

Les second et troisième points nous renvoient aux deux aspects précédemment discutés. Ce qui est nouveau ici sont l'importance attachée aux opérations de *segmentation* et de *mise en relation hypertextuelle* des segments ainsi délimités. Cela nous paraît à peu de chose près la même problématique et les mêmes objectifs que ceux de (Cerbah), avec également en vue le document technique (mais avec des méthodes différentes).

Tout se passe comme si une *certaine structuration du texte*, en terme de segments interreliés, était une composante à part entière de la « perception sémantique » d'un document par un lecteur. Cette perception étant révélée par les *parcours de lecture*. Les dispositifs proposés par ces auteurs tentent alors de repérer automatiquement cette structure pour aider le lecteur dans sa découverte.

En fait, il nous semble que cette démarche est aussi un des aspects du travail de (Charlet *et al.*), puisque les index *pointent* sur le texte (structuration), renvoyant à des segments dont la définition, « l'empan », (segmentation) est considéré par les auteurs comme une tâche importante et difficile : « la difficulté étant de sélectionner les [occurrences d'un terme] les plus pertinentes et de définir la taille de l'empan de texte auquel il est pertinent de renvoyer ». On pourrait aussi voir dans un domaine non représenté dans CIDE 7, le résumé automatique, une occurrence de l'approche « structurelle ». Nous pensons ici particulièrement au résumé « par extraction », dans lequel il s'agit de repérer des segments de texte pertinents pour constituer un

résumé — et tout en gardant un lien sur le document original, dans une perspective de navigation intra-documentaire [MIN 03]. Est-ce là une question et une approche « sémantique » ? Nous y reviendrons dans la quatrième partie.

Finalement il nous semble que cette préoccupation « structurelle » est aussi au cœur d'un article très différent, à savoir (Boukottaya, Vanoirbeek). L'article pose le problème de correspondances entre XML-schémas, de manière à « échanger des données XML entre applications Web autonomes et hétérogènes ». Ce qui est ici significatif de notre propos est la *fonction sémantique de la structuration* même, qui nous semble exprimée par les auteurs, lorsqu'elles parlent « d'*information sémantique* nichée dans la structure du document » (« *semantic information* nested within the document structure »). « La sémantique est d'abord capturée à travers l'explicitation de la signification du nom des éléments, et ensuite à travers l'analyse du point de vue du concepteur du XML-schéma, exprimée par l'organisation logique du contenu XML (…) ».

3. Méthodes

La question que nous posons maintenant est la suivante : Y a-t-il des méthodes spécifiques d'une approche sémantique ? Et/ou des regards particuliers sur des méthodes « plus générales » ? Ici, plus que jamais, la réflexion est prospective, l'enjeu étant de réfléchir sans *a priori* aux moyens à mettre en œuvre pour réaliser des objectifs tels que présentés ci-dessus. Nous distinguerons trois « types » de méthodes, en remarquant d'emblée que les applications les combinent en général.

3.1 Méthodes sémiotiques propres aux différents médias

C'est évidemment la première caractéristique possible d'une approche sémantique. Rappelons que le terme « sémiotique » désigne l'étude des divers systèmes de signes, quels qu'ils soient : ici donc le texte, l'image, la vidéo, le document sonore, et sans oublier la dimension « hyperdocument ». Une « approche sémantique » va donc souvent se référer à une connaissance relativement élaborée du « fonctionnement » de ces différents « mode sémiotiques ».

Un certain nombre de travaux de CIDE 7 déjà mentionnés entrent dans cette « rubrique » : (Besson *et al.*) mettent en œuvre des méthodes d'analyse et d'interprétation d'image : (Vinet) développe un modèle sémiotique complet du document musical ; (Cerbah), (Smolczewska, Lallich-Boidin), (Charlet), (Crestan *et al.*) réfèrent à des modèles linguistiques, avec une composante sémantique forte, du lexique (Cerbah) au discours (Smolczewska, Lallich-Boidin), (Lessard *et al.*). Nous avons déjà eu l'occasion (section 2.2) d'insister sur la dimension sémantique de l'Extraction d'Information et de mentionner nos propres recherches pour développer des méthodes en rapport — au delà d'ailleurs de l'EI *stricto sensu*, jusqu'à des applications en Recherche d'Information et en structuration de documents composites [DUP 02].

D'autres auteurs font une référence peut-être encore plus explicite à une *théorie* sémantique particulière. C'est le cas de (Valette) avec la sémantique différentielle de François Rastier, qui trouve là une application particulièrement originale et stimulante au « web », débouchant sur une approche non triviale — en gros « anti-ontologique », et intégrant des facettes variées, non exclusivement linguistiques au sens usuel du terme. (Perlerin, Ferrari) se réfère également à la sémantique différentielle avec une application à la détection des métaphores et un prolongement plus large sur la conception d'outils d'exploration de textes. Enfin (Dutoit *et al.*) proposent une réflexion personnelle très « amont » sur le thème « forme et sens », autour des applications industrielles développées par ces auteurs en ingénierie linguistique.

Toutes ces tentatives illustrent une voie de recherche qui nous semble fondamentale, et susceptible (à plus ou moins long terme, il est vrai) de contribuer significativement à repousser certaines limites de l'ingénierie documentaire évoquées en introduction.

3.2 Ingénierie des connaissances

Ce second aspect a en fait déjà été introduit dans le « schéma » proposé en section 2.1. Il est aussi, à l'évidence, porté par la communauté du « Web Sémantique » et ses « ontologies » et autres formats de description de contenu (RDF, Topic maps etc.). Toute personne familière avec l'IA reconnaît dans ces formats des avatars de formalismes de représentation des connaissances développés dans les années 80 — avec une tentative d'application « en vraie grandeur » particulièrement réjouissante.

Il s'agit donc ici d'insister sur la nécessaire prise en considération du facteur « ingénierie des connaissances » dans une approche sémantique du document. Cette conception va à l'encontre de bien des idées reçues et de bien des pratiques courantes. L'idée communément admise est effet plutôt que seuls des « traitements de surface », « de la forme » sont possible, pour une double raison de temps de développement et de temps de calcul.

Il y aurait à s'interroger sur la prégnance de cet *a priori* méthodologique. Tradition de la RD ? Méconnaissance des acquis de l'ingénierie des connaissances, en termes de méthodes et d'outils ? Tradition linguistique volontiers formaliste (le « primat de la syntaxe ») ? On pourrait engager un débat salutaire… Mais il nous semble que l'histoire est en train de trancher. Le « Web sémantique » gèrera des connaissances ou ne sera pas. Les combinaisons Ingénierie des connaissances / Ingénierie linguistique sont devenues une réalité solide (voir par exemple tout ce qui touche à l'extraction de terminologie) [CHA 00]. L'EI a de longue date bien mis en évidence, et les besoins impératifs en ingénierie des connaissances, et la possibilité de développer des méthodes « légères » (« shallow knowledge ») appropriées [PIA 97].

3.3 Méthodes de structuration du document

La dernière « facette » considérée ici concerne la structuration du document. Il s'agit donc au départ de méthodes on ne peut plus « généralistes » et « ingénieuriales », mais qui trouvent ici des applications et, peut-être, un « regard » particuliers. On pense ici typiquement aux technologies XML et hypermédia.

En fait, presque tous les articles de CIDE 7 utilisent les premières, et beaucoup les secondes (dans le but de faciliter la navigation ou autre « arpentage » de bases documentaires). Aux articles déjà cités ajoutons (Blanchon, Boitet), dans le domaine de la traduction automatique interactive. Le système proposé enrichit le texte par des annotations portant sur les segments reconnus comme ambigus par le logiciel de traduction. Ces annotations sont en quelque sorte dynamiques, et ouvrent une boîte de dialogue permettant à l'utilisateur de choisir parmi plusieurs traductions proposées. Cette contribution nous paraît significative d'une idée somme toute assez simple, mais peut-être fructueuse : l'enrichissement du texte apportant des nouveaux « éléments de sens » et susceptible d'aider le lecteur à sa propre « interprétation ». Et insistons encore, sur (Boukottaya, Vanoirbeek) qui, on l'a vu, proposent une appréhension proprement sémantique d'XML lui-même, à travers la notion de matching / correspondance de XML schemas.

Se dessine ainsi, pensons-nous, un domaine à explorer : quels outils de structuration / annotation développer à l'appui d'approches sémantique ? Et comment, en retour, donner une assise sémantique à ces outils ?

4. Conclusion : qu'est-ce qu'une « approche sémantique » du document numérique ?

Le parcours que nous venons d'opérer, autour des contributions à CIDE 7, d'un ensemble de travaux actuels qualifiables de « sémantiques » à un titre ou un autre laisse apparaître une belle variété d'objectifs et de méthodes. La question se pose alors de l'*unité* de ces problématiques. Par ailleurs, il semble évident que cette diversité questionne la notion même de sémantique. Le sujet est trop ancien et parcouru de courants philosophiques, linguistiques, sémiologiques… trop divers pour espérer proposer une réponse « claire et définitive » ! Pour autant il nous paraît tout à fait pertinent de nous poser la question du « fait sémantique » en regard des nouvelles pratiques à l'œuvre dans le document numérique : à la fois pour éclairer ces dernières, et pour apporter peut-être un regard nouveau, ou renouvelé, sur le « fait » en lui-même. Nous tenterons donc dans cette dernière section de repérer quelques lignes de force transverses émergeant, nous semble-t-il, des travaux examinés[5], sans manquer de se poser la question critique de l'apport concret,

5 « Émergence » sous un certain regard, cela va sans dire, conditionnée par certaines positions de principe développées notamment dans [ENJ 96] et [GAI 01].

pratique, d'un point de vue sémantique ainsi qualifié au développement des technologies documentaires.

4.1 Dimensions d'une sémantique du document

Trois « dimensions » nous paraissent particulièrement caractériser un « regard sémantique » sur le document. Précisons bien encore ici qu'il ne s'agit pas de « découper » un ensemble d'objectifs ou de méthodes qui seraient disjointes d'autres pratiques documentaires, mais de cerner certaines manières de les aborder.

1. Une dimension Document-Connaissances

Ce point a déjà été présenté en section 2.1. Une caractéristique majeure, peut-être même « la » caractéristique essentielle et quasiment définitoire du « fait sémantique » est de rapporter une donnée (perçue ou déjà construite comme *signe*) à un *espace de référence* accepté ou posé à un moment déterminé par le lecteur. C'est cet espace que nous appellerons ici « connaissances » selon la tradition en Intelligence Artificielle et parce que ce terme marque bien l'idée d'une certaine « stabilité » de l'espace en question. Par ailleurs le terme convient bien dans la mesure où beaucoup de documents ont une fonction informationnelle. Il pourrait être problématique ou partiel pour des documents artistiques, par exemple – tels que les documents musicaux – et demander alors un certain élargissement, respectant l'idée générale de référence stabilisée.

Cette notion de connaissance nous paraît également reprendre, de manière plus appropriée, l'idée de « référence au monde réel » constitutive de nombre de sémantiques formelles du langage, et reprise ici à propos de l'image par (Besson *et al.*) : en vérité, l'idée d'une telle référence directe au monde « tel qu'il est » paraît surprenante à propos de documents images dont on sait les transformations numériques (quand ce n'est les travestissements) qu'il peuvent subir. La *médiation* par une notion de « connaissance », de « représentation du monde » paraît nécessaire.

Quoi qu'il en soit, le fait de porter attention à un « niveau Connaissance » extérieur en quelque sorte aux documents eux-mêmes, et dont l'élaboration fait partie de l'ingénierie documentaire, paraît bien émerger de nombre de travaux analysés ci-dessus.

2. Une dimension Document-Document

Il y a deux facettes à cette dimension. La première est de noter que de nombreux traitements vont se traduire *in fine* par la création de nouveaux documents, *enrichissant* d'une manière ou d'une autre le document (ou l'ensemble de documents) traité. On pense là par exemple à des index, qui restent évidemment liés aux documents, notamment au moyen d'hyperliens. Mais aussi aux annotations diverses des documents telles que les entités nommées d'un (Crestan), ou les annotations de désambiguïsation de (Blanchon, Boitet). Un autre exemple typique est celui du résumé automatique.

Le schéma est ici celui d'un *ajout de nouvelles informations*, intégré au document ou, ce qui revient au même, notées dans un nouveau document relié au document traité.

La seconde dimension (non exclusive de la première) concerne l'idée de *structuration* du document. Découvrir l'organisation d'un document, sa structure, semble faire partie intégrante de l'activité de « lecture », comme relevé dans la section 2.3. C'est aussi une perspective clairement et avec force énoncée dans (Vignaux). C'est toute la problématique de l'hyperdocument qui se dessine ici en perspective.

On pourrait donc évoquer ici une sémantique à la Peirce, reprise notamment par U. Eco [ECO 85] présentant le sémiotique comme renvoi de signe à signe[6]. Avec probablement des actualisations importantes liées aux technologies employées. Une direction importante, pensons-nous, pour « penser le document numérique ».

3. Une dimension Humain-Document

Finalement, il convient évidemment de ne pas oublier l'utilisateur humain qui prend connaissance du document. Remarquons que cette problématique n'est pas toujours présente : ainsi dans le modèle « classique » de l'Extraction d'Information, les traitements visent à constituer des bases de données à partir des faits extraits des textes traités : un support d'information qui n'a donc plus rien à voir avec les documents initiaux. Mais à côté de ces approches, il en existe d'autres dont une préoccupation importante est précisément la prise en compte de l'*appropriation du document par le lecteur,* et des moyens d'y aider. Il est frappant que l'idée de « navigation documentaire » soit présente en quelque sorte « en perspective » dans de nombreux articles de CIDE 7, quel que soit le centre d'intérêt principal. Mais c'est évidemment encore (Vignaux) qui développe avec le plus de force et de détails ce point de vue.

L'étude en temps que tels des modes d'appropriation du document et des moyens, appuyés sur les technologies numériques, d'y aider, constitue donc bien une des facettes d'une approche sémantique.

4.2 « Qu'est-ce qu'on gagne ? »

Nous pensons avoir ainsi dégagé quelques points de convergence forts qui constituent (ou : participent de) l'unité d'approches revendiquées comme sémantiques. Le lecteur sceptique pourra néanmoins se demander ce qu'apporte un tel regard sur le document. N'est-ce qu'un habillage théorique particulier sans conséquence pratique ? Nous pensons que non, et voudrions pour conclure relever un certain nombre de points sur lesquels l'apport méthodologique nous paraît important.

6 « Un signe, s'adresse à quelqu'un (…) crée (…) un signe équivalent, ou peut-être un signe plus développé » (Peirce).

1) **L'identification du (des) « problème(s) » et de son (leur) ampleur.**

   S'il est effectivement question *in fine* de « faire du sens » pour l'utilisateur, alors, qu'on le veuille ou non, on est confronté aux questions complexes qui caractérisent « le sémantique ». Le savoir peut éviter des impasses, par exemple la recherche de progrès *exclusivement* dans des techniques d'analyse de données basées sur des « formes pures » — ou des déconvenues prévisibles : par exemple, on peut penser que les travaux actuels sur le « Question Answering » mésestiment gravement l'ampleur des problèmes à traiter.

2) **Un point de vue unificateur entre différents médias, entre différentes tâches pouvant se combiner.**

   De plus en plus nous aurons à traiter de documents multimédia. Si l'on en reste aux technologies, aux procédés de calcul, les traitements risquent de longtemps diverger. Un point de vue sémantique — ou, en l'occurrence, sémiotique— peut nous permettre de penser l'*intégration* des différentes informations et supports.

3) **Un décloisonnement des méthodes**

   Il est trop souvent convenu d'opposer méthodes numériques et linguistiques, linguistique et ingénierie des connaissances, reconnaissance et interprétation d'image, traitement du signal et niveau symbolique… Alors même que de plus en plus de travaux mêlent ces différents niveaux (voir ici (Cerbah), (Charlet), (Vinet) par exemple). La reconnaissance d'objectifs communs « de haut niveau » peut laisser la place à l'intégration de ces techniques et méthodes.

4) **Le développement de méthodes sémiotiques spécifiques**

   Nous pensons qu'un investissement « de fond » en relation avec des théories relativement approfondies des différents « modes sémiotiques » (langue, image, son, vidéo…) est une des voies pour progresser, qu'il s'agisse de « traitement du contenu » ou « d'appropriation humaine du document »

5) **Un « retour théorique »**

   Inversement, le développement d'une telle approche du document numérique est de nature à renouveler nos conceptions du fait sémantique, ne serait-ce que parce que s'offre ainsi un champ d'expérimentation et d'objectivation totalement nouveau. La « linguistique de corpus » (ici dans sa composante sémantique) en a déjà pris conscience depuis quelques temps et indique en quelque sorte une voie prometteuse.

5. Références bibliographiques

[BIL 03] Bilhaut F., Charnois T., Enjalbert P., Mathet Y., « Passage extraction in geographical documents », *Proc. Intelligent Information Systems 2003, New Trends in Intelligent Information Processing ans Web Mining*, Zakopane, Poland, 1-4 Juin 2003, pp. 121-130.

[CHA 00] Charlet J. (éd.) « Ingénierie des connaissances », *Eyrolles*, 2000.

[CIR 99] Ciravegna, F. *et al.*, « FACILE: Classifying Texts Integrating Pattern matching and Information Extraction », *Proceedings of IJCAI'99*, pp. 890-895, 1999.

[DUP 02] Dupont M., Vuillaume J.-M., Victorri B., Enjalbert P., Mathet Y., « Nouvelles tendances en extraction d'informations », *Techniques et Sciences Informatiques*, vol 21 n°1/2002, 2002, pp. 37-64, 2002.

[ECO 85] Eco U., « Lector in Fabula », *Le livre de Poche, Coll. "Essais"*, n°4098, 1985.

[ENJ 96] Enjalbert P., « De l'interprétation (sens, structures et processus) », *Intellectica*, vol 23, n° 2, pp. 79-120, 1996.

[GAI 01] Gaio M., « Traitements de l'information géographique : représentations et structures », *Mémoire d'Habilitation à Diriger les Recherches*, Université de Caen, 2001.

[MIN 03] Minel J.-L., « Filtrage sémantique. Du résumé automatique à la fouille de textes », *Hermès*, 2003.

[PIA 97] Piacenza M.-T., (éd.), « Information Extraction », *Springer Verlag*, 1997.

[POI 03] Poibeau T., « Extraction automatique d'information », Hermès, 2003.

[VIN 02] Vinet H., Herrara P., Pachet F., « The CUIDADO Project », *Proc. Int. Conf. On Music Information Retrieval*, IRCAM, Paris, 2002, pp. 197-203.

[VOO 01] Voorhees E. «Overview of the TREC 2001 Question Answering Track », http://trec.nist.gov/pubs/trec10/t10_proceedings.html, 2001.

Du corpus à l'hypertexte

Georges Vignaux

Laboratoire Communication et Politique – CNRS FRE 2813
Equipe Hypertextes et Textualité électronique
27 rue Damesme, 75013 Paris - France

georges.vignaux@damesme.cnrs.fr

Résumé :

Le noyau du projet CoLiSciences consiste en la mise en ligne d'un grand corpus balisé dans la perspective de consultations hypertextuelles multiples. Il s'agit d'un corpus des écrits des naturalistes et biologistes de langue françaises du XIX^{e} siècle (environ 6 000 pages). Revisiter ainsi le concept d'hypertexte apparaît fondamental, y compris dans le contexte éducatif actuel (secondaire et supérieur) où la recherche d'informations, leur organisation et leur réorganisation, dans le cadre d'activités interdisciplinaires, occupent une place croissante.

Mots-clés : Hypertexte, Internet, lecture, corpus, histoire des sciences.

1. La nécessité de contenus originaux accessibles sur internet

1.1 Quels contenus ?

Tout contenu porté sur le réseau Internet serait-il par nécessité intéressant sinon important ? Rien n'est moins sûr. On peut s'interroger sur les durées d'usage et de vie de tel ou tel site Internet offrant du contenu et sans plus. Il est donc fondamental de développer des réflexions appuyées d'expérimentations et qui testeraient différentes catégories de contenus et notamment la pertinence de grands corpus aujourd'hui de plus en plus réclamés. Les grandes questions sont ici les suivantes :

- Qu'est-ce qu'un corpus et quels sont ses critères de légitimité ?
- Comment construire un corpus et où arrêter les frontières de constitution de celui-ci ?
- Pour quels objectifs ? Quelles applications ? Et quels lectorats uniques ou multiples ?

1.2 Indexations, parcours

Il existe une croyance, laquelle suppose une sorte d'objectivité *a priori* des contenus. Il y a du mirage là-dedans. En effet, les visées de neutralité classiques dans la constitution de thesaurus ne sont jamais qu'apparentes. Citons quelques problèmes récurrents :

- Quels sont les mots clés choisis pour organiser l'accès à un corpus ?
- Quels modes de croisement va-t-on établir entre ces mots clés ?
- Quels types d'ordre va-t-on alors ainsi imposer dans l'articulation entre entrées et unités lexicales ?
- Quels croisements entre entités lexicales va-t-on stabiliser en vue d'établir des données sémantiques de niveau plus élevé ?

1.3 Quelles lectures ?

Les nouveaux usages du texte qui naissent de l'informatique sont à scruter dans deux directions :

D'une part, sur un plan quantitatif, il est nécessaire d'aborder les problèmes induits par la puissance fonctionnelle de l'outil électronique – essentiellement : comment délimiter, dans l'espace des textes interconnectés, un sous-ensemble homogène et susceptible d'une exploitation raisonnée.

D'autre part, au plan qualitatif, menant de front une réflexion théorique et une évaluation pratique, il importe de s'interroger sur le rendement sémantique, cognitif et informationnel des fonctionnalités de lecture qui caractérisent l'hypertextualité. S'intéressant très précisément aux logiques d'accès, d'affichage et d'enregistrement des données, il s'agit de mettre au jour les pratiques de lecture que ces trois logiques, dans leurs interactions, déterminent. Et, plus avant, il s'agit d'optimiser des systèmes de parcours textuels en fonction des différentes sortes de données, des différentes finalités de lectures et des différents types de lectorats. La question de l'hypertextualité ne peut donc être dissociée des interrogations sur les contenus et les modalités de lecture.

2. Hypertextes numérisés et histoire des idées

La mise à disposition de contenus évolués et ciblés sur Internet est reconnue de plus en plus souhaitable voire nécessaire. La question des modes de construction de ces corpus numérisés (choix des domaines, types de délimitations, etc.) demeure capitale car encore peu explicitée et ne faisant guère l'objet d'analyses approfondies. L'interrogation sur la notion d'hypertexte et les types de lecture ainsi induits sur Internet est encore plus cruciale et pourtant, les réflexions sur les « parcours de navigation » restent parcellaires, souvent pauvres en développements.

Les réflexions de notre équipe sont ainsi fortement orientées vers l'étude des formes internes de la textualité électronique et sur les opérations sémantiques et cognitives que celles-ci favorisent, cela dans le contexte concret de mise en place d'une plate-forme expérimentale d'offres de ressources, de développement ou d'adaptation d'outils et de services multimédias.

C'est dans ce contexte que s'inscrit notre programme *Hypertextes numérisés et histoire des idées : la naissance d'une science moderne du vivant, naturalistes et biologistes français au XIX^{e} siècle*. Ce programme scientifique est concrétisé par le *projet CoLiSciences* (*Corpus de littérature scientifique*) qui aboutit aujourd'hui à la mise en ligne d'un grand corpus balisé dans la perspective de consultations hypertextuelles multiples (http://www.colisciences.net).

3. La construction hypertextuelle

L'hypertextualité est généralement définie par trois propriétés, à savoir :

- *La fragmentation du contenu* : un hypertexte est un agrégat d'éléments d'information, de taille réduite, entretenant des connexions diverses ;
- *L'informatisation* : le contenu d'un hypertexte est installé sur support électronique ;
- *La non linéarité des lectures* : le contenu et la matérialité d'un hypertexte n'imposent aucune directive de lecture et les fonctionnalités du dispositif informatique permettent d'effectuer des parcours multiples dans le fonds enregistré.

Mais cette vision demande à être nuancée. C'est que les connaissances inscrites dans le réseau hypertextuel ne peuvent être conçues comme des explicitations conformes du contenu des documents, explicitations sur lesquelles toutes les navigations consistantes devraient se régler, mais plutôt comme des explicitations *possibles* de ces documents. Le système hypertextuel introduit une interface dans les rapports d'interprétation : entre le document et le lecteur, il établit une couche intermédiaire de « connaissances » qui guide les approches du lecteur, mais qui ne constitue en aucun cas une grille de référence absolue. La couche « connaissance »

est de ce point de vue, indépendante de la couche « document » : elle est projetée sur les textes dont elle constitue en définitive, une interprétation, un « éclairage » choisi comme voie d'accès[1].

La structure fondamentale de l'hypertextualité est donc une *structure dynamique* : le réseau des connaissances qui reflète l'interprétation des documents à un moment donné du processus de lecture, va orienter des parcours qui vont susciter de nouvelles interprétations, c'est-à-dire donner lieu à de nouvelles connaissances qui, à un moment ultérieur, pourront à leur tour jouer le rôle de guide de lecture et ainsi de suite...

4. Le Projet CoLiSciences (Corpus de littératures scientifiques)

4.1 Objectifs

- Construire des contenus numérisés et spécifiés en vue d'usages scientifiques, culturels et pédagogiques,
- Élaborer des outils d'aide à l'acquisition de savoirs organisés sous forme d'hypertextes,
- Établir un diagnostic expérimental du rapport entre outils multimédias et contextes de transmission et d'acquisition.

4.2 Le projet

Le noyau du projet consiste aujourd'hui en la mise en ligne d'un grand corpus balisé dans la perspective de consultations hypertextuelles multiples. Il s'agit d'un corpus des écrits des naturalistes et biologistes de langue française du XIX^{e} siècle.

Réaliser un tel corpus (environ 6 000 pages) pose immédiatement la question de la pertinence du choix des auteurs sélectionnés. Pour la même période, mais dans le monde anglo-saxon, nous aurions évidemment choisi Charles Darwin, à cause du « continent de savoirs » qu'il inaugure et développe. En France, six figures « emblématiques » s'imposent au cours de ce siècle, durant lequel la biologie prend son essor : Jean-Baptiste Lamarck, Georges Cuvier, Étienne Geoffroy Saint-Hilaire, Isidore Geoffroy Saint-Hilaire, Louis Pasteur et Claude Bernard. Le premier pour sa

[1] « Une des questions centrales tient à ce que l'hypertextualisation d'un fonds imprimé se factorise en deux temps d'opération. A savoir : (i) la construction d'une représentation du texte-source conformément au format d'enregistrement électronique, et (ii) la conception de modules de présentation et de consultation adaptés aux données ainsi consignées. » (D. Piotrowski, « De l'hypertextualité en général et de CoLiSciences en particulier).

théorie transformiste et pour sa zoologie des Invertébrés ; les deux suivants, pour la célèbre querelle qui les opposa au sujet des plans d'organisation des êtres vivants[2], et pour leur rôle dans la réflexion sur l'ordre et la diversité du vivant ; le quatrième pour sa tératologie ; les derniers, pour les bouleversements théoriques et expérimentaux de la chimie biologique et de la physiologie, ainsi que de la microbiologie. Déjà, dans cette sommaire opposition, se dessine une dichotomie entre biologie en tant qu'histoire naturelle et biologie en tant que mécanisme. Le champ de la biologie est ainsi partagé suivant cette ligne de séparation. Les grandes oppositions théoriques et métaphysiques sur la nature du vivant (matérialisme, vitalisme, déterminisme, nécessité, finalisme, etc.) alimentent les réflexions de ces chercheurs et des nombreux commentateurs.

Les dimensions du projet

- *La dimension stratégique* : il importe de construire un certain nombre de modèles de fonctionnalités transversales, permettant de tester les différentes formes d'indexation des données selon les contenus et les fonctionnalités requises (notices historiques et biographiques, bibliographies, classifications terminologiques, historiques, épistémologiques et sémantiques).
- *La dimension de l'offre* : elle consiste dans cette mise à disposition d'un corpus de données originales, numérisées, balisées et indexées.

Les étapes méthodologiques

- La numérisation et le balisage de textes,
- La construction d'hypertextes structurant le corpus, en fonction de repérages fondés sur les parcours de lecture (navigations dans les textes) et l'appropriation de connaissances à la faveur des liens proposés au lecteur et modifiables par lui (annotations, commentaires, coupures, reconstitutions).

Les niveaux procéduraux

- *Des niveaux manipulatoires* : les formes simples qui doivent permettre des acquisitions premières (notices, définitions, descriptions, explications, illustrations).
- *Des niveaux référentiels* : le corpus comme outil interrogeable pour l'apprentissage de l'histoire des grandes questions des sciences du vivant

[2] Ce n'est qu'environ 150 ans plus tard que les plus récentes avancées de la génétique du développement ont permis de valider les hypothèses de Geoffroy Saint-Hilaire, d'une considérable hardiesse en leur temps.

(théorie de l'évolution, etc.) et l'aide à la réflexion sur des concepts de base (les « grandes notions »).

- *Des niveaux érudits* : les lectures, annotations et commentaires critiques sur des questions établies au XIX^e^ siècle et toujours réactualisées (exemple : la problématique de « l'instinct » ou celle de « l'amour maternel » (inné ou acquis) du XIX^e^ à nos jours).

Développements

Plusieurs types de développements sont ici prévisibles :

- La construction de sites en ligne offrant des textes susceptibles de nombreuses manipulations sur la matière même des contenus (usages pédagogiques, épistémologiques et de recherche) ;
- Un bilan diagnostic des différentes modalités hypertextuelles ;
- Une prospective des hypertextes pédagogiques, littéraires et scientifiques du futur.

5. CoLiSciences : un système hypertextuel sur le Web[3]

Notre démarche étant la construction d'un site offrant des textes susceptibles de nombreuses manipulations sur la matière même des contenus, il nous fallait trouver un système hypertexte qui s'appuie sur l'architecture, mais il fallait aussi que ce système prenne suffisamment de recul par rapport à l'interface Web pour offrir des parcours novateurs qui permettent des lectures différentes et qui soient dégagés des contraintes de HTML.

Les ouvrages numérisés devant être encodés sous format XML, il convenait de trouver une architecture qui simplifie les étapes à franchir entre ces documents XML et l'objet final affiché dans la fenêtre d'un lecteur Internet. Idéalement, celle-ci devrait permettre de conserver ces documents XML comme nœuds du réseau hypertextuel.

Cette première analyse étant faite, il est apparu que l'architecture capable de supporter ces contraintes pourrait se composer de la manière suivante :

> Un automate hypertexte serait la réelle interface entre le lecteur et les documents, cet automate serait capable de comprendre les requêtes de l'utilisateur, de les analyser et en fonction de celles-ci de construire un document, à partir des documents de sa base de connaissance. Le terme base de connaissance est employé à la place de base de données. Les bases de données relationnelles sont le fondement de la quasi totalité des sites dynamiques

[3] Les développements de ce paragraphe doivent beaucoup à Marc Augier, notre informaticien, architecte du site, et professeur au CERAM à Nice.

commerciaux. Notre contexte est très différent, nos besoins et nos attentes sont donc eux aussi différents, aussi nous avons choisi de ne pas utiliser une base de données relationnelles, mais plutôt de conserver la base de connaissance au format XML.

Ce choix nous offre beaucoup plus de souplesse :

- Mise en ligne de nouveaux ouvrages sans phase de transformation préalable en vue de l'importation dans une base de données.
- Les documents sont conservés sous un format qui permet de les retravailler directement, on peut soit les afficher grâce à une feuille de style CSS, soit les traduire grâce à XSL.
- XML est implémenté dans PHP, d'où une manipulation simplifiée.

Le contexte technologique

Pour assurer la pérennité de nos choix, il est vite apparu que nous devions être indépendants des sociétés d'éditions de logiciel. Typiquement, le problème de tout système informatique est celui de la maintenance et du suivi. Si nous adoptions la solution d'un éditeur, il nous fallait ensuite continuer à suivre ses recommandations en termes de renouvellement de logiciel et risquer de se trouver dans le futur devant soit la disparition de l'éditeur, soit la disparition du logiciel du catalogue de l'éditeur, soit encore face à des coûts de maintenance subitement prohibitifs.

Dans ce cas, il nous faudrait être capable de migrer notre solution vers une autre, mais choisir une solution « propriétaire », cela veut souvent dire être enfermé dans cette solution, et courir le risque de ne pas pouvoir assurer la migration vers un système plus fiable ou moins coûteux.

Pour toutes ces raisons nous avons fait le choix de n'utiliser que des logiciels libres. Pour en savoir plus sur ce que cache cette expression il suffit d'aller lire sa définition sur le site http://www.gnu.org/.

L'expression « logiciel libre » fait référence à la liberté pour les utilisateurs d'exécuter, de copier, de distribuer, d'étudier, de modifier et d'améliorer le logiciel.

Notre plate-forme technique de base s'est donc constituée autour de Linux (Mandrake), Apache, PHP et MySQL. En plus de la référence au logiciel libre, on peut justifier ce choix de plusieurs manières :

- Techniquement cette solution est reconnue et éprouvée, les statistiques semblent montrer que le couple Linux+Apache équipe au moins la moitié des serveurs Internet.
- Economiquement, ce choix est bien entendu absolument sans comparaison.

Le choix de PHP comme langage de script pour dynamiser la création des pages nous a ouvert plusieurs portes intéressantes : d'une part on peut trouver de nombreuses librairies de codes et même des applications distribuées sous licence GNU, d'autre part PHP possède des librairies de fonctions traitant directement les documents XML (phpdom) et donc nous pouvons plus simplement lire et analyser les documents XML.

La base MySQL sert en fait à l'administration du site et contient aussi certaines définitions comme le glossaire.

6. Les ambitions du site CoLiSciences

Nos ambitions sont au nombre de quatre :

- *Culturelles et patrimoniales* : Il s'agit de collecter et mettre à disposition un grand corpus des ouvrages des principaux biologistes et naturalistes du 19e siècle, en langue française (près de 6 000 pages déjà offertes).
- *Intellectuelles et épistémiques* : Le choix de ces textes permet de retracer une « histoire des idées », à savoir le développement durant cette période, d'une science moderne du vivant, articulée en plusieurs grands domaines : l'anatomie, l'anthropologie physique, la classification des espèces, les théories de l'évolution et de la sélection, la physiologie, l'éthologie, etc.
- *Scientifiques au sens de la modélisation sémantique* : L'architecture du site traduit partiellement les réflexions de l'équipe centrées sur la problématique des hypertextes.

  On peut définir l'hypertexte comme un système interactif qui permet de construire et de gérer des liens sémantiques entre des objets repérables dans un ensemble de documents. Ici, le lecteur peut, entre autres, à partir du texte, accéder à : 1) un glossaire des termes scientifiques et techniques, 2) un répertoire des notions, 3) un dictionnaire des savants cités dans chaque texte. Des *parcours de lecture* lui sont proposés grâce à l'établissement de *liens hypertextuels* exprimant les relations sémantiques et conceptuelles que les notions entretiennent entre elles au travers des textes.
- *Cognitives et pédagogiques* : Une de nos problématiques centrales est celle de la lecture et de la navigation dans une double perspective : 1) Les modalités de la lecture vont-elles radicalement changer avec le support électronique ? Quelles spécificités nouvelles seront introduites dans l'acte de lire ? 2) Réciproquement, comment spécifier ces nouvelles conditions de l'offre de lecture pour l'apprentissage ? Comment maîtriser une hypertextualité largement déployée au travers des parcours offerts ?

Plusieurs types de parcours de lecture sont ainsi rendus possibles :

- ⇨ *Un parcours érudit* : lecture des textes et des notices historiques et critiques permettant d'éclairer les textes ;
- ⇨ *Un parcours paratextuel et historique* : les auteurs, les biographies, les bibliographies ;
- ⇨ *Un parcours épistémologique* : approche de la genèse des principales notions scientifiques avec leurs notices explicatives ;
- ⇨ *Un parcours cognitif* : les différentes compréhensions des œuvres et des « révolutions » scientifiques apportées par celles-ci ;
- ⇨ *Un parcours didactique* : l'introduction à des méthodes de pensée et à des formes de conceptualisation historiques et spécifiques.

7. Lecture sur écran et parcours sémantiques

7.1 Hypertexte et réseau de connaissances

En 1936, l'écrivain H.G. Wells décrivait déjà l'organisation d'un réseau nerveux qui servirait à tisser les liens entre les acteurs intellectuels mondiaux grâce à un canal universel de communication et d'échange. Un peu plus tard, en 1945, V. Bush définissait le principe de l'hypertexte :

> *« L'esprit humain opère par association. On ne peut pas espérer dupliquer pleinement ce processus mental artificiellement […] Mais il serait possible de dépasser la puissance et la permanence par le stockage de l'information aux fins de mieux stocker les informations et les mettre en interaction ».*

L'hypertexte est donc d'abord, un système d'organisation et de classification des connaissances. Il modélise des fonctionnements qui tentent de s'approcher de ceux du cerveau, tels les mécanismes associatifs de la mémoire ou l'emploi de métaphores visuelles pour accéder à un concept (principe de l'hyper-image), etc. A la différence du texte traditionnel composé d'une suite séquentielle de paragraphes, les documents qui composent l'hypertexte ne sont pas reliés les uns aux autres de manière continue, mais visent à s'organiser selon un réseau hiérarchique cognitivement ordonné.

7.2 La co-construction des parcours de lecture

En raison des renvois multiples qui le constituent, l'hypertexte n'a d'une certaine façon, ni début ni fin. C'est une structure évolutive à laquelle peuvent être ajoutés selon les besoins, de nouveaux liens et documents. L'écriture hypertextuelle nécessite donc la décomposition du domaine de savoir considéré en unités d'information interdépendantes. On sépare pour permettre le développement de nouvelles mises en relation. On ne peut cependant jamais anticiper les orientations de lecture qu'adopteront les différents utilisateurs. Selon le contexte, chacun peut décider de son cheminement dans la gamme des parcours possibles. En fonction des degrés de liberté qui lui sont accordés par la structure hypertexte, l'utilisateur peut élaborer son propre programme de lecture par concaténations d'un fragment à un autre, par ajouts et retraits de nœuds et de liens. Ainsi, la constitution de données hypertextuelles autorise un processus de recherche constructiviste au cours duquel les hypothèses s'élaborent et se reconfigurent selon l'effet des informations rencontrées chemin faisant par l'internaute. L'*hyperlien* permet de la sorte, la co-construction d'un « scénario » sous la forme d'enchaînements grâce auxquels le lecteur construit sa propre *logique de parcours*.

En résumé, l'hypertexte n'est pas seulement à lire, mais aussi à écrire. Il ne s'agit pas d'un moyen de connaissance au sens ordinaire, mais bien d'un *système de métaconnaissances* : on actualise les programmes conçus par d'autres pour construire ses propres parcours de sens, ses circulations dans des architectures cognitives, ses consultations de banques d'informations, etc.

7.3 Navigation

L'hypertexte permet un processus d'appropriation des relations spatiales et cognitives entre différentes informations : c'est un véritable mode d'exploration. Prenons l'exemple d'une découverte urbaine : on peut visiter une ville à partir d'un plan avec un programme de lieux où se rendre. Mais en présence d'une intersection ou d'un carrefour inattendus, certains visiteurs se laisseront guider par le contexte et changeront de direction : c'est la déambulation qui crée alors une nouvelle topographie de l'exploration. De même, lors d'une requête dans un hypertexte, tout élément de réponse induit une forme supérieure de questionnement et le lecteur collabore avec le programme pour générer un processus évolutif. L'interface assume ainsi un rôle de *miroir* en obligeant le chercheur à reformuler sa problématique et à en éliminer les aspects non pertinents.

7.4 Reconstructions du sens et parcours

La psychosociologie des interactions montre que les situations de communication comme l'entretien ou l'audition d'une émission impliquent chez le récepteur des opérations mentales telles que le filtrage des informations en fonction de ce qu'il juge pertinent, le remodelage de ces informations selon ses propres

présupposés et éventuellement l'ajout d'unités de sens destinées à « compléter les blancs ». Les études comportementales ont montré qu'un grand nombre de lecteurs sélectionnaient les pages à lire et à écarter en fonction des *toutes premières lignes*. C'est pourquoi, dans la mise en page hypermédia, la présence de liens répartis de façon équilibrée dans la page est un facteur puissant de dynamisation de la lecture. Le lien hypertexte permet au lecteur de découvrir quelle réalité se dissimule derrière un titre ou un document et de créer une connexion vers un nouveau sous-ensemble cognitif. En conséquence, le concepteur d'hypertexte doit savoir contrôler le contenu de chaque document hyperlié en le rapportant non pas à une progression linéaire mais à un parcours qui peut être aléatoire, fragmenté, décontextualisé. Enfin, un problème structurel est celui de la *surcharge cognitive* : il faut éviter au lecteur de se perdre dans l'arborescence du site.

8. « Visiter » CoLiSciences

Sur la page d'accueil, on va trouver cinq entrées :

- *présentation du site,*
- *acteurs et soutiens,*
- *accès corpus,*
- *le projet CoLiSciences,*
- *mode d'emploi.*

Explications :

- « ***Présentation du site*** » résume les quatre grandes ambitions du programme CoLiSciences : culturelles et patrimoniales, épistémologiques, scientifiques, cognitives et pédagogiques.

- « ***Acteurs et soutiens*** » présente les membres de l'équipe « Hypertextes et textualité électronique » en charge du programme, nos partenaires et les soutiens institutionnels et financiers du programme.

- « ***Accès corpus*** » : il s'agit là du « cœur » du site : on y accède aux textes constitutifs du corpus sous quatre entrées : auteurs, texte en fac similé, texte en mode numérisé, notions et relations sémantiques dans le texte.

- « ***Le projet CoLiSciences*** » : on trouvera là un menu permettant d'accéder à plusieurs types de textes en pdf rangés par dossiers :
  - Présentations du projet et du corpus,
  - Réflexions sur l'hypertexte,

- Réflexions sur la numérisation,
- Modèles d'analyse sémantique et cognitive inspirant la démarche,
- Projets en réponse aux appels d'offre.

- L'entrée « ***mode d'emploi*** » :

  ⇨ Les points d'entrée du site et ce qu'ils offrent :

  Dès l'ouverture de l'*accès corpus* en page d'accueil, quatre onglets d'entrée s'offrent au visiteur : *les auteurs, les disciplines, les domaines, les notions*.

  - Si on clique sur *auteurs*, on y trouve la liste de tous les ouvrages du corpus rangés par noms d'auteurs.
  - Les *disciplines* : on trouvera là les ouvrages rangés selon le contenu et l'ancrage disciplinaires (exemples : la physiologie cellulaire, l'anatomie, etc.).
  - Les *domaines* : ceux à l'intérieur desquels on peut ranger les ouvrages selon les types d'objets dont ils traitent (exemples : l'évolution, la tératologie, les insectes, etc.).
  - Si on clique sur l'onglet *notions*, on va se voir offrir la liste des notions (trente-sept notions) et la possibilité de rechercher chacune d'elles comme notion principale en relation avec une notion secondaire ou réciproquement, en tant que notion secondaire associée à telle ou telle notion principale, en même temps qu'apparaissent à l'écran si on le souhaite (requête : *sélection de parcours notionnels*) les textes des paragraphes concernés.

  ⇨ Où aller et dans quel ordre :

  Sont ici proposés quelques schémas de parcours ; il en existe bien d'autres.

  En cliquant sur le titre d'un ouvrage dans la liste des ouvrages par auteur, le texte de cet ouvrage va apparaître à l'écran selon trois types de présentations :

  - Le *fac similé* : tout le texte est disponible en mode image, c'est-à-dire l'image de l'édition originale de l'ouvrage, y compris en *vue d'ensemble*.
  - Le *texte* : le texte est alors disponible en mode texte et présenté par paragraphes ; les termes techniques et les noms de savants sont surlignés ; on accède en cliquant à leur définition dans un *glossaire* d'une part, ou dans un *dictionnaire des savants cités* d'autre part (onglets au bas de l'écran).

- L'entrée fac similé et l'entrée texte sont synchronisées.
- L'entrée *notions et relations* offre deux opportunités de recherche : découvrir les notions et relations sémantiques présentes dans chaque paragraphe et dégager une *sélection de parcours notionnels*, telle qu'explicitée précédemment ; on peut aussi afficher sous forme d'icônes tous les paragraphes où telle ou telle notion apparaît (« vue aérienne »).

9. Les parcours dans CoLiSciences

On pourra ainsi spécifier plusieurs types de parcours témoignant du statut de l'hypertexte comme « lieu de créations » :

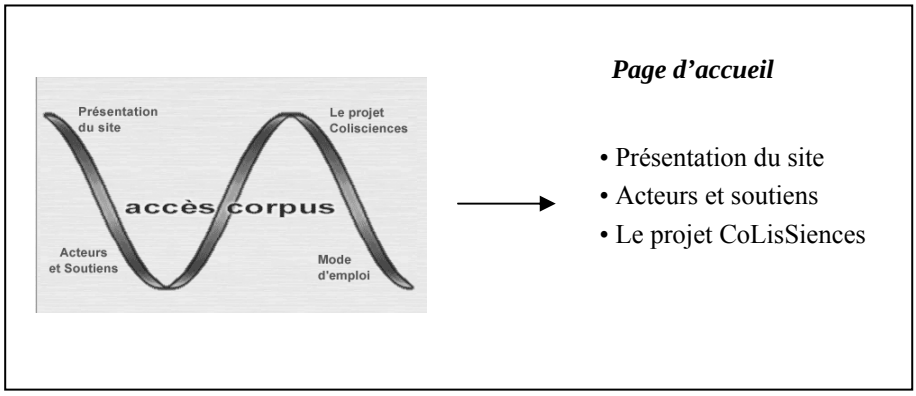


Figure 1 : Parcours informatif (pour « situer » le site)

Page d'accueil : les textes en pdf

- réflexions sur l'hypertexte
- modèles sémantiques
- mode d'emploi du site

L'ensemble des paratextes à télécharger
au format PDF (adobe acrobat reader)

Présentation du site
* Présentation du site
* Soutiens institutionnels

Le projet CoLiSciences :
* Le projet CoLiSciences
* Acteurs et Partenaires du programme

Réflexions sur l'hypertexte :
* Marc Augier : Construction d'un système hypertextuel sur le Web
* Eric Bruillard : Hypertexte. De l'accès aux documents à leur structuration
* David Piotrowsky : De l'hypertextualité en général et de CoLiSciences en particu
* Georges Vignaux : L'hypertexte

Réflexions sur la numérisation :
* Abdel Belaïd : Reconnaissance automatique de l'écriture et du document
* Marie-Elise Fréon : Conversion de documents

Cognition et langage :
* Georges Vignaux : Processus d'ajustement et opérations langagières
* Georges Vignaux : Opérations langagières, opérations cognitives
* Georges Vignaux : Un essai d'analyse "La femme a le coeur plus tendre que l'h

Mode d'emploi :
* Mode d'emploi
* Typologie des relations (Les relations sémantiques et cognitives appliquées au
* Glossaire et notions : leurs rôles
* Les savants cités

Les notions :
* Marc Silberstein : Experimentum crucis (expérience cruciale)
* Georges Vignaux : Analogie
* Georges Vignaux : Empirisme
* Georges Vignaux : Raisonnement

(Cliquer sur les textes sur les textes * afin de les télécharger au format PDF)

Figure 2 : Parcours instructif

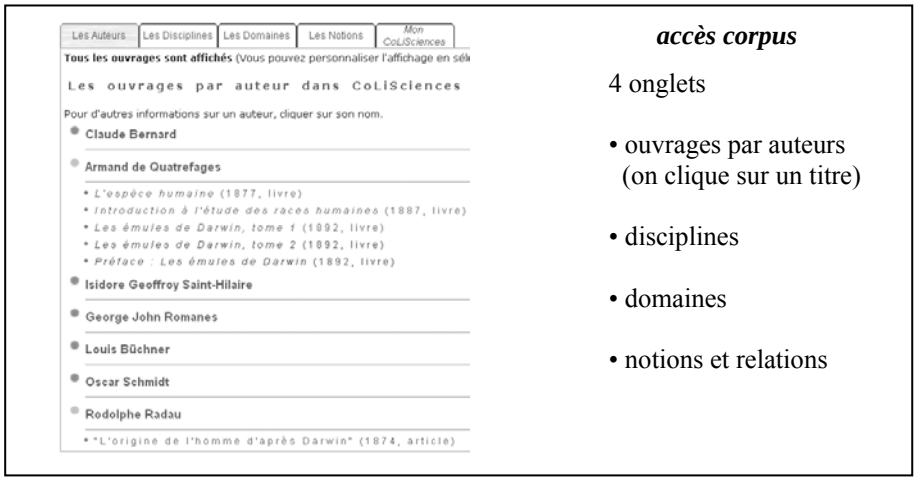


Figure 3 : Parcours exploratoires

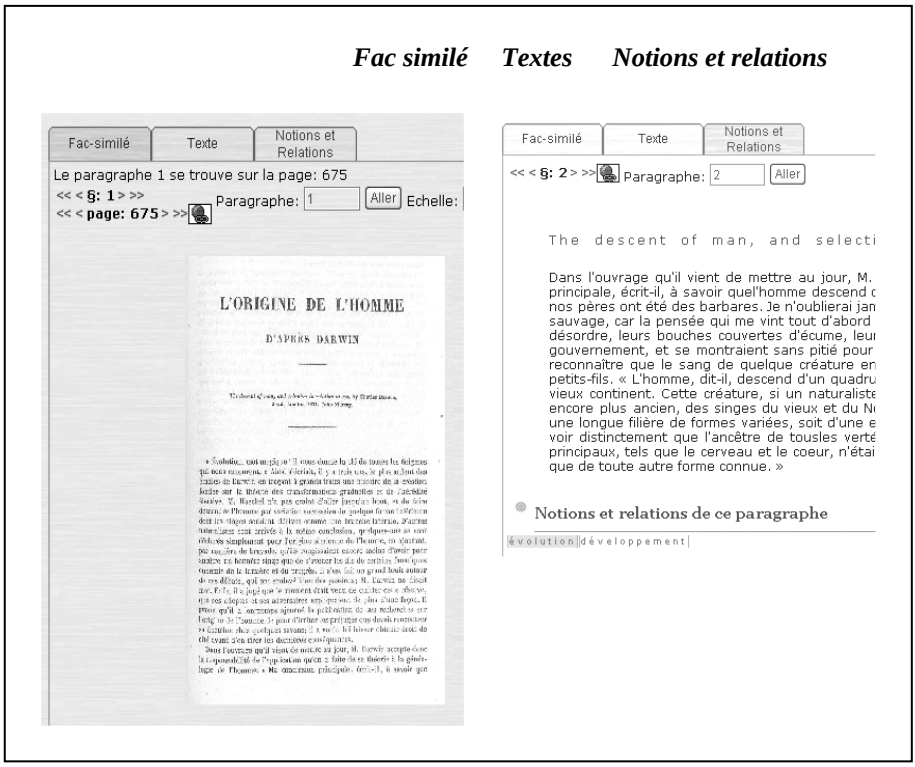


Figure 4 : Parcours érudit (histoire des idées)

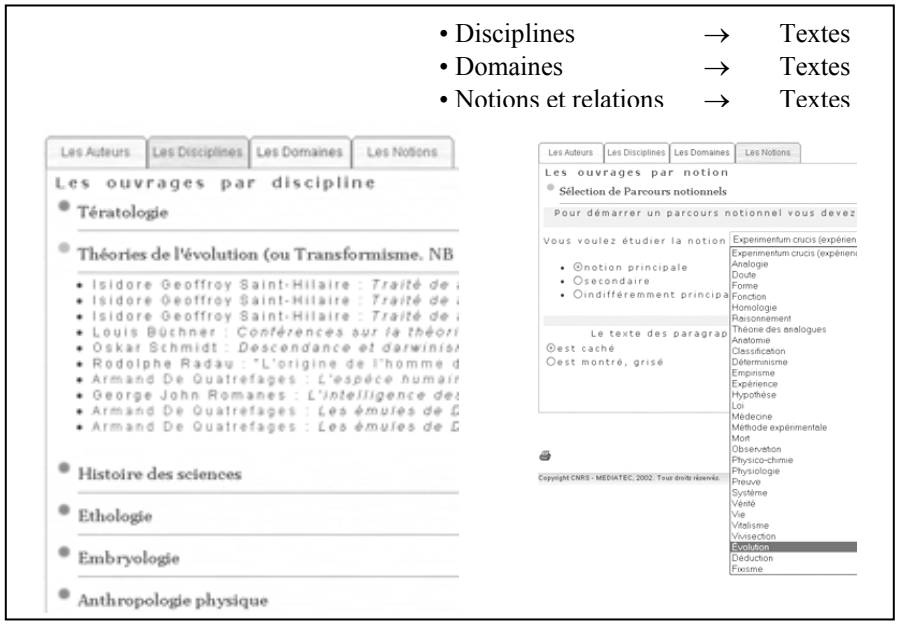


Figure 5 : Parcours cognitifs (placements de connaissances)

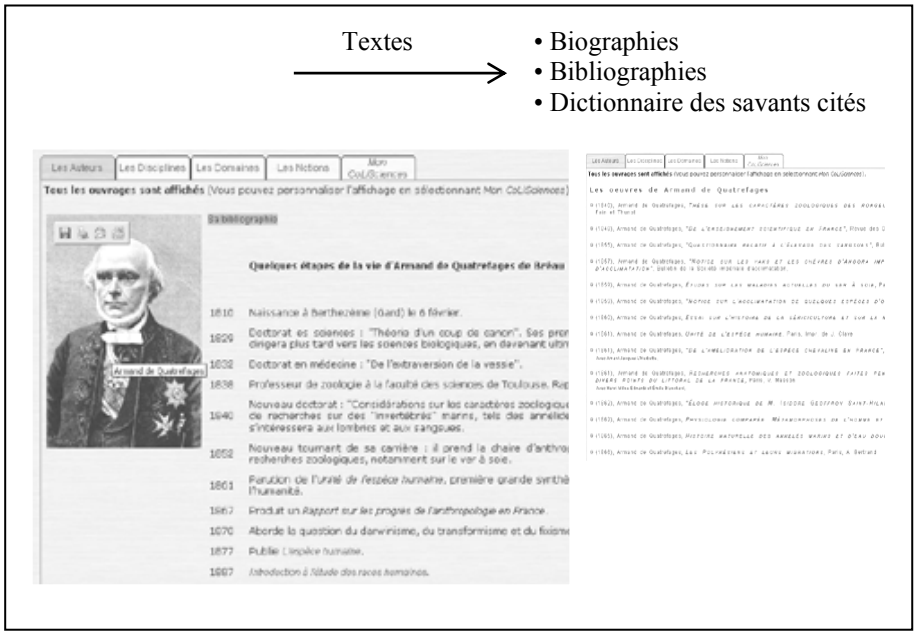


Figure 6 : Parcours encyclopédiques

10. CoLiSciences : un outil pour l'étude des processus d'appropriation

Considérant la complexité intrinsèque de ce qui est proposé, à la fois en termes de contenu et d'exploitation de celui-ci, ainsi que l'absence d'une tradition d'exercisation dans ce domaine, il paraît impossible de trouver des formes d'utilisation éducative suffisamment proches de ce qui peut être fait dans les formations institutionnelles. Cela conduit soit à imaginer des activités simples centrées sur des points précis soit des scénarios plus sophistiqués. L'idée centrale est donc celle de *la création de parcours*. On peut d'abord regarder du côté des experts, essentiellement des enseignants de SVT (Sciences de la Vie et de la Terre), des historiens des sciences et des philosophes. Ces experts sont conduits à bâtir des cours, notamment autour de notions ou de problèmes abordés ou traités dans le corpus indexé dans le cadre du programme CoLiSciences. Comparer les chemins construits, correspondant en gros à des visites guidées, et décrire comment sont utilisés les outils proposés (par exemple les relations entre les notions aidant à construire *une généalogie de ces notions en termes d'histoire des idées*) est un objectif essentiel. Les observations des experts peuvent fournir des indications sur les usages possibles par des plus novices et les chemins collectés enrichissent le corpus par l'adjonction de parcours sémantiques finalisés.

Revisiter le concept d'hypertexte apparaît donc fondamental, y compris dans le contexte éducatif actuel (secondaire et supérieur) où la recherche d'informations, leur organisation et leur réorganisation, dans le cadre d'activités interdisciplinaires, occupent une place croissante.

11. Références

Références sur les fondateurs de l'hypertexte

Bush et le Memex

- Brève présentation de Vannevar Bush : http://www.iath.virginia.edu/elab/hfl0034.html
- *As we may think*: http://www.theatlantic.com/unbound/flashbks/computer/bushf.htm ou http://www.isg.sfu.ca/~duchier/misc/vbush/
- Une image du Memex : http://www.kerryr.net/pioneers/memex_pic.htm

Douglas Engelbart

- Brève présentation : http://www.iath.virginia.edu/elab/hfl0035.html
- Démonstration en 1968 : http://sloan.stanford.edu/MouseSite/1968Demo.html
- Interview pour les enfants : http://www.superkids.com/aweb/pages/features/mouse/mouse.html
- Le travail coopératif : http://www.bootstrap.org/
- *Augmenting Human Intellect: A Conceptual Framework* (http://www.histech.rwth-aachen.de/www/quellen/engelbart/ahi62index.html)

Bush et Engelbart (pour l'histoire des interfaces graphiques).

- http://www2.kenyon.edu/people/adamsal/gui/bush_engelbart.htm

Ted Nelson et l'hypertexte :

- Brève présentation : http://www.iath.virginia.edu/elab/hfl0155.html
- Projet Xanadu : http://xanadu.com/
- Sa home page : http://www.sfc.keio.ac.jp/~ted/
- Ted Nelson et le mot hypertexte en 1965 : http://iberia.vassar.edu/~mijoyce/Ted_sed.html
- Memex et au-delà : http://www.cs.brown.edu/memex/

Autres références

Agosti, M. Crestani, F. Melucci, M. (1996), "Design and implementation of a tool for the automatic construction of hypertexts for information retrieval", *Information Processing & Management*, 1996, 32, 4, 459-476

Arents, H. Bogaerts, W. (1993), "Concept-based retrieval of hypermedia information: from term indexing to semantic hyperindexing", *Information Processing & Management*, 1993, 29, 3, 373-386

Amélineau,, C. & Giovanni, L. (1996), "Utilisation pédagogique de l'outil logiciel hypertexte avec un public d'adultes illettrés en stage d'insertion sociale et professionnelle". In Bruillard, E., Baldner, J.M. & Baron, G.L. (dirs.). pp. 175-182.

Baddeley, A. (1986), *Working Memory*. New York, Oxford University Press.

Baddeley, A. (1990), *Human Memory. Theory and Practice*, Londres, LEA.

Baldner Jean-Marie, Bruillard Éric (2000), *L'usage des manuels scolaires et des ressources technologiques dans la classe*. Rapport de la première année de la recherche INRP 40124, 121 p. Première partie reprise en Point de Vue dans *Sciences et Techniques Educatives*, vol. 7, n° 2, Hermès, pp. 443-480.

Balpe, J.P., Lelu, A., Papy, F. & Saleh, (1996), Techniques avancées pour l'hypertexte. Paris : Hermès.

Baron Georges-Louis, Bruillard Éric (2001), Une didactique de l'informatique ? *Revue Française de Pédagogie*, n° 135, pp. 163-172.

Bates, M. (1989), "The design of browsing and berrypicking techniques for online search interface", *Online Review*, 1989, 13, 5, 407-423

Benoit, J., Fayol, M. (1989), « La catégorisation des types de textes », *Pratiques*, 62, 71-85.

Bosak Jon, et Bray Tim, "XML and the Second-Generation Web", *Scientific American*, May 1999.

Bruillard Éric (1997), *Les machines à enseigner*. Éditions Hermès, Paris, 320 p.

Bruillard Éric, Grandbastien Monique (eds.) (2001), *Éducation et informatique. Hommage à Martial Vivet*. Sciences et Techniques éducatives, vol. 7, n° 1, Hermès Science, 300 p.

Bruillard Éric, de La Passardière Brigitte et Baron Georges-Louis (eds.) (1998), *Le livre électronique*. Sciences et Techniques Educatives, vol. 5, n° 4, Hermès Science.

Bruillard Éric, de La Passardière Brigitte. (1998), « Fonctionnalités hypertextuelles dans les environnements d'apprentissage », in Tricot A. et Rouet J.-F. (dir.) *Les hypermédias,*

approches cognitives et ergonomiques, Hermès, Paris, pp. 95-122 (correspondant à un numéro spécial de la revue Hypertextes et Hypermédias).

Cattenat, A., Paul, G. (1992), « Thésaurus ou non ? ou comment interroger des bases de données textuelles sans être savant », *Les bases de données avancées* I. Saleh (ed), 1992, Hermes, 11-22

Chanier, T. (1988), « Hypertexte, hypermédia et apprentissage dans des systèmes d'information et de communication ». Étude de Linguistique Appliquée (ELA). nº 110, avril-juin. pp 137-146.

Chanier, T. (1996), "Learning a Second Language for Specific Purposes within a Hypermedia Framework". Computer-Assisted Language Learning (CALL), vol. 9, 1. pp 3-43.

Chanier, T. & Selva, T. (1997), "Visual representations in lexical learning environments: application to the ALEXIA system". Conference Computer Assisted Language Learning, Exeter, September. A paraître.

Collombet-Sankey, N. (1997), "Surfing the net to acquire communicative and cultural knowledge". In Debski, R., Gassin, J. & Smith, M. (dirs.), pp. 143-158.

Chartier, R., « Bibliothèques sans murs », dans *L'ordre des livres, lecteurs, auteurs, bibliothèques en Europe entre XIV^e^ et XVIII^e^ siècles*, Aix-en-Provence, Alinéa, 1992, pp. 69-94, – et J.M. Goulemot, « En guise de conclusion : les bibliothèques imaginaires (fictions romanesques et utopies) », dans *Histoire des bibliothèques françaises*, Paris, Promodis-Editions du Cercle de la Librairie, tome II, « Les bibliothèques sous l'Ancien Régime », C. Jolly ed., 1989, pp. 500-511.

Chartier, R., « Du codex à l'écran : les trajectoires de l'écrit », 1994, www.info.unicaen.fr: bnum:jelec/Solaris/d01/1chartier.html.

Chartier, R., « Révolutions et modèles de lecture », XV^e^-XX^e^ siècles. *Le Français aujourd'hui*, décembre 1995, n° 112, pp. 6-15.

Cochrane, P., Johnson, E. (1996), *Visual Dewey : DDC in Hypertextual Browser for the Library User*, *Advances in Knowledge Organization*, 5, 95-106.

Crinon, J., Legros, D., Pachet, S. & Vigne, H. (1996), « Étude des effets de deux modes de navigation dans un logiciel d'aide à la réécriture ». In Bruillard, E., Baldner, J.M. & Baron, G.L. (dirs.). pp. 73-84.

Culioli, A. (1989), "Representation, referential processes, and regulation. Language Activity as Form Production and Recognition". Genève : Fondations Archives Jean Piaget, *Cahier n° 10, Language and Cognition*.

Denhiere, G., Baudet, S. (1992), *Lecture, compréhension de texte et science cognitive*. Paris, PUF.

Fayet-Scribe, S. (1997), *Chronologie des supports, des dispositifs spatiaux, des outils de repérage de l'information*. http://www.info.unicaen.fr/bnum/jelec/Solaris/d04/4fayet_0intro.html

Fayol, M. (1997), *Des idées au texte*. Paris, PUF.

Fayol, M. *& al.* (2000), *Maîtriser la lecture*, Paris, Odile Jacob.

Ferrand Nathalie (éd.) (1997), *Banque de données et hypertextes pour l'étude du roman*. Paris, PUF.

Fuchs, C. (1994*)*, *Paraphrase et énonciation*, Paris-Gap, Ophrys.

Fuchs, C., Robert, S. (1997), *Diversité des langues et représentations cognitives*, Paris-Gap, Ophrys.

Ganascia Jean-Gabriel (2001), « Du néo-structuralisme supposé de l'hypertextualité », *Diogène*, n° 16, 9-24.

Grandbastien Monique (à paraître, 2002), « Quelques questions à propos de l'indexation et de la recherche de ressources pédagogiques sur le WEB » in *Les technologies en éducation : perspectives de recherche et questions vives*, Baron Georges-Louis, Bruillard Éric (eds), INRP, Paris.

Grize, J.B. (1990), *Logique et Langage*. Paris-Gap : Éditions Ophrys.

Jackendoff, R. (1987), *Consciousness and the Computational Mind*. Londres : Bradford Book ; Cambridge, Mass. : MIT Press.

Le Ny, J. F. (1989), *Science cognitive et compréhension du langage*. Paris : Presses Universitaires de France.

Mangenot, F. (1996), *Les aides logicielles à l'écriture*. Paris : Centre National de Documentation Pédagogique (CNDP).

Lebrave, Jean-Louis (1994), « Hypertextes, Mémoires, Écritures », *Genesis*, n° 5.

Lebrave, Jean-Louis (1997), « Hypertexte et édition génétique : Flaubert », in Ferrand Nathalie (éd.).

McAleese, R., Duncan, E. B. (1987), "The Graphical Representation of "terrain" and "street" Knowledge in an Interface to a Database System", *Proceedings of the 11th International Online Information Meeting*, 443-456.

McDonaald, S., Stevenson, R. (1996), "Disorientation in hypertext : the effects of three text structures on navigation performance", *Applied Ergonomics*, 1996, 27, 1, 61-68.

McKnight, C. Dillon, A. Richardson, J. (1990), "A comparison of linear and hypertext formats in informtion retrieval", *Hypertext : State of the Art*, R. McAleese & C. Green (eds.), University of Aberdeen, Intellect, Oxford, England, 1990, 10-19.

Marshall C., Shipman F., Coombs J. (1994), VIKI : spatial hypertext supporting emergent structure in *Proceedings ECHT'94*, pp. 13-23.

Marshall C.C., Rogers R.A. (1992), Two Years before the Mist : Experiences with Aquanet, Proc. 4th ACM Conference on Hypertext, ECHT'92, Milan, pp. 53-62.

Moscovici Serge, Vignaux Georges (1994), « Le concept de thêmata », in *Pratiques et transformations des représentations sociales* (éd. C. Guimelli), Delachaux et Niestlé, 1994.

Nanard, M., Paolini, P., (eds.), Milano 11/30-12/4 1992, 131-140, SAVOY J. (1994), "A learning scheme for information retrieval in hypertext", *Information Processing & Management*, 1994, 30, 4, 515-533

Nanard, M. (1995), « Les hypertextes : au-delà des liens, la connaissance ». *Sciences et Techniques Éducatives* (STE), vol 2, 1. pp. 31-59.

Nauer, E., Lamirel, J.C. (1997), « Environnement d'investigation sur WWW : assistance à l'utilisateur par des connaissances fédérées », *Hypertextes et hypermédias*, 1997, 1, 2-3-4, 101-113

Nielsen, J., *Designing Web Usability*.

Les conclusions d'une étude de Jakob Nielsen, éditée par Sun Microsystems : http://www.sun.com/980713/webwriting/wftw9.html ; L'étude de l'Université de Stanford, précitée, est complètement accessible ici : http://www.poynter.org/eyetrack 2000/index.htm.

« Quelques recommandations pour la rédaction de contenus Web ». Dans cet article de l'Ergonome, fort bien documenté comme à leur habitude, il est question de lisibilité des textes à l'écran : http://www.lergonome.org/dev/pages/article_11.asp

Normand Sylvie, Bruillard Éric (2001), « Que révèlent les discours de futurs enseignants sur leur compréhension du fonctionnement des applications informatiques ». Point de vue, *Sciences et Techniques éducatives*, vol. 8, n° 3-4, Hermès Science, pp. 435-445.

Piotrowski David (1996), *Lexicographie et informatique : autour de l'informatisation du TLF*, Actes du Colloque de Nancy, Paris, Didier-Erudition.

Piotrowski David (1997), « Lexicographie et formes opératoires de l'hypertextualité », *Sémiotiques*, n° 12.

Piotrowski David, Silberstein Marc (2001), « Le prototype HyperCB : Principes, architecture et fonctionnalités d'un hypertexte », à paraître.

Rabardel, P. (1995), *Les activités avec instruments, de l'outil au système technique : une approche cognitive*. Paris : Armand Colin.

Rouet, J.F. & Tricot, A. (1995), « Recherche d'informations dans les systèmes hypertextes : des représentations de la tâche à un modèle de l'activité cognitive », *Sciences et Techniques Éducatives* (STE), vol 2, 3. pp. 307-331.

Vignaux, G. (1988), *Le Discours, acteur du monde. Énonciation, argumentation et cognition*. Paris-Gap, Éditions Ophrys.

Vignaux Georges (1992), *Les sciences cognitives : une introduction*, Paris, La Découverte (Le livre de poche, 1994).

Vignaux Georges (1996), « Hypertextes, dictionnaires : approche sémantique, perspective cognitive », in Piotrowski, D. (éd.) (1996).

Vignaux Georges, Piotrowski David, Kieu Quien (1998), « Lexicographie et Hypertextes », Actes du 4e Colloque Hypermédias et apprentissages, Poitiers, MSHS.

Vignaux Georges (2000), « L'hypothèse du livre électronique », *Les cahiers de médiologie*, n° 10.

Vignaux Georges (2001), *Le démon du classement*, Paris, Seuil.

Vignaux Georges (2003), *Du signe au virtuel*, Paris, Seuil.

Description des contenus musicaux et applications

Hugues Vinet

IRCAM-CNRS – STMS, 1 place Igor Stravinsky
75004 PARIS - France

vinet@ircam.fr

Résumé :

Ce texte présente sous forme résumée la teneur de cette conférence invitée, dont l'objet est de dresser un état de l'art synthétique des recherches en matière de description des contenus musicaux et sonores, d'extraction automatisée de ces descripteurs à partir des informations musicales, et de leurs applications, existantes ou potentielles. Ce propos s'appuie à cet effet sur des exemples de travaux récents, en particulier issus des recherches de l'Ircam et du projet européen CUIDADO (Content-based Unified Interfaces and Descriptors for Audio/music Databases available Online) qui s'est achevé fin 2003.

Mots-clés : Musique, Son, MPEG7, Indexation, Bases de données multimédia, Apprentissage automatique, Technologies cognitives, Théorie de l'information.

Abstract:

This paper presents a sum-up of this invited conference, which proposes a synthetic state-of-the-art of current research on the description of musical and audio contents and on the automatic extraction of these descriptors from musical information, and their existing and potential applications. It is based on recent examples taken from research performed at IRCAM and in the framework of the EU-funded CUIDADO (Content-based Unified Interfaces and Descriptors for Audio/music Databases available Online) project, completed end 2003.

Keywords: Music, Sound, MPEG7, Indexing, Multimedia Databases, Machine Learning, Cognitive Technologies, Information Theory.

1. Problématique

Les enjeux économiques et culturels liés à la diffusion de contenus musicaux numérisés sont considérables et les usages liés aux développements technologiques récents donnent lieu à un bouleversement de l'industrie de distribution de la musique enregistrée et des modèles économiques sous-jacents. Dans ce contexte, les problématiques de description des contenus musicaux fédèrent depuis quelques années une communauté croissante de chercheurs dans le domaine de la recherche d'informations musicales (*Music Information Retrieval*), issus de disciplines diverses (bibliothéconomie, analyse statistique et apprentissage, traitement du signal, cognition musicale et musicologie, informatique musicale, etc.). La notion de descripteur se réfère ici à la terminologie Mpeg7 et s'inscrit dans le cadre du développement de nouvelles applications de gestion et de manipulation des informations musicales *par le contenu*, c'est-à-dire dans lesquelles l'utilisateur accède à ces contenus à travers des représentations intermédiaires, dites *métadonnées*. Le cas typique d'application concernée est la recherche et la navigation par contenu dans des bases de données d'enregistrement musicaux. On distinguera en particulier descripteurs de haut niveau, formalisant les structures de connaissances liées à l'application, de descripteurs de bas niveau, non nécessairement accessibles à l'utilisateur, mais pouvant dans certains cas être calculés automatiquement à partir des contenus musicaux. Les verrous scientifiques et technologiques concernent notamment, dans le cas général :

- La formalisation des structures de connaissances et leur médiation à l'utilisateur sous la forme d'interfaces homme-machine, ces deux aspects devant être conçus dans leur interdépendance en vue de réalisation des fonctions visées,
- L'automatisation de l'extraction de ces informations à partir des contenus eux-mêmes, via la constitution de descripteurs de bas niveau adaptés,
- La mise en œuvre de ces différentes structures de description et interfaces utilisateur dans des applications en vraie grandeur,
- La prise en compte des usages de telles applications, en particulier du point de vue du caractère évolutif et personnalisable des bases de connaissances et des corpus, qui renvoie au premier point.

2. Les niveaux de représentation des informations musicales

Le problème étant posé, il convient d'abord de préciser les spécificités des contenus musicaux et sonores. Nous proposons une caractérisation des représentations numériques des informations musicales, dans les différentes applications existantes, selon quatre types, organisés en niveaux d'abstraction croissants, les niveaux *physique*, *signal*, *symbolique* et *cognitif*, ces notions se

référant à des classes de données bien précises. Du point de vue de la théorie de l'information, il est aisé de montrer que cette organisation en niveaux rend compte de quantités d'information décroissantes, et que la conversion des données correspondant au passage d'un niveau à un autre procède, respectivement dans les sens ascendant et descendant, des *processus génériques d'analyse et de synthèse*, liés à la mise en œuvre d'une réduction (extraction) ou d'une augmentation (génération) de la quantité d'information.

Cette structuration met en particulier en évidence un premier stade de problématiques scientifiques relatives aux relations entretenues entre données de types signal et symbolique : analyse pour l'extraction de structures musicales à partir du signal (de l'enregistrement aux notes), et réciproquement synthèse pour le passage de la notation au signal. Du point de vue de l'analyse, l'état de l'art actuel en traitement de signal ne permet pas aujourd'hui l'extraction automatique des représentations symboliques dans le cas général, notamment dans le cas d'enregistrements polyphonique. Les méthodes, dites d'*alignement*, assurant une indexation automatique de données symboliques (MIDI) à partir des enregistrements correspondants, atteignent cependant un taux de fiabilité satisfaisant et commencent notamment à être utilisées pour les applications de *comparaison d'interprétations*.

Dans un deuxième stade, une part importante de descripteurs de haut niveau relevant par définition du niveau cognitif, l'extraction d'informations à ce niveau à partir des niveaux signal et/ou symbolique, à travers notamment la mise en œuvre de techniques d'apprentissage, constitue une part importante des recherches en matière d'extraction automatisée.

3. Descripteurs globaux et applications

Une première classe de descripteurs de haut niveau, la plus simple à aborder, concerne des descripteurs relatifs à l'intégralité du contenu sonore considéré, notamment dans son déroulement temporel, que nous désignons ici par « descripteurs globaux ». Ceux-ci sont adaptées à des fonctions de recherche et de navigation *inter-documents*, ainsi que de manipulation globale des contenus, et concernent notamment les applications suivantes, chacune associée à des structures de descripteurs de haut niveau, de possibilités d'extraction automatique, et d'heuristiques de manipulation spécifiques :

- Classification et navigation par le contenu dans les *bases de données d'échantillons sonores*[1]. Différents types de descriptions de haut niveau sont mis en œuvre à cet effet, notamment :
  - Les taxonomies de sources sonores, avec la possibilité de d'inférer les classes correspondantes, à l'aide de méthodes de classification automatique,

[1] Sons isolés utilisés en production sonore et musicale.

  - Des descripteurs verbaux sous différentes formes : adjectifs pour qualifier globalement le son (*métallique, sombre, sourd*, etc.), onomatopées (*cling, bip*…), locutions telles que sujet/verbe/ complément se référant à l'action de production sonore, voire verbalisations libres sans contrainte structurelle,
  - Les espaces de timbres, issus d'études perceptives, également calculables à partir des signaux audio,
  - Des descripteurs morphologiques, relatifs à une description phénoménologique des sons dans la lignée des travaux de Pierre Schaeffer, dont de nombreux attributs peuvent également être extraits automatiquement des signaux audio : profils d'intensité et de hauteur, grain, sons itératifs, etc.

- *Identification des morceaux de musique* par l'utilisation de descriptions caractéristiques sous forme condensée, ou empreintes digitales (*fingerprinting*),
- *Navigation dans des bases de données musicales et génération de listes de morceaux* (playlists) selon des critères de contenu : informations documentaires et/ou éditoriales, taxonomies de genres, descripteurs extraits du signal (tempo, timbre, etc.), mesures de similarité « culturelle », etc.
- Méthodes d'*édition et traitement des sons par le contenu*, c'est-à-dire sur la base d'une spécification des attributs du son à produire (effet), et non de la description de l'algorithme de production (cause).

4. Description et indexation intra-documentaire

Une seconde classe de descripteurs de haut niveau, plus complexe à appréhender, concerne les structures d'indexation intra-documentaires. La difficulté principale vient du fait qu'il n'existe pas de manière unique d'indexer les contenus musicaux à un niveau pertinent de description (pour ne pas parler de sémantique, notion relativement inappropriée aux contenus musicaux). Même dans le contexte restreint de la musique traditionnelle, liée à une notation symbolique, s'il est possible, du moins théoriquement, d'indexer les notes constituant l'information musicale, la donnée de notes isolées, comparables aux lettres isolées de l'information textuelle, ne suffit pas pour construire des structures de description pertinentes. Un niveau de description plus élevé est nécessaire : celui des motifs musicaux, c'est-à-dire de structures obtenues par agrégation de notes isolées en séquence (ligne mélodique) et/ou en superposition (polyphonie), mais il n'existe pas de manière univoque de produire de tel motifs, qui seraient l'équivalent d'unités lexicales, à partir des informations musicales. Les approches existantes, encore peu développées, sont donc valides pour des types de corpus limités, et mettent en œuvre des heuristiques et points de vue d'analyse particuliers. On peut notamment citer :

- Le calcul de *résumé automatique*, c'est-à-dire la segmentation temporelle de l'ensemble du morceau en une suite d'états en nombre limité. Cette méthode est adaptée à certains corpus musicaux conçus selon un découpage clair en parties distinctes (Introduction, refrain, couplet, etc.) et des méthodes existent pour le calcul automatisé de ces états par analyse du signal,
- L'analyse de *profils mélodiques*, qui convient pour des mélodies simples (chansons avec accompagnement), et donnent lieu à différentes heuristiques de recherche, notamment la recherche par chantonnement (*query par humming*).
- *La séparation* automatique des sources sonores, visant à décomposer une polyphonie en voix ayant chacune sa propre variation,
- Des méthodes récentes, qui combinent de manière systématique différents critères de similarité pertinents (mélodiques, harmoniques, rythmiques), pour inférer l'extraction et la comparaison de structures motiviques. Ces méthodes fonctionnent dans le cas de morceaux conçus selon des règles de contrepoint.

Selon les cas, l'analyse est effectuée à partir du signal ou des représentations symboliques. L'application de ces descriptions intra-documentaires concerne des modes de recherche et de navigation à l'intérieur d'un morceau de musique, mais également entre documents différents, par exemple sur la base d'une comparaison de motifs musicaux entre morceaux. Devant la difficulté, voire l'impossibilité théorique, dans le cas général, de systématiser les procédures d'extraction automatique, la réalisation de procédures d'indexation et d'annotation manuelle doit être prise en compte, avec la possibilité de constituer des structures motiviques quelconques comme entités de référence, pouvant être alors reliées entre elles, catégorisées, selon tel ou tel point de vue d'analyse.

L'enjeu culturel de ces descriptions intra-documentaires est important : elles permettent de dépasser les interfaces actuelles de manipulation des contenus musicaux enregistrés, se limitant à des commandes de types lecture, arrêt, morceau suivant, en offrant au mélomane une présentation analytique du contenu musical, vecteur d'une nouvelle intelligibilité des œuvres.

Session 1

Structuration de documents

Exploitation de ressources lexicales pour la mise en hypertexte

Farid Cerbah

Dassault Aviation – DPR/ESA – *78, quai Marcel Dassault*
92552 Saint-Cloud cedex 300 - France

farid.cerbah@dassault-aviation.fr

Résumé :

Les nouveaux formats et outils documentaires intègrent des fonctionnalités permettant la création manuelle de liens, mais encore peu de services d'aide active à la pose de ces liens. Le coût rédactionnel peut dès lors s'avérer colossal, en particulier dans les domaines techniques où les fonds documentaires sont fortement hypertextuels. Nous proposons dans cet article une méthode de la mise en hypertexte s'appuyant sur l'estimation de similarités lexicales entre sources et cibles potentielles d'hyperliens. Nous présentons des expérimentations menées sur une documentation technique de taille conséquente, et nous mettons en évidence l'apport de différents types de ressources linguistiques générales.

Mots-clés : Mise en hypertexte, exploitation de ressources lexicales en RI, similarité lexicale, ressources sémantiques.

1. Introduction

La notion d'hypertexte est aujourd'hui largement banalisée dans le domaine de la rédaction technique. Elle permet de proposer des parcours de lecture plus conformes aux modes d'utilisation des documentations techniques qui se prêtent rarement à une lecture linéaire [3, 17]. Mais, le coût rédactionnel à consentir pour hypertextualiser une documentation peut s'avérer colossal (plusieurs centaines de milliers de liens pour une documentation de maintenance aéronautique).

Les nouveaux formats et outils documentaires intègrent des fonctionnalités supportant la création manuelle de liens, mais encore peu de services d'aide active à l'identification d'hyperliens potentiellement pertinents. Nous proposons dans cet

article une méthode de la mise en hypertexte s'appuyant sur l'estimation de similarités linguistiques entre sources et cibles potentielles d'hyperliens. Cette méthode a été expérimentée sur une documentation technique de taille significative. Outre la définition d'une méthode formalisée et générique de la mise en hypertexte, l'objectif de ce travail est d'évaluer l'apport de ressources linguistiques générales. Il s'agit d'estimer le niveau de performance d'une approche qui évite de solliciter des ressources terminologiques exigeant un effort d'acquisition important.

Nous commencerons par situer notre problématique dans le cadre de la documentation structurée (§ 2), et dans une perspective d'extension d'un environnement de production documentaire (§ 3). Nous aborderons ensuite la question de l'identification des ressources lexicales requises (§ 4) avant de consacrer un long développement à la définition d'une méthode formalisée (§ 5) et à son évaluation (§ 6). Nous conclurons sur quelques perspectives.

2. Hypertextualisation et documentation structurée

Face à la complexité croissante des objets techniques à documenter, la plupart des secteurs industriels ont évolué vers une conception plus structurante de la documentation technique qui favorise la modularisation pour gagner en réutilisabilité et maintenabilité. Les nouvelles pratiques rédactionnelles incitent fortement à fragmenter les fonds documentaire en unités autonomes au contenu clairement spécifié et fortement interconnectées au moyen d'hyperliens. Cette structuration renforcée du contenu s'exprime dans un langage de balise, SGML ou XML, qui rend possible la génération quasi-immédiate d'hyperliens d'ordre structurel. Il est cependant clair que nombre d'hyperliens ne peuvent être posés sans une interprétation plus ou moins profonde du contenu textuel balisé. L'insertion de ces liens est aujourd'hui entièrement à la charge des rédacteurs techniques.

Plusieurs raisons donnent à penser que le cadre de la documentation structurée offre un terrain particulièrement favorable à la mise en place d'une approche de la mise en hypertexte fondée sur l'analyse du contenu :

- L'organisation textuelle et paratextuelle des documentations produites est mieux explicitée, de sorte qu'il est possible de répartir avec une certaine précision les unités documentaires en genres. Par exemple, dans une documentation aéronautique, les unités procédurales sont bien distinguées des unités descriptives. Il est de fait envisageable d'adapter les mécanismes d'analyse et les ressources linguistiques impliquées aux genres des unités à traiter[1].
- Il est raisonnable de considérer que les extrémités des hyperliens sont des éléments de structure. C'est une donnée essentielle pour la mise en

[1] La nécessité de rendre les traitements d'analyse plus sensibles à la diversité des genres textuels est mise en exergue par plusieurs auteurs (cf. [5, 11, 1]..

hypertexte, en ce sens que le processus automatisé aura à mettre en relation des fragments dont la nature sémantique et les frontières sont bien identifiées. Il n'est pas nécessaire de procéder à une opération délicate de découpage en segments potentiellement pertinents à partir d'un flot textuel faiblement structuré.

- On sait que les traitements linguistiques sont extrêmement coûteux. La connaissance de la structure des modules documentaires permet de mieux caractériser les éléments susceptibles d'être mis en relation, et dans une optique d'automatisation, de réduire l'espace de recherche à des couples potentiellement pertinents compte tenu de leur nature sémantique.
- Enfin, il est possible de décider de la pertinence d'un lien sur la base d'une estimation de la similarité linguistique entre des zones bien localisées structurellement de l'élément-source et de l'élément-cible, de sorte qu'il n'est pas nécessaire d'analyser la totalité du contenu des éléments à lier.

Ces hypothèses qui ont guidé notre démarche peuvent être mises en perspective sur le corpus qui nous sert de base d'expérimentation dans cette étude. Il s'agit d'une partie de la documentation de maintenance d'un avion civil, composée d'environ 2 000 modules documentaires[2] encodés en XML, majoritairement procéduraux, mais qui comptent également une part non négligeable de modules descriptifs. On trouve dans ce fonds documentaire une grande diversité de liens. Nombre d'entre eux ont pour cibles des entrées dans des catalogues graphiques ou des référentiels de composants, ou d'équipements qui sont aussi encodés en XML. Cependant, les liens les plus nombreux sont établis entre les instructions et les modules procéduraux qui les décrivent sous une forme expansée. Nous parlerons de type *expansif* pour désigner ces liens. Comme dans la plupart des documentations de maintenance et d'utilisation, ces liens sont fondamentaux en ce sens qu'ils matérialisent une part essentielle de la macrostructure logique de la documentation.

Dans ce corpus, on observe que lorsqu'un lien expansif est posé, il existe une forte similarité linguistique entre l'instruction et un élément bien localisé structurellement dans le module cible, à savoir l'élément XML de type **Désignation** qui au niveau présentationnel prend la forme d'un titre principal. Le contenu textuel de cet élément décrit la macroaction à réaliser sous une forme nominalisée et entretient un rapport paraphrastique plus ou moins immédiat avec l'instruction source exprimée sous une forme impérative ou déclarative. Des exemples sont donnés en figures 1 et 3.

C'est là une caractéristique déterminante dans une perspective d'automatisation du processus de mise en hypertexte : pour décider de poser un lien de type expansif, l'analyse peut être focalisée sur la partie fortement pertinente de la cible que constitue l'élément **Désignation**, et éviter ainsi une exploration intégrale du module.

[2] L'ensemble de la documentation compte plus de 20 000 modules.

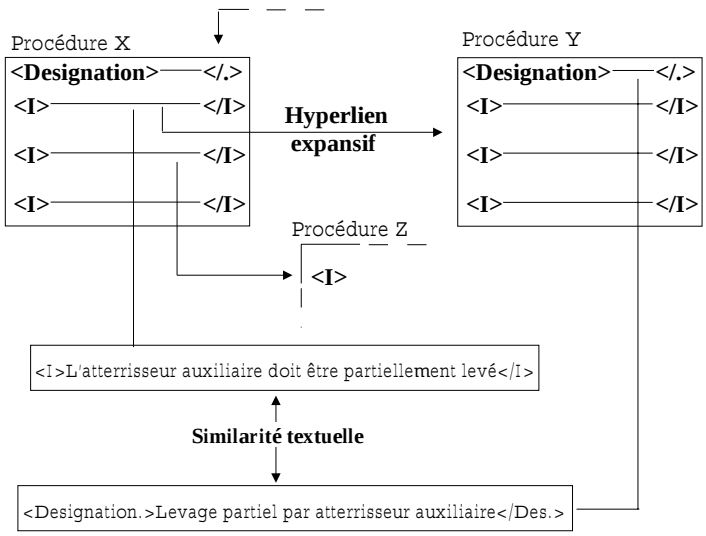


Figure 1 : Exemple de lien expansif d'une instruction vers un module documentaire

3. Un système d'aide à la pose d'hyperliens

Avant d'entrer plus en détail dans la problématique d'analyse linguistique, il nous parait utile de donner quelques indications sur la façon d'inscrire un système d'aide active à la pose d'hyperliens dans un environnement d'élaboration documentaire. Un tel système doit être interactif. Il ne s'agit en aucun cas de concevoir un système totalement automatique qui poserait ses liens sans intervention du rédacteur. Cela conduirait à exiger du système des niveaux de performance très élevés, voire irréalistes compte tenu de la complexité du problème et de l'état de maturité des techniques sollicitées. Le but est de réduire automatiquement et avec robustesse l'espace des solutions. Le choix de la solution finale reste du ressort de l'utilisateur. Cette conception interactive des services d'ingénierie linguistique a permis de mener à un niveau (quasi) opérationnel des applications impliquant une analyse du contenu, telles que la recherche/extraction d'information et l'aide à la traduction. La pose d'hyperliens peut aussi s'inscrire dans cette conception. Ainsi, pour poser un hyperlien vers un module documentaire, le système se contentera de repérer les 10 modules les plus plausibles parmi les 20 000 modules que compte la base documentaire. Ces 10 modules sont ensuite présentés au rédacteur à qui revient la décision finale.

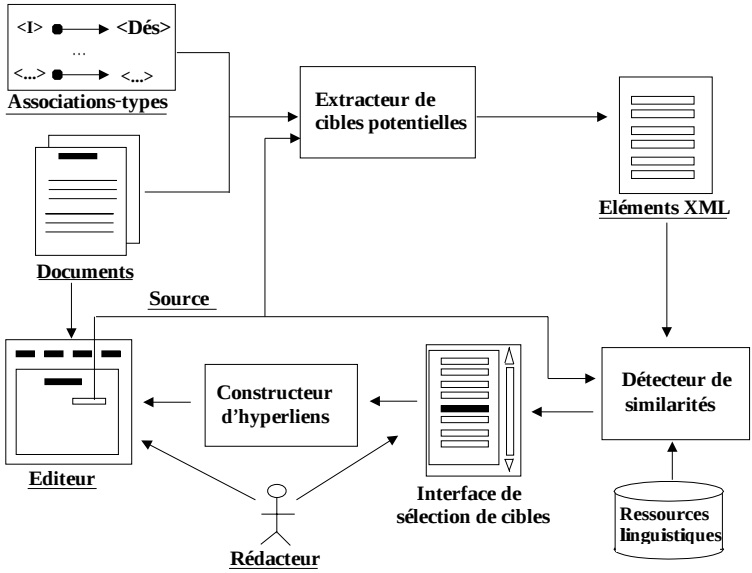


Figure 2 : Un système générique d'aide à la pose d'hyperliens fondé sur la détection de similarités lexicales

La figure 2 donne une vue globale d'un système intégrant une aide active à la pose de liens. On distingue dans ce système les composants suivants :

- **Extracteur de cibles potentielles**. A partir d'un élément de document sélectionné par le rédacteur dans un éditeur XML, ce composant collecte les éléments structurels constituant des cibles potentielles conformément à des associations-types préalablement définies.
- **Détecteur de similarités**. C'est le composant linguistique du système. Il assure au minimum une analyse morphosyntaxique et une lemmatisation de l'ensemble des éléments préfiltrés. Les mécanismes d'estimation de similarités sont ensuite mis en oeuvre pour élaborer une liste de cibles classées par ordre de pertinence. Comme dans les systèmes de recherche documentaire, un module d'indexation peut être utilisé pour accélérer la recherche de liens au moment où elle est invoquée par l'utilisateur.
- **Interface de sélection**. Une interface permet au rédacteur de procéder au choix final de la bonne cible qui doit être située dans la partie haute de la liste proposée.
- **Constructeur d'hyperliens**. La dernière phase consiste à matérialiser les hyperliens, en général par insertion dans le document d'un élément de structure localisant la cible.

4. Quelles ressources linguistiques ?

---

1. Effectuer un essai de détection incendie. [**Lem**]
- **Détection incendie. Essai**
- Détection de fuite. Essai

2. La batterie de servitudes doit être déconnectée. [**Lem+Dériv**]
- **Batterie de servitudes. Déconnexion/Connexion**
- Batterie de servitudes. Contrôle de la tension

3. Vérifier la tension de la batterie essentielle. [**Lem+Dériv+Syn]**
- **Batterie essentielle. Contrôle de la tension**
- Batterie essentielle. Dépose

4. Déposer l'axe de liaison bielle - atterrisseur. [**Lem+Dériv+Term**]
- **Jambe. Dépose**
- Atterrisseur auxiliaire. Dépose

- **Lem** : Lemmatisation, **Dériv** : Dérivation, **Syn** : Synonymie, **Term** : Terminologie

---

Figure 3 : Différentes relations lexicales attestées entre sources et cibles d'hyperliens.

Sont indiquées pour chaque exemple l'instruction-source (en italique), la bonne cible (en gras), et une cible concurrente. Nous précisons également les types des ressources lexicales nécessaires pour distinguer la bonne cible.

Une des questions centrales qui se posent dans notre problématique est l'identification des ressources linguistiques requises pour rendre compte des rapports de similarité entre les fragments à lier.

La nature très technique des textes à produire peut laisser penser que les options lexicales offertes aux rédacteurs sont considérablement restreintes, si bien que l'écart entre deux expressions d'une même idée est souvent ténu. On pourrait dès lors se suffire d'une méthode d'identification de recouvrements d'unités lexicales strictement identiques pour atteindre un niveau de performance satisfaisant. Mais une analyse même sommaire de notre corpus fait apparaître une variabilité lexicale suffisamment importante pour mettre en difficulté une méthode aussi rudimentaire. Nous verrons par la suite que l'inadéquation d'une telle méthode est confirmée par nos expérimentations.

Les exemples de la figure 3 aident à appréhender l'impact des différents types de ressources lexicales sur les mécanismes d'estimation de similarités. L'exemple 1 est un cas pour ainsi dire « idéal » où l'on observe un nombre suffisant de correspondances lexicales strictes avec la cible attendue. Il n'est pas utile de recourir à une source d'information autre que les lemmes pour isoler la bonne cible de ses concurrentes qui présentent aussi un recouvrement partiel avec la source.

En revanche, dans l'exemple 2, il faut identifier le rapport de dérivation morphologique qui lie d*éconnectée* à *déconnexion* pour faire émerger la bonne cible. Avec l'exemple 3, la discrimination de la cible passe par la combinaison d'informations dérivationnelles et synonymiques qui permettent d'associer le verbe *vérifier* au nom *contrôle*.

Enfin, l'exemple 4 est un cas plus difficile où un rapport synonymique d'ordre terminologique doit être identifié entre le terme complexe *axe de liaison bielle - atterrisseur* et le terme simple jambe.

Le traitement des cas 1, 2 et 3 ne fait appel qu'à des ressources linguistiques générales, alors que la résolution du cas 4 requiert l'acquisition de ressources terminologiques spécifiques au domaine technique considéré voire même à la documentation traitée. Dans cette étude, nous nous contentons d'évaluer le gain apporté par les ressources lexicales générales sans recourir à des ressources terminologiques. Nous reviendrons cependant dans notre discussion consacrée aux perspectives sur l'apport potentiel de ces ressources spécifiques.

Nous avons utilisé dans nos expérimentations un réseau de synonymes extraits du dictionnaire en ligne du Crisco[3] de l'université de Caen. Ce dictionnaire est le fruit d'un travail lexicographique d'ampleur qui a permis d'intégrer de manière homogène dans une unique source informatisée plusieurs dictionnaires éditoriaux. La notion de synonymie prend dans ce réseau lexical un sens très large qui couvre les notions d'hyperonymie et d'hyponymie.

En recherche d'information, l'intérêt des ressources synonymiques est loin de faire l'unanimité. Des études ont montré que le gain en rappel apporté par les synonymes s'accompagne inéluctablement d'une perte de précision prohibitive (cf. [18]). En d'autres termes, les synonymes permettent d'identifier plus de documents pertinents mais noyés dans tout autant – voire davantage – de documents hors sujet. Il convient donc d'être prudent quant à l'implication de ces ressources dans l'estimation de la similarité lexicale. Dans le cadre de l'hypertextualisation, nos expérimentations tendent à confirmer qu'une utilisation systématique des synonymes dégradent sérieusement les performances. Cependant, nous montrons qu'un effet mineur mais positif peut résulter d'un usage plus contraint, motivé par des considérations spécifiques aux genres des textes traités.

[3] http://elsap1.unicaen.fr

5. Un modèle formel de la mise en hypertexte

Une analogie peut être établie entre la problématique classique de la recherche documentaire et la mise en hypertexte telle nous l'envisageons dans cette étude. La première consiste à identifier un ensemble de documents-réponses à partir d'une requête, alors que la seconde vise à identifier un ensemble de cibles d'hyperliens à partir d'une source désignée. Dans les deux cas, il est possible de recourir à une mesure de similarité textuelle pour estimer la pertinence des fragments de texte à identifier par rapport au fragment donné en entrée (la requête ou la source de lien).

Nous définissons ici un modèle de la mise en hypertexte en nous inscrivant dans le cadre de formalisation standard de la RI [16, 2]. Pour impliquer des connaissances externes dans l'estimation de la similarité lexicale, nous introduisons en § 5.2 un mécanisme de repondération des cibles.

5.1 Rappel du modèle vectoriel

Dans ce modèle de référence, l'estimation de la similarité entre deux textes est assimilée à un calcul de distance entre deux vecteurs définis dans un espace où chaque dimension représente une unité lexicale (*terme d'indexation*).

Chaque fragment de texte manipulé – source ou cible dans notre contexte – se voit attribuer une représentation dans l'espace à N dimensions construit à partir de l'ensemble des unités lexicales considérées U = (u_1, . . ., u_N).

La représentation d'une cible *c* prend ainsi la forme d'un vecteur C = (c_1, . . ., c_N) dans lequel la c_i représente le poids de l'unité u_i dans la cible *c*.

De même, une source *s* est représentée par un vecteur S = (s_1, . . ., s_N).

Différents types de pondérations sont envisageables pour assigner des valeurs aux composantes de *S* et *C*. Nous utilisons ici une pondération du type *TD.IDF* :

$$c_i \,(ou\ s_i) = tf_i \cdot log_{10}\left(\frac{M}{df_i}\right)$$

où tf_i est le nombre d'occurrences de l'unité u_i dans la cible (ou source), *M* le nombre total de documents, et df_i le nombre de documents dans lesquels u_i apparaît.

Cette pondération permet d'intégrer, comme facteur réducteur de la pertinence, la « dispersion » de l'unité dans l'ensemble de la base documentaire, mais sans lui accorder trop d'importance.

La mesure du cosinus de l'angle entre les deux vecteurs est une estimation possible de la proximité lexicale :

$$M_{cos}(S,C) = \frac{S \cdot C}{\| S \| \cdot \| C \|} = \frac{\sum_{i=1}^{N} s_i c_i}{\sqrt{\sum_{i=1}^{N} s_i^2 \sum_{i=1}^{N} c_i^2}} \qquad (5.1)$$

Cette mesure introduit une normalisation des vecteurs qui évite de privilégier les textes longs.

En définitive, la similarité entre la source et la cible est estimée par $M_{cos}(S,C)$, et la meilleure cible est celle qui maximise cette mesure.

Les *unités lexicales* correspondent dans la pratique à des formes canoniques, simples ou complexes, obtenues par analyse des formes fléchies attestées dans les textes. Il peut s'agir aussi bien de formes lemmatisées que de formes radicales. Cependant, la prise en compte de relations lexicales d'ordre sémantique nous conduit naturellement à ne manipuler que des unités lemmatisées pouvant constituer des entrées de dictionnaires généraux ou spécialisés.

5.2 Intégration de connaissances par repondération

Les mécanismes de repondération ont pour effet d'enrichir la représentation vectorielle de chaque cible par la prise en compte de relations identifiées entre ses unités et celles de la source. Plus précisément, deux phases de repondération sont appliquées successivement. La première concerne les relations de dérivation morphologique, la seconde les relations de synonymie.

Le principe général de la méthode est relativement simple. Supposons qu'une unité de la source soit absente de la cible mais qu'elle y soit quand même représentée par l'un de ses dérivés morphologiques (ou synonymes). Le dérivé (ou synonyme) est alors remplacé dans la représentation de la cible par l'unité apparentée dans la source, avec une pondération éventuellement ajustée. Avant de porter un jugement de pertinence, il s'agit de modifier la représentation de la cible pour la rapprocher au mieux de la source, en exploitant des connaissances externes.

Nous donnons ici une description formelle de la repondération de cibles dans le cadre du modèle vectoriel.

Avant confrontation à une source donnée, une cible potentielle subit un changement de représentation défini par :

$$C' = \alpha \cdot C + \beta \cdot D + \gamma \cdot E$$

- D = (d_1, . . ., d_N) est un vecteur défini pour représenter la repondération due à la présence de dérivations morphologiques.
- E = (e_1, . . ., e_N) représente la repondération due aux synonymes.

- $\alpha, \beta, \gamma \in [0,1]$ sont des paramètres qui permettent d'ajuster l'importance donnée aux différentes informations exploitées dans la représentation des fragments à comparer.

Les composantes du nouveau vecteur sont donc définies par :

$$c'_i = \alpha \cdot c_i + \beta \cdot d_i + \gamma \cdot e_i$$

⇨ **Repondération dérivationnelle**

Le vecteur *D* permet d'ajuster la représentation de la cible de manière à assigner un poids positif à toute unité présente uniquement dans la source mais qui est en relation de dérivation avec une unité présente dans la cible. *D* a aussi pour effet d'annuler le poids de ces dérivés :

$$- \left.\begin{array}{l} d_i = s_i \\ d_j = -c_j \end{array}\right\} \text{ pour tout couple } (i,j), 1 \le i, j \le N \text{ tel que } \begin{cases} c_i = 0 \\ s_i \neq 0 \\ mderiv(u_i, u_j) \end{cases}$$

- Toutes les autres composantes de *D* prennent une valeur nulle.

mderiv est un prédicat qui est satisfait si les deux unités lexicales passées en argument sont en relation de dérivation morphologique.

⇨ **Repondération synonymique**

La construction de *E* est analogue à celle de *D*, à la différence que cette seconde repondération est également contrainte par les valeurs déjà attribuées aux composantes de *D* :

$$- \left.\begin{array}{l} e_i = s_i \\ e_j = -c_j \end{array}\right\} \text{ pour tout couple } (i,j), 1 \le i, j \le N \text{ tel que } \begin{cases} c_i = 0 \\ s_i \neq 0 \\ d_i = 0 \\ syn(u_i, u_j) \end{cases}$$

- Toutes les autres composantes de *E* prennent une valeur nulle.

syn est un prédicat qui est satisfait si les deux unités lexicales passées en argument sont en relation de synonymie. La relation peut être transcatégorielle (dans le sens où elle peut concerner des unités de catégories grammaticales différentes).

Le vecteur *E* permet d'ajuster la représentation de la cible de manière à assigner un poids positif à toute unité présente uniquement dans la source mais qui est

en relation de synonymie avec une unité présente dans la cible. *E* annule le poids des synonymes identifiés.

Du point de vue de son expression formelle, notre méthode présente quelques similitudes avec certains modèles d'expansion de requêtes, en particulier avec les modèles d'ajustement de la requête par prise en compte du retour utilisateur (*relevance feedback*) [10, 9]. Ces modèles font intervenir une repondération de la requête pour intégrer un jugement de pertinence externe. Mais notre approche s'en distingue sur des points fondamentaux. L'information externe est d'ordre lexical, et intégrée automatiquement. De plus, la repondération dans notre modèle s'applique sur les cibles et non sur la source (équivalent de la requête).

C'est là une différence essentielle. En recherche d'information, l'introduction de connaissances passe généralement par des mécanismes d'expansion de requêtes. Il s'agit d'enrichir la requête en la complétant avec des unités liées d'une façon ou d'une autre aux unités déjà présentes (transposé dans le contexte de la mise en hypertexte, cela consisterait à étendre la source). L'expansion est globale, au sens où elle doit inclure toutes les unités complémentaires susceptibles de rapprocher la requête des documents pertinents, et elle est effectuée en une seule opération préalable à l'invocation de la recherche. Une requête courte est ainsi transformée en une requête bien plus prolixe, ce qui n'est pas sans conséquences négatives sur l'estimation de la pertinence. Dans le cadre formel du modèle vectoriel, où la requête et le document ont même statut, rien n'interdit d'appliquer l'expansion aux documents plutôt qu'aux requêtes, et de la restreindre pour n'introduire que des unités en relation avec des unités communes. Cela conduit à remplacer une expansion globale par autant d'expansions locales que la base compte de documents. Sur le plan algorithmique, on comprend bien les réserves que peut susciter cette approche dans une perspective classique de recherche documentaire où il s'agit d'attaquer des bases extrêmement volumineuses. Par souci d'efficacité, on considère que les représentations des documents doivent être élaborées lors d'une phase préalable d'indexation, et ne peuvent en aucun cas être modifiées pendant les phases de recherche. Avec la mise en hypertexte, les contraintes d'efficacité sont d'un autre ordre, et notre méthode de repondération des cibles est tout à fait implémentable dans une optique opérationnelle.

6. Expérimentations

Nous avons mis en place une chaîne de traitement qui nous a permis de mener des expérimentations d'ampleur significative.

Le corpus que nous exploitons est composé de 1 927 modules pour un total de 450 000 mots. Il s'agit d'une documentation déjà hypertextualisée. Le système est donc évalué dans sa capacité à retrouver des liens déjà posés par les rédacteurs techniques.

Ces expérimentations portent exclusivement sur les liens expansifs qui associent les instructions à leurs formes expansées. La détection de similarités vise à confronter les instructions (éléments <**i**>} en figure 1) à des zones de fort potentiel de pertinence, bien localisées dans les cibles (éléments <**Désignation**> en figure 1). On évite ainsi, comme évoqué en § 2, une exploration intégrale du contenu textuel des cibles[4].

6.1 La chaîne de traitement

Le traitement d'un ensemble de sources comporte les étapes suivantes :

- **Analyse XML**. Tous les modules sont analysés pour extraire les instructions accompagnées de leurs hyperliens. Ces instructions constituent l'ensemble des éléments qui joueront le rôle de source d'hyperliens. Est aussi constitué à ce stade l'ensemble des éléments réprésentant les cibles potentielles pour les calculs de similarités.
- **Etiquetage morpho-syntaxique**. Les éléments textuels extraits sont traités par MultAna, un outil de désambiguïsation morpho-syntaxique bâti comme une extension de l'analyseur morphologique MMORPH [15]. Après cette opération, les unités manipulées sont toutes sous une forme lemmatisée.
- **Collecte des unités lexicales et calcul des effectifs**. Les textes étiquetés sont explorés pour élaborer l'espace des unités lexicales, et recenser les effectifs requis par le calcul des pondérations.
- **Construction des représentations vectorielles**. Les représentations « standards » des cibles sont construites avant la phase d'identification des liens, mais celles qui sont effectivement utilisées dans les comparaisons ne sont obtenues qu'après repondération (cf. § 5.2), lors du traitement d'une source particulière.
- **Identifications des liens**. La représentation de chaque source à traiter est élaborée, puis confrontée aux cibles repondérées. Une liste ordonnée de cibles est associée à chaque source.

6.2 Ressources exploitées

Nous avons utilisé un réseau de dérivations morphologiques qui couvre toutes les unités identifiées dans les cibles. Les dérivations **N ↔ V** sont les plus exploitées, mais les dérivations **Adv ↔ Adj** sont aussi mises à contribution.

[4] Nous avons cependant réalisé plusieurs essais en exploitant l'intégralité du contenu des cibles. Les résultats sont nettement inférieurs à ceux obtenus en focalisant la recherche sur des zones bien déterminées.

Les ressources synonymiques issues du dictionnaire du Crisco ont été utilisées sans pré-filtrage qui viserait à exclure les synonymes d'usage trop littéraire. Il faut savoir qu'un tel filtrage serait délicat à réaliser

Nous avons déjà évoqué en § 4 les difficultés que pose l'intégration de connaissances synonymiques. Nos premières expérimentations ont consisté à invoquer une recherche de synonymes sur toutes les unités candidates à la repondération synonymique selon les principes définis en 5.2. Il en résulte un effondrement drastique des performances. Cela s'explique en grande partie par les reformulations imposées aux unités terminologiques qui se prêtent peu à la variation synonymique.

Comme indiqué en § 2, il nous est cependant permis dans notre contexte spécialisé et fortement structuré de faire intervenir des considérations relatives aux genres textuels du matériau manipulé pour définir un usage plus subtil de ces ressources sémantiques. Une analyse du corpus fait en effet apparaître que la variation synonymique dans les textes de genre procédural touche avant tout les unités prédicatives. En particulier, les verbes exprimant l'action principale dans les instructions sont les premiers concernés[5]. Cela nous conduit à restreindre la relation générale *syn* définie en 5.2 à la relation syn_{PP} qui n'est satisfaite que si elle met en relation de synonymie deux unités prédicatives[6].

Nous précisons qu'il s'agit d'une contrainte globale dont la mise en oeuvre est aisée du fait qu'elle n'exige aucune personnalisation manuelle des ressources.

6.3 Evaluation

La chaîne de traitement a été appliquée sur 7 721 sources de liens. Nous avons testé un grand nombre de configurations déterminées par les valeurs des paramètres α, β et γ. La mise à zéro d'un paramètre permet d'éviter l'implication de la ressource correspondante.

La distinction Précision/Rappel n'a ici pas grande signification, en ce sens que, dans la grande majorité des cas, chaque source ne compte qu'une seule cible pertinente. Dans le protocole d'évaluation que nous adoptons, nous considérons que la bonne cible a été localisée si elle figure dans les k premières propositions. Le tableau 1 décrit les performances obtenues avec quatre configurations représentatives.

---

[5] *Brancher ↔ connecter/connexion, vider} ↔ vidanger/vidange, réparer ↔ réfection, vérifier ↔ inspecter/inspection, ...*

[6] Est considéré comme unité prédicative tout verbe ou tout nom dérivé morphologique d'un verbe.

α / β / γ	**(1)** 1 / 0 / 0	**(2)** 1 / 1 / 0	**(3)** 1 / 1 / 1	**(4)** 1 / 1 / 0,5
k = 1	24,40 %	56,10 %	49,10 %	57,11 %
k = 5	54,58 %	73,33 %	70,44 %	73,51 %

Tableau : Résultats des expérimentations sur la pose de 7 721 hyperliens.

On notera tout d'abord les faibles performances atteintes par la configuration (1) réduite à l'exploitation des lemmes (α à 1, β et γ à 0). Cela s'explique par la forte proportion d'unités en relations de dérivation morphologique qui ne peuvent pas être appariées. La configuration (2) exploitant le réseau de dérivations α et β à 1, γ à 0) apporte dès lors un gain substantiel.

En regard des acquis en recherche documentaire, un effet positif des dérivations morphologiques était prévisible. Les mécanismes classiques de racinisation (*stemming*) constituent aussi une forme bien établie d'intégration indirecte (et approximative) d'informations dérivationnelles. En revanche, l'importance de l'écart de performance avec la configuration sans dérivations est plus surprenant. Elle s'explique certainement par la prééminence des verbes et de leurs formes nominalisées dans ces documentations techniques à dominante procédurale. Dans un fonds documentaire général, la capacité à mettre en correspondance les verbes et leurs nominalisations n'a pas aussi souvent un impact déterminant dans l'estimation de la pertinence.

Les synonymes, utilisés ici de manière contrainte (cf. 6.2) dans les configurations (3) et (4), ont certes un effet discret, mais ils n'engendrent plus un effondrement des performances. C'est, il nous semble, un résultat important compte tenu de l'envergure de ces expérimentations. La configuration (4) est la plus performante des quatre, avec un apport minime par rapport à la configuration (2). Les synonymes ne sont décisifs que dans peu de cas bien identifiés (l'exemple 3 du tableau est un cas typique), mais une analyse détaillée des résultats fait apparaître qu'ils contribuent à renforcer le score d'un nombre important de bonnes cibles. De plus, ils ne sont que très rarement la cause directe d'échecs (on compte très peu de cas où il y a réussite avec la configuration (2) mais pas avec la (4)).

La perte de performance observée avec la configuration (3) nous révèle que les synonymes gagnent à être considérés comme une source d'appoint (γ = 0,5) : une identité stricte ou un rapport de dérivation doit prendre plus d'importance qu'une relation synonymique.

7. Conclusion et perspectives

L'apport limité mais positif des synonymes est en accord avec des travaux récents sur l'exploitation de ressources lexicales d'ordre sémantique en recherche d'information [7, 13]. Pour affiner cette première appréciation, il nous paraît nécessaire d'expérimenter ce modèle sur d'autres corpus de même genre. Il n'est pas exclu qu'une telle source externe puisse apporter un gain plus significatif sur une documentation technique de variabilité lexicale plus élevée.

Bien que la méthode définie ici soit restreinte aux relations de dérivation morphologique et de synonymie, il est possible de l'étendre à d'autres types de relations.

En particulier, les cooccurrences lexicales ont déjà été exploitées dans le cadre de la mise en hypertexte de documentations techniques [14]. La méthode proposée dans ces travaux repose aussi sur un calcul de similarité, et introduit une pondération sensible aux profils de cooccurrence des unités lexicales. L'intégration dans notre modèle de cooccurrences lexicales selon cette méthode est immédiate du fait qu'elle n'affecte que le choix de la fonction de pondération.

Il nous semble toutefois qu'une amélioration substantielle de la méthode proposée ici ne peut faire l'économie de l'implication de ressources d'ordre terminologique voire « ontologique » (cf. [4]). Dans des corpus aussi techniques, de nombreux phénomènes de variation lexicale échappent à une analyse fondée exclusivement sur des ressources générales, et une part importante des échecs est imputable à des relations terminologiques non identifiées[7]. Mais la constitution de telles ressources exige un effort d'acquisition autrement plus important que l'adaptation de ressources générales, même si des techniques d'identification de variations terminologiques [12, 8, 6] peuvent aider à réduire de manière significative la part de travail manuel.

8. Références

[1] AUSSENAC-GILLES, N., and CONDAMINES, A., Rapport de l'action spécifique ASSTICCOT, Rapport interne IRIT/2003-23-R, 2003.

[2] BAEZA-YATES, R., and RIBERTO-NETO, B., *Modern Information Retrieval*, Addison-Wesley ed., 1999.

[3] BALPE, J.-P., LELU, A., PAPY, F., and SALEH, I., *Techniques avancées pour l'hypertexte*, Hermes, 1996.

[7] Quelques exemples où les relations d'ordre terminologique sont déterminantes pour rapprocher une source de la bonne cible :
- Abréviation : Poser la **CI**. → **Centrale à inertie**. Pose.
- Synonymie : **Saisiner** l'avion. → **Amarrage** normal. Procédure.
- Méronymie : Régler les **butées internes**. → Cinématique **verrière**. Réglage.
- Hyperonymie : **Rentrer** l'échelle intégrée. → Echelle intégrée. **Manoeuvre**.

[4] BAZIZ, M., AUSSENAC-GILLES, N., and BOURGHANEM, M., Désambiguïsation et expansion de requêtes dans un SRI, *Ingénierie des systèmes d'information*, 8(4), 2003.

[5] BIBER, D., Using register-diversified corpora for general language studies, *Computational Linguistics*, 19(2), 1993.

[6] DAILLE, B., ROYAUTE, J., and POLANCO, X., Evaluation d'une plate-forme d'indexation de termes complexes, *T.A.L.*, 41(2), 2000.

[7] DE LOUPY, C., and EL-BÈZE, M., Managing synonymy and polysemy in a document retrieval system using wordnet, In *LREC '02*, 2002.

[8] HAMON, T., and NAZARENKO, A., Detection of synonymy links between terms: Experiments and results, In *Recent Advances in Computational Terminology*, John Benjamins, 2001.

[9] HARMAN, D., Relevance feedback revisited, In *Proc. of 15th ACM SIGIR Conference on Research and Development in Information Retrieval*, 1992.

[10] HUST, A., KLINK, S., JUNKER, M., and DENGEL, A., Query reformulation in collaborative information retrieval, In *Proc. of IKS 2002*, 2002.

[11] ILLOUZ, G., HABERT, B., FLEURY, S., FOLK, H., HEIDEN, S., and LAFON, P., Maîtriser les déluges de données hétérogènes, In *Atelier Corpus et TAL, Pour une réflexion méthodologique, TALN 99*, 1999.

[12] JACQUEMIN, C., *Spotting and discovering terms through NLP*, MIT Press, Cambridge, 2001.

[13] J. GONZALO, PENAS, A., and VERDEGO, F., Indexing with wordnet synsets can improve text retrieval, In *Proceedings of Workshop on Usage of WordNet for Natural Language Processing*, 1998.

[14] NAKAGAWA, H., MORI, T., OMORI, N., and OKAMURA, J., Hypertext authoring for linking relevant segments of related instruction manuals, In *Proceedings of COLING 98*, 1998.

[15] PETITPIERRE, D., and RUSSELL, G., MMORPH – The Multext Morphology Program, Tech. rep., Multext Deliverable 2.3.1, 1995.

[16] SALTON, G., and MCGILL, M.-J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New-York, 1983.

[17] VIGNAUX, G., L'hypothèse du livre électronique, *Les cahiers de médiologie*, (10), 2000.

[18] VOORHEES, E.-M., Using wordnet to disambiguate word senses for text retrieval, In *Proceedings of 16th ACM SIGIR Conference on Research and Development in Information Retrieval*, 1993.

Validation par prototypage d'un modèle de segmentation des documents techniques composites

Agnieszka Smolczewska, Geneviève Lallich-Boidin

URSIDOC / DOCSI
Université Claude Bernard Lyon 1 - France

prenom.nom@univ-lyon1.fr

Résumé :

Dans le présent article nous nous intéressons à la problématique de la mise en place d'un dispositif informatique de segmentation de documents techniques composites sous format électronique. Un modèle théorique pour la segmentation de ce type de documents a été défini récemment sans être testé informatiquement. Dans le modèle, l'identification et la délimitation des segments sémantiquement cohérents et autonomes reposent sur un ensemble de différentes connaissances dispositionnelles, linguistiques et typographiques, véhiculées par le document technique. Notre apport consiste à implémenter ce modèle et à le valider sur un corpus de documents techniques électroniques. A travers une analyse des résultats obtenus, nous évaluons les hypothèses à la base de sa définition, nous mettons en évidence les limites de cette démarche et nous dégageons des pistes de recherche pouvant mener à son amélioration.

Mots-clés : document technique, segmentation automatique, analyse linguistique de surface.

Abstract:

This paper describes the implementation of a segmentation system for digital composite technical documents. A theoretical model for segmentation of this kind of documents has been recently defined but not tested. This model identifies and delimits semantically coherent and autonomous segments by exploiting positional, linguistic and typographical information carried by technical documents. Our contribution consists in implementing this model

and validating it on corpus of digital technical documents. The result analysis let us assess model assumptions and underline its limits. Finally, we suggest some research directions to ameliorate it.

Keywords : technical document, automatic segmentation, linguistic surface form analysis.

1. Introduction

Avec l'avènement de la documentation électronique et le développement de ses outils de production, de stockage et de consultation, se pose inévitablement le problème de l'accès pertinent à ce format de documents. L'hétérogénéité des ressources électroniques disponibles, nombreuses et en évolution incessante, impose un ajustement des modes d'accès à ces ressources en fonction des besoins d'information spécifiques de leurs utilisateurs. Cette adaptation s'avère d'autant plus indispensable dans le domaine de la documentation technique. Etant donné qu'un technicien mène une recherche d'information dans un cadre professionnel et pour un besoin de type opérationnel, la pertinence des éléments donnés en réponse se doit d'être en totale adéquation avec ses besoins et attentes.

Bien que la manière de lire la documentation technique s'apparente à la manière de lire sur l'écran, la majorité des documents techniques électroniques actuels obéissent à une mise en page de type document papier, et par conséquent, ne tirent pas profit des possibilités d'exploitation proposées par le format informatique. Ce constat semble confirmer l'hypothèse de l'existence d'un décalage entre la logique de la production de ce type de document, qui est exclusivement linéaire, et la logique de son usage, essentiellement consultatif et ponctuel.

Faciliter l'accès à l'information technique numérique est au centre de notre problématique de recherche, puisque nous situons notre travail dans une optique de conception de systèmes de mise à disposition de ce type d'information.

Nous cherchons à définir un modèle de restructuration et d'enrichissement de l'information technique qui constituera la base de la construction, d'une représentation du contenu du document technique à partir de son texte intégral. Cette représentation, en intégrant des contraintes et besoins spécifiques liés au contexte de l'utilisation de la documentation technique, devrait faciliter l'exploitation de cette documentation par un technicien expert. Sa construction est le résultat de plusieurs étapes intermédiaires telles que :

- La segmentation du document en unités d'information autonomes ;
- La caractérisation du contenu de chaque unité ;
- Le filtrage des unités pertinentes par rapport aux unités non pertinentes pour l'utilisateur ;
- La construction de liens entre les unités exprimant des parcours de lecture possibles.

A chacune de ces étapes, la représentation est tenue de répondre aux besoins d'information de son utilisateur en tenant compte des caractéristiques particulières de la documentation technique.

Dans cet article nous nous intéressons à la première étape de ces processus, et, plus particulièrement, à la mise en place d'un prototype de segmentation de documents techniques composites. Un modèle théorique de segmentation de ce type de documents a été défini par T. Ouerfelli dans le cadre de son travail de thèse [Oue01]. Notre apport consiste à implémenter ce modèle et à le valider informatiquement sur un corpus. A travers une analyse des résultats obtenus, nous mettons en évidence les limites de cette démarche et nous dégageons des pistes de recherche qui pourraient mener à son amélioration [Mar00].

2. Les spécificités d'usage du document technique

Dans le cadre du présent travail, le terme « document technique » désigne tout document qui véhicule des savoirs et des savoir-faire propres à un champ technique particulier. Nous adhérons à la définition présentée dans [Heu02], qui décrit ce type de documents comme « *documents utilitaires spécialisés dont la fonction principale est de communiquer, à des spécialistes ou à des non spécialistes, des informations techniques relatives à une procédure, ou fonctionnement et/ou l'utilisation d'un processus ou d'un dispositif, en vue de permettre une utilisation immédiate ou différée de ces informations pour accomplir une tâche professionnelle ou de la vie quotidienne* ». Ces documents sont qualifiés de procéduraux lorsqu'ils communiquent des procédures (i.e. ensemble d'opérations et/ou d'actions à exécuter dans le but d'atteindre un objectif donné) et/ou lorsque les informations qu'ils véhiculent prennent la forme d'instructions ou de consignes destinées à être exécutées (consigne de sécurité, mode d'emploi, notice d'utilisation) et de non procéduraux lorsqu'ils visent avant tout à informer (rapports, lettres techniques, descriptions de processus ou de dispositif, etc.) [HG02].

Les auteurs qui ont observé l'utilisation d'une documentation technique s'accordent à dire qu'elle se manifeste selon deux modes. Le premier, représenté par la lecture exhaustive et linéaire du document, ayant pour but d'acquérir un aperçu général sur le dispositif concerné [Wri97], ne se fait que rarement, et souvent lors de la première utilisation du document. Une fois la connaissance générale sur le dispositif acquise, l'utilisateur se sert du document technique que ponctuellement, pour répondre à un besoin d'information précis à un moment donné. Ainsi, le second mode d'utilisation, le plus fréquent, concerne la lecture sélective ou la consultation. Ce type d'usage doit permettre à l'utilisateur d'accéder directement à l'information exacte lui permettant de résoudre un problème particulier pour lequel il n'a pas a priori d'informations suffisantes. Il s'apparente à une recherche d'information de type « résolution de problèmes » [Gir89] et il ne s'effectue pas de manière linéaire car l'objectif de l'utilisateur est de trouver une information précise le plus rapidement possible [Heu94].

L'objet sur lequel se fait la recherche dans le domaine technique est le document dans son intégralité car l'information recherchée peut être véhiculée par plusieurs médias comme le texte, les figures, les tableaux etc. En revanche, une demande d'information dans un document technique n'est pas liée à la volonté de tout savoir sur le dispositif ou le système concerné par le document, mais plutôt de répondre à un besoin précis d'une information sur un des aspects de l'objet dont traite le manuel technique. C. Paganelli [Pag97] définit ce type de demande comme une *demande partielle*. Sa réflexion s'appuie sur les travaux de Eymard [Eym92] selon lesquels une recherche d'information qui émane d'un spécialiste passe par une question de nature partielle. Par opposition aux questions totales qui appellent une réponse en oui/non, les demandes d'information partielles nécessitent comme réponse une valeur à sélectionner dans une liste de valeurs possibles et utilisent le plus souvent les interrogatifs de type qui, quand, comment. Les travaux de Paganelli montrent que les demandes d'information de l'utilisateur de la documentation technique ne sont pas des demandes sur un thème général, par exemple « le répertoire téléphonique », mais des demandes précises qui ne portent que sur une partie de ce thème, par exemple sur un procès ou une de ses caractéristiques, du type « Comment créer un répertoire ? ». De cette manière, l'utilisateur vise essentiellement une recherche sélective et partielle sur un élément précis du document.

3. Travaux apparentés

L'objectif de la première étape de la construction d'un système de mise à disposition des documents techniques est de partitionner le document en composantes élémentaires de ce système, des unités d'information, qui serviront de base à l'indexation et qui seront utilisées comme morceaux réponses à la requête de l'utilisateur. Ce processus de segmentation se voulant automatique se pose alors la question de la méthode permettant d'identifier les points de rupture dans le continuum sémantique caractéristique de tout texte cohérent.

Comme nous le rappelle [Ada94] toute suite de phrases ne constitue pas un texte, même si toute phrase de cette suite analysée séparément, est correcte grammaticalement. Dans la perspective de la linguistique textuelle, le critère qui assure qu'une suite d'énoncés forme une séquence ou un texte cohérent, indépendamment de la diversité et hétérogénéité des ses composants, est désigné par le terme de *cohésion textuelle*[1]. Appliquer cette notion à la tâche de segmentation

[1] Les notions méthodologiques que la linguistique textuelle propose pour en rendre compte sont quelque peu instables et peuvent varier selon les auteurs. Ainsi, ce que Lundquist [Lun80] appelle la *cohérence* et Adam, la *connexité*, semble correspondre à ce que d'autres auteurs comme Halliday et Hasan [HH76] désignent par le terme de *cohésion*, Bronckart appelle *mécanismes de textualisation* [Bro97], et Ghiglione et Landré appellent *liens référentiels* [GLBM98]. Sans vouloir trancher ce débat terminologique, il nous semble que

revient à identifier des marques de la cohésion et/ou de la décohésion[2] existant entre les différents « items » qui participent à la composition du discours et qui correspondent, respectivement, aux points de continuité et/ou de rupture sémantiques. La cohésion textuelle peut s'exprimer sous la forme d'un ensemble d'expressions « lexico-grammaticales » telles que la coréférence, les connecteurs ou la cohésion lexicale. L'organisation logique du document ainsi que certains marqueurs linguistiques marquent la décohésion textuelle. Suivant que l'on se focalise sur l'une ou l'autre catégorie d'indices de cohésion textuelle, des approches différentes se dégagent.

Dans l'approche structurelle, le processus de segmentation repose sur des moyens graphiques d'organisation physique et logique du document. Le choix de cette méthode pose inévitablement la question de la granularité de ce découpage, où « *l'unité d'information ne peut pas être inférieure à la phrase et peut être constituée par une des unités textuelles conventionnelles (paragraphes, chapitres, etc.*) » [Bal90]. C'est la méthode de segmentation appliquée dans les systèmes de recherche d'information technique ELF DTG et Info Banque [Pag97], ainsi que dans un système d'hypertextualisation des documents techniques [Pap95]. Dans ces systèmes, les segments textuels issus du découpage correspondent à des unités logiques minimales synthétisées par un titre apparaissant dans le sommaire. Aujourd'hui, nous savons que ce choix de granularité ne satisfait pas les attentes des utilisateurs de la documentation technique. L'expérimentation de Paganelli [Pag97], visant à connaître les représentations de l'unité de réponse pertinente pour des experts en situation de recherche d'information technique, a montré que les unités logiques dont le titre apparaît dans le sommaire, sont de taille trop grande pour être considérées comme pertinentes. L'utilisateur n'ayant pas la possibilité d'effectuer une recherche plus précise à l'intérieur de ce type d'unités, ne peut accéder à l'information qui l'intéresse que par une consultation séquentielle du texte. Le choix de la structure logique comme moyen de son partitionnement est également à la base du découpage des documents scientifiques effectué dans le cadre du projet Profil-Doc où la segmentation consiste à identifier les différentes unités documentaires caractérisées par leur fonction parmi les 14 catégories suivantes : *résumé, introduction, description du contexte, description du thème, environnement, développement, expérimentation, résultats, discussion, conclusion, bibliographie, table de matière, annexes* [Mic99].

La segmentation du texte dans le cadre de l'approche linguistique correspond à la détection des indices linguistiques de la cohésion et de la décohésion textuelle. Ainsi, Adam [Ada92] observe qu'il est envisageable d'identifier et de décrire des « séquences textuelles » discoursives du type: narratif, descriptif, argumentatif, explicatif et dialogal, à partir des marques linguistiques qui les caractérisent. Chez

---

le terme *cohésion* présente l'avantage d'être moins large que celui de *cohérence*, et plus communément admis que la *connexité*.

[2] Par opposition au terme de *cohésion* nous employons le terme de *décohésion* pour désigner l'absence de relation interphrastique qui nous indique une rupture thématique dans le continuum sémantique du texte.

[Jac02] et [FGMP01], le processus du découpage du texte en unités thématiquement homogènes s'effectue selon la nature sémantique des cadres discursifs qui apparaissent à la surface du texte. Les cadres discursifs, des structures thématiques fortement cohésives, définis par [Cha95] comme *les circonstances dans lesquelles il faut envisager un certain état ou une série d'événements*, sont introduits par un ensemble des marques linguistiques de structuration argumentative. Les introducteurs des cadres, qui ont pour rôle d'assurer l'unicité et la continuité textuelle, sont exploités, afin de segmenter le texte en cadres de type organisationnel [Jac02] et thématique [FGMP01]. L'exploitation des indices de cohésion est également à la base de travaux de [Mar00] qui fait appel à un ensemble des connecteurs afin de délimiter des unités élémentaires d'un texte, et de reconstruire ses structures rhétoriques.

L'approche statistique, en partant de l'hypothèse que dans un texte, un changement de thème entraîne un changement de vocabulaire, cherche à détecter les traces de la cohésion lexicale sur la base de la distribution des unités lexicales. La détection des points de coupure se fait en mesurant la « distance » (différence de distribution) entre deux séquences de mots à cheval sur une frontière potentielle. Ainsi, dans TextTiling [Hea97] deux séquences de texte adjacentes sont plus ou moins « proches » suivant le nombre de mots identiques qu'elles contiennent. A chaque « bloc » (séquence de mots de longueur fixe) est associé un vecteur de descripteurs contenant le nombre d'occurrences de chaque mot dans la séquence. Le regroupement ou la séparation entre les deux séquences s'effectue en fonction du produit scalaire de leurs vecteurs. La valeur du produit scalaire de vecteurs associés aux séquences peut varier entre 0 (deux séquences totalement différentes) et 1 (séquences contiennent exactement les mêmes mots). Le résultat de l'analyse est présenté sous la forme d'une courbe où les creux représentent les endroits de changement de thème. De nombreuses méthodes sont assimilables au TextTiling, avec des différences concernant le pré-traitement et la métrique. Ainsi, chez [FGM98] et [HG02] la pondération pour les vecteurs est de type *tf*idf* [SSCM96]. En ce qui concerne le pré-traitement des descripteurs, Hearst envisage un filtrage des mots grammaticaux par rapport aux mots informatifs, [KKMK98] lemmatise les termes, et [BN00] fait appel à une analyse lexicale et morphologique. Afin de calculer la mesure de similarité entre des descripteurs liés sémantiquement sans être identiques (synonymes, hyperonymes, termes proches sémantiquement etc.) certains systèmes font appel à une source de connaissance externe au document, tel qu'un réseau sémantique construit sur un thésaurus [Koz93] ou un réseau lexical bâti sur un corpus [FGM98]. Chez d'autres, la mesure de cohésion n'est pas calculée en termes d'unités lexicales mais en termes de concepts. Ainsi, le module LCP (*Local Component Analysis*) de [PC97] fournit pour chaque séquence, une liste de mots et phrases qui représente les concepts présents dans la séquence originale, et le calcul de similarité se fait sur ces listes de concepts. Dans les méthodes plus récentes, la décision de segmenter dépend également de la position de la frontière probable par rapport au paragraphe [KKMK98], par rapport à la structure logique du document (titres et paragraphes) [BN00] et de la présence de certains connecteurs [BN00].

Les méthodes cherchant à analyser la cohésion lexicale ne sont pas toujours très précises dans la reconnaissance de changements thématiques et peuvent ainsi ignorer les limites de certains segments. En revanche, les méthodes qui basent la segmentation sur les connaissances structurelles permettent de fragmenter le document d'une manière précise et non ambiguë, bien que les segments issus de ce type de découpage puissent ne pas correspondre aux besoins d'information spécifiques des utilisateurs. Quant aux méthodes exploitant les formes linguistiques de surface, elles s'avèrent précises et fiables, à condition que les marques linguistiques qu'elles cherchent à détecter soient présentes de manière régulière dans le texte. En effet, on constate que les résultats les plus satisfaisants sont obtenus par les méthodes combinant plusieurs approches. Le travail que nous présentons dans ce texte s'inscrit dans cette tendance.

4. Les fondements théoriques

Le modèle de fragmentation des documents techniques que nous avons implémenté, a été développé au cours du travail de thèse de T. [Oue01]. Ses hypothèses de travail postulent que l'unité sémantique d'un document technique est le résultat de « l'union » d'un ensemble de blocs d'information distincts et relativement autonomes sémantiquement. La forte structuration physique et logique caractéristique du document technique, permettant à l'utilisateur d'accéder rapidement à l'information pertinente, constitue en partie le reflet de cette union. Il est ainsi possible d'« éclater » un document en ses unités d'information, les Unités Documentaires, en détectant des points de rupture entre elles. A la différence du continuum thématique, les points de rupture entre les unités documentaires ne sont pas marqués à la surface du texte. En exploitant de manière pertinente des indicateurs de continuité thématique de nature dispositionnelle et linguistique, on peut remonter aux unités documentaires en les reconstruisant à partir de leurs composants minimaux : des paragraphes typographiques (pour le texte), et des figures, tableaux etc. (pour la représentation graphique de l'information). L'identification des limites entre les différentes unités documentaires repose ainsi sur un ensemble de différentes connaissances dispositionnelles, linguistiques et typographiques, véhiculées par le document. Cette démarche nous permet de ne pas faire appel aux méthodes de type TextTiling, qui en indexant pour segmenter, ne correspondent pas à notre objectif principal de segmenter dans le but d'indexer.

Fidèle à ces hypothèses, le modèle propose le découpage du document en deux étapes. Dans un premier temps, on segmente le document en composants minimaux typographiques tels que des paragraphes, des figures et des tableaux. La notion de paragraphe, qui est défini comme « *une surface textuelle minimale permettant l'émergence d'un propos, d'un thème et un bloc de texte délimité par deux alinéas* », permet d'englober différents types d'objets caractéristiques de l'organisation logique et s'applique aussi bien à un titre, qu'à ce qu'on appelle habituellement une énumération ou une liste d'éléments.

Dans un deuxième temps, on regroupe ces éléments de base en blocs de texte appelés *Unités Documentaires*[3]. Une UD est définie par T. Ouerfelli comme « *une suite de formes (textuelles ou non textuelles) dont l'ensemble traite d'un point qui leur est commun* » [Oue01]. C'est un objet abstrait, artificiel et composé de différents constituants de la structure logique. Notons qu'il se définit avant tout en fonction de ses propriétés telles que:

- L'autonomie sémantique par rapport au reste du document et aux autres Unités Documentaires ;
- La cohérence interne du point de vue syntaxique et sémantique ;
- La possibilité d'être facilement repérable et isolable dans le corps du texte grâce à un ensemble d'indicateurs formels de surface.

Le regroupement des composants logiques en Unités Documentaires se réalise à travers l'analyse d'un ensemble des connaissances dispositionnelles, linguistiques et typographiques véhiculées par le document. Ainsi, on procède d'abord à l'analyse de la position de chaque composant par rapport aux composants qui l'entourent directement. A titre d'exemple, une figure s'inscrira toujours dans la continuité de l'élément qui la précède, à la différence d'un titre non indexé par le sommaire, qui sera toujours automatiquement lié à l'élément le suivant et détaché de l'élément le précédant. En revanche, un titre indexé par le sommaire ne sera retenu dans la composition d'une UD qu'à condition d'être suivi d'une liste d'éléments ou d'un paragraphe textuel débutant par un marqueur de continuité.

Les marqueurs de continuité exploités dans le modèle représentent un ensemble d'indices repérables sur la surface du texte qui appartiennent soit au système de la langue, soit aux modalités particulières de transcription graphique. Comme indicateurs de continuité linguistiques, la méthode intègre certains marqueurs d'intégration linéaire (*alors, ensuite, aussi, de plus, d'autre part*, etc.), certains connecteurs (*pour cela, pour ce faire, par ailleurs, également, en particulier*, etc.), des pronoms démonstratifs (*ce, cette, ces, ceci*) et personnels (*il, elle*). Toutes ces unités, pour remplir ce rôle, doivent être situées au début du paragraphe textuel. Des déictiques spatiaux de renvoi anaphorique (*ci-dessus, plus avant*, etc.) et cataphorique (*au dessous, ci-après*, etc.) font également partie de cet ensemble. Concernant les marques de continuité typographique, seul le signe deux points (:) fermant le paragraphe textuel est retenu comme représentatif du continuum sémantique.

Ajoutons que selon ce modèle seul le paragraphe textuel est suffisamment indépendant pour former à lui seul une UD. Les autres composants entrent dans la composition d'une UD en s'associant aux unités adjacentes en fonction de leur positionnement dans la structure logique du document ou de la présence des indices

[3] UD dans la suite de ce texte.

formels de continuité. A défaut de telles informations, la méthode considère qu'il existe une rupture dans le continuum thématique et recommande la segmentation.

5. L'implémentation

Le modèle théorique présenté formalise l'ensemble des ces règles sous la forme d'un algorithme. Nous l'avons adapté au processus informatique et implémenté afin de valider son automatisation et de relever ses limitations dans la tâche d'identification et de délimitation des UD.

Le prototype développé a été testé sur un corpus, constitué de manuels d'utilisation de différents dispositifs techniques (une console de sons, une pelliculeuse) et de logiciels (de traitement de texte et de gestion de sécurité d'accès aux ordinateurs en réseau). Ce type de documents, accompagnant la vente du produit final, appartiennent à la documentation externe de l'entreprise et par conséquent, ne font pas l'objet de clauses de confidentialité. Avec le développement de la Toile, de plus en plus souvent les entreprises rendent disponible cette documentation au moyen de téléchargements de fichiers. En ce qui concerne le format de notre corpus, la nécessité d'identifier les éléments spécifiques de la structure logique (des titres, des listes, des paragraphes, etc.) nous a orienté vers les formats de documents structurés tels que SGML, XML ou HTML. Les deux premiers formats étant essentiellement utilisés en interne, nous avons finalement porté notre choix sur le format HTML.

Par le terme « documents structurés » nous désignons des documents numériques dans lesquels des conventions additionnelles (*marques ou balises*), permettent de représenter la structure hiérarchique du document et de rendre explicite la nature de l'information contenue. La présence de ce balisage dit *générique*[4], exprimé sous la forme de paires de balises ouvrantes (<titre>) et fermantes (</titre>), facilite la tâche d'identification de la nature ou de la fonction intrinsèque des différents éléments logiques composant le document.

Afin d'accéder aux portions du document encadrées par des balises, et remédier à certaines entorses d'HTML au principe des documents structurés[5], nous avons choisi la solution de baser notre prototype sur un parseur d'HTML[6] implémenté en Perl. Le parseur, en parcourant le document, reconstruit la structure d'arbre représenté par le code HTML. Dans cet arbre chaque noeud correspond à un élément de la structure logique.

---

[4] Par opposition aux balisages orientés à la mise en page que l'on retrouve dans l'approche traditionnelle du traitement de documents.

[5] Par exemple, la possibilité d'omettre la balise fermante `</p>` pour les paragraphes ou l'existence de balises qui ne servent qu'à la visualisation du document (comme `<hr>`, qui dessine une ligne horizontale) et qui n'ont pas, par conséquent, de contenu.

[6] Il s'agit de module HTML : TreeBuilder disponible à partir de la version 5.6.1. de Perl.

L'utilisation de ce parseur, conçu pour parcourir une structure de type arbre, a rendu nécessaire une adaptation de l'algorithme original, défini à partir d'une logique de parcours linéaire. L'algorithme actuel, dans un premier temps, recherche pour chaque élément de l'arbre balisé par <titre>, <paragraphe>, <liste>, <figure>, <tableau>, son premier « frère droit » (c'est-à-dire le premier élément à droite du nœud courant) entouré par une des balises listées. Une fois cette paire identifiée, il vérifie la présence des marques de continuité entre ces deux éléments. A défaut des tels indices, une marque de discontinuité est insérée. Le processus est appliqué à tous les éléments marqués par les balises recherchées. Les modifications apportées à l'algorithme original permettent de réduire le nombre d'éléments à traiter simultanément (deux au lieu de trois) et de commencer l'analyse directement avec le premier élément rencontré.

Il est important de souligner que grâce au choix de travailler sur des documents structurés (même faiblement) et à l'utilisation d'un outil adapté, indépendant du style de codage HTML et de l'absence de certaines balises, la tâche d'implémentation a été considérablement simplifiée et notre système de segmentation a finalement gagné en robustesse, fiabilité et généralité.

6. L'analyse des résultats

La mise à l'épreuve de notre prototype de segmentation sur un corpus de documents techniques a permis d'identifier un certain nombre de faiblesses à différents niveaux de la définition du modèle.

6.1 La composition des UD selon le critère dispositionnel

La limite principale du modèle résulte du choix de regrouper les composants atomiques d'une UD d'une manière exclusivement linéaire. Ainsi, la décision d'intégrer un composant dans une UD selon un critère dispositionnel, dépend systématiquement du composant qui le précède ou le suit directement. Ce type de regroupement séquentiel s'avère valide et recevable dans la plupart des cas concernant l'assemblage des composants textuels, comme titres ou paragraphes. En revanche, il ne permet pas de prendre en compte le cas où, pour constituer une UD, il est nécessaire de regrouper deux composants non adjacents. Une telle situation, est fréquente pour les unités d'information graphiques (figures, tableaux). Elle peut également s'appliquer à certaines unités textuelles, comme le montre l'exemple qui suit :

---

1. Remplacement du moteur C3i

Démontage

- Extraire la goupille qui relie le vérin au couvercle.
- Retirer la face avant du centrifugeur en ôtant les cinq vis et les connexions électriques suivantes : interrupteur principal, limandes, alimentation de la carte afficheur et câble de mise à la terre. Soulever le joint de cuve et dévisser les trois vis de serrage de la chambre de centrifugation.
- Déconnecter les connecteurs du moteur, du tachymètre, du capteur balourd, de la carte uP + puissance, ainsi que la cosse faston de mise à la terre moteur, et la connexion du thermoswitch.
- Retirer le joint de fond de cuve.
- Soulever et retirer la cuve.
- (…)

Montage

- Suivre la procédure de démontage en sens inverse.

En respectant les recommandations du modèle, cet extrait est fragmenté en deux UD. La première commence au titre « Remplacement du moteur C3i » pour se terminer au dernier élément de la liste précédant le titre « Montage », alors que la seconde intègre le titre « Montage » ainsi que l'item de la liste qui le suit. Le titre « Remplacement du moteur C3i », à cause du traitement strictement linéaire des composants, n'est intégré qu'à la première UD. Et pourtant, sa présence est indispensable à l'interprétation de la deuxième unité, car il nous permet d'identifier le dispositif technique auquel la tâche de montage s'applique.

L'exemple ci-dessus est d'autant plus intéressant qu'il illustre le principal défaut de l'approche linguistique basée essentiellement sur la reconnaissance des indices de surface. Selon la méthode testée, les deux segments « Démontage » et « Montage » sont distincts et relativement indépendants. Pourtant, l'item « Suivre la procédure de démontage en sens inverse » de l'UD « Montage » renvoie explicitement au contenu de l'UD « Démontage » et pour cette raison, la compréhension de cette unité est dépendante voire indissociable de la présence des informations contenues dans le « Démontage ». Il n'est pas évident de remédier à cette limitation du modèle, sans faire appel à une analyse sémantique approfondie (à condition qu'une telle analyse soit possible et pertinente), qui pourrait mettre en évidence que ces deux UD sont sémantiquement inséparables.

Le regroupement des UD d'une manière linéaire, exclut également le traitement des références croisées et des références multiples. Dans le premier cas il s'agit de traiter des composants atomiques qui ne sont pas systématiquement liés par leur position aux paragraphes qui les référencent. C'est une situation très

caractéristique des composants comme les figures ou les tableaux, mais qui peut également concerner les renvois référentiels entre les paragraphes textuels :

Comme indiqué dans la procédure de remplacement du moteur (4.1), extraire la goupille qui relie le vérin au couvercle.

Quant aux références multiples, l'assemblage des UD d'une manière séquentielle ne permet pas de prendre en compte le fait que les unités telles que figures, tableaux peuvent intégrer la composition de plusieurs UD.

Mode Performance/mode GM

Les effets EFX, Chorus et Reverb peuvent être configurés individuellement en mode Performance comme en mode GM. Le réglage de l'intensité des effets est défini sur chaque Partie (**Fig. 1**) ; cependant, cette intensité est également définie par le réglage du niveau de départ de chacun des Tones (**Fig. 2**). Les réglages d'effets du Patch affecté à chacune des Parties seront ignorés, mais l'effet de la section EFX appliqué à un Patch affecté à une Partie donnée peut également être appliqué à la totalité de la Performance.

Ainsi, dans le document d'où vient cet exemple, la figure (Fig. 1), qui suit directement le paragraphe qui la mentionne, est attachée automatiquement à la fin de ce paragraphe. En revanche, la figure (Fig. 2), ancrée plus loin dans le document, va être liée au composant qu'elle suit, même si elle ne devrait pas faire partie de cette UD.

6.2 La composition d'une UD en fonction des formes linguistiques de surface

En ce qui concerne les indices linguistiques de continuité permettant de regrouper des unités logiques en UD, les tests sur le corpus nous ont permis d'en enrichir la liste. Afin que la couverture assurée par cette liste s'avère complète pour le corpus testé nous avons modifié ou ajouté certaines règles.

La première, formalise les paradigmes commençant par un déterminant défini (*la, le, les*) suivi, dans son contexte immédiat, soit par un adjectif numéral ordinal (*premier, second*, etc.), soit par un adjectif cardinal (*deux, trois*, etc.), soit par

l'adjectif *dernier*, et dans son contexte proche, par des adjectifs comme *suivant, précédent*. Certains de ces paradigmes acceptent des variations en genre (*la première*) et en nombre (*les premiers*), des insertions (*les règles suivantes*) et des compositions (*les deux règles précédentes*). A l'exception du paradigme formé avec l'adjectif suivant, tous les autres, en exprimant un renvoi anaphorique, doivent se situer dans la première phrase de l'unité analysée. La règle exposée couvre les cas représentés par les deux exemples suivants:

L'écran principal, appelé « Visualiser les tâches globales » sert à visualiser la liste des tâches d'un utilisateur. En fonction des droits de l'utilisateur, l'entête de cette page peut contenir l'action « Créer une nouvelle tâche » qui permet d'accéder au formulaire d'ajout d'une tâche globale. Le corps de l'écran principal est divisé en deux parties:

- la liste des tâches déléguées ;
- la liste des tâches à faire.

La première liste contient les tâches que l'utilisateur a déléguées à quelqu'un d'autre, et **la seconde**, les tâches non déléguées ou celles qui ont été déléguées à l'utilisateur.

Restriction sur la structure des mots de passe: Si 'O', non seulement les 7 derniers mots de passe sont inutilisables lors d'une mise à jour, mais **les deux règles suivantes** doivent aussi être suivies:

- Un nombre minimum de caractères différents doivent être utilisés.
- La différence avec les mots de passe précédents doit porter sur un nombre de caractères suffisant.

Le nombre minimum de caractères requis dans les **deux règles précédentes** est dépendant des longueurs des mots de passe utilisés.

Chacun de ces deux extraits est composé de trois unités logiques (deux paragraphes et une énumération). Une énumération est toujours systématiquement attachée au composant qui la précède. Dans le cas de ces deux exemples, cet enchaînement sémantique est également exprimé par les deux points fermant le premier paragraphe. En revanche, en ce qui concerne la détection de la continuité

entre l'énumération et le paragraphe qui la suit, elle ne devient possible qu'après l'intégration de la règle décrite.

L'analyse des résultats issus du découpage automatique du corpus nous a permis d'identifier une autre catégorie de marques de surface, permettant de signaler que le segment discursif, dans lequel elles sont présentes, est à intégrer dans une UD. Ainsi, ont été omis dans la définition originale du modèle certains adverbes et certaines locutions adverbiales tels que *d'abord, premièrement, deuxièmement, ensuite, finalement, enfin*. Comme l'illustre l'exemple ci-dessous, ces unités constituent un point repérable dans une succession par référence à un « avant » et à un « après » dans le texte, et pour cette raison, peuvent être significatives dans la tâche d'identification du continuum thématique entre segments adjacents.

Opérations diverses

Pour pouvoir entre autres utiliser les fonctions de changement de session à l'intérieur de SESAME et utiliser le module STREAMER, il est nécessaire de charger initialement la structure d'accounts MPE dans la base SESAME à l'aide de la fonction 1 dans le module STREAMER (appelé par la commande "/STRM" dans SESAME ou sous MPE en lançant le programme LOADPASS.PUB.SESAME (en étant connecté dans MANAGER.SYS).

Enfin, pour que toutes les UDCs cataloguées au niveau MPE, soient reconnues par SESAME, il suffit de lancer le programme PLOADUDC.PUB.SESAME à partir de manager.sys.

Un travail de formalisation de cette catégorie de marques linguistiques constitue le noyau du projet décrit par A. Jackiewicz [Jac02]. Dans le cadre de ces travaux, la détection de ces marques, appelées marqueurs d'intégration linéaire, sert la tâche d'identification et de délimitation automatique des cadres organisationnels [Cha97], les structures discursives de type série (*d'une part/d'autre part ou premièrement/deuxièmement*) qui se tissent à travers le texte. Nous pensons que les résultats de cette étude peuvent alimenter notre réflexion, et leur prochaine intégration à notre prototype pourrait considérablement améliorer le processus de repérage des connexions existant entre les composants atomiques d'une UD.

Une réflexion linguistique plus approfondie serait également nécessaire pour améliorer la formalisation du rôle joué dans l'identification du continuum sémantique par les pronoms démonstratifs (*ce, cette, ceci, ces*). Le modèle intègre ces unités comme des indices de renvois anaphoriques à condition qu'ils soient situés au début du paragraphe et, en conséquence, seulement quand ils commencent par une majuscule. Nous ne remettons pas en question l'hypothèse qui permet de considérer les démonstratifs comme des expressions de la cohésion textuelle et de

les faire contribuer à la reconnaissance de la continuité thématique. En revanche, conditionner leur traitement et exploitation au seul cas évoqué, nous semble être un choix réducteur et discutable. Ainsi, dans les exemples qui suivent, les démonstratifs soulignés, bien qu'ils ne soient pas placés à l'initial de leurs paragraphes, remplissent indiscutablement le rôle d'indices de continuité.

Dans la partie haute de la fenêtre de votre navigateur se trouve un bandeau bleuté.

Il comporte des outils vous permettant de gérer votre connexion et vos paramètres utilisateurs. En mode standard, à gauche de **ce** bandeau, le logo de Mioga vous permet de revenir à la page d'accueil de votre communauté en ligne. Cette action ne remet pas en cause l'authentification en cours.

Si l'accès au dossier courant est autorisé en modification, trois liens sont proposés :

- Créer un dossier …
- Créer un fichier …
- Envoyer un fichier qui permet de transférer un fichier depuis le disque dur vers le serveur Mioga, dans le dossier courant.

Quand le presse papier n'est pas vide, c'est en bas de **cette** zone qu'il sera visualisé avec le nom du fichier ou dossier sauvegardé et deux icônes. La première de la forme d'une poubelle …

Afin de pouvoir couvrir de tels cas, il semble logique d'intégrer à la méthode une règle qui, en limitant la détection d'un déterminant à la première phrase du paragraphe, l'autorise à être précédé par d'autres formes linguistiques ou signes typographiques. Une étude linguistique examinant la distribution des démonstratifs dans une phrase, réalisée sur un corpus d'une taille plus significative, devrait nous permettre de recenser leurs paradigmes possibles, de formaliser ces connaissances, et de les intégrer au modèle sous la forme de règles contextuelles.

La dernière modification importante apportée au modèle concerne les règles formalisant le rôle des déictiques spatiaux, les unités linguistiques qui n'ont pas de valeur en langue, mais dans l'instant d'énonciation. L'exemple qui suit nous prouve

que les déictiques ne peuvent fonctionner que dans le cas où cette situation d'énonciation (la situation de consultation du document) est évidente pour le lecteur.

4.7.2. Import

ATTENTION : cette fonctionnalité efface toutes les données liées aux applications propres à chaque groupe. Cette fonctionnalité ne doit pas être accessible à n'importe qui. L'administrateur porte la responsabilité de l'accès à cette fonctionnalité.

Le fichier à importer provient de la fonction export décrite **ci-dessus**. Les modes disponibles sont SET, ADD et SYNCHRO.

Dans le modèle de T. Ouerfelli, les déictiques spatiaux sont considérés comme des marques de continuité, à condition qu'ils se trouvent dans la première ou dernière ligne d'un paragraphe. Etant donné le rôle qu'ils jouent dans le texte, il nous semble plus approprié de les traiter comme tels indépendamment de la place qu'ils occupent dans le paragraphe.

7. Conclusion

Nous nous sommes intéressées à la mise en œuvre d'un prototype informatique de segmentation des documents techniques composites, basé sur un modèle théorique. Cette démarche d'implémentation a été entreprise afin de vérifier que les fondements théoriques à l'origine de la définition de ce modèle passent le test d'une automatisation.

Evaluer un modèle par implémentation soulève inévitablement la question de la validation quantitative de la méthode implémentée. Confrontées à la difficulté de constituer un corpus de documents de référence découpés manuellement, nous n'avons pas engagé cette démarche. En outre, confronter nos résultats à un corpus découpé automatiquement n'est recevable qu'à condition de disposer des documents segmentés, selon une méthode similaire à la nôtre. Or, aujourd'hui, nous n'en avons pas pris connaissance.

Cependant, les tests réalisés sur un corpus ont montré que le système fragmente le document technique en unités documentaires qui sont conformes à leur définition dans le modèle. Ceci nous permet d'affirmer que les fondements théoriques du modèle sont valides et recevables dans la tâche de segmentation de ce type de document. Nous pouvons en conclure que les différentes sources d'information présentes à la surface du texte technique, que ce soit des informations explicites pour

le lecteur (formes linguistiques et structurelles) ou bien des informations implicites (disposition), peuvent être efficacement exploitées et combinées pour mener à bien la tâche de segmentation automatique. L'exploitation judicieuse de ces indices de surface permet d'obtenir des segments de textes sémantiquement cohérents et relativement autonomes, qui reflètent l'articulation sémantique des différents thèmes traités dans le document, et d'approcher ainsi le sens par la forme.

8. Références bibliographiques

[Ada92] J. M. Adam, *Les textes, types et prototypes*, Nathan, Paris, 1992.

[Ada94] J. M. Adam, *Le texte narratif : traité d'analyse pragmatique et textuelle*, Nathan, Paris, 2ème édition, 1994.

[Bal90] J. P. Balpe, *Hyperdocuments, hypertextes, hypermédias*, Eyrolles, Paris, 1990.

[BN00] B. K. Boguraev and M. S. Neff, Lexical Cohesion, Discourse Segmentation and Document Summarization, In *Actes RIAO'00 : Recherche d'Information Assistée par Ordinateur*, Paris, 2000.

[Bro97] J. P. Bronckart, *Activités langagières, textes et discours*, Neuchâtel, Delachaux et Niestlé, 1997.

[Cha95] M. Charolles, Cohésion, cohérence et pertinence du discours, *Travaux de linguistique*, (29):125-15, 1995.

[Cha97] M. Charolles, L'encadrement du discours - Univers, champs, domaines et espace, *Cahier de recherche linguistique*, (6), 1997.

[Eym92] G. Eymard, *Traitement documentaire des sommaires : Des mots-clés à l'extraction de connaissances. Application à une documentation technique*, Thèse de doctorat en sciences de l'information et de la communication, Université des Sciences Sociales, Grenoble, 1992.

[FGM98] O. Ferret, B. Grau, and N. Masson, Thematic segmentation of texts : two methods for two kinds of texts, In *Actes ACL-Coling'98*, pp. 392-396, Montréal, Canada, 1998.

[FGMP01] O. Ferret, B. Grau, J. L. Minel, and S. Porhiel, Repérage de structures thématiques dans des textes, In *Actes TALN 2001*, pp. 163-172, Tours, 2001.

[Gir89] B. Girard, *La production de documents techniques assisté par ordinateur*, Hermès, Paris, 1989.

[GLBM98] R. Ghiglione, R. Landre, R. Bromberg, and P. Molette, *L'analyse automatique des contenus*, Dunod, Paris, 1998.

[Hea97] M. Hearst, Texttiling: segmenting text into multi-paragraph subtopic passages, *Computational Linguistics*, 23(1):33 - 64, 1997.

[Heu94] L. Heurley, *Traitement de textes procéduraux : étude de psycholinguistique cognitive des processus de production et de compréhension*, Thèse de doctorat en psychologie, Université de Bourgogne, 1994.

[Heu02] L. Heurley, Psychologie de la production et de l'utilisation de documents techniques, *Psychologie française*, 47(1), march 2002.

[HG02] N. Hernandez and B. Grau, Analyse thématique du discours : segmentation, structuration, description et représentation, In *Actes CIDE 5 2002*, pp. 277-288, Hammamet, Tunisie, 2002.

[HH76] M. A. K. Halliday and R. Hasan, *Cohesion in English*, Longman, Londres, 1976.

[Jac02] A. Jackiewicz, Repérage et délimitation des cadres organisationnels pour la segmentation automatique des textes, In *Actes CIFT'02 : Colloque International sur la Fouille de Textes*, pp. 95-105, Hammamet, Tunisie, 2002.

[KKMK98] M.-Y. Kan, J. L. Klavans, and K. Mc Keown, Linear segmentation and segment significance, In *Actes ACL-Coling*, Montréal, Canada, 1998.

[Koz93] H. Kozima, Text segmentation based on similarity between words, In *Proceeding of 31th Annual Meeting of the Association for Computational Linguistics*, Colombus, Ohio, USA, 1993.

[Lun80] L. Lundquist, *La cohérence textuelle : syntaxe, sémantique, pragmatique*, Nyt Nordisk Forlag Arnold Busck, Copenhague, 1980.

[Mar00] D. Marcu, The rhetorical parsing of unrestricted texts: A surface-based approach, *Computational Linguistics*, 26(3), 2000.

[Mic99] C. Michel, *Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogène*, Thèse de doctorat en sciences de l'information et de la communication, Université de Lyon 1, 1999.

[Oue01] T. Ouerfelli, *La segmentation des documents techniques composites dans une perspective d'indexation : vers la définition d'un modèle dans une optique d'automatisation*, Thèse de doctorat en sciences de l'information et de la communication, Université de Grenoble 3, 2001.

[Pag97] C. Paganelli, *La recherche d'information dans des bases de documents techniques en texte intégral. Etude de l'activité des utilisateurs*, Thèse de doctorat en sciences de l'information et de la communication, Université de Grenoble 3, 1997.

[Pap95] F. Papy, *Hypertextualisation automatique de documents techniques*, Thèse de doctorat en sciences de l'information et de la communication, Université de Paris 8, 1995.

[PC97] J. Ponte and W. B. Croft, Text segmentation by topic, In *Proceeding of the European Conference on Research and Advanced Technology for Digital Libraries*, pp. 113 - 125, 1997.

[SSCM96] G. Salton, A. Singhal, Buckley C., and M. Mitra, Automatic text decomposition using text segments and text themes, In *Actes Hypertext'96, Seventh ACM Conference on Hypertext*, pp. 53-65, Washington, USA, 1996.

[Wri97] P. Wright, Presenting technical information : a survey de research findings, *Instructional science*, 6:93-134, 1997.

Schema matching for Semantic Reuse of XML documents

Aida Boukottaya, Christine Vanoirbeek

Media Research Group
EPFL (Swiss Federal Institute of Technology
1015 Lausanne, Switzerland

aida.boukottaya@epfl.ch
christine.vanoirbeek@epfl.ch

Résumé :

La structuration des données à l'aide du langage XML démontre clairement les avantages qui peuvent en être tirés dans un but de publication. Cependant, le concept de document structuré s'intègre dans un cadre plus général : le document XML inclut désormais une dimension sémantique reflétée par la conception de modèles de documents qu'ils soient définis via des DTDs ou, plus récemment via des Schémas XML, décrivant la structure sémantique des contenus. L'adoption de tels modèles a favorisé la production et la disponibilité des ressources, toutefois structurées de manière hétérogène. Dans un tel contexte, partager et réutiliser des ressources existantes s'avère d'un intérêt incontestable, à la fois en termes d'interopérabilité entre les applications et entre les différents auteurs. La difficulté pour établir des transformations entre documents XML hétérogènes, utilisant des langages comme XSLT, réside dans le fait que les correspondances structurelles et sémantiques entre ces documents doivent être soigneusement spécifiées par un expert humain. Le présent article propose un algorithme visant à simplifier et automatiser cette tâche, en analysant la sémantique des documents XML inférée à partir de leur structure logique (les éléments et leur organisation) ainsi que par l'analyse des mécanismes de typage introduit par les schémas XML.

Mots clés : La sémantique des documents structurés, la réutilisation des documents XML, calcul des correspondances sémantiques entre documents XML.

Abstract:

Structuring data with XML clearly proves the benefits to be taken for publishing purposes. However the concept of XML document appeared to be of much wider interest. XML document currently acts as a major component within global information systems: it becomes a dynamic and rich component that includes *semantic information* nested within the document structure. As the number of applications that utilize XML documents grows, the importance of XML documents reuse increases greatly. A serious obstacle for translating directly between two XML documents, using languages like XSLT, is that a mapping between the two XML representations needs to be carefully specified by a human expert. In this paper, we introduce a novel *schema matching algorithm* to help the process of exchanging XML data between autonomous and heterogeneous Web applications. The latter algorithm is based on semantic information nested within XML structures. Semantic is first captured through the explicitation of element names meanings, and second through the analysis of XML Schema's designer point of view expressed by the logical organisation of XML content and additional semantic information given by means of features such as user defined types, subtyping mechanisms, substitution groups, etc.

Keywords: XML document semantics, XML data reuse, Semantic matching.

1. Introduction and Motivation

The WWW is a great success with respect to the amount of stored documents and the number of users. Until now, the web has been designed for direct human processing and lacks machine understandable semantics. The Web community is attempting to address this limitation by designing the semantic Web. The semantic Web aims to extend the current World Wide Web by providing data that embeds semantic information processable by computers and developing tools that automatically interpret this information in order to perform specific tasks which are up to now manual. XML mark-up language [XML 98] has been proposed as a first step to this end. An important feature of XML language is the separation between the logical and physical structures of a document which clearly offers an easy and efficient way to manipulate data. In order to make the Web interoperable, most current research efforts focus on proposing several knowledge representation languages built upon XML for describing data meaning. Well-known examples include RDF and RDF Schema, as well as ontologies description languages: DAML+ OIL and OWL, the recent W3C standard.

A significant problem (also noticed by [Alon 03] and [Patel 02]) with the proliferation of such knowledge representation languages, is the semantic discontinuity between WWW languages (XML/ XML Schema [XML Sch 01]) and the semantic web languages (RDF/ RDFS/ DAML + OIL). This discontinuity results from a difference in their respective modeling foundations. RDF, RDFS, as well as other semantic web languages, are focused on identifying the "*domain structure*" based on directed graph model. In contrast, most existing Web data sources export their data into XML which tends to focus on "*document structure*". Structuring data based on a tree model clearly proves the benefits to be taken for publishing purposes. However the concept of XML document appeared to be of much wider interest. XML document currently acts as a major component within global information systems: it becomes a dynamic and rich component that includes *semantic information* nested within the document structure. Most Web applications and proposed tools relay on this structure. In this context, it would be desirable for the semantic web to focus on how to interoperate with existing data sources and thus to be able to map not only different domain structures but also heterogeneous document structures. XML data reuse is defined as the problem of migrating the contents of data sources to an instances of a given target schema. This is typically attained in real world by writing translators (custom code) encoded on a case-by-case basis using specific languages such as XSLT [XSLT 99]. Writing custom code is time consuming and generally needs programming skills.

In this paper, we introduce a novel *schema matching algorithm* to help the process of exchanging XML data between autonomous and heterogeneous Web applications. Schema matching is defined as the task of finding correspondences, so called *mappings* between two heterogeneous schemas. Many XML schema matching algorithms have been proposed in the literature, but none really encompass the overall matching process (matching discovery, matching presentation and matching execution). The goal of this work is to contribute to the development of a framework that provides a generic view and figure onto the overall matching process based on semantic information nested within XML structures. Semantic is first captured through the explicitation of element names meanings, and second through the analysis of XML Schema's designer point of view expressed by the logical organisation of XML content and additional semantic information given by means of features such as user defined types, subtyping mechanisms, substitution groups, etc.

The paper is organized in the following way. In section 2, we outline the state of the art in the field of XML data reuse and sharing. Section 3 describes our proposed approach for modeling XML schemas. In Section 4 we present the source-to-target mapping algebra and point a set of primitive transformation operations. Section 5 deals with the matching process. We describe our prototype system in section 6. Finally, we conclude the paper with a discussion and future work.

2. Research background and related work

2.1 XML Data integration

Several related research projects were conducted in the area of *data integration systems*. Research in these systems refers the problem of reconciling and combining data residing in distributed and heterogeneous data sources and providing the user with a unified view, namely mediated schema (or global schema). Efforts to develop XML data integration systems are still ongoing [Castano 02], [Lee 02]. Two significant problems have to be noticed with such systems. The first limitation is that these systems are centralized systems that exploit mappings between a single mediated schema and schemas of data sources. However, the decentralized nature of the Web makes existing information integration systems and approaches inappropriate because it is unrealistic to assume that a single mediator server can be deployed in a distributed environment such as the Web. Second the goal of schema integration systems is evaluate queries on the global schema. Whenever a user poses a query in terms of global schema, the data integration system uses a query reformulation procedure to translate the query into sub-queries that can be executed in sources. Query reformulation uses mapping (which is either specified by a user or calculated within the integration process) between local schemas and the global schema. In contrast, in our interoperability problem, there is not freedom in designing a target schema that matches the sources and mappings are not given in advance, but have to be created. Hence, we must cope with both structures and constraints of the source and target schemas.

2.2 XML Data translation

Research on XML data translation has been mostly focused on translation languages rather than on automating the generation of transformation programs. Several simpler and highly declarative transformation languages [Krishnamurthi 00], [Tang 01] have been introduced as solutions to avoid programming. Special graphical tools have been also proposed to assist the specification of the transformations [Pietriga 01], [XSLWIZ 01]. See [Vernet 02] for more examples of transformation languages and tools. These languages and tools are very useful in describing and specifying transformations. However, they still require developers to manually indicate mappings for each source and target pair. Manual mapping is time consuming and thus especially unacceptable for applications where the information sources change frequently.

2.3 XML Schema matching

An alternative strategy that is used for reconciling XML data is based on schema matching techniques to automate the mapping process [Su 01]. Schema

matching is the task of finding correspondences between two heterogeneous schemas. TranScm [Milo 98] uses schema matching to derive translation between schema instances. The approach is based on a set of predefined "rules" that describe the common transformations. Rules are checked in a fixed order based on their priorities. However since TranScm system aims to provide a general approach, new rules and priority assignment to these rules should be provided to deal with XML schemas features. Clio project [Popa 02] [Miller 00] has already created a tool that migrates data from a relational data source to a target which is either relational or XML; it does not perform XML-to-XML data translation. Other work has considered *schema matching*, the automatic detection of schemas similarities. A recent survey of automatic schema matching [Rahm 01] classifies approaches respecting to the schemas information (element naming, structure, data types, integrity constraints, etc.) used to discover schemas similarities and auxiliary information (generally domain specific common terminologies or thesaurus). It is handled in some systems [Doan 01] through machine learning techniques to evaluate data instances and train matchers, and then predict element similarities by combining their results. Other systems [Madhavan 01] combine name and structural matching to predict element similarity based on the similarity of the name and data type of their components. These works present several limitations (described in detail in a previous paper [Boukottaya 04]). First, they deal essentially with *mapping discovery*. The result is a similarity coefficient in [0,1], expressing the linguistic and structural closeness between two schema elements. With such result the problem of data translation is partially solved, we still need to specify the operation needed to transform a source instance into a target one. Second, the majority of current schema matching algorithms ignore the semantic dimension of XML such as generalization / specialization relationships expressed by means of substitution groups and sub-typing mechanisms.

3. XML Schema data model

XML Schema retains the ability of DTDs to structure XML data via a flexible way based on a tree data model. At the same time, XML Schema standard introduces other features such as user defined types, subtyping through extension and restriction mechanisms, substitution groups. To capture all XML schema features, we need a rich data model with complex modelling constructs and constraints. The widely adopted data model to describe XML Schema features is a graph based model. [Hardt 02] proposes a schema-based querying approach where XML schema is described as a graph where nodes (also called concepts) are XML Schema types and edges are relationships between concepts. Two kinds of relationships have been considered: containment and generalization/specialization. [Feng 02] provides a graph based model for XML Schema through four major components: a set of *atomic* and *complex* nodes, a set of directed edges, representing relationships between nodes, a set of labels denoting different types of relationships,

including *aggregation, generalization, association*, and *of-property* relationships; and finally a set of constraints defined over nodes and edges. The problem of graph matching has been studied in the literature in the context of graph isomorphism and weighted graph matching [Galil 86], [Ullmann 76] and it was proved that it is NP-Complete [Gold 96]. To deal with XML Schema features, and avoid the problem of graph matching, we introduce a new XML Schema model based on three different views: the *semantic view*, the *logical view* and the *constraint view*. We do not aim at a complete formalization of all XML Schema details, but rather at capturing its essential modeling features as required for XML schema matching.

3.1 The semantic view

XML Schema standard allows a designer to express additional semantics beyond the information contained in XML documents. The semantic view aims to capture such information. The semantic view describes essentially types and conceptual abstractions expressed by means of user defined types, generalization / specialization relationships and substitution group mechanisms. From modelling perspective a semantic view can be seen as XML Schema type (or class) hierarchy.

⇨ **Definition 1: (Semantic view)**

A semantic is a labelled graph ($\mathbb{r}_t$, $\mathbb{T}$, $\mathbb{R}$) where $\mathbb{T}$ is a set of type names. Each type $T \in \mathbb{T}$ is either simple or complex, and either abstract or concrete leading to two partitions: $\mathbb{T} = \mathbb{T}_s \cup \mathbb{T}_c$ and $\mathbb{T} = \mathbb{T}_a \cup \mathbb{T}_{na}$. $\mathbb{R}$ is a set of binary relationships between types. $R \in \mathbb{R}$ is either an association relationship $\mathcal{A}$, or a subtyping relationship $\mathcal{S}$. Two kinds of subtyping relationships exist: restriction $\mathcal{R}$ or extension $\mathcal{E}$, leading to the partition $\mathcal{S} = \mathcal{R} \cup \mathcal{E}$. $\mathbb{r}_t$ is the root type.

Figure 1 illustrates a portion of a semantic view related to the schema (*publication.xsd*) described in the annexe. Types PUBLICATION and ARTICLE are abstract types. The figure shows several subtyping relationships, as example, the one between PUBLICATION and BOOK, conceptually means that BOOK is a more specific concept than PUBLICATION.

3.2 The logical view

The logical view is based on traditional tree data model (used for DTDs) describing the logical structure of XML elements and attributes. A binding of elements and attributes (belonging to the logical view) to types (belonging to the semantic view) has to be provided. More formally:

⇨ **Definition 2: (Logical view)**

A logical view is a labelled tree ($\mathbb{r}$, $\mathbb{N}$, $\mathbb{E}$) where $\mathbb{N}$ is a set nodes names, a node is either atomic or complex, leading to a partition $\mathbb{N} = \mathbb{N}_a \cup \mathbb{N}_c$. XML

Schema Leaf elements $\mathbb{L}$ and attributes $\mathbb{A}$ are considered as atomic nodes, thus $\mathbb{N}_a = \mathbb{L} \cup \mathbb{A}$. $\mathbb{E}$ is a set of edges representing *containment* relationship between nodes. r is the root of the tree.

⇨ **Definition 3: (Binding elements and attributes to types)**

A binding elements and attributes to types is binary relationships $\mathbb{B}$, where $\mathbb{B} \subseteq (\mathbb{N} \times (\mathbb{T}_s \cup \mathbb{T}_c)) \cup (\mathbb{A} \times \mathbb{T}_s)$.

Figure 2 illustrates the logical view of the schema *publication.xsd*. For the clarity of the example, we do use the same names for elements and types in both semantic and logical view. An example of binding of elements to types is element *book* (figure 2) is of type BOOK (figure 1).

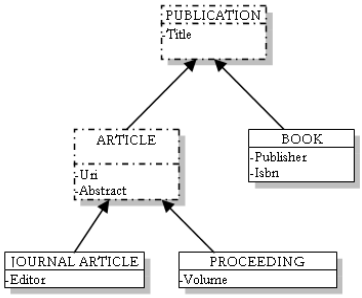


Figure 1 : A semantic view Example

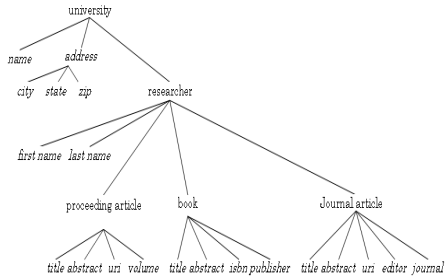


Figure 2 : A logical view Example

3.3 The constraint view

In addition to typing and structural organisation, XML schema enables users to express different kinds of constraints. These constraints can be defined over both nodes and edges. They explicate various requirements such as domain, cardinality, uniqueness, and referential integrity. Basically, the constraint view describes the following kinds of constraints.

⇨ **Definition 4: (constraint view)**

A constraint view is a set of constraint expressions. We distinguish three types of constraints namely *constraints over edges* and *constraints over a set of edges* and *constraints over an element*.

- ***Constraints over edges***: are typically *cardinality constraints* that specify the number of instances of one node that may relate to a single instance of another node. For example an element researcher can have several

publication. This constraint is expressed in XML Schema syntax by the use of "*maxOccurs=unbounded*"

- ***Constraints over a set of edges***: are typically ordered composition ("sequence" in XML schema syntax), unordered composition ("all") or exclusive disjunction ("choice").
- ***Constraints over a node***: are uniqueness, referential integrity and domain constraints.
  - *Uniqueness*: When a node appears more then once in the document, the uniqueness constraint requires each of these appearances to have a unique content.
  - *Referential integrity*: The referential integrity constraint on a node *n* requires that there must exist another corresponding referential node *n'* where both nodes *n* and *n'* must have the same content.
  - *Domain constraints*: are very broad, they include: restrict the legal range of numerical values by giving the maximal/minimal values; limit the length of string values; specify the permissible members for a constructional content (such as *list* or *union*)

4. Matching Algebra

We propose an algebra for source-to-target mapping to capture the required operations needed to express such mapping. We propose a set of primitive operations that are the building blocks that would enable schemas transformations. These primitive operations can be composed together to represent larger transformations. We classify such operations into two categories:

⇨ **Conceptual operations:**

Conceptual operations are operations that capture semantic heterogeneities between schemas. Semantic heterogeneities deal essentially with differences in structure (*how are the data organized?*) and differences in interpretation (*what do the data mean?*). We propose four operators:

- *Merge*: Two distinct entities (element/attributes) are merged into one entity. Elements like *street*, *country* and *state* can be separate in source schema and merged as *address* element in target schema. Then the values of *address* in the target schema match a concatenation of values from source schema elements.
- *Split*: This is the reverse operation of merge. The values of target schema match then a decomposition of a source schema element.
- *Rename*: change element and attribute names.
- *Connect*: The connect operation is one to one mapping that maps two equivalent entities without any transformation.

⇨ **Implementation operations:**

The richness of XML schema language gives rise to a larger variety of possibilities to implement the same concepts. For example, *dates* may be represented as *strings* in one schema, or in another schema as instances of the primitive type "*Date*". These conflicts are very difficult to resolve since they require extra information that only the user can provide. By analysing XML Schemas, we provide a set of common functions within a library that the user can easily modify or extend. Examples of such functions deal with type compatibility problem. Imagine that the target type is included in the source type. We suggest functions like *rounding a real to an integer, truncating string to a given length*, etc.

5. Matching techniques

5.1 Linguistic Analysis

The linguistic analysis aims at making explicit the meaning of used element names and establishes semantic relationships between them based on WordNet [Miller 95], an electronic lexical database, which organizes English words into synonym and hypernym sets. Our linguistic analysis is inspired essentially from Hirst and St-Onge's work [Hirst 98]. The idea behind Hirst and St-Onge's measure of semantic relatedness is that two concepts are semantically close if their WordNet synsets are connected by a path that is not too long and that does not change direction too often. WordNet relations are classified as *upward*, *downward*, or *horizontal*. Upward relations connect more specific concepts to more general ones. For example, *is-a* is an upward relation while *contains* is considered to be a downward relation. Horizontal links are very specific specializations including antonymy and synonymy. Since we are focusing on matching similar words, we restrict horizontal relations to synonymy. The Hirst and St-Onge measure has four levels of relatedness: *extra strong, strong, medium strong* and *weak*. An extra strong relation is based on the surface form of the words and therefore does not apply in our case since we are measuring the relatedness of word senses. Two words have a strong relation if one horizontal link exists between their respective synsets. We restrict our algorithm to these three scenarios: the first occurs when there is a synset common to two words. The second occurs when there is a synonymy relation between synsets of each word. The third occurs if one of the words is a compound word and the other word is a part of the compound, and if there is any kind of synset relation (except antonomy) between the two words. The medium strong relation is determined by a set of allowable paths between words that are described by Hirst and St-Onge's algorithm. If such a path exists between two words then we have a medium strong relation between them. Based on these Hirst and St-Onge's relations

we identify five kinds of semantic relationships between words, namely *equivalent* (≡), *Border than* (⊇), *Narrower than* (⊆), *related to* (~), and *disjoint* (⊥).

5.2 Semantic view analysis

The semantic view can be viewed as a rich model that allows a designer to express additional semantics beyond the information contained in XML documents. These additional information traduces the XML Schema's designer point of view, thus can successfully give a clue to which schema elements match.

- ⇨ **Abstract Types:**

  XML schema introduces the notion of abstract types. Like object oriented modeling, abstract types may not have direct instances, but their concrete subtypes may. Our XML Schema example described two abstract types PUBLICATION and ARTICLE. This information is not described in the related tree structure (logical view). Let consider that our XML Schema example is the source schema (called S) and we want to match it with a target schema T where each *researcher* have an element *publication* in the related logical view. The fact that *ProceedingArticle*, *JournalArticle* and *book* elements in S are bounded to PROCEEDINGARTICLE, JOURNAL, ARTICLE and BOOK types which are (direct or indirect) subtypes of the abstract type PUBLICATION allow us to match *ProceedingArticle*, *JournalArticle* and *book* elements in S with *publication* element in T.

- ⇨ **Subtyping:**

  XML Schema allows subtyping through extension and restriction. Restriction of a simple or complex type leads to the same kind of type, whereas extension of a simple type or complex type always yields to a complex type. Conceptually, extension and restriction between two types (in the XML Schema sense) indicate that these types are semantically related. This information can successfully help the matching algorithm to infer new semantic relationships. Let us consider a source schema where *publication*, *book* and *article* are schema elements, of type PUBLICATION, BOOK and ARTICLE, respectively. Consider also a target schema having an element *publication* of type PUBLICATION. WordNet querying gives that a source schema's element *publication* is equivalent to the target element *publication*. The fact that BOOK and ARTICLE are subtypes of PUBLICATION in the source schema, allow us to infer that *book* and *article* in the source are narrower than *publication* in the target.

- ⇨ **Substitution Group:**

  XML Schema allows defining a group of substitutable elements (called a *substitutionGroup*) by declaring an element (called the head) and then

declaring other elements which state that they are substitutable for the head element. The type of every element in the substitutionGroup must be the same as, or derived from, the type of the head element. Two substitutable elements are conceptually at the same level. Let us consider a source schema S where elements *book* and *monograph* are substitutable and a target schema T where element *book* is equivalent (by WordNet) to the source element *book*. Since in the source schema elements *book* and *monograph* are substitutable, an equivalence relationship between *monograph* and target schema element *book* and can be inferred.

5.3 Structural analysis

The two previous matching techniques identify matching pairs based on their semantics given by used labels and semantic view details. This is not enough since *access paths* for retrieving data from the source have to be provided. For this we make use of the structure described in the source and target logical views. When comparing structural properties of a target and a source schema, we apply a two step strategy. First, we compute correspondences between complex elements of the target and source schema. Each complex element is composed of a set of elements representing its *context*. The comparison is based on the comparison of such contexts to identify a set of possible *compatible* elements. Then, with the guide of compatible elements between the target and source schemas, we compute the similarity based on the global structure of logical views.

5.3.1 Leaf Context comparison

Based on the data model introduced in section 3, we partition schema elements into atomic and complex nodes. We further introduce the notion of leaf context of complex schema elements based on the following definition:

⇨ **Definition 5: (Leaf context)**

The leaf context of a node n ∈ Nc is the set of atomic nodes that are n's children

$$L_{ctx}(n)_{n \in Nc} = \cup_{Na} \{n' \mid \exists e \in \mathbb{E} \text{ where } nen'\}$$

Examples:

L_{ctx} (university) = {name, location}
L_{ctx} (address) = {city, state, zip}
L_{ctx} (researcher) = {Firstname, Lastname}

After leaf context construction for complex elements, as well as in target and source schema, we compare their leaf contexts between the schemas

based on linguistic and semantic view analysis. In particular, we make a guess about how to map a source schema into a target schema by determining a set of *compatible elements* that are highly similar. Two complex elements are considered as compatible if the two following conditions hold (1) they are similar by WordNet and semantic view analysis and (2) They have similar leaf contexts. Let us consider the two schemas given by figure 2 and 3. The schema of figure 2 is considered as the target schema (T) and the one of figure 3 as the source schema (S). For example we guess that the elements *book* in both schema are compatible since they are equivalent (by WordNet analysis) they have the same leaf context (title, abstract, isbn, publisher). Identifying compatible elements aims to reduce the context scope and thus limits the search of mappings to appropriate, smaller search space. It is important to notice that leaf context comparison can introduce wrong mappings. For example, the element *address* in the target schema is mapped to the element *address* in the source schema, which is wrong since one represent the address of a university and the other represent the address of an author. The aim of the next matching step is to identify such wrong mapping based on the analysis of the global structure of source and target schemas.

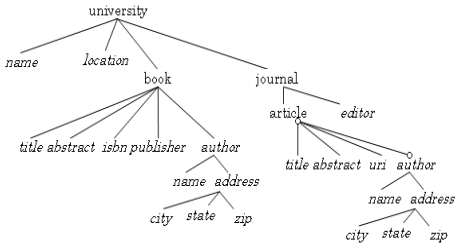


Figure 3 : Source schema logical view

5.3.2 Global structure analysis

Once compatible elements are identified, we need to analyse their relationships within the global schema structure in order to provide access paths enabling data retrieval from sources. To this end, we introduce the notion of ancestor context. Our global structure analysis can then be seen as ancestor context comparison.

⇨ **Definition 6: (Ancestor context)**

The *ancestor context* of a node $n \in \mathbb{N}_c$ is given by the path from the root to n.

$$A_{ctx}(n) = (\mathbb{r}_t, i, j, \ldots, n)$$

Examples:

A_{ctx} (university) = (university)
A_{ctx} (book) = (university, researcher, book)
A_{ctx} (researcher) = (university, researcher)

When comparing ancestor contexts of compatible elements, we assume that elements that are connected by a path are semantically related. By analysing the ancestor context of element *address* in the target schema (which is *university*), *address* target element is mapped to the source element *location* and not *address*. Several issues are considered when comparing ancestor context:

- ***Partial mapping:***

  Since our main goal is to populate a target schema from a source schema, where both schemas are heterogeneous and given in advance, all target elements may not be matched with some source elements. A notion of *partial matching* has to be introduced to overcome this problem. Imagine that the element *journal* and the sub-tree rooted at journal don't exist in the source schema S. The target element *journalarticle* is not matched.

- ***Flexible mapping*:**

  Since source and target schema are autonomous, they may structure the same information in different ways. For example, in the target schema, details about publications are grouped by authors; in contrast the same details are grouped by the kind of publication (article, book, etc.) in the source schema. In such cases rigid matching doesn't work and some *flexibility* to the matching has to be added. Contexts: (university, researcher, book) in the target schema, and (university, book, author) are considered equivalent under flexible matching semantics.

5.4 Constraint analysis

The last matching technique checks constraint compatibility between two matched elements. For this, we adopt the same four cases outlined in [Xu 03]: (1) The constraints on source schema and target schema are equivalent, schema entities are matched and nothing further need be done. (2) The constraints of source schema imply the constraints of target schema but not vice versa, entities are matched and

nothing further need be done. (3) The constraints of target schema imply the constraints of source schema but not vice versa. In this case, we can select only source elements that satisfy target constraint. (4) Neither the constraints of target schema imply the constraints of source schema nor vice versa. This is a combination of (2) and (3) and thus, since there is nothing to do for (2), we act as explained in (3).

6. The Prototype System

The prototype system, we are developing, incorporating all the ideas discussed in the paper, consists of three *modules*: *Modelling toolkit*, *Matcher engine*, and *execution engine*. Figure 4 outlines the whole system architecture.

⇨ **Modelling toolkit**

For the generation of semantic and logical views from XML schemas, we developed a toolkit that is composed essentially of two graphical tools: *Semantic view editor and viewer*, *Logical view editor and viewer*. These graphical tools will encourage the user to either add semantics or extend the generated views.

⇨ **Matcher engine**

We develop a *Matcher engine* to perform schema matching. The matcher engine uses additional modules: an interface for querying WordNet and a graphical user interface that allows the users to validate mappings generated by the system, and provide further domain constraints. We also provide a mapping evolution module. This module focuses on storing the generated mappings and on their synchronization with the changes in source and target schemas. Keeping mappings evolution allows their reusability. The goal is to avoid reapplying the mapping process every time schemas change.

⇨ **Execution engine**

The execution engine is responsible for parsing mapping results and generating automatically XSLT transformation scripts. It also permits the translation of data instances (XML files) valid against a source schema to instances valid against a target schema.

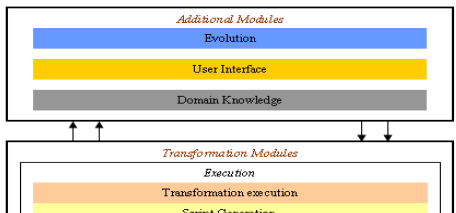


Figure 4 : Prototype architecture

7. Conclusion

Due to the extensive use of XML markup language in several domains, there has been a great interest on proposing rich data models (XML Schemas) that reflect document semantics and structure. The existence of such rich schemas has made a large amount of heterogeneously XML documents widely available. In this framework, XML documents reuse and sharing is of major concern. Currently, to make XML documents interoperable, the burden falls on the human to first analyse both the semantics and the structure of the source and target XML documents, and second to manual coding the transformations using specific languages such as XSLT. This work proposes a framework to reuse of XML documents based on schema matching techniques. We have specially focused on two fundamental problems: *How to deal with all features of XML schema during the matching process and which matching techniques to use to discover semantic relationships between schema entities*. For this end, we first propose a data model for XML Schemas that capture all XML Schemas features based on three views: *Semantic, logical* and *constraints* views. Second we propose several matching techniques to identify similar schemas entities. In the future, we intend first to finalise the matching prototype system and to use mapping results to generate transformation scripts (essentially XSLT scripts).

8. References

[XML 98] Extensible Markup Language (XML) Version 1.0. World Wide Web Consortium. http://www.w3.org/TR/REC-xml.

[XML Sch01] W3C Recommendation, "XML schema Primer", W3 Consortium, available at http://www.w3.org/TR/xmlschema-0, 2001.

[Alon 03] Alon Y. Halevy, Zachary G. Ives, Peter Mork, Igor Tatarinov, Piazza : Data Management Infrastructure for semantic Web Application. World Wide Web Conference, May 2003.

[Patel 02] P. Patel-Schneider and J. Simeon. Building the Semantic Web on XML. In Int'l Semantic Web Conference' 02, June 2002.

[XSLT 99] W3C Recommendation. XSL Transformations XSLT Version 1.0, Available at http://www.w3.org/TR/xslt (June 2002).

[Castano 02] S. Castano, A. Ferrara, G.S. Kuruvilla Ottathycal, V. De Antonellis (2002): A Disciplined Approach for the Integration of Heterogeneous XML Datasources. DEXA Workshops 2002: 103-110.

[Lee 02] M. Li Lee, L. Huai Yang, W. Hsu. Xclust: Clustering XML Schemas for Effective Integration. Proceedings of the 2002.

[Krishnamurthi 00] S. Krishnamurthi, K. Gray, and P. Grauke. (2000) Transformation-by-example for XML. Proceeding of the Second International Workshop on Practical Aspects of Declarative Languages (PADL'00), Lecture Notes in Computer Science, vol. 1735, pp. 249-262.

[Tang 01] X. Tang and F. Tompa. (2001). Specifying transformations for structured documents. In proceeding of 4th International Worshop on Web and Databases (WebDB 2001), pp. 67-72.

[Pietriga 01] E. Pietriga, J-Y. Vion-Dury, and V.Quint.(2001). Vxt: a visual approach to XML transformations. Proceeding of the ACM Symposium On Document Engineering.

[XSLWIZ 01] XSLWIZ. (2001). http://www.induslogic.com/products/xslwiz.html.

[Vernet 02] A. Vernet. (2002) XML transformation languages. http://www.scdi.org/~avernet/misc/xml-transformation.

[Su 01] H. Su, H. Kuno, E.A. Rundensteiner. (2001) Automating the transformation of XML Documents. Proceeding of the ACM Symposium On Document Engineering.

[Milo 98] T. Milo and S. Zohar. Using Schema Matching to Simplify Heterogeneous Data Translation. In Proc. Of the Int'l Conf VLDB' 98, pp. 122-133, 1998.

[Popa 02] L. Popa, Y. Velegrakis, R. J. Miller, M. A. Hernandez, and R. Fagin. Translating Web Data. In Proc. VLDB' 02, pp. 598-609, 2002.

[Miller 00] R. J. Miller, L. M. Hass and M. A. Hernandez. Schema Mapping as Query Discovery. In Proc. VLBD'00, pp. 77-88, 2000.

[Rahm 01] E. Rahm and P.A. Bernstein. (2001). On Matching Schemas Automatically, Technical Report, Dep. Of Comp science, Univ of Leipzig.

[Doan 01] A. Doan, P. Domingos, A.Y. Halevy (2001). Reconciling schemas of disparate data sources: A machine learning approach. In SIGMOD'01.

[Madhavan 01] J. Madhavan, P.A. Bernstein and E. Rahm (2001). Generic Schema matching with Cupid. Proc. 27th Int. Conf. On Very Large Data Bases (VLDB).

[Boukottaya 04] A. Boukottaya, C. Vanoirbeek, F. Paganelli, O. Abou Khaled Automating XML document Transformations: A conceptual modelling based approach. The First Asia-Pacific Conference on Conceptual Modelling Dunedin, New Zealand, January, 2004.

[Hardt 02] M. G. Hardt. Querying concepts- An approach to retrieve XML data by means of their data types. WLP - Workshop Logische Programmierung, TU Dresden, 2002.

[Feng 02] L. Feng, E. Chang, and T. Dillon (2002). A Semantic Network- Based Design Methodology for XML Documents. ACM Transactions on Information Systems (TOIS) Volume 20 , Issue 4, pp. 390-421.

[Galil 86] Z. Galil. Efficient algorithms for finding maximum matching in graphs. ACM Computing Surveys, pp. 23-38, 1986.

[Ullmann 76] J.R. Ullmann. An algorithm for subgraph isomorphism. Journal of the Association for Computing Machinery, pp. 31-42, 1976.

[Gold 96] S. Gold, A. Rangarajan. A graduated assignement algorithm for graph matching. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 377-388, 1996.

[Miller 95] A.G. Miller (1995). WordNet: A lexical Database for English. ACM 38 (11), pp. 39-41.

[Hirst 98] "Lexical chains as representations of context for the detection and correction of malapropisms". In: Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press, 1998.

[Xu 03] Li Xu, David W. Embley: Discovering Direct and Indirect Matches for Schema Elements. DASFAA 2003: 39-46.

Annexe : (*publication.xsd*)

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema" elementFormDefault="qualified"
attributeFormDefault="unqualified">
<xs:element name="university" type="UNIVERSITY"/>
<xs:complexType name="UNIVERSITY">
    <xs:sequence>
      <xs:element name="address" type="ADDRESS"/>
      <xs:element name="researcher" type="RESEARCHER"/>
    </xs:sequence>
        <xs:attribute name="name" type="xs:string" use="required"/>
    </xs:complexType>
    <xs:complexType name="ADDRESS">
      <xs:sequence>
         <xs:element name="city" type="xs:string"/>
         <xs:element name="state" type="xs:string"/>
         <xs:element name="zip" type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
    <xs:complexType name="RESEARCHER">
      <xs:sequence>
         <xs:element name="Firstname" type="xs:string"/>
         <xs:element name="Lastname" type="xs:string"/>
         <xs:element name="publication" type="PUBLICATION" abstract="true" maxOccurs
          ="unbounded"/>
      </xs:sequence>
    </xs:complexType>
    <xs:complexType name="PUBLICATION" abstract="true">
      <xs:sequence>
         <xs:element name="title" type="xs:string"/>
      </xs:sequence>
    </xs:complexType>
    <xs:complexType name="BOOK">
      <xs:complexContent>
         <xs:extension base="PUBLICATION">
             <xs:sequence>
                 <xs:element name="isbn" type="xs:string"/>
                 <xs:element name="publisher" type="xs:string"/>
             </xs:sequence>
         </xs:extension>
      </xs:complexContent>
    </xs:complexType>
    <xs:complexType name="ARTICLE" abstract="true">
      <xs:complexContent>
         <xs:extension base="PUBLICATION">
             <xs:sequence>
                 <xs:element name="abstract" type="xs:string" minOccurs="0"/>
                 <xs:element name="uri" type="xs:string" minOccurs="0"/>
             </xs:sequence>
         </xs:extension>
      </xs:complexContent>
    </xs:complexType>
    <xs:complexType name="JOURNALARTICLE">
      <xs:complexContent>
         <xs:extension base="ARTICLE">
             <xs:sequence>
                 <xs:element name="editor" type="xs:string" minOccurs="0"/>
             </xs:sequence>
         </xs:extension>
      </xs:complexContent>
    </xs:complexType>
    <xs:complexType name="PROCEEDINGARTICLE">
      <xs:complexContent>
         <xs:extension base="ARTICLE">
             <xs:sequence>
                 <xs:element name="volume" type="xs:string" minOccurs="0"/>
             </xs:sequence>
         </xs:extension>
      </xs:complexContent>
    </xs:complexType>
</xs:schema>
```

Session 2

Analyse textuelle

Pour une recherche semi automatisée des topoï narratifs

Greg Lessard[1], Stéfan Sinclair[2], Max Vernet[1], François Rouget[1], Elisabeth Zawisza[1], Émile Fromet de Rosnay[1], Élise Blumet[1]

[1] *Etudes françaises, Queen's University*
Kingston, Ontario, K7L 3N6, Canada

{lessard,vernetm,zawiszae}@qsilver.queensu.ca
{0drlef,2eb19}@qlink.queensu.ca ; fr2@post.queensu.ca

[2] *Humanities Computing, University of Alberta*
Edmonton, AB T6G 2E6, Canada

Stefan.Sinclair@UAlberta.ca

Résumé :

Les lecteurs humains cultivés sont capables de reconnaître dans les textes narratifs l'existence d'un ensemble de *topoï romanesques* : de courtes séquences textuelles qui décrivent une action qui se répète plusieurs fois au long du corpus littéraire français. Nous décrivons un projet (TopoSCan) qui crée des textes balisés en XML où sont captés les topoï identifiés par une équipe de lecteurs. L'analyse de ces textes balisés permet d'identifier des caractéristiques morphosyntaxiques et lexicales utilisées par la suite pour élaborer des critères algorithmiques de recherche de topoï. Les exemples trouvés par la machine sur la base de ces algorithmes sont par la suite proposés à un lecteur expert pour validation. Ceux qui sont acceptés alimentent la classe des exemples et aident à raffiner l'algorithme de recherche. Lecteur humain et machine fonctionnent ainsi en symbiose.

Mot-clés : topoï narratifs, analyse informatisée de textes, systèmes-experts, littérature française.

Abstract:

Cultivated human readers are capable of recognizing in narrative texts the existence of a set of *narrative topoï*: short textual sequences describing an action which is repeated several times throughout the French literary corpus. We describe here a project (TopoSCan) which creates texts marked up in XML in which are captured all the topoï identified by a team of readers. The analysis of these marked-up texts allows us to identify morpho-syntactic and lexical characteristics which permit the elaboration of algorithmic criteria for discovering potential new instances of topoï. The examples found by the machine through use of these algorithms are presented to an expert reader for validation. Those which are accepted enrich the class of examples and help in further refining the search algorithm. In this way, human and machine function in a symbiotic relationship.

Keywords: narrative topoï, computer-aided text analysis, expert systems, French literature.

1. Introduction

Depuis quelques années, les progrès en informatique ont fait apparaître un ensemble d'outils (comme les moteurs de recherche) qui facilitent la navigation dans les masses d'information qui nous entourent. Cependant, l'aide apportée par l'informatique se manifeste de façon inégale. Les outils traditionnels fonctionnent relativement bien dans une perspective *sémasiologique*, là où il est question d'identifier les documents où figurent telle ou telle forme ou ensemble de formes. Par contre, leur utilité est limitée dès qu'on envisage les données dans une perspective *onomasiologique*, c'est-à-dire, là où on recherche les formes et par conséquent les documents qui partagent un même concept. En même temps, la *granularité* de la plupart des travaux reste le plus souvent au niveau des documents entiers : l'être humain doit par la suite retrouver les éléments particuliers d'un document qui l'intéressent.

Bien entendu, des chercheurs de plusieurs sortes se sont déjà penchés sur ces questions: qu'on pense aux travaux qui visent la production automatique de résumés ou la recherche des patrons sous-jacents (l'exploration de données). Cependant, notre perspective sur ces questions est un peu spéciale: nous nous intéressons aux **topoï narratifs** dans un corpus littéraire français pour la période entre 1200 et 1800[1]. Nous désignons ainsi de courtes séquences textuelles (entre une phrase et un paragraphe) récurrentes (dans la pratique, qui reviendraient au moins trois fois dans

[1] Bien qu'il porte sur l'anglais, et malgré sa visée surtout syntaxique, le projet PROPBank a des ressemblances avec le nôtre. Pour plus de détails, voir le texte 'From Treebank to PropBank' (http://www.cis.upenn.edu/~ace/LREC02-propbank.pdf).

le corpus) envisagées non pas du point de vue formel, mais du point de vue de leur contenu. Un exemple : tout au long de la littérature française, on peut relever des passages où un père donne des conseils à son fils. Les formes utilisées pour décrire cette situation peuvent varier d'un texte à l'autre, mais un lecteur cultivé (c'est-à-dire, ayant lu et réfléchi sur ses lectures) serait en mesure de se prononcer (non sans hésitation, il est vrai, mais avec une certaine assurance) sur l'appartenance de telle ou telle séquence textuelle à la classe en question. Bien entendu, la classe des topoï narratifs est floue ; il faudrait la distinguer des thèmes (qui s'étendent le plus souvent sur des textes entiers) et aussi des items lexicaux isolés[2]. D'après notre usage, un topoï narratif se laisse représenter par une relation fonctionnelle entre un ensemble d'arguments. Pour reprendre l'exemple du père et du fils, la structure de base serait du genre PERE[agent] CONSEILLER[action] FILS[patient][3].

Nos travaux (c'est-à-dire, ceux de l'équipe TopoSCan) ont leur origine dans ceux de la SATOR (la Société d'Analyse de la Topique Romanesque) qui depuis 1986 tient des colloques et mène des recherches dans ce domaine, entre autres des études détaillées sur divers topoï narratifs, et la production d'une banque d'occurrences de topoï (www.satorbase.org[4]). L'équipe TopoSCan a ajouté deux aspects fondamentaux à la perspective satorienne. D'un côté, la banque des topoï produite par la SATOR ne contient que des exemples isolés (c'est-à-dire, des séquences sorties de leur contexte). D'un autre, les méthodes artisanales utilisées avaient comme conséquence que les travaux de la SATOR étaient lents et assez coûteux [SINCL05]. L'équipe TopoSCan a donc pris la décision de s'intéresser aux textes entiers et de tenter de baliser toutes les occurrences de topoï narratifs dans chaque texte. En même temps, il a été décidé que la seule solution raisonnable consistait à définir un partenariat entre chercheurs et logiciels. Nous avons donc adopté une variante de l'approche classique des systèmes-experts: un spécialiste se sert de ses connaissances et de son intuition pour relever dans une oeuvre romanesque toutes les occurrences de topoï qu'il arrive à déceler. Par la suite, une version numérisée du texte à l'étude est balisée dans un métalangage de type XML afin d'identifier chaque occurrence et d'y rattacher un identificateur métalinguistique[5]. Dans une troisième étape, l'analyse linguistique des occurrences donne lieu à un ensemble de constatations sur la structure formelle des topoï

---

[2] La SATOR (Société d'Analyse de la Topique Romanesque) a déjà réfléchi à ce problème lors d'une demi-journée théorique du colloque de Montpellier en 1997. Voir les Actes de ce colloque: *Homo narrativus*, Montpellier, Presses de l'Université Paul-Valéry, 2001. Voir aussi <http://alor.univ-montp3.fr/SATOR/programme.html#objet> pour une tentative de définition du topos et des notions voisines.

[3] Dans ce qui suit, les éléments de topoï seront représentés en lettres majuscules.

[4] La base de données a été développée par un des auteurs (Sinclair).

[5] Le format des documents est une version simplifié du XHTML avec des extensions pour les données topiques. Nous avons choisi ce format pour simplifier la tâche des chercheurs littéraires (nous estimions que l'apprentissage d'autres systèmes, comme le Text Encoding Initiative (TEI) était trop onéreux). Le schéma (W3C XML Schema) est disponible à partir du site web http://tapor.ualberta.ca/TopoSCan/

narratifs en général et sur les éléments lexicaux habituellement associés au topoï en général ou en particulier. Formalisées et entrées dans un logiciel de recherche, ces constatations nous permettent de faire lire par la machine d'autres textes non encore balisés afin de proposer à un lecteur expert de nouveaux candidats pour inclusion dans la banque des occurrences. Ceux de ces exemples qui sont acceptés par le lecteur sont versés dans les patrons de recherche qui sont ainsi renforcés, et ainsi de suite, en boule de neige. Par contre, les candidats proposés par la machine mais refusés par le lecteur humain nous incitent à revoir nos patrons. Ainsi, la machine joue un rôle de prothèse à l'intelligence humaine. Dans ce qui suit, nous présenterons les problèmes posés par la recherche des topoï, nous fournirons des précisions sur le métalangage utilisé et une description des caractéristiques du corpus, et nous décrirons les outils informatiques que nous utilisons, ainsi que quelques-uns des résultats que nous avons obtenus.

Derrière nos travaux il y a une hypothèse de base. Contrairement à ceux qui considèrent que l'identification des points de convergence entre deux textes sera réservée aux seuls lecteurs spécialistes, ou qu'elle sera possible seulement au terme d'une analyse sémantique profonde, nous croyons qu'il existe une relation solide et directe entre forme et signification et qu'une spécification formelle suffisamment détaillée nous permettra d'identifier au moins une bonne partie des variantes d'au moins un certain nombre de topoï narratifs. C'est cette hypothèse que nous essayerons de tester dans ce qui suit. Nous commencerons par présenter le métalangage que nous avons adopté.

2. Le métalangage

Les topoï narratifs peuvent être conçus comme des actions qui impliquent un ou plusieurs acteurs et éventuellement quelques précisions sur les circonstances. Nous avons donc décidé de les représenter au moyen d'une grille fonctionnelle en format XML. L'exemple suivant, tiré de la version balisée du roman *Manon Lescaut* de l'abbé Prévost[6], illustre le résultat. Dans ce qui suit, le texte du roman est en italique, les balises se trouvent entre les crochets habituels, et les indications topiques sont en gras.

> *Comme je n'en avais pas à perdre, je repris la parole pour lui dire que j'étais fort touché de toutes ses bontés, mais que, la liberté étant le plus cher de tous les biens, surtout pour moi à qui on la ravissait injustement, j'étais résolu de me la procurer cette nuit même, à quelque prix que ce fût ;*

[6] Le titre complet du roman, écrit en 1731 par l'abbé Prévost, est : *Manon Lescaut (Histoire du chevalier des Grieux et de Manon Lescaut)*. Le texte est tiré des *Mémoires d'un homme de qualité*, t. 7. Nous nous sommes servis de la version électronique du texte disponible sur le site ABU (http://abu.cnam.fr/)

<zone>
<infos>
<contexte/>
<descripteur id="d229" etat="acceptable">
<agent>**héros**</agent>
<action>**menacer**</action>
<patient>**personnage**</patient>
<circonstance>**avec arme pour évasion** </circonstance>
<date>**2004-01-31**</date>
<chercheur>**TopoSCan**</chercheur>
<notes>**évasion 2/4**</notes>
</descripteur>
</infos>
<cotexte>
et de peur qu'il ne lui prît envie d'élever la voix pour appeler du secours,
<citation ref="d229">
je lui fis voir une honnête raison de silence, que je tenais sous mon juste- au-corps. Un pistolet !
</citation>
me dit-il. Quoi ! mon fils, vous voulez m'ôter la vie, pour reconnaître la considération que j'ai eue pour vous ?
</cotexte>
</zone>

A Dieu ne plaise, lui répondis-je. Vous avez trop d'esprit et de raison pour me mettre dans cette nécessité; mais je veux être libre, et j'y suis si résolu que, si mon projet manque par votre faute, c'est fait de vous absolument.

On remarquera que chaque occurrence d'un topos se trouve à l'intérieur d'une <**zone**>, divisée elle-même en deux sous-ensembles : d'abord les <**infos**> (c'est-à-dire, les précisions éventuelles sur le <**contexte**> – la place qu'occupe la séquence dans le plan du roman ; les descripteurs narratifs – <**agent**>, <**action**>, <**thème**>, <**patient**>, <**circonstance**> ; l'identité du chercheur; la date du travail; des notes) et ensuite la <**citation**> (le noyau textuel qui véhicule le topos, entouré de son <**cotexte**>, c'est-à-dire la séquence textuelle plus large autour de la citation). La structure de certains des champs est figée (p.ex. la date), d'autres champs sont libres (p.ex. les notes), et d'autres encore, comme la désignation des topoï, font l'objet d'un ensemble de conventions définies par l'équipe TopoSCan. Pour ce qui est de l'attribution du statut de topos à tel ou tel passage, nous avons adopté comme principe de travail la lecture d'un texte par au moins un spécialiste littéraire, suivi d'une contre lecture par au moins un autre spécialiste et discussion des cas de désaccord. Les problèmes de définition et les questions de principe sont soumis à

toute l'équipe en réunion, de même que les questions soulevées par le balisage des textes[7].

Une fois un texte entier balisé de cette façon, il est possible de l'interroger[8] afin de sortir des désignateurs narratifs particuliers (p.ex. toutes les citations où l'agent a la valeur HEROS) ou des ensembles (p.ex. tous les contextes rattachés à la séquence HEROS MENACER PERSONNAGE). Entre autres choses, cela rend possible l'analyse détaillée des unités lexicales et des séquences associées à tel ou tel topos. Pour reprendre l'exemple ci-dessus, on constate que le verbe *menacer* n'est nulle part présent; par contre, il est clair que la séquence *raison de silence* et le mot *pistolet* sont associés au concept de menace.

3. Le Corpus

Dans ce qui suit, nous baserons notre analyse sur deux textes balisés par notre équipe : *Histoire du chevalier des Grieux et de Manon Lescaut* de Prévost (1731), désigné par la suite au moyen de l'étiquette **Manon**, et *Mémoires du comte de Comminges* de Mme de Tencin (1735)[9], désigné par la suite **Comminges**. Du point de vue de la forme et des thèmes abordés, les deux ouvrages relèvent de la même écriture romanesque, en vogue entre 1700 et 1750. Tous les deux sont écrits sous forme de *Mémoires* qui ont pour objet la peinture et l'analyse des sentiments. Divers procédés topiques (textuels et paratextuels) de la nouvelle convention mettent en place le personnage principal qui raconte lui-même sa vie. L'intérêt de sa narration, et de la formule, repose sur la différence entre le présent du récit et le passé du vécu, sur un dédoublement du je qui s'observe et s'écrit après coup, sur une perspective subjective unique de l'individu qui s'analyse. Exploitée à fond par Prévost et Tencin, la technique de *Mémoires* crée ainsi l'illusion du vrai qui permet au public des Lumières d'y retrouver ses préoccupations psychologiques et matérielles. Plusieurs topoï, révélateurs de l'écriture de l'intime en ascension à l'époque, peignent dans les deux romans les heurs et les malheurs des gens sensibles saisis sur le fond de leurs rapports familiaux et sociaux.

Les ressemblances entre les deux textes se manifestent dans l'étiquetage proposé par les lecteurs, comme l'illustrent les listes suivantes qui reproduisent par ordre de fréquence décroissante, les agents, actions, thèmes et patients les plus fréquents.

---

[7] Nous avons développé une transformation (XSLT) qui permet d'isoler les endroits problématiques dans les textes XML - il s'agit là d'un autre avantage de travailler dans XML.

[8] Pour interroger nos fichiers nous nous servons le plus souvent des fonctions de XPath offertes par XMLSpy, l'éditeur XML principal du projet; pour des renseignements sur Xpath, voir <http://www.w3.org/TR/xpath>.

[9] Pour nos travaux, nous avons utilisé une version électronique du texte tirée de la banque Gallica (http://gallica.bnf.fr). L'édition originale était celle de Paris : d'Hautel, 1812.

Agents

Manon : personnage (202), héros (95), amant (36), personnages (24), ami (24), amante (23), narrateur (22), père (17), héroïne (13), fils (8), amants (7), femme (5), apparences (5), amour (5), valet (4), frère (3), rival (2), prisonnière (2), étranger (2)...

Comminges : héros (56), amant (54), personnage (52), femme (28), père (19), mère (16), amante (16), domestique (15), personnages (13), mari (13), fils (11), amants (8), rival (6), amour (6), confident (5), cousins (4), blessure (3), religieux (2), parents (2), frères (2), fille (2), beau-frère (2), rencontre (1), pères (1)...

Actions

Manon : pleurer (23), demander (21), refuser (14), conseiller (13), révéler (12), donner (12), acheter (12), souper (11), offrir (11), promettre (9), trouver (8), secourir (7), se (7), fuir (7), chercher (7), cacher (7), arriver (7), reconnaître (6)...

Comminges : pleurer (16), écrire (14), cacher (14), conseiller (9), se (8), demander (8), révéler (7), voyager (6), secourir (5), refuser (5), informer (5), faire (5), tomber (4), rencontrer (4), reconnaître (4), promettre (4), passer (4), envoyer (4), enfermer (4), avouer (4), suivre (3), s'isoler (3), partir (3), offrir (3), observer (3), choisir (3)...

Thèmes

Manon : secours (19), lettre (15), infidélité (11), identité (10), services (9), argent (8), renseignements (6), rendez-vous (6), logement (6), sentiments (5), condition (5), vérité (4), secours (4), pardon (4), conseil (4), amour (4), secret (3), renseignement (3), récit (3), passion (3), mort (3), mariage (3), charmes (3), vengeance (2), souper (2), rupture (2), rencontre (2), prison (2), présence (2)...

Comminges : identité (10), sentiments (7), secours (6), amour (6), reproches (5), lettre (5), lettre (5), portrait (4), mariage (4), nuit (3), mariage (3), lettres (3), information (3), désobéissance (3), départ (3), son (2), procès (2), pardon (2), obéissance (2), mort (2), malheurs (2), maladie (2), lettre (2), informations (2), haine (2), fuite (2), autorité (2), vie (1), trahison (1), tombeau (1)...

Patients

Manon : ami (20), amante (13), fils (11), rival (5), père (4), héros (4), femme (4), amants (3), amant (3), personnage (2), maîtresse (2), homme (2), frère (2), autorité (2), voyageurs (1), voyageur (1), soeur (1), protecteur (1), prostituée (1)...

Comminges : fils (22), amante (14), maître (9), rival (7), amant (7), femme (6), domestique (6), père (3), mère (3), maîtresse (3), épouse (3), à (3), mari (2), malheureux (2), amant (2), adversaire (2), personnage (1), père (1)...

Malgré les caractéristiques qu'ils partagent, les deux textes présentent un certain nombre de différences. Le texte de *Manon* est environ deux fois plus long que *Comminges*. Par contre, les phrases contenues dans *Comminges* sont en moyenne deux fois plus longues que celles contenues dans *Manon*. Par conséquent, chaque phrase de *Comminges* a plus de chances de contenir un élément topique. En effet, comme l'illustre le tableau suivant, plus de la moitié des phrases dans *Comminges* ont été vues comme topiques par nos lecteurs, contre le cinquième dans le cas de *Manon*.

Attribution par lecteurs	*Manon*		*Comminges*	
	N	*%*	*N*	*%*
Phrases jugées comme topiques	762	22 %	345	51 %
Phrases jugées non topiques	2 752	78 %	325	49 %
Total	3 514	100 %	670	100 %

4. Outils informatiques

Une fois qu'ils ont été lus et relus par les experts littéraires, les textes à étudier sont balisés au moyen de l'éditeur XMLSPY. L'analyse préliminaire et les recherches essentielles sur les textes balisés sont menées au moyen des outils Unix traditionnels (surtout AWK, uniq, sort) et de l'éditeur IVI[10]. Par la suite, l'essentiel du travail de navigation et d'analyse est assuré par un outil spécialisé (TopoDétecte) conçu par un membre de notre équipe (S. Sinclair) sur la base d'un outil antérieur, HyperPo [SINCL03]. Comme c'est le cas pour son prédécesseur, l'interface de TopoDétecte passe par un navigateur (Mozilla, etc.) qui interroge un serveur où est fait l'essentiel du travail d'analyse. Le logiciel lit un texte balisé en XML et le transforme en une base de données dans laquelle chaque unité lexicale est accompagnée d'une gamme d'informations, y compris sa position dans le texte, la position des autres occurrences de la même forme, sa fréquence relative, etc. Pour les besoins de nos recherches, les trois éléments principaux du logiciel sont les suivants :

[10] Pour plus de détails, voir http://www.cs.queensu.ca/CompLing/.

a) Un ensemble de menus dynamiques qui permettent au chercheur d'attribuer une valeur positive ou négative à certains traits formels, y compris :

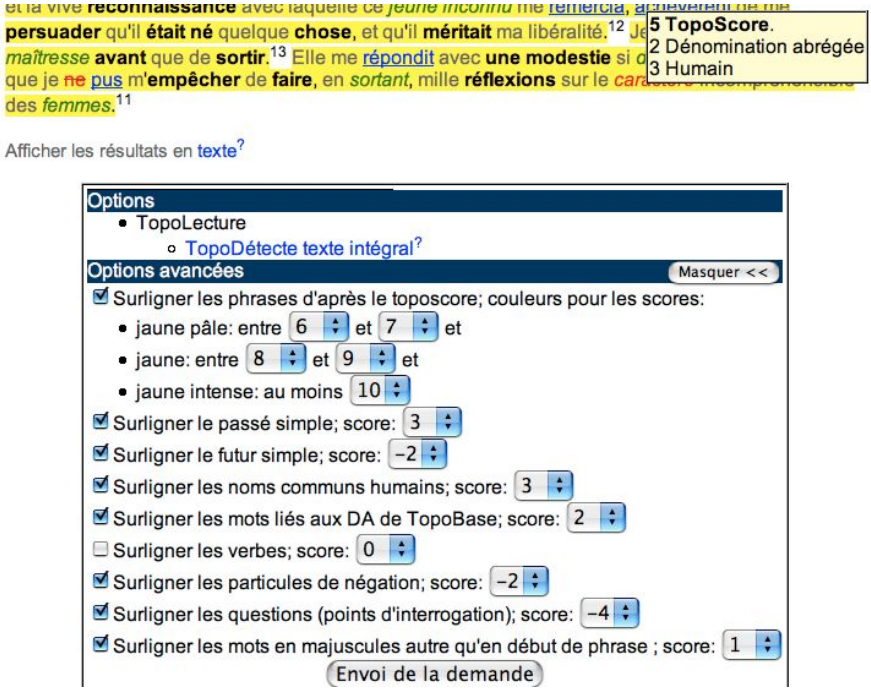
   - la présence des différents temps verbaux, en particulier le passé simple et le futur simple (l'identification de ces formes est assurée par la consultation d'une liste établie par J. Véronis[11]) ;
   - la présence d'un nom humain (l'identification de ces formes est assurée par la consultation d'une liste inédite produite par G. Lessard et M.E. Surridge) ;
   - la présence d'un nom propre (identifié par la présence d'une lettre majuscule ailleurs qu'en début de phrase) ;
   - la présence de la forme négative *ne* ;
   - la présence d'un marqueur interrogatif (le point d'interrogation).

b) L'accès à divers lexiques topogènes : c'est-à-dire des listes de formes lexicales associées à la présence d'un topos narratif. Nos lexiques topogènes ont été constitués de diverses façons (voir ci-dessous).

c) Un système d'affichage qui lit un texte balisé, phrase par phrase, et pour chaque mot de la phrase, propose une valeur positive ou négative selon les critères énumérés ci-dessus en a) et b), et pour chaque phrase, une valeur totale basée sur la valeur de chaque mot de la phrase, pondérée par la longueur de la phrase. Les phrases ayant une valeur au-dessus d'un seuil fixé par le chercheur seront surlignées en couleur. Dans nos travaux, ce seuil est établi en fonction du taux d'occurrences de topoï identifiées par le lecteur humain.

La figure suivante illustre le menu de sélection au premier plan, et en arrière-plan, quelques lignes d'un texte affichées par le navigateur :

[11] Pour plus de détails, voir http://www.up.univ-mrs.fr/~veronis/donnees/.



5. Analyse et résultats préliminaires

Dans cette section, nous présenterons et discuterons quelques expériences que nous avons menées pour mesurer l'influence de différents facteurs sur le taux d'accord entre les topoï identifiés « à la main » par un lecteur expert, et ceux proposés par la machine. Ces travaux ont été rendus possibles par l'existence de TopoCompare, une variante du logiciel de base décrit en 4 : là où un lecteur humain aura déjà balisé un texte, ce logiciel nous permet de fixer le poids associé à différents paramètres, de confronter les citations proposées par le lecteur humain et celles proposées par la machine, de calculer le taux d'accord ou de désaccord entre les deux, et d'afficher les points de convergence et de divergence. Dans ce qui suit, nous considérerons comme réussites tous les cas où une phrase retenue comme topique par la machine (c'est-à-dire, cotée au-dessus du score préétabli au début de l'expérience) englobe une citation proposée par le lecteur humain (attribution positive correcte), ainsi que toutes les phrases considérées comme non topiques et par le lecteur humain et par la machine (attribution négative correcte). Seront considérés comme échecs tous les cas où la machine propose un candidat non retenu par le lecteur humain (attribution positive incorrecte), ainsi que tous les cas où la machine ne retient pas une phrase comprenant une citation proposée par l'être

humain (attribution négative incorrecte). Il est à noter que ce test est relativement exigeant dans la mesure où il écarte la possibilité qu'une occurrence proposée par la machine représente un passage que les lecteurs humains auraient dû baliser.

Dans ce qui suit, nous examinerons trois sortes de critères utilisés pour identifier des topoï. Certains sont relativement abstraits, d'autres plus spécifiques. Nous nous pencherons en particulier sur le taux d'efficacité des différents critères, ainsi que sur leur applicabilité à plus d'un texte.

5.1 Un test basé sur des critères morphosyntaxiques et lexicaux généraux

Au niveau le plus général, l'analyse globale d'un ensemble de textes et la lecture d'études narratologiques et linguistiques nous ont amenés à proposer un certain nombre de critères généraux qui à notre avis seraient associés à la présence de topoï narratifs. Entre autres : (a) dans les textes narratifs traditionnels, c'est le passé simple qui véhicule les événements d'une histoire, tandis que l'imparfait a tendance à présenter des descriptions et d'autres éléments de deuxième plan ; le présent se trouve surtout dans les dialogues et dans les constatations générales (bien que le présent dit historique complique ce tableau), et d'autres temps comme le futur se trouvent ailleurs ; (b) la présence d'une forme interrogative diminue en général les chances d'être devant un topos puisqu'il s'agit souvent d'une recherche d'informations plutôt qu'un événement narré ; (c) les agents topiques ont fortement tendance à appartenir à la classe des êtres humains (sauf dans les contes de fées, etc.) de sorte que la présence d'une unité lexicale humaine peut donc être considérée comme un indice positif. Dans un premier temps, nous avons utilisé ces constatations pour fixer les paramètres ainsi : passé simple = +3, nom humain = +3, futur simple = -3 ; nom propre = +1, forme négative = +1, point d'interrogation = -3. Appliqué aux versions balisées de *Manon Lescaut* et des *Mémoires du comte de Comminges*, cela donne les résultats suivants :

Attribution de statut topique par la machine (critères généraux)				
	Manon Lescaut		*Comminges*	
	N	%	N	%
Attribution positive correcte	306	9 %	219	33 %
Attribution positive incorrecte	611	17 %	140	21 %
Attribution négative correcte	2 141	61 %	186	28 %
Attribution négative incorrecte	456	13 %	126	19 %
	3 514	100 %	670	100 %
% d'attributions correctes	**70 %**		**60 %**	

On constate que ces critères donnent plus d'attributions positives correctes qu'incorrectes dans le cas de *Comminges* (33 % contre 21 %) mais que le taux d'efficacité est nettement plus bas dans le cas de *Manon*. Par contre, la proportion d'attributions négatives correctes est supérieure dans le cas de *Manon* (61 % contre 13 %) que dans *Comminges*. Dans l'ensemble pourtant, ces quelques critères généraux fournissent un rendement intéressant (70 % des attributions sont correctes dans le cas de *Manon* et 60 % le sont dans le cas de *Comminges*.

5.2 Un test basé sur un lexique topique général

Par la suite, nous avons testé l'efficacité d'un autre critère général plus étroitement relié au domaine narratif. Les exemples compris dans la banque SatorBase comprennent un ensemble de descripteurs narratifs du genre : AMOUR REND CREDULE, COURTISER PAR RUSE, GARDER SECRET PAR HONTE. Au total, ces 1 023 descripteurs se répartissent entre 1 253 unités lexicales distinctes. Ensemble, ces formes représentent en quelque sorte une liste préliminaire des concepts fréquents dans les topoï narratifs en général. La liste de ces formes a été enrichie par l'utilisation de EuroWordnet[12] , la version française de WordNet, pour obtenir les synonymes, hyponymes et hyperonymes des formes de base. Cet ensemble lexical fournit un patron de recherche complexe que nous avons appliqué aux deux textes à l'étude ici, sans retenir les autres facteurs généraux énumérés ci-dessus. Plus précisément, nous avons attribué une valeur positive de +5 à toutes les unités lexicales d'une phrase identiques au lexique topique général que nous venons de décrire. Appliqué aux deux textes, ce critère donne les résultats suivants :

Attribution de statut topique par la machine (critères lexicaux généraux)				
	Manon Lescaut		*Comminges*	
	N	%	N	%
Attribution positive correcte	226	6 %	209	31 %
Attribution positive incorrecte	512	15 %	138	21 %
Attribution négative correcte	2 240	64 %	187	28 %
Attribution négative incorrecte	536	15 %	136	20 %
	3514	100 %	670	100 %
% d'attributions correctes	**70 %**		**59 %**	

[12] Pour des renseignements sur EuroWordnet, voir http://www.elda.fr/catalogue/en/text/M0015.html.

Les résultats obtenus au moyen de ce critère lexical général sont très proches de ceux obtenus au moyen des critères formels généraux comme on peut le constater en comparant les deux tableaux. Pourtant, nous avons constaté que le mécanisme qui établit les liens sémantiques est sans doute trop général et qu'une version plus raffinée fournirait de meilleurs résultats.

5.3 Un test basé sur le vocabulaire topique de Manon Lescaut

Le lexique topogène utilisé dans l'expérience précédente est basé sur une gamme importante de textes. Il manque donc de précision. Pour tester une autre approche, nous avons utilisé des outils Unix pour établir la liste des formes qui figurent à l'intérieur et à l'extérieur des balises <citation> et </citation> dans le texte de *Manon Lescaut*. Les résultats peuvent être divisées en trois classes, qui épuisent les 5 947 formes comprises dans le roman :

- Les formes qui figurent à l'intérieur d'une citation, mais non pas à l'extérieur : 620 formes ou 10 % du total, y compris des mots intuitivement associés à la thématique du roman à l'étude (*abandon, amoureusement, désespéraient, gentilhomme, languissaient, noblement, pleurâmes, pleurait, racontai, regrettais, séparâmes, toucha, troubler*) mais aussi des formes communes dont la présence uniquement dans une citation semble surtout le fruit du hasard et des dimensions réduites du corpus (*associer, boue, choisi, conduis, finis, immense, menton, pont*) ;
- Celles qui figurent à l'extérieur d'une citation, mais non pas à l'intérieur : 3 362 formes, ou 57 % du total, y compris des formes générales (*acheter, ajuster, avenir, dépourvu, entiers, façon, irai, mena*) mais aussi des formes spécifiques à l'un ou l'autre des topoï romanesques (*adorais, agitation, émotion, baiser, caresse, folie, languissais, pleurez*) ;
- Celles qui figurent dans au moins une citation et également à l'extérieur des citations : 1 965 formes, ou 33 % du total, y compris (*affligé, aimable, baisers* (mais non pas *baiser*), *choses, déguisement, elle, faveurs, hélas, libertinage, offrir*).

Nous avons analysé nos deux textes au moyen de ce lexique spécifique. Plus précisément, chaque occurrence d'une forme spécifique aux citations qui apparaît dans une phrase candidate reçoit une pondération de +5. Aucun autre critère n'est utilisé. Appliqué aux deux romans, cela donne les résultats suivants :

Attribution de statut topique par la machine (vocabulaire de citations de Manon)				
	Manon Lescaut		*Comminges*	
	N	%	N	%
Attribution positive correcte	323	9 %	77	11 %
Attribution positive incorrecte	85	2 %	46	7 %
Attribution négative correcte	2 667	76 %	279	42 %
Attribution négative incorrecte	439	12 %	268	40 %
	3 514	100 %	670	100 %
% d'attributions correctes	**85 %**		**53 %**	

Compte tenu de l'origine de ce vocabulaire, le taux élevé d'attributions correctes dans le premier roman (85 %), le plus grand nombre d'attributions positives correctes qu'incorrectes (9 % contre 2 %) et le petit nombre d'attributions incorrectes (15 % en tout) ne devraient pas nous surprendre. Toutefois, cet avantage n'est pas facilement transféré à l'autre roman, où le taux d'attributions correctes dépasse à peine les 50 %, tout en restant positif et pour les attributions positives et pour les attributions négatives. Il reste que les facteurs lexicaux peuvent jouer un rôle significatif dans l'identification des topoï. En outre, puisque notre but est non seulement de faire proposer par la machine des candidats au statut général de topos sans égard au descripteur narratif, mais aussi d'identifier des exemples potentiels de topoï spécifiques, ce facteur lexical est essentiel. Pour mieux cerner les enjeux à ce niveau, nous avons mené une expérience complémentaire, qui consistait à partir des formes lexicales qui figurent à l'intérieur des citations porteuses d'un topoï narratif spécifique (toujours dans la version balisée de *Manon Lescaut*), afin de mieux déceler leurs caractéristiques. Pour réduire la classe des formes, nous avons retenu seulement les noms, verbes et adjectifs. Or, il devient vite évident que les topoï présentent une grande diversité en ce qui concerne la classe des formes qui les véhiculent. Dans certains cas, un petit nombre de formes revient souvent, constituant ainsi un noyau dur et précis. C'est le cas, par exemple, du topos narratif PLEURER, qui présente 23 occurrences dans le roman et qui est caractérisé par la présence relativement fréquente des formes *pleur.*, larme.**. Plus précisément, on retrouve dans la liste : *larmes (13 occ.), pleurs (4), versant (3), tomber (3), ai (2), yeux (2) visage (2), ruisseau (2), quelques (2), point (2), pleurer (2), fois (2), couler (2), avez (2), amèrement (2), amour (2)* et plusieurs mots à une seule occurrence ayant un lien clair avec la notion de pleurer (*chagriner, malheureuse, pleurait, pleurâmes, pleurer, sanglotant*). Par contre, un topoï comme ACHETER (12 occurrences dans le roman) présente une grande diversité lexicale, de sorte qu'il est difficile d'y identifier un noyau lexical dur. Par fréquence décroissante, la liste comprend les

formes suivantes : *pistoles (4 occ.), dix (4), ont (3), louis (3), dis (3), quelque (2), peu (2), pas (2), or (2), mille (2), homme (2), est (2), bourse (2)*. On constate que peu de formes ont une relation unique avec le topos : *écu, offre, payer, sou*. Le défi consistera à déterminer l'effet de cette diversité sur l'efficacité relative des mécanismes de recherche.

5.4 Un test basé sur tous les critères utilisés ensemble

Enfin, nous avons combiné les trois ensembles de critères, ce qui donne les résultats suivants :

Attribution de statut topique par la machine (les trois critères ensemble)				
	Manon Lescaut		*Comminges*	
	N	%	N	%
Attribution positive correcte	231	7 %	207	31 %
Attribution positive incorrecte	503	14 %	136	20 %
Attribution négative correcte	2 249	64 %	189	28 %
Attribution négative incorrecte	531	15 %	138	21 %
	3 514	100 %	670	100 %
% d'attributions correctes	**71 %**		**59 %**	

Ensemble, les trois sortes de critères (les critères formels, le lexique topique général, et le lexique topique spécifique) fournissent des résultats raisonnables. Le taux d'attributions correctes se situe entre 59 % et 71 % et dans tous les cas sauf les attributions positives dans *Manon* on trouve plus de jugements corrects que de jugements incorrects. Par contre, on ne voit pas de gain important par rapport aux critères utilisés seuls. Il se peut que la repondération des différents critères ou peut-être l'application des critères les uns après les autres nous permettra d'améliorer nos résultats. C'est ce que nous travaillerons prochainement. Cependant, compte tenu du petit nombre de critères utilisés, nous sommes relativement satisfaits des résultats.

6. Conclusions

Le projet TopoSCan est en cours depuis deux ans. Les membres de l'équipe ont passé une bonne partie de cette période à établir les nécessaires ponts interdisciplinaires entre les perspectives littéraire, narratologique, linguistique, et informatique, à constituer un métalangage approprié, et à construire le formalisme XML et les logiciels nécessaires. Jusqu'à présent, nous avons balisé une dizaine de textes. C'est donc dire que les résultats présentés ici sont préliminaires et avant tout d'ordre méthodologique. Par contre, le fait que les topoï potentiels proposés par la machine ne manquent pas de cohérence et que la machine propose un taux de réussite bien au-delà de 50 % nous incite à penser que nous sommes dans la bonne voie.

En terminant, nous voudrions toutefois soulever deux questions, l'une méthodologique, l'autre épistémologique. D'abord la méthodologie : jusqu'à présent, nous avons exploré deux approches pour la détection des topoï : d'un côté, l'utilisation du contenu lexical d'un ensemble de citations, et de l'autre, l'utilisation d'un nombre limité de marques morphologiques et syntaxiques. Les deux approches partagent cependant un point faible commun : dans les deux cas, le critère est appliqué à un mot à la fois. Il n'y a aucune tentative pour tenir compte des relations entre les mots d'une phrase. Et pourtant, il est clair que les collocations représentent un puissant mécanisme d'individualisation. Un exemple parmi d'autres : si dans le passage reproduit dans la section 2 on prend la séquence *faire le silence* et le mot *pistolet*, et si on les insère dans le moteur de recherche Google, le résultat se limite à trois pages, dont deux fois le texte du roman *Manon Lescaut* et un article où est cité le passage même de la citation. À l'avenir, il nous faudra explorer davantage cette voie. À un autre niveau, il faudrait tenir compte des dépendances syntaxiques qui tournent autour des verbes, et surtout des classes de verbes qui partagent la même structure d'arguments. Des travaux antérieurs comme celui de Levin [LEV93] nous seront utiles dans cette perspective. Finalement, jusqu'à présent, nous avons pris la phrase comme horizon d'attente pour la détection des topoï ; or, l'examen du corpus montre clairement que les citations peuvent s'étendre sur plusieurs phrases ou se limiter à une partie de phrase. Il faudra à l'avenir modifier nos mesures pour tenir compte de ce fait.

Du point de vue épistémologique, il faudrait retenir un fait essentiel : la classe des topoï identifiés dans un texte par un lecteur humain (ou même par une équipe de lecteurs) est un ensemble ouvert et relativement flou qui dépend des dimensions du corpus, des différentes stratégies narratives, et des lectures antérieures de ces spécialistes ainsi que le contexte théorique dans lequel ils travaillent. Par conséquent, nous restons bien conscients que nous posons le pied dans un terrain marécageux. Par contre, les avantages sont certains : ou bien nous arriverons à produire un outil qui facilitera la recherche des topoï, ou bien nous constaterons que l'identification des topoï exige une analyse sémantique poussée ou qu'elle est tout simplement au-delà des capacités de l'informatique.

7. Références bibliographiques

[LEV93] Beth Levin, English *Verb Classes and Alternations: A Preliminary Investigation*, Chicago: University of Chicago Press, 1993.

[SINCL03] Stéfan Sinclair, "Computer-Assisted Reading: Reconceiving Text Analysis", *Literary and Linguistic Computing*, vol. 18, n° 2, 2003, pp. 175-184.

[SINCL05] Stéfan Sinclair, "*Trois arguments pour l'alimentation de SatorBase*", Étrange topos étranger, Greg Lessard, François Rouget et Max Vernet (ed.), Sainte-Foy : Presses de l'Université Laval [à paraître].

Expériences lexicométriques sur les cooccurrences

Jean-Marc Leblanc

CEDITEC – Université de Paris 12 Val de Marne
94 Créteil cedex - France
leblanc.jeanmarc@free.fr

Résumé :

À partir d'un corpus de discours institutionnels fortement ritualisés (messages de vœux des présidents de la 5e république, 1959-2001), diverses fonctions cooccurrentielles de Lexico3, Weblex, Hyperbase, Alceste sont successivement utilisées et comparées. Elles mettent en évidence, par les calculs, les tris, et les présentations graphiques des faits de cooccurrences différents, susceptibles d'interprétations complémentaires concernant l'ethos discursif de chaque président. La comparaison permet par ailleurs d'insister sur la nécessité d'une expérimentation approfondie dans le traitement des données textuelles avant toute entreprise herméneutique.

Mots-clés : cooccurrences, lexicogrammes, discours institutionnel.

Abstract:

The functional comparison of various approaches to cooccurrence in some robust standard textual analyzers makes it possible to develop an experimental method in the courses of textual data processing intended for researchers in social sciences, which is too often absent from the field. Using a corpus of highly ritualized institutional speeches (New year addresses of the presidents of the French 5th Republic, 1958-2003), various functions to measure cooccurrence with Lexico3, Weblex, Hyperbase, Alceste are used in turn and compared. Different cooccurrence data are thus highlighted through figures, sorting, and graphic display, each leading to specific interpretations. The comparison shows the absolute necessity of in-depth experimentation in the analysis of textual data before any hermeneutic venture.

Keywords: cooccurrences, lexicogrammes, political discourse.

1. Méthode expérimentale et corpus des vœux présidentiels

Le calcul des cooccurrences, ou mesure probabilisée des attirances entre formes dans un contexte donné, qui repose sur une notion bien décrite dans le domaine, a connu plusieurs types d'implémentations et de nombreuses exploitations ces dernières années. Mesure des cooccurrences binaires [Lafon, 1984], recherche des lexicogrammes, [Tournier, 2003, Heiden, 2003], voisinages [Labbé, 1990], cooccurrences des formes spécifiques [Salem, Martinez, 2003]. Le site Textopol[1] qui offre un accès ergonomique et initie à divers logiciels facilite une démarche expérimentale, cumulative et comparée visant à mettre en évidence les propriétés de chaque méthodologie.

Nous examinons ici le statut du *je* présidentiel réduit à sa forme graphique en explorant ses espaces cooccurrentiels dans les messages de vœux aux Français sous la Cinquième République (1959-2001). Nous mobilisons différents outils statistiques pour mettre en lumière les réseaux lexicaux mais aussi sémantiques et thématiques qui gravitent autour d'une marque personnelle structurante dans ce genre institutionnel. Les réseaux de cooccurrence, saisis par ces approches différentes apportent un éclairage sur la construction de l'ethos présidentiel [Amossy, 1999, Maingueneau, 2003] qui se manifeste au sein d'un genre de discours politique fortement codifié.

1.1 Les outils

Deux types d'outils sont mis en œuvre et éprouvés :

- Des Logiciels lexicométriques dits « classiques » (Lexico 3, Hyperbase). Travaillant sur la base d'un tableau lexical, après réorganisation de la séquence textuelle et segmentation en unités minimales (ici la forme graphique), ces outils introduisent la notion de partition, sur laquelle portent des analyses contrastives et des mesures de ventilation du stock lexical dans les sous parties du corpus (bornes chronologiques, locuteurs…). Les fonctions documentaires (concordances, contextes), statistiques (spécificités), analyses multidimensionnelles (Analyse factorielle des correspondances, arborées), constituent les fonctionnalités essentielles de ces outils.
- Des « Cooccurrenceurs » (Weblex, Alceste) où la mesure des voisinages joue un rôle essentiel.

Weblex présente des caractéristiques similaires aux logiciels de type lexicométrique mais intègre des fonctionnalités évoluées de recherche de cooccurrences reposant sur un modèle probabiliste. (Cooccurrences, lexicogrammes

[1] Le site Textopol est réalisé dans le cadre du Céditec (Ea 3119), Université de Paris 12 Val de Marne (http://textopol.free.fr).

simples et récursifs, associés ou non à une forme pôle dont le principe sera détaillé ultérieurement).

Dans la méthodologie Alceste dont la perspective diffère sensiblement, l'algorithme ne repose pas sur une segmentation pré-établie mais constitue des classes d'énoncés indépendamment des grandes divisions du corpus. Cette démarche inductive, fondée sur une analyse statistique distributionnelle met en évidence les grandes articulations du corpus, ses « mondes lexicaux », en classant les énoncés du texte en fonction de la distribution de leur vocabulaire. Le texte, considéré comme un ensemble d'énoncés est découpé en unités des contexte (plus ou moins la phrase). Le logiciel effectue un repérage des unités lexicales, identifiées au moyen d'un dictionnaire, puis procède à une lemmatisation. Les énoncés sont alors triés en fonction de la présence / absence des formes qui les composent puis classés selon la méthode de classification descendante hiérarchique. On obtient des classes de mots les plus représentatifs de ces énoncés, triés selon leur coefficient d'association à la classe par la méthode du khi2.

1.2 Les paramètres du corpus

Nous nous intéresserons aux interventions produites de décembre 1959 à décembre 2001 soit 43 discours représentant un volume textuel de 41 125 occurrences pour 5 203 formes. Ces discours, qui forment par ailleurs une série textuelle chronologique2 sont abordés ici dans leur dimension synchronique. Bien qu'étant attentifs, lorsqu'il s'agit de retourner au texte, aux emplois individuels, nous menons l'expérience en considérant le corpus dans son ensemble.

Locuteur	Nombre Discours	Nombre occurrences	Nombre formes	Longueur Moyenne
De Gaulle	10	11 498	2 407	1 150
Pompidou	5	2 850	890	570
Giscard	7	6 066	1 360	866
Mitterrand	14	11 991	2 521	856
Chirac	7	8 720	1 799	1 245

Tableau 1 : Principales caractéristiques de la partition locuteur

[2] Nous reprenons ici la définition qu'en donnent [Lebart, Salem 1997] : « Corpus homogènes constitués par des textes produits en des situations d'énonciation similaires, si possible par un même locuteur, individuel ou collectif, et présentant des caractéristiques lexicométriques comparables. » Autre caractéristique essentielle, l'étalement dans le temps permettant de mettre en évidence des variations chronologiques.

Les représentations graphiques produites ci-dessus mettent l'accent sur le relatif déséquilibre des sous parties. Le matériau est beaucoup plus riche pour de Gaulle et Mitterrand que pour les autres locuteurs, les cinq présidents n'ayant pas assumé leur charge sur les mêmes durées.

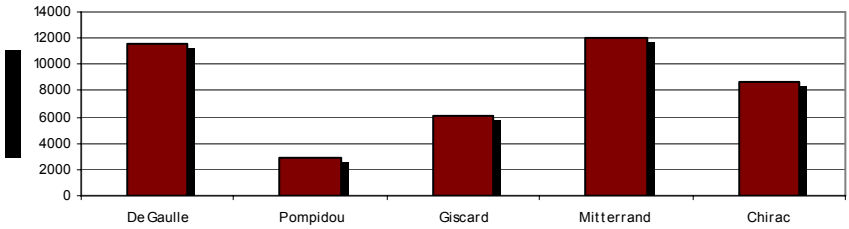


Tableau 2: Distribution des fréquences

En outre la longueur des messages est très variable selon les locuteurs. De Gaulle et Chirac consacrent en moyenne 1 150 et 1 245 occurrences à leurs discours tandis que Mitterrand et Chirac ne leur accordent que 856 et 866 occurrences. Les messages de Pompidou sont les plus brefs avec 570 occurrences. Ces données restent cependant comparables et autorisent des conclusions fiables.

1.3 Les marques énonciatives

Les spécificités[3] par partie des pronoms personnels et adjectifs possessifs mettent en lumière des profils énonciatifs très contrastés qui caractérisent ce discours pourtant très ritualisé (Tableau 3).

[3] La méthode des spécificités permet de porter un jugement sur la répartition des formes dans les parties d'un corpus. Ce jugement s'exprime en termes de sur-emploi (spécificité positive) et de sous-emploi (spécificité négative). Selon le modèle hypergéométrique, une forme est notée spécifiquement positive si sa fréquence dans une partie est supérieure à la fréquence théorique attendue, et spécifiquement négative si cette fréquence est inférieure au seuil retenu. Ces fréquences probabilisées s'appuient sur la comparaison de quatre données : le nombre des occurrences du corpus, le nombre des occurrences dans la partie, la fréquence de chaque forme dans le corpus, et la fréquence de chaque forme dans la partie. Les indices indiquent le degré de spécificité de chaque forme et représentent la valeur absolue de l'exposant de probabilité. Un exposant de valeur 2 exprime une probabilité de l'ordre du centième, 3 du millième… L'absence d'exposant indique que l'usage ne présente pas de caractéristique remarquable. On dira que la forme est banale pour la partie considérée.

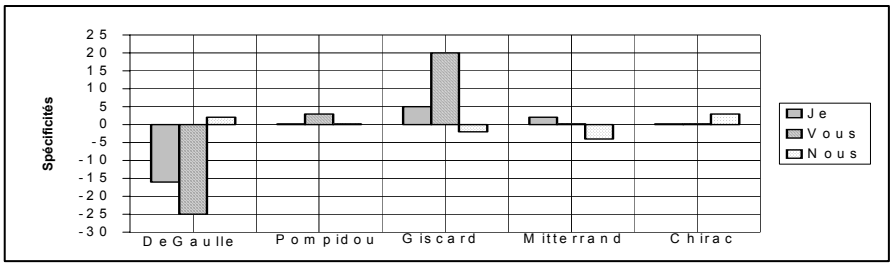


Tableau 3 : Histogramme des spécificités de je, vous, nous sur la partition locuteur.

À l'énonciation fortement personnalisée, centrée sur le *je* de Giscard (*je*, +5), multipliant aussi les marques en direction des Français (*vous*, +20), on oppose la prise de distance chez de Gaulle par le rejet des marques de la première personne du singulier (*je*, -16) et de la seconde du pluriel (*vous*, -25), la prise en charge de l'énoncé étant assurée par un *nous* dont le référent est la France (*nous*, +2).

	De Gaulle	**Pompidou**	**V.G.E.**	**Mitterrand**	**Chirac**
nous	+E02	-	-E02	-E04	+E03
je	-E16	-	+E05	+E02	-
j'	-E06	-	-	+E05	-
vous	-E25	+E03	+E20	-	-
on	-E02	-E03	-	+E10	-E03
notre	+E03	-	-E02	-E05	+E02
nos	+E02	-E03	-E04	-	+E03
mes	-E08	-	+E03	+E03	-
votre	-E09	+E03	+E11	-E02	-E02
vos	-E06	-	+E11	-E03	-
moi		-	-	+E02	-E02
me	-E02	-	+E02	+E02	-E03
m'	-E03	-	-	+E02	-

Tableau 4 : Spécificités des principaux pronoms personnels et adjectifs possessifs.

2. Approches des cooccurrences de JE

2.1 Topographie textuelle et cooccurrents spécifiques (Lexico 3)

L'outil carte des sections établit la distribution de la forme personnelle dans la linéarité du texte, délimité en paragraphes.

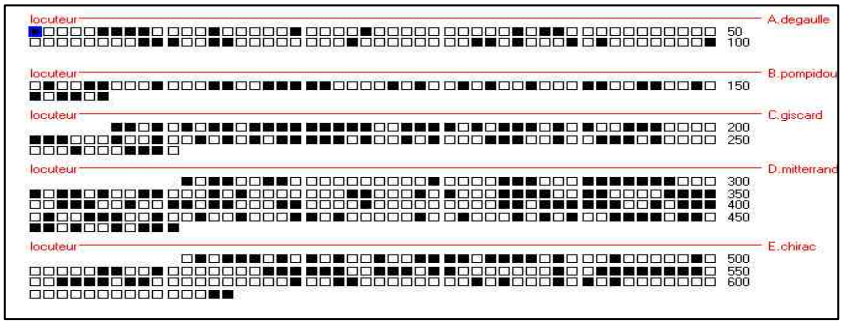


Tableau 5 : Carte des sections (paragraphe) de la forme JE sur la partition locuteur

Dans cette « topographie textuelle », [Lamalle, Salem, 2002] chaque rectangle du tableau 5 matérialise une section de texte, les unités colorées indiquant les paragraphes attestant au moins une fois la forme recherchée. Au moyen de cette cartographie, on appréhende des usages, des positionnements énonciatifs en termes de fréquences mais aussi de rythme, de cadence. Entre de Gaulle et Giscard par exemple, on note deux configurations : de longues successions de paragraphes contenant *je* chez Giscard, de brèves interruption, ou rares îlots chez De Gaulle.

Le calcul des cooccurrents spécifiques met en évidence à partir des sections délimitées par cette cartographie les formes spécifiques des paragraphes attestant le *je*. La liste produite porte sur l'ensemble du corpus et ne présente que les formes dont la valeur absolue de l'indice de spécificité est supérieure à 2. Les seuils sont de 5 %, la fréquence minimale des formes considérées est de deux occurrences. Ce calcul ne diffère pas du modèle de spécificité évoqué précédemment si ce n'est que les parties sur lesquelles porte la comparaison ne sont plus matérialisées par une partition en locuteurs mais constituées selon la présence ou l'absence du pronom personnel. Le diagnostic de spécificité est alors établi sur la base d'une partition binaire : l'ensemble des sections dans lesquelles la forme analysée est présente par rapport à l'ensemble du corpus. Les coefficients indiqués au tableau qui suit correspondent donc à des indices de spécificité. Une spécificité positive signifie qu'une forme considérée a tendance à apparaître de façon plus importante que le modèle théorique ne le laissait prévoir dans les contextes du pôle analysé, par rapport aux autres sections du corpus, une spécificité négative indiquera un rejet ou un sous-emploi. En d'autres termes, ce calcul appliqué aux sections permet de

repérer les fréquences remarquables au voisinage de la forme pôle. Le calcul des cooccurrents spécifiques ne fournit pas d'information quant aux cooccurrents droits ou gauches du pôle choisi et ne comporte pas d'autre notion de distance entre les termes co-présents que celle fixée par la longueur des sections.

Forme	Frq. Tot.	Fréquence	Coeff.	Forme	Frq. Tot.	Fréquence	Coeff.
je	344	344	51	j	88	51	3
souhaite	65	64	23	grandeur	7	7	3
vous	326	227	23	soir	42	28	3
voeux	80	62	10	adresser	7	7	3
mes	102	75	10	ma	20	15	3
sais	19	19	8	nom	30	21	3
voudrais	19	19	8	mon	29	19	3
pense	22	21	7	amis	11	10	3
suis	26	24	7	fraternité	19	14	3
forme	17	16	5	seuls	14	12	3
vive	60	42	5	france	302	150	3
heureuse	22	19	5	vivent	10	9	3
bonne	76	51	5	m	23	17	3
veux	11	11	5	vois	6	6	3
crois	11	11	5	ministre	6	6	3
que	677	336	5	famille	25	17	3
dire	48	35	5	très	27	18	3
dis	12	12	5	fais	6	6	3
compatriotes	62	43	5	françaises	41	27	3
ai	41	31	5	doivent	14	1	-3
année	205	110	4	quel	12	0	-3
vos	39	27	4	la	1397	546	-3
chers	55	37	4	algérie	21	3	-3
me	22	17	4	économique	46	10	-3
français	142	80	4	qu	313	108	-3
votre	59	38	4	peut	50	11	-3
chacune	25	20	4	europe	99	28	-3
mer	19	14	3	part	32	3	-5
				nous	655	217	-7

Tableau 6 : Cooccurrents spécifiques de « je » sur la totalité du corpus « vœux »

Les spécificités positives montrent la forte proportion de verbes, gravitant autour du référent du locuteur. (Tableau 6). Verbes marquant la volition (*souhaite, voudrais, forme* [le vœu], *veux*), le jugement (*pense, crois*), factifs (*fais* +3), verbes d'état et auxiliaires (*suis, ai*), énonciatifs (*dis*), verbes marquant la connaissance (*sais, vois*), quelques infinitifs (*dire, adresser*), constituent l'essentiel du système verbal restitué par la recherche des cooccurrents spécifiques. On note aussi de façon plus inattendue la présence d'un verbe à la troisième personne du pluriel : *vivent* (+3), dont on trouve la réalisation dans de fréquentes adresses aux Français *qui vivent à l'étranger* (de Gaulle, 1967), *qui vivent dans la solitude* (V.G.E., 1978), *qui vivent dans la peine* (Mitterrand, 1986), *qui vivent dans la difficulté quotidienne* (Mitterrand, 1988).

Ces messages sont donc particulièrement marqués par des verbes de « circonstance », (*souhaiter, adresser former*), par des volitifs et des verbes exprimant la connaissance. Cependant, cette interprétation sémantique a priori doit être corrigée par l'examen des contextes.

Une analyse approfondie indique que la forme *voudrais* est intimement liée au référent de l'interlocuteur, sur employée chez les locuteurs qui précisément multiplient les marques énonciatives en direction des Français. La valeur n'est donc que rarement purement volitive, les emplois étant essentiellement métadiscursifs,

modalisateurs, intervenant dans des annonces de plan où bien souvent le locuteur s'adresse à une certaine catégorie de Français (*Je voudrais d'abord exprimer ma sympathie à toutes celles et à tous ceux qui vivent ces derniers jours de 1999 dans l'épreuve* [Chirac, 1999]). Giscard et Pompidou qui entretiennent un lien plus étroit avec les Français emploient cette forme dans une modalité directive qui intensifie la relation (*Je voudrais que vous sentiez, que vous compreniez*…). Les contextes de *veux* montrent également une tendance vers des emplois métadiscursifs ou explicatifs (*Je veux dire*), même si la volition apparaît parfois chez Chirac et Mitterrand dans une faible mesure.

Quant aux verbes exprimant le jugement, on remarque que *penser* intervient essentiellement dans des énoncés énumératifs (*Je pense aux artisans, je pense aux agriculteurs, je pense à certaines petites entreprises*) mais bien souvent affectifs et empathiques, liés à l'évènementiel (*Et je pense aussi à nos compatriotes de Toulouse…* [Chirac, 2001]) ou plus généralement destinés à adresser un geste en direction des Français les plus démunis, évocation qui devient systématique à partir de Pompidou (*Je pense spécialement à ceux de nos aînés qui vont franchir seuls le cap du nouvel an* [Chirac, 2000]), (*Je pense à celles et à ceux d'entre vous qui connaissent le deuil, les chagrins, le poids de la maladie et de la solitude, qui souffrent du chômage* [Mitterrand, 1981]).

Parmi les verbes exprimant la conscience et la connaissance, l'examen des contextes montre que la forme *sais* entre essentiellement dans des modalités allocutives. Les emplois sont avant tout des renforçateurs d'empathie, plus particulièrement chez Chirac, parfois constitutifs d'un procédé argumentatif. Cette marque d'empathie introduit dans de nombreux cas chez Chirac une relance incitative et mobilisatrice, que l'on peut synthétiser dans le tableau 7.

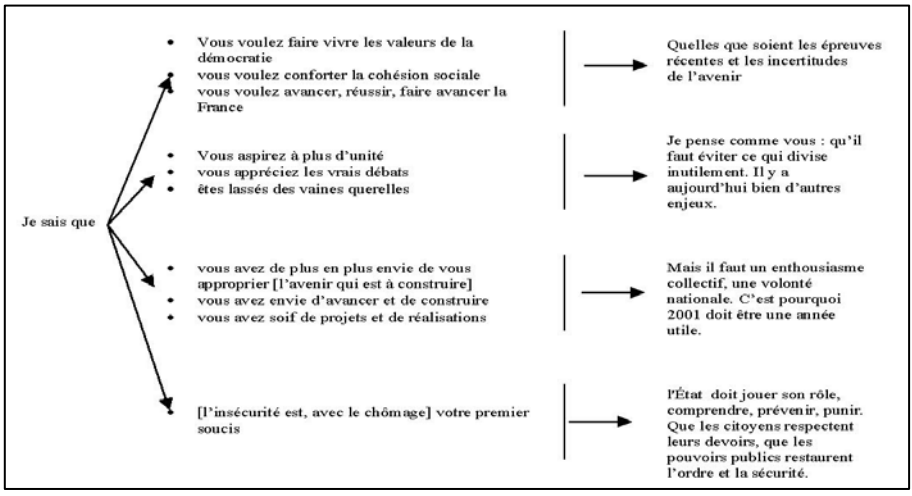


Tableau 7 : Marques de l'empathie chez J. Chirac et relances incitatives.

Le champ lexical du rituel, verbes, substantifs, formules d'adresse est largement représenté parmi les formes qui gravitent autour du *je* présidentiel : *compatriotes, chers, mer (compatriotes d'outre mer), Français, Françaises, vive, année, bonne, heureuse, vœux*.

Parmi les spécificités négatives - assez minoritaires et dont les indices se révèlent moins élevés que pour les sur-emplois - nous retiendrons le diagnostic porté sur la première personne du pluriel (*nous*, -7). Ainsi, le schéma *je/vous*, dont on relève les traces parmi les pronoms personnels et adjectifs possessifs se construit par le rejet de la première personne du pluriel.

2.2 La notion d'environnement thématique (Hyperbase)

C'est à partir de la recherche des contextes qu'il nous est offert d'étudier l'environnement thématique sous Hyperbase. Cette recherche procède d'un calcul de spécificité particulier. Plutôt que de porter un jugement sur la fréquence d'apparition d'une forme dans une sous-partie du corpus par rapport aux autres et à l'ensemble, on calcule les spécificités des mots qui se trouvent au voisinage d'une forme pôle.

En d'autres termes - nous empruntons cette formulation à Etienne Brunet - : « Ce programme de repérage thématique fait le décompte de tous les mots situés dans le même paragraphe que le ou les mots-pôles et mesure leur spécificité, c'est-à-dire la plus ou moins grand attirance que le mot-pôle exerce sur eux. ».

Le choix des partitions n'est pas sans incidence sur la recherche thématique. Aussi adopterons-nous la partition locuteur qui présente l'avantage de ne pas morceler le corpus par trop ainsi qu'un découpage en paragraphes qui constitue un échantillon contextuel suffisamment large. Les formes relevées comme constitutives de « l'environnement thématique » de *je* sont classées selon leur rang de significativité, c'est-à-dire dans l'ordre décroissant de l'écart réduit, qui s'étend ici de la valeur 25,79 à 2. L'écart réduit peut être considéré comme une approximation – satisfaisante pour des corpus de grande ampleur – du modèle hypergéométrique utilisé dans le calcul des spécificités.

La loi normale est d'usage dans les versions antérieures d'Hyperbase, appliquée aux spécificités et chaque fois que la distribution d'une forme devait être pondérée pour tenir compte des étendues inégales des textes comparés. Toute fréquence était alors traduite en écart réduit. La puissance de calcul des ordinateurs actuels autorise désormais l'application de la loi hypergéométrique (utilisée également sous Lexico et Weblex), plus appropriée car s'appliquant à des données discrètes (la loi normale traitant des valeurs continues) et d'une plus grande exactitude sur des corpus de faible étendue.

Les résultats sont toutefois convertis afin de conserver l'échelle de l'écart réduit et la représentation adoptée dans les précédentes versions d'Hyperbase. L'écart réduit s'interprète en termes de déficits ou d'excédents, sa valeur est donc positive ou négative. Une valeur avoisinant 2 ou -2 est considérée comme négligeable. Le calcul de l'environnement thématique ne tient pas compte des déficits. Ainsi, contrairement aux cooccurrents spécifiques de Lexico, aucun diagnostic n'est porté sur les sous-emplois.

La seconde colonne du tableau 8 présente l'effectif de la forme considérée sur l'ensemble du corpus, la dernière colonne indique le nombre d'occurrences dans la sous partie, constituée à partir des contextes de *je*. On note que la fréquence de l'extrait est parfois supérieure à l'effectif total. Le cas le plus flagrant est celui de la forme que nous avons choisie comme pôle de référence : 344 occurrences au total pour 347 formes dans l'extrait. Cet artefact s'explique par la comptabilisation des graphies qui apparaissent plusieurs fois au sein d'un même paragraphe. Il conviendrait donc de revoir la segmentation du corpus afin de neutraliser ce comptage. Toutefois de multiples vérifications, ainsi que l'utilisation comparée des plusieurs outils nous autorise à considérer ces résultats comme fiables. Enfin, pour faciliter la lecture, et la comparaison avec les listes produites par Lexico nous avons porté les indices de spécificité relevés par cet outil. (Colonne *Sp*).

L'essentiel des faits que nous avions abordés par l'entrée des cooccurrents spécifiques est restitué. Les premiers rangs sont identiques à ceux observés au moyen de la méthode de cooccurrents spécifiques, ce qui conforte nos premières conclusions sur le rituel et la thématique des vœux. Parmi les verbes : *souhaite* dont on a déjà examiné les contextes et établi qu'il était essentiellement employé dans la formulation de vœux, de même *forme*, *adresser*, mais aussi *limiter* (*je ne voudrais pas me limiter à vous présenter mes vœux*…). Les usages de la première personne sont essentiellement conditionnés par le genre discursif y compris chez les locuteurs dont l'énonciation est peu personnalisée. Le *Je* est à la fois dialogique et familier, rituel et circonstanciel.

On s'interrogera sur la procédure d'élagage appliquée sous Hyperbase - dont on maîtrise peu les paramètres mais dont la présentation synthétique est propre à faire émerger les faits saillants du corpus mais aussi plus anecdotiques, qui apparaissaient sous Lexico parmi de nombreuses autres formes, dotées généralement d'indices très faibles de spécificité. On soulignera cependant l'absence remarquable des formes *chers* (indice positif de 4 sous Lexico), *compatriotes* (+5), *France* (+3), *Vive* (+5) *bonne* (+5) qui entrent dans la réalisation de formules d'adresse ou qui appartiennent à la dimension rituelle du discours (*Vive la France, bonne année*…).

Ces outils fournissent des représentations des faits de cooccurrences, propres à orienter l'analyse dans des directions parfois différentes, d'où la nécessité de les interroger systématiquement.

Ecart	Corpus	Extrait	Mot	Sp.	Ecart	Corpus	Extrait	Mot	Sp.
25.79	344	347	JE	51	2.74	4	4	LÉGISLATIVES	2
11.80	323	213	VOUS	23	2.74	4	4	AUGMENTÉ	2
11.06	65	65	SOUHAITE	23	2.71	29	17	MON	3
7.23	79	58	VŒUX	10	2.65	11	8	RAISONS	2
6.46	19	20	VOUDRAIS	8	2.61	19	12	FRATERNITÉ	3
6.17	26	24	SUIS	7	2.60	127	58	AUX	1
5.99	22	21	PENSE	7	2.60	122	56	AVEC	2
5.98	19	19	SAIS	8	2.60	17	11	ADRESSE	2
5.56	48	35	DIRE	5	2.50	6	5	UNS	2
5.50	41	31	AI	5	2.50	6	5	MINISTRE	3
5.44	674	301	QUE	5	2.50	6	5	EXPRIME	1
5.15	17	16	FORME	5	2.50	6	5	DÉPARTEMENTS	2
4.75	12	12	DIS	5	2.50	6	5	AVAIS	2
4.55	11	11	VEUX	5	2.40	178	77	PAS	2
4.55	11	11	CROIS	5	2.40	33	18	GOUVERNEMENT	1
4.34	25	19	CHACUNE	4	2.40	8	6	COMPRENDRE	2
4.25	59	36	VOTRE	4	2.35	202	86	ANNÉE	4
4.20	22	17	ME	4	2.35	18	11	AVEZ	2
4.17	101	55	MES	10	2.33	14	9	SEULS	3
4.03	42	27	SOIR	3	2.33	14	9	CHOSES	2
3.68	30	20	NOM	3	2.22	41	21	FRANÇAISES	3
3.63	7	7	ADRESSER	3	2.17	90	41	AUSSI	2
3.59	3	4	LIMITER	2	2.16	21	12	POSSIBLE	1
3.55	101	52	CEUX	2	2.14	30	16	ABORD	2
3.52	39	24	VOS	4	2.13	37	19	AN	2
3.49	27	18	CŒUR	2	2.13	5	4	VRAIMENT	1
3.36	6	6	VOIS	3	2.13	5	4	VENU	1
3.36	6	6	FAIS	3	2.13	5	4	STRASBOURG	1
3.30	22	15	BONHEUR	2	2.13	5	4	RÉPONDRE	1
3.28	11	9	AMIS	3	2.13	5	4	PROGRESSER	1
3.08	25	16	FAMILLE	3	2.13	5	4	OFFRE	1
3.08	23	15	M'	3	2.13	5	4	KOWEÏT	2
3.07	5	5	MESSAGE	2	2.13	5	4	GARANT	1
3.07	5	5	MARS	2	2.13	5	4	FRAPPÉ	1
3.07	5	5	FIER	2	2.13	5	4	FIDÈLE	1
3.07	5	5	FAIBLES	1	2.13	5	4	ADRESSENT	1
3.07	5	5	DISAIS	2	2.12	19	11	MER	3
3.05	83	42	J'	3	2.09	17	10	ÉTRANGER	2
3.01	10	8	VIVENT	3	2.04	7	5	PRÉSIDENT	1
2.88	26	16	TRÈS	3	2.04	7	5	PARLER	1
2.85	22	14	HEUREUSE	5	2.04	7	5	MAJORITÉ	2
2.85	20	13	MA	3	2.04	7	5	FRATERNELLE	1
2.84	7	6	GRANDEUR	3	2.02	11	7	SOUFFRENT	1
2.84	7	6	DEMANDE	2	2.02	11	7	PROFESSIONNEL	2
2.77	142	65	FRANÇAIS	4	2.02	9	6	SOLITUDE	1
2.74	4	4	VAIS	2	2.02	9	6	RÉUSSIR	1
2.74	4	4	TIRE	2	2.02	9	6	DÉPIT	1
2.74	4	4	TIENS	2	2.02	9	6	DÉBUT	2
2.74	4	4	REÇU	2	2.00	24	13	TROP	2
2.74	4	4	QUICONQUE	2	2.00	24	13	BESOIN	1

Tableau 8 : Environnement thématique de la forme « je » - Hyperbase. Partition locuteur, contexte paragraphe

2.3 Lexicogrammes simples, lexicogrammes récursifs (Weblex)

Le calcul de cooccurrences implémenté dans Weblex repose sur le modèle développé par Pierre Lafon [Lafon, 1984]. On se reportera à [Heiden 2004] et [Tournier, 2003] pour des approfondissements méthodologiques.

Cette méthode permet de porter sur le lexique présidentiel deux éclairages complémentaires.

Le lexicogramme simple, associé à une forme affiche les formes « les plus cooccurrentes » d'une forme pôle. « Le lexicogramme d'un mot s'interprète comme une synthèse des cooccurrents gauches et droits d'un mot, à l'intérieur de toutes les phrases où il apparaît ». [Heiden, 2004]. Le tableau 9 fournit ainsi une dimension supplémentaire qui ne pouvait être appréhendée avec les outils précédents. Il

présente les principaux cooccurrents gauches et droits de la forme pivot *je* dans l'ensemble du corpus au seuils Fréquence et co-fréquence minimale de 3 occurrences, probabilité de 5 % (soit 5.0e-2), distance moyenne 1 000 occurrences.

Pour chaque cooccurrent la Fréquence totale de la forme dans le corpus (**F**), sa cofréquence avec la forme pôle, c'est-à-dire le nombre de rencontres attestées (**CF**), le diagnostic de probabilité de la rencontre (**P**) et la distance moyenne ($\mathbf{d_m}$) fournissent les indications quantitatives et statistiques de ces rencontres (tableau 9).

Nous constatons au premier regard un déséquilibre entre les cooccurrents gauches et les cooccurrents droits qui dominent très largement. Ce phénomène ne surprendra pas en raison de la nature de la forme pivot qui implique par essence de plus fréquentes relations sur sa droite que sur sa gauche. Les cooccurrents gauches restituent la dimension rituelle du discours. Cinq d'entre eux sont constitutifs de formules d'adresse (*mes chers compatriotes de métropole et d'outre mer*). La position du substantif « vœux » ne surprendra pas d'avantage (*les vœux que je forme, que je vous adresse*…) ni la présence du substantif *cœur* dont les contextes attestent la réalisation (*C'est de tout cœur que je*…).

Les verbes, généralement post-posés au pronom personnel figurent tout naturellement à droite du pivot. On y retrouve ceux que nous avions recensés au moyen des autres outils mais la hiérarchie est quelque peu différente. *Souhaite* se trouve au premier rang (tri par probabilités) comme c'était le cas avec les autres analyses. Sa distance moyenne avec la forme pôle indique qu'il s'agit sans doute souvent de collocations ou qu'un terme peut s'interposer, probablement une marque de l'interlocuteur (*je vous souhaite*). Les onze premières formes à droite sont des verbes dont la morphologie indique qu'ils sont fléchis à la première personne du singulier, et qui sont peu distants de la forme pivot. On note que le temps est très majoritairement le présent de l'indicatif. Après la catégorie verbale viennent des adjectifs et des substantifs qui ancrent le discours dans le présent, ou sont issus de la thématique des vœux (*heureuse, vœux, soir, bonne, année, adresse*)… Peu de termes politiques émergent de ces lexicogrammes, ainsi que nous l'avions déjà observé au moyen des autres outils cooccurrentiels. Ceci confirme nos premières observations : le *je* est essentiellement mobilisé par le genre discursif. Une dernière remarque concerne la distance moyenne des cooccurrents, plus importante sur la gauche de la forme pivot que sur sa droite où l'on observe quelques collocations. Qu'en est il de la présence forte de l'interlocuteur que nous avions cru déceler dans les analyses précédentes ? Parmi les verbes, aucun ne semble être conjugué à la deuxième personne du pluriel, ainsi que nous l'avions déjà noté. On ne s'étonnera pas de l'absence des marques personnelles renvoyant à l'interlocuteur dont on a noté la forte spécificité sous Lexico, confirmée par Hyperbase. Nous avons en effet choisi de conserver l'élagage des formes outils du vocabulaire afin de porter un regard différent sur le corpus, en ne considérant que les seuls mots pleins.

Une expérience menée sur ces mêmes lexicogrammes sans suppression des formes outils a cependant confirmé l'attirance importante des marques de la première du singulier et de la seconde du pluriel.

je (346)

cooccurrents gauches					**cooccurrents droits**				
	f	**cf**	**p**	**d_m**		**f**	**cf**	**p**	**d_m**
chers	55	21	4e-06	7.3	souhaite	65	64	6e-57	0.2
métropole	18	10	3e-05	11.1	pense	22	21	9e-18	0.0
compatriotes	62	20	1e-04	7.2	voudrais	19	19	3e-17	0.1
françaises	41	15	2e-04	10.7	sais	19	18	4e-15	0.6
outre-mer	14	7	1e-03	8.6	forme	19	16	8e-12	0.0
vœux	81	21	2e-03	4.9	crois	11	11	3e-10	0.1
cœur	27	9	8e-03	4.3	veux	11	11	3e-10	0.1
vois	6	3	4e-02	17.0	dis	12	11	3e-09	0.7
					dire	39	19	1e-07	3.3
					adresser	7	7	9e-07	6.9
					fais	6	6	7e-06	0.8
					heureuse	22	12	7e-06	13.5
					vœux	81	26	9e-06	7.3
					soir	42	17	1e-05	6.8
					bonne	75	24	2e-05	8.5
					demande	7	6	4e-05	2.2
					disais	5	5	5e-05	0.8
					vois	6	5	3e-04	0.0
					vais	4	4	4e-04	0.0
					année	202	42	7e-04	10.1
					adresse	17	8	9e-04	1.0
					fier	5	4	2e-03	6.5
					faibles	5	4	2e-03	20.8
					salue	3	3	3e-03	0.7
					dirai	3	3	3e-03	1.0
					répète	3	3	3e-03	1.3
					souviens	3	3	3e-03	1.3
					assure	3	3	3e-03	14.0
					sûr	12	6	3e-03	1.0
					rendre	9	5	4e-03	7.6
					promis	6	4	4e-03	2.0
					espère	6	4	4e-03	5.8
					propose	6	4	4e-03	6.5
					dit	14	6	8e-03	3.8
					fraternelle	7	4	9e-03	17.8
					demandé	4	3	9e-03	2.0
					satisfaction	4	3	9e-03	12.0
					tire	4	3	9e-03	19.3
					nom	30	9	2e-02	6.4
					vivre	30	9	2e-02	13.1
					parle	8	4	2e-02	2.2
					constate	5	3	2e-02	0.0
					ardents	5	3	2e-02	3.0

Tableau 9 : Lexicogramme de la forme je dans le corpus vœux.
Seuils : f 3, cf 3, p 5.0E-2, dm 1000.0

Le lexicogramme récursif associé à une forme révèle des réseaux de cooccurrences sur le principe d'un enchaînement de cooccurrents de cooccurrents En partant de la source du lexicogramme *(je*) il en établit les principales attirances. Partant de ces nouvelles formes, d'autres cooccurrents sont mis en évidence selon le principe que chaque nouveau cooccurrent est à son tour pris comme source. La représentation qui en résulte synthétise l'ensemble des connexions qui prennent naissance autour de la forme pôle. Le lexicogramme récursif de la première personne du singulier, a été établi avec élagage des formes outils aux seuils précédemment adoptés (Tableau 10).

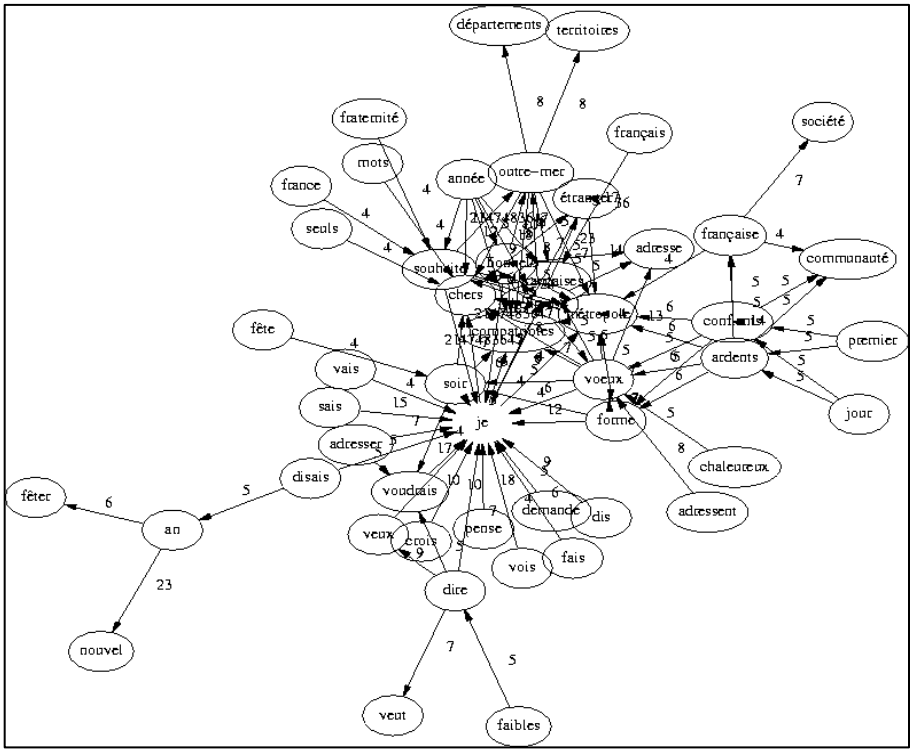


Tableau 10 : Lexicogramme récursif de la forme je sur le corpus voeux2. Élagage des formes outils, numéraires et hapax. Seuils : f 3, cf 3, p 5.0E-2, dm 1000.0

La thématique des vœux ressort nettement de cette représentation. La connexion de verbes gravitant autour du *je* est manifeste : *forme, adresser, sais, demande, dis…*, verbes qui nous sont désormais familiers. Dans cette configuration, l'ordre des termes n'est pas restitué bien qu'il soit possible de le faire. Cette

perspective spatiale permet de saisir des faits qui ne seraient pas nécessairement apparus aussi clairement sur des listes hiérarchisées. Ainsi les formules d'adresse sont-elles matérialisées par les nœuds *départements, territoires, outre-mer*. Les qualifiants de *vœux, ardents et confiants* que le général de Gaulle *adresse* à la France sont clairement identifiés, de même que les emplois métadiscursifs (*je veux dire, je voudrais dire, je voudrais adresser*). On se gardera cependant d'interpréter trop hâtivement ces connexions. Suivons par exemple la « branche » qui relie le pronom personnel à *disais*. Le verbe est lui-même relié à *an* qui forme un nœud à partir duquel deux nouveaux arcs partent vers *fêter* et *nouvel*. Nul doute que le substantif *an* entre en cooccurrence avec le verbe et l'adjectif. Des réalisations comme « *fêter le nouvel an* » ne seront pas difficiles à attester. Pas de doute non plus sur la relation *je-disais* vérifiée en contexte. Il en va tout autrement de réalisations poly-cooccurrentes qui attesteraient *je, disais, fêter, nouvel an* dans un même contexte. Ceci tient au caractère de récursivité des lexicogrammes. Le lexicogramme récursif établit bien les cooccurrents de *je* parmi lesquels le verbe *disais*. Cependant considérant ce verbe à son tour comme source une nouvelle cooccurrence sera établie comme ici *an*. *Disais* est bien cooccurrent de *an*, et donc de *je*. Mais *an* ne renvoie pas nécessairement dans les mêmes contextes à *fêter* et *nouvel*. Il s'agit donc de ne pas considérer le chemin que nous venons de retracer comme un « squelette de phrase » [Martinez, 2003]. Illustrons notre propos à l'aide de quelques contextes.

Il y a 3 occurrences de "an"[]*"disais"%c dans le corpus voeux

Pompidou, 1971 une fête. Ils oublient la présence de l'hiver pour ne pressentir que le prochain printemps. Ils veulent croire que vieillir est un moyen de marcher vers le meilleur. Cela s'appelle l'espérance. Avons-nous, en tant que Français, des raisons d' espérer ? Eh bien, oui, n'en déplaise à tous les spécialistes de la triste figure. Il y a un **an, jour pour je vous disais** : "Nous ne sommes pas les plus forts, mais nous comptons et nous sommes respectés". L'année 1971 n'en a-t-elle pas apporté quelques preuves ? Les visites amicales que nous ont faites tant de chefs d'État et de gouvernement étrangers, une délégation chinoise, le premier responsable soviétique, l'entrevue que j'ai eue en terre européenne.

Pompidou, 1971 l'élargissement de la Communauté européenne et la crise monétaire internationale. A Berlin, aux Nations unies, son action a été visible et utile. Il n'y a pas lieu d'en tirer vanité. Mais, pourquoi le dissimuler, notre pays, indépendant, pacifique et sûr de lui, n'a pas déchu du rang où l'avait placé le général de Gaulle. Il y a un **an, je vous disais** encore : "Nous ne sommes pas les plus riches, mais nous sommes parmi les plus heureux. Il suffit de regarder autour de nous". Or, aujourd'hui, il suffit d'écouter la voix des commentateurs étrangers, qu'ils soient Anglais, Américains ou Russes, pour apprendre que la situation de la France est appréciée par tous et enviée par beaucoup.

Pompidou, 1973 onde dans ses biens, dans sa situation, dans ses libertés. Je suis convaincu que vous en avez conscience et c'est pourquoi c'est en pleine confiance et de tout coeur que je vous dis ce soir : Bonne année ! Que 1973 apporte à vous tous un peu plus de bonheur. De mon mieux, soyez-en certains, j'y aiderai. Françaises, Français, il y a un **an, en vous offrant mes voeux pour l'année 1973**, **je vous disais** que ce serait une année d'expansion exceptionnelle et de grands progrès dans divers domaines. Eh bien , c'est ce qui s'est passé. Les chiffres le prouvent et les observateurs sérieux, les plus rigoureux, le reconnaissent. Et pourtant, il faut admettre que l'année se termine dans une atmosphère moins sereine et que les perspectives sont p

2.4 Distributions statistiques et distributions linguistiques (Alceste)

L'expérimentation menée ici repose sur une utilisation particulière d'Alceste. Il ne s'agit pas de faire émerger les structures saillantes du corpus, d'en identifier les classes d'énoncés ou « mondes lexicaux » [Reinert, 1993, 1998] comme le fait la procédure par défaut. On utilise ici le tri croisé sur une forme pour mettre en évidence les cooccurrents de *je*, sur l'ensemble du corpus, afin de recouper cette analyse aux examens précédemment réalisés.

L'analyse en tri croisé, qui peut porter sur une forme ou sur une variable consiste à croiser forme ou variable avec l'ensemble du corpus ce qui aura pour effet de scinder le corpus en deux parties, celle où la forme est attestée et l'autre où elle n'apparaît pas. Contrairement aux analyses usuelles qui reposent sur le principe de la classification descendante, 100 % des U.C.E. (unités de contexte élémentaire) sont classées. Ici, l'ensemble des U.C.E. est pris en compte, sur la base d'une classification ascendante.

Nous produisons dans le tableau 11 les formes les plus caractéristiques des deux classes obtenues, ordonnées selon le Khi2 décroissant[4].

Classe 1 : 255 U.C.E (27%)			
Forme réduite	Khi2	Forme réduite	Khi2
je	969	votre	12,07
vous	148,97	soir+	12,07
souhait<	148,06	bonne+	11,34
suis	60	redire.	11,25
dire.	45,77	tiens	11,25
mes	44,38	present+er	11,25
voeu+	38,95	echang+er	11,25
dire+	38,14	metropole+	10,98
forme+	37,03	savoir.	10,9
me	27,54	m	10,52
annee+	27,44	soiree+	10,12
vouloir.	27,02	avais	10,12
heur+eux	23,87	coeur+	9,97
compatriote+	23,61	promettre.	9,86
francais+	23,31	parl+er	9,43
adress+er	22,45	année_1974	9,34
cher+	19,82	mon	9,34
ai	18,98	pas	9,32
nom+	17,59	ce	9,27
outre	15,11	ma	8,67
qu+	14,12	mer+	8,67
enjeu+	14,07	recu+	8,43
adresse+	13,15	aim+er	8,43
*loc_giscard	12,76	demand+er	8,09
propos+er	12,36	rappel+er	7,63

Classe 2 : 714 U.C.E (73%)			
Forme réduite	Khi2	Forme réduite	Khi2
*loc_dg	52,38	trouv+er	4,1
nous	14,84	ailleurs	3,97
econom+	10,6	leur	3,97
année_1963	10,31	guerre+	3,93
devoir.	8,67	moderne+	3,73
année_1965	7,29	solid+e	3,61
année_1961	7,27	rapport+	3,61
année_1968	6,24	acquerir	3,61
année_1960	6,16	soi	3,61
moyen+	5,56	quel	3,61
developpement+	5,5	troubl+er	3,61
*année_1966	5,12	scientifi<	3,61
arme+	5,07	ensemble+	3,55
cooperati+f	5,07	*année_2000	3,55
techn+	5,07	*année_1962	3,44
peuple+	5,06	mondia+	3,44
face+	4,71	moment+	3,38
organisat+ion	4,71	cas	3,24
*année_2001	4,57	dehors	3,24
avons	4,52	telle	3,24
elle	4,46	multipli+er	3,24
jusqu+	4,34	monetaire+	3,24
plan+	4,34	but+	3,24
conflit+	4,34	route+	3,24
inflation<	4,34	aid+er	3,01

Tableau 11 : Croisement de la forme je avec le corpus. Les 50 premières formes caractéristiques

[4] La distance du Khi^2 est une pondération de la distance euclidienne. Dans le cas présent ce coefficient est un indicateur du degré d'appartenance d'une forme réduite à une classe. La forme la plus constitutive des énoncés où figure *je* est la seconde personne du pluriel. La forme la plus représentative de la classe opposée est la variable *De Gaulle* puis la première personne du pluriel.

Ces classes nous donnent confirmation de certains faits observés précédemment.

- Sur l'aspect pronominal, la dimension dialogique du discours est confortée, ainsi que le rejet de la première personne du pluriel : *vous* figure au second rang des formes significatives de la classe 1, les marques de la première du pluriel à un rang identique dans la classe opposée.
- Sur l'aspect énonciatif : Parmi les énoncés qui s'opposent au *Je* la variable loc_dg est la première forme significative de la deuxième classe. C'est-à-dire que la plupart des énoncés du locuteur de Gaulle se trouve classé dans la catégorie la plus éloignée du *je*.
- Sur la chronologie : Les années 63, 65, 61, 68, 60, 66, 2001 et 2000 sont dotées d'indices de Khi2 relativement importants et sont donc fortement constitutives de la classe 2. Ainsi cette classification binaire et somme toute assez brutale porte un éclairage chronologique sur les emplois pronominaux puisque, recoupant cette information par les spécificités par partie, il ressort que la forme *je* est sous employée dans les messages correspondants par rapport à l'ensemble du corpus. Ainsi résumons nous notre démarche par ce va et vient entre les différents outils, les résultats de l'un nous engageant à interroger l'autre.

Dans la première classe en revanche, c'est la variable Giscard qui est la plus représentative. Nous avions constaté qu'il était celui chez qui la personnalisation du discours était la plus sensible. Ses messages par ailleurs sont ceux qui laissent la part la plus importante à la thématique des vœux. Ce fait se vérifie, là encore par la méthode des spécificités. Cette thématique des vœux est précisément très représentée dans la première classe, à travers des substantifs : *vœux, année, soirée*, des adjectifs : *heureux, bonne* mais aussi des verbes : *souhait<, forme, adresser, présenter, échanger*.

Le rituel et les formules d'adresse sont par ailleurs constitutifs de cette première classe : *Compatriote +, cher +, métropole +*. On notera que le lexique est plus intimiste, plus affectif dans cette même classe, la classe 2 recensant un lexique plutôt politique, économique et social : *économie, développement, organisation, conflit*…

Ceci nous amène à nous interroger sur la nature des relations entretenues entre les marques de la première personne et les différentes thématiques des messages : le *je* rejette-t-il dans ces messages la dimension politique au profit d'un lexique plus circonstanciel ou rituel, cette dimension politique serait-elle alors développée autour d'autres référents ?

3. Conclusion

Le croisement des approches et des outils statistiques dont la finalité première n'est pas nécessairement la recherche localisée de phénomènes cooccurrentiels révèle des résultats convergents pour les faits les plus saillants. Cette démarche expérimentale et comparative a permis de réunir un faisceau sur lequel l'analyse interprétative peut s'appuyer plus fortement, que sur les faits révélés par un outil unique.

Le croisement des perspectives valide la démarche lexicométrique sur un corpus de faible étendue tel que celui des vœux présidentiels. L'application de paramètres et de seuils, plus ou moins intuitifs mais souvent différents, le choix des fenêtres contextuelles, l'élagage ou non de certaines formes de vocabulaire, sont autant de facteurs qui incitent à confronter ces outils avant toute phase interprétative. Spécificités positives et négatives, topographie textuelle (Lexico 3), présentation synthétique (Hyperbase), ou ordonnée (Weblex lexicogrammes simples), réseaux d'affinités et récursivité (Lexicogrammes récursifs), partition binaire et classification ascendante (Alceste), fournissent des perceptions différentes et complémentaires des faits de cooccurrences.

4. Bibliographie

Amossy R. (1999), *Images de soi dans le discours, la construction de l'ethos*, Lausanne, Delachaux et Niestlé.

Heiden S. (2004), *Interface hypertextuelle à un espace de cooccurrences: implémentation dans Weblex*, JADT 2004 : 7èmes.

Labbé D. (1990), *Le vocabulaire de François Mitterrand*, Paris, Presses de la Fondation Nationale des Sciences Politiques.

Lafon P. (1984), *Dépouillements et Statistiques en Lexicométrie*, Slatkine-Champion.

Lamalle C., Salem A. (2002), *Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels*, JADT 2002 : 6émes

Lebart L., Salem A. (1994), Statistique Textuelle, Paris, Dunod.

Maingueneau D. (2003), « Discours éphémères et non-éphémères : deux gestions de l'ethos ? », In J. Härmä dir., *Le langage des médias: discours éphémères* ?, Paris, l'Harmattan, pp. 67-82.

Martinez W. (2003), *Contribution à une méthodologie de l'analyse des cooccurrences lexicales multiples dans les corpus textuels*, Thèse de troisième cycle, Sorbonne Nouvelle Paris 3, décembre 2003.

Reinert M. (1993), Les « mondes lexicaux » et leur logique, Langage et société 66.

Reinert M. (1998), Quelques interrogations à propos de l'« objet » d'une analyse de discours de type statistique et de la réponse « Alceste », *Langage et société* (1998).

Reinert M. (1990), Système Alceste : Une méthodologie d'analyse des données textuelles, JADT 1990.

Tournier M. (2003), *De France à Je. La traversée des emplois Cooccurrences et connexions*, In Des sources du sens, École Normale Supérieure Lettres Sciences Humaines Lyon, Collection Langages.

Session 3

Organisation et exploration de corpus documentaires multimédia

Expression du point de vue des lecteurs dans les bibliothèques numériques spécialisées

Aurélien Bénel

LIRIS (Lyon) & Réseau transdisciplinaire ARTCADHi - France

Aurelien.Benel@insa-lyon.fr

Résumé :

Le système *Porphyry* s'adresse à des communautés d'experts appelés à travailler sur des corpus documentaires numérisés. Il est fondé sur l'enrichissement itératif du corpus par des structures hypermédias. Ces structures sont construites par les experts en fonction de leurs problématiques et de leurs spécialisations. Cette modélisation « à base de points de vue » est contrairement aux systèmes classiques « à base de connaissances » : dynamique : des hypothèses peuvent être réfutées, d'autres ajoutées ; plurielle : les points de vue ne sont pas forcément arborescents, de plus, plusieurs points de vue peuvent se croiser en un même objet documentaire.
Des outils spécifiques sont offerts pour découvrir, de manière interactive, le corpus à travers ses différentes structures.

Mots-clés : Bibliothèque numérique, Assistance à l'interprétation, Collaboration, Annotation, Hypermédia.

Abstract:

The targeted audience of the *Porphyry* system is experts communities working with digitized documents. The corpus is iteratively augmented by hypermedia structures. These structures are built by the experts depending on their interests and their specializations. This "point-of-view based" modeling, contrary to the classical knowledge-based modeling is: dynamic: some hypotheses can be falsified, others can be added , plural: points of view do not have to be tree-like, many points of view can describe the same document object.
Special tools are offered to discover interactively the corpus through its different structures.

Keywords: Digital Libraries, Interpretation Assistance, Collaboration, Annotation, Hypermedia.

> *Our ineptitude in getting at the record is largely caused by the artificiality of systems of indexing. When data of any sort are placed in storage, they are filed alphabetically or numerically, and information is found (when it is) by tracing it down from subclass to subclass. It can be in only one place, unless duplicates are used; one has to have rules as to which path will locate it, and the rules are cumbersome.*
>
> Vannevar Bush, As we may think, 6.

1. Introduction

Nos travaux s'insèrent dans un projet du réseau ARTCAHi[1] visant à offrir aux chercheurs en Sciences Humaines des assistants à la construction du sens dans les bibliothèques numériques[2]. Les bibliothèques considérées sont des bibliothèques spécialisées, destinées à des experts. Dans un tel cadre, limiter la description des documents à une indexation, unique, fixe et effectuée par un tiers, reviendrait à nier l'expertise des lecteurs.

Autrement dit, dans notre approche, la description structurée d'objets documentaires (textes courts, images...) permettra non seulement la *rédaction* (description d'une section parmi un document) et l'*indexation* (description d'un document parmi une collection), mais également l'*annotation* (description d'un fragment parmi un document « sur mesure »).

Nous étudierons dans un premier temps comment la question de la description de documents (de manière structurée) est ordinairement traitée. Dans un deuxième temps, nous proposerons un modèle basé sur la notion de « point de vue », accompagné d'un algorithme de filtre permettant « d'arpenter » l'espace documentaire. Enfin, dans un dernier temps, nous analyserons un cas d'application de ce modèle.

2. Décrire de manière structurée des objets documentaires

2.1 Des arbres qui cachent … la bibliothèque

L'exergue de cet article rappelle que l'organisation traditionnelle des bibliothèques est basée sur l'idée que chaque livre traiterait d'un sujet unique, sujet qui lui-même serait situé sans ambiguïté possible dans une hiérarchie universelle.

---

[1] ARTCADHi : Atelier de Recherche Transdisciplinaire sur la Construction du sens en Archéologie et dans les autres Disciplines Historiques, <http://www.porphyry.org>.

[2] Programme Société de l'Information (2001-2004), CNRS STIC-SHS, « Assistance dans la gestion de ressources intertextuelles multiformes. Production et intégration interactives de parcours interprétatifs », Collaboration EFA–LIRIS–MOM–ENST-Br.

L'auteur (Vannevar Bush) dénonce ici le caractère artificiel et inutilisable de ce modèle arborescent instauré dans les bibliothèques par Dewey (1859-1952).

Les racines du problème remontent sans doute à Aristote (IV^e^ s. av. J.-C.). Celui-ci nous déclare en effet :

> Quand une chose est attribuée à une autre comme à son sujet, tout ce qui est affirmé du prédicat devra être aussi affirmé du sujet […]. Si les genres sont différents et non subordonnés les uns aux autres, leurs différences seront elles-mêmes autres spécifiquement […]. Par contre, dans les genres subordonnés les uns aux autres, rien n'empêche que leurs différences soient les mêmes, car les genres plus élevés sont prédicats des genres moins élevés, de sorte que toutes les différences du prédicat seront aussi des différences du sujet. [Aristote-300, Catégories 3].

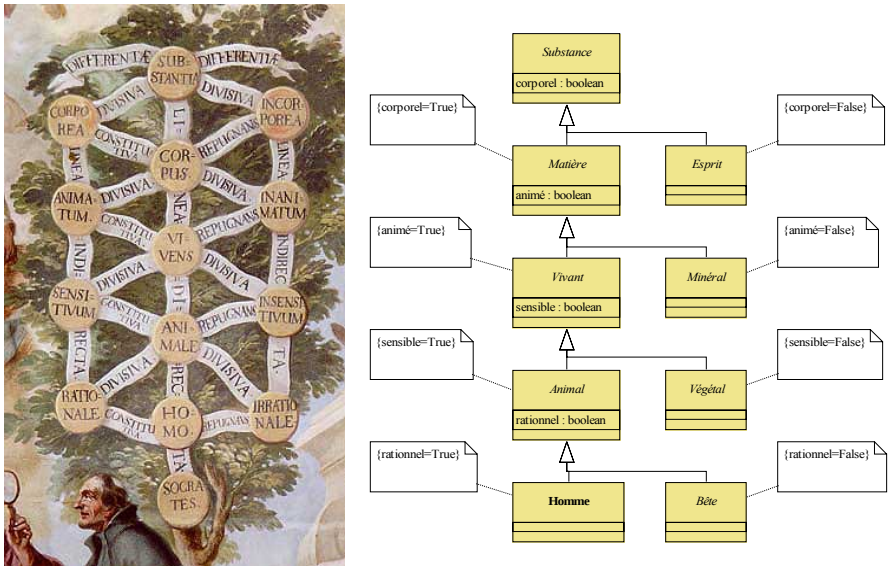


Figure 1 : « L'arbre de Porphyre », fresque du XVIII^e^ s.[3] *et diagramme de classe UML*

Ces trois principes sont interprétés par Porphyre (III-IV^e^ s. ap. J.-C.) comme définissant un arbre dans lequel à chaque nœud un genre (*genus*) se divise en sous-genres en fonction d'une différence (*differentia*). Chacun de ces trois principes peut donc se lire de manière appliquée dans « l'arbre de Porphyre » (cf. figure 1) :

[3] Détail d'une fresque du monastère de Schussenried (Allemagne) ; Artiste : Hermann, 1757 ; Photographe : J. Garrett, 2000 ; Disponible sur Internet : http://www.library.northwestern.edu/collections/garrett>

1. Les hommes sont des animaux. Les animaux sont sensibles. Donc les hommes sont sensibles.
2. « Rationnel » est une propriété « d'Homme » qui ne subordonne pas « Végétal ». Donc, « Rationnel » n'est pas une propriété de « Végétal ».
3. L'Homme est :
   - un animal rationnel,
   - un être vivant sensible et rationnel,
   - une matière animée, sensible et rationnelle,
   - une substance corporelle, animée, sensible et rationnelle.

On comprend sans peine qu'une organisation des connaissances d'une telle esthétique et d'une telle efficacité soit devenue le parangon de la pensée occidentale. Cependant, si cette méthode est tout à fait valide pour parler de classes, elle ne devrait en aucun cas être utilisée pour des instances, celles-ci pouvant souvent être placées dans plusieurs classes contradictoires. En effet, on pense tout de suite au célèbre exemple de Nixon potentiellement pacifiste en tant que quaker et belliciste en tant que républicain.

2.2 Alternatives

La méthode arborescente étant inutilisable pour classer des instances, *a fortiori* elle l'est également pour des livres dont la description pourra contenir des classes, des instances et des liens entre instances. Aussi, depuis longtemps, des alternatives au modèle de l'indexation hiérarchique ont été proposées. La plus connue en sciences de l'information est celle de Ranganathan (1872-1972) appelée aussi « indexation par facettes », mais les plus ambitieuses sont sans doute celles basés sur les graphes conceptuels de John F. Sowa [MechkourEtAl95, Martin96, Genest00]. Dans ces dernières, chaque document est décrit par un graphe (distinct) comprenant des objets (éventuellement génériques) et des liens entre ces objets (cf. figure 2). Ces objets et ces liens, sont des instances d'un modèle du domaine (cf. figure 3).

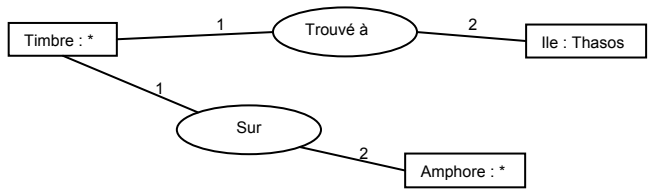


Figure 2 : Indexation à l'aide des graphes conceptuels d'une monographie traitant des timbres amphoriques thasiens

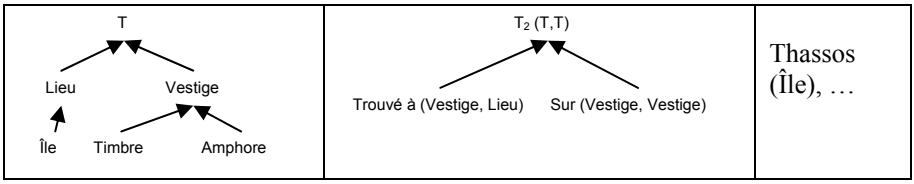


Figure 3 : Modèle du domaine nécessaire à l'indexation de la figure précédente : concepts, relations et instances

Malheureusement, malgré leur puissance d'expression, ces alternatives sont, comme nous allons le voir, assez peu adaptées au cas qui nous occupe : celui d'une modélisation dynamique effectuée *par les usagers* de la bibliothèque.

2.3 Réfutation

Prenons un exemple en archéologie. Philippe Bruneau [Bruneau76], en réaction aux premières « banques de données archéologiques », faisait remarquer l'impossibilité de décrire « objectivement » une photographie du type de la figure 4. Etait-on en présence de la représentation d'une mosaïque noire sur fond blanc ou blanche sur fond noir ? Plus grave encore, l'auteur nous faisait même douter du bien fondé d'une telle typologie.

Figure 4 : Mosaïque noire sur fond blanc ou blanche sur fond noir ? [Bruneau76]

Dans un tel cas, nous devrons disposer d'un modèle permettant d'exprimer qu'un premier point de vue affirme qu'il s'agit d'une mosaïque noire sur blanc, qu'un second affirme l'inverse, et qu'un troisième propose une typologie toute autre.

Les deux premiers points de vue étant contradictoires, notre « modèle de connaissance » devra être beaucoup plus permissif que la normale :

- Les structures seront non hiérarchiques (graphes orientés acycliques),
- Il n'y aura pas de négation (donc pas de principe de tiers-exclu, ni de principe de non-contradiction),
- Les points de vue ne seront pas dépendants les uns des autres, si ce n'est par l'intermédiaire des corpus décrits.

Du fait que le troisième des points de vue remette en cause la typologie utilisée dans les deux premiers, nous ne pourrons plus considérer qu'il existe *un* modèle *fixe* du domaine, mais plutôt des modèles *hypothétiques* et *transitoires*, évoluant de pair avec leurs instances. La séparation des classes et des instances en deux espaces apparaît par conséquent inutile. De manière plus générale, l'aspect dynamique de la modélisation empêchera un typage trop fort des primitives.

Les descriptions n'étant plus normées, il sera impossible de connaître *a priori* leur forme. Les interactions homme-machine ne devront donc pas suivre le modèle question-réponse mais plutôt celui de la navigation. La recherche de documents se fera de manière itérative et ira de pair avec une découverte de la structure du corpus. Dans une telle approche, la description du document sera un sous-graphe de la description du corpus. En ce sens, nous nous rapprocherons un peu des techniques qui visent à agréger des graphes disjoints afin de donner une vision d'ensemble [Chalendar97, PredigerEtWille99, BurrowEtEklund94, EklundEtCole02].

Enfin, le fait que les experts ne soient pas des professionnels de la modélisation, nous encourage à proposer un modèle dont l'utilisation pour des descriptions simples sera assez intuitive, et dans lequel, il sera possible, moyennant une formation, d'établir des descriptions plus précises.

3. Les réseaux de description

3.1 Un modèle à base de points de vue

Notre modèle appelé « réseau de description » se présente sous la forme d'un graphe orienté acyclique (cf. figure 5). Les nœuds sont appelés des « descripteurs » et les arcs des « spécialisations ». Un arc orienté entre les descripteurs *A* et *B* se lit : « tout objet documentaire décrit par *B* l'est aussi par *A* ».

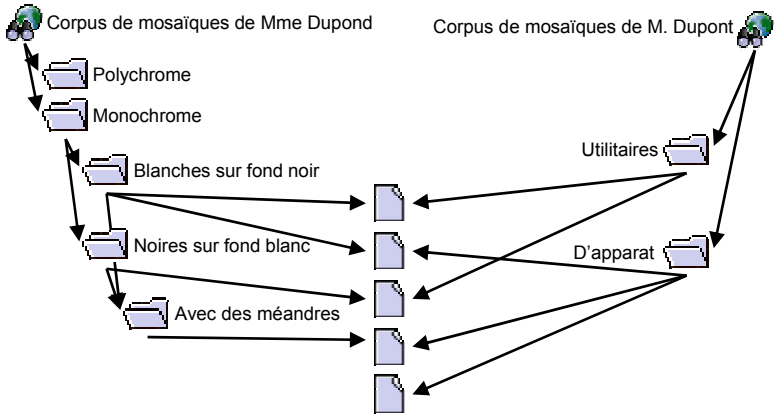


Figure 5 : Exemple de réseau de description

Il est important de mentionner que seul les nœuds et les arcs sont utiles pour le système. Mais de sorte que les usagers puissent interpréter le réseau, nous associons à chaque nœud une étiquette et à chaque arc son historique. Du moment que la définition formelle des arcs est respectée, l'usager est libre d'utiliser ces arcs pour modéliser des taxinomies, des méréonymies, des instanciations…


Parmi les descripteurs (cf. figure 6), certains ne sont pas généralisables : on les appelle des « facettes ». Chacun correspond à un point de vue indépendant.


D'autres ne sont pas spécialisables, on les appelle des « identifiants ». Chacun fait référence à un objet documentaire unique.


Les autres sont appelés « descripteurs ordinaires ».

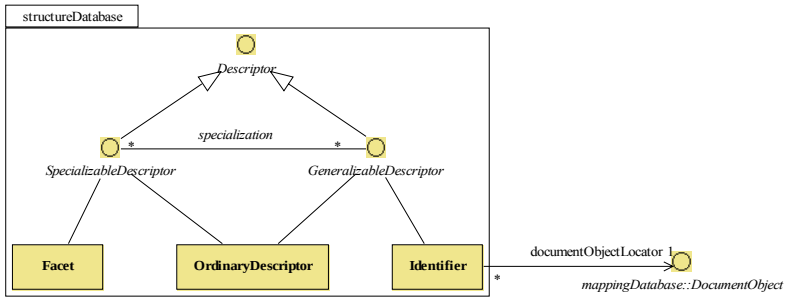


Figure 6 : Réseau de description (diagramme de classe UML)

Plus formellement, nous avons affaire aux ensembles suivants :

- *Descriptor*,
- *SpecializableDescriptor*,
- *GeneralizableDescriptor*,
- *Identifier*,
- *Facet*,
- *DocumentObject*.

Ces ensembles satisfont axiomatiquement les relations suivantes :

$Descriptor = SpecializableDescriptor \cup GeneralizableDescriptor$
$Identifier \subset GeneralizableDescriptor$
$Facet \subset SpecializableDescriptor$

Nous allons maintenant définir par des spécifications algébriques :

- Le schéma des données à stocker (primitives),
- Les contraintes supplémentaires que ces données doivent respecter (contraintes),
- Les requêtes complexes qui seront effectués sur ces données (définitions).

Primitive : *specialization*(_,_) : $SpecializableDescriptor \times GeneralizableDescriptor \rightarrow Boolean$

Cette relation exprime qu'il existe une spécialisation du premier descripteur vers le second.

Primitive : *_.getDOI* : $Identifier \rightarrow DocumentObject$

Cette fonction permet d'obtenir le document correspondant à un identifiant donné.

Définition : *describes*(_,_) : $Descriptor \times GeneralizableDescriptor \rightarrow Boolean$

Cette relation est construite de sorte qu'elle soit réflexive et qu'elle constitue la fermeture transitive de la relation « spécialization ».

$describes(x,x)$
$describes(x,y) \leftarrow specialization(z,y) \wedge describes(x,z)$

Contrainte : « Acyclicité »

Aucun cycle ne doit exister dans le réseau de description.

$\perp \leftarrow specialization(x,y) \wedge describes(y,x)$

Contrainte : « Enracinement »

Un descripteur ne doit pas appartenir à plusieurs facettes.

$\perp \leftarrow f_1 \in Facet \wedge f_2 \in Facet \wedge f_1 \neq f_2 \wedge describes(f_1,x) \wedge describes(f_2,x)$

3.2 Arpenter l'espace documentaire

Les réseaux de description présentés ci-dessus permettent à chaque expert d'exposer son point de vue, sa théorie sur une partie de la discipline. Le but, ici, n'est pas de glorifier une subjectivité débridée, mais d'autoriser le débat, pour viser l'intersubjectivité. Il s'agit donc de pouvoir comparer entre eux ces points de vue.

Le mécanisme que nous avons offert aux utilisateurs est un filtre de graphes. Il permet par induction totalisante de trouver des rapports entre descripteurs, non-dits au niveau des modèles, mais apparaissant dans leurs usages. Pour reprendre notre exemple de typologies de mosaïques, le système nous indiquerait que lorsque tel auteur décrit les mosaïques comme blanche sur fond noir, un autre les décrit « toujours » (ou « parfois », ou « jamais ») comme des mosaïques noires sur fond blanc. De plus, en ne présentant à un moment donné qu'une partie du réseau de description (dont la taille augmentera au fur et à mesure de son utilisation), nous espérons aider l'expert à arpenter l'espace documentaire.

Nous allons maintenant voir plus précisément comment ce filtre de graphe est obtenu.

Définition : $\_.getCorpus : Descriptor \rightarrow DocumentObject^n$

Par cette fonction, on obtient par récursivité l'ensemble des objets documentaires décrits directement ou indirectement par un descripteur donné.

$x.getCorpus = \{y\} \leftarrow x \in Identifier \wedge x.getDOI = y$

$x.\ getCorpus = \{z \mid specialization(x,y) \wedge z \in y.\ getCorpus\}$
$\leftarrow x \in SpecializableDescriptor$

Définition : $\_.getCorpus : Descriptorn \rightarrow DocumentObject^n$

On généralise la fonction homonyme à une sélection de plusieurs descripteurs. L'intersection des corpus signifie qu'être décrit par une sélection de descripteurs revient à être décrit à la fois par chacun d'eux.

$\{d_0,..d_n\}.getCorpus = d_0.getCorpus \cap ..d_n.getCorpus$

Axiome

Pour une sélection donnée, l'état d'un descripteur prendra sa valeur parmi « connu », « possible » ou « impossible ».

$State = \{KNOWN, POSSIBLE, IMPOSSIBLE\}$

Définition : $\_.getState(\_) : Descriptor \times DocumentObject^n \rightarrow State$

Il s'agit maintenant d'attribuer un état (*connu*, *possible*, *impossible*) à un descripteur pour un corpus *C* sélectionné.

Plus précisément, on dira que le descripteur est *connu* s'il décrit *tous* les objets documentaires de *C*.

$$x.getState(C) = KNOWN \leftarrow C \neq \emptyset \wedge C \subseteq x.getCorpus$$

Il sera *impossible* si le descripteur ne décrit *aucun* objet de *C*.

$$x.getState(C) = IMPOSSIBLE \leftarrow C \cap x.getCorpus = \emptyset$$

Il sera *possible* si le descripteur décrit *quelques* objets de *C* (mais pas tous).

$$x.getState(C) = POSSIBLE \leftarrow x.getState(C) \neq IMPOSSIBLE \wedge x.getState(C) \neq KNOWN$$

Définition : $\_.getFilter(\_) : Descriptor \times DocumentObject^n \rightarrow (Descriptor \times State)^n$

On filtre le réseau en descendant récursivement dans les descripteurs connus et en s'arrêtant aux descripteurs possibles et impossibles.

$$x.getFilter(C) = \{(x,IMPOSSIBLE)\} \leftarrow x.getState(C) = IMPOSSIBLE$$
$$x.getFilter(C) = \{(x,POSSIBLE)\} \leftarrow x.getState(C) = POSSIBLE$$
$$x.getFilter(C) = \{(x,KNOWN)\} \cup \{(z,s) \mid specialization(x,y) \wedge (z,s) \in y.getFilter(C)\} \leftarrow x.getState(C) = KNOWN$$

Définition : $\_.getDescriptionContext(\_) : Facet^n \times Descriptor^{n \times n} \rightarrow (Descriptor \times State)^{n \times n}$

Voici comment obtenir le contexte de description pour un ensemble de facettes et de sélections associées. On calcule d'abord le corpus global par intersection des corpus de chaque facette. Ensuite, on applique à chaque facette le filtre correpondant au corpus global.

$$\{f_0,..f_n\}.getDescriptionContext(\{S_0,..S_n\}) = \{f_0.getFilter(C), .. f_n.getFilter(C)\} \leftarrow C = f_0.getCorpus(S_0) \cap .. fn.getCorpus(S_n)$$

Les figures 7 et 8 illustrent l'obtention du contexte de description et ses optimisations.

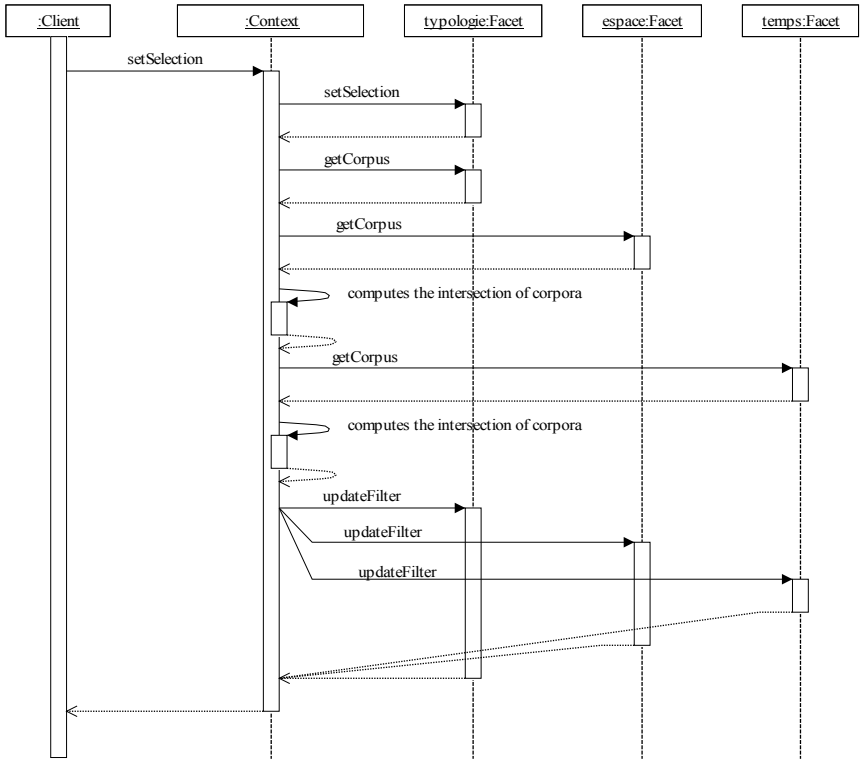


Figure 7 : Mise à jour des filtres dans les facettes « typologie », « espace » et « temps » après changement de sélection dans la facette typologie (diagramme de séquence UML)

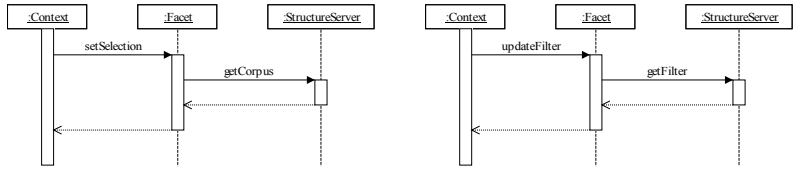


Figure 8 : Requêtes envoyées au serveur lors de la mise à jour des filtres (cf. figure précédente)

4. Etude de cas

Nous avons souhaité éprouver notre prototype, appelé *Porphyry*, en nous mettant « à la place » d'un archéologue. Nous sommes conscients de la portée très relative d'une telle expérience. Toutefois, il nous semble que les défauts du modèle qui pourraient apparaître dans notre usage du prototype devraient *a fortiori* causer des problèmes aux archéologues et avoir ainsi valeur de réfutation.

Cette étude de cas portera sur les recherches d'Andrea Iacovella concernant la nécropole occidentale de Mégara Hyblaea (Sicile). Dans un premier temps, nous essaierons de nous mettre dans la situation du chercheur en présentant ses objectifs et méthodes. Dans un second temps, nous verrons les problèmes rencontrés avec la précédente version du prototype et surtout la difficulté de faire une description avancée sans être guidé.

4.1 « Fouiller » un rapport de fouille

A la croisée de l'archéologie, de l'historiographie[4] et des sciences cognitives, les travaux d'Andrea Iacovella visent à analyser le discours des archéologues [OrsiEtCavallari1892] afin d'en extraire de nouvelles conclusions archéologiques. En quelque sorte, il s'agit de refaire, virtuellement, une fouille effectuée au siècle dernier.

Après avoir développé dans sa thèse une approche quantitative (à l'aide de statistiques descriptives), Andrea Iacovella souhaitait passer à une approche plus qualitative. Dans cette perspective, une première « modélisation cognitive » d'une vingtaine de descriptions de sépultures avait été effectuée [Dubois99]. Les modèles prenaient la forme de diagrammes d'instance UML étiquetés avec les termes de Paolo Orsi (traduits de l'Italien en Français). Pour notre part, nous avons intégré dans *Porphyry* le rapport de fouille (sous forme de fac-similés de pages) et avons « traduit » les modèles UML en réseaux de description.

Nous avons considéré un certain nombre de facettes pour décrire le corpus (cf. Figure 9) : une première pour sa structure typographique en colonne, une seconde pour sa structure spatiale (sépultures) et trois autres pour la typologie des tombes, des ossements et du matériel trouvés dans ces sépultures. Volontairement, nous avons adopté un modèle simple dans lequel la composition de descripteurs ne se fait qu'au niveau du fragment (intersection du découpage par colonne et par sépulture) et où les adjectifs n'étaient pas pris en compte. Nous plaçant dans la perspective d'une modélisation dynamique, il semblait en effet naturel de commencer par des modèles « naïfs » et de les affiner par la suite.

[4] Historiographie : Etude de l'écriture de l'Histoire.

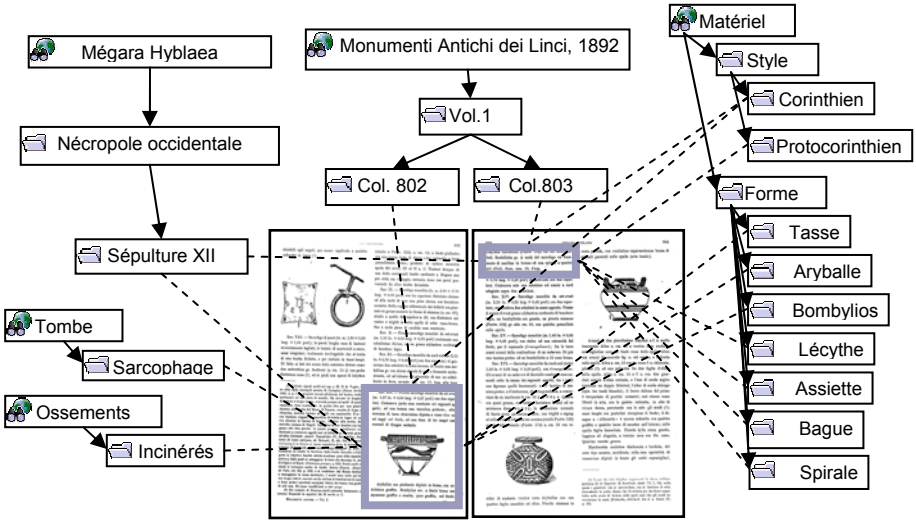


Figure 9 : Explication par le lecteur d'une section concernant une sépulture donnée.

4.2 Retour d'expérience

Le filtre de graphe appliqué à ce réseau de description permit d'observer un certain nombre de propriétés en résonance avec les préoccupations d'Andrea Iacovella. Dans la Figure 10, il apparaît que des ossements d'enfants ont été identifiées dans la sépulture IV (entre autres) mais pas dans la sépulture V. De plus, ces sépultures d'enfant présentent toutes des traces d'incinération et aucune ne contient de masque féminin ou d'aiguille. L'archéologue pourrait alors se demander si le matériel de la sépulture est déterminé par l'age du défunt. A l'inverse, l'historiographe, pourrait se demander si, dans le cas d'incinérations (donc en l'absence de squelette), ce n'est pas le matériel qui permit au fouilleur de déterminer l'age du défunt.

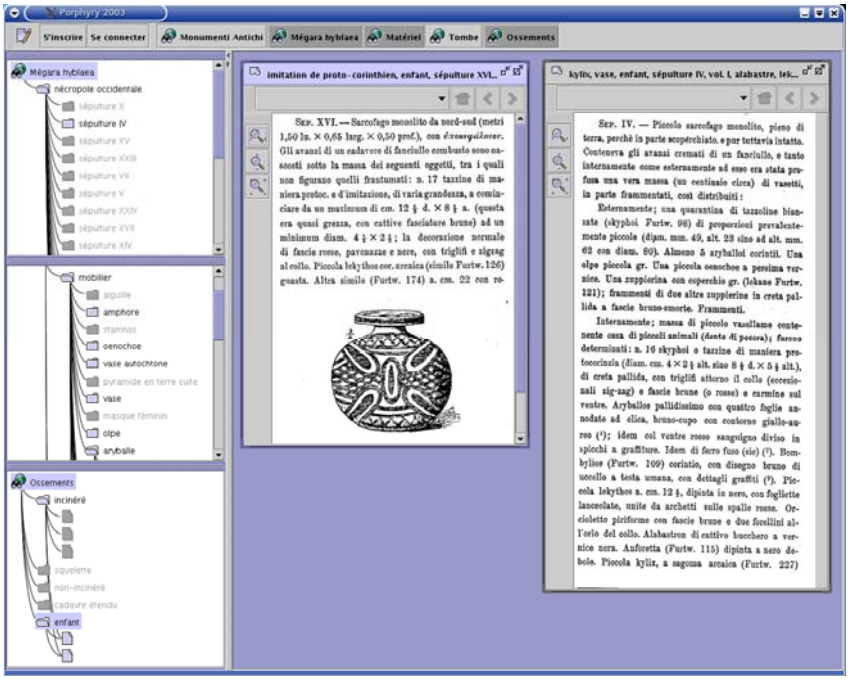


Figure 10 : Feuilletage du corpus enrichi par le lecteur (avec filtrage des réseaux de description en fonction de la sélection « ossements/enfants »)

Pour approfondir cette étude, l'archéologue pourrait s'intéresser à la répartition spatiale des tombes d'enfants par rapport aux tombes d'adultes [Iacovella97]. Il faudrait alors ajouter au corpus une carte de la nécropole, créer un fragment pour chaque emplacement de sépulture, et associer ce fragment au descripteur correspondant à la sépulture. Ensuite, le fait d'afficher ces fragments en contexte avec leur source pourrait dessiner automatiquement la position des fragments dans leur source et donc la carte de répartition des tombes d'enfants. Cette dernière fonction n'existant pas dans la version du prototype utilisée lors du test, elle fut ajoutée depuis.

Continuons à nous mettre à la place de l'archéologue. Celui-ci pourrait par exemple se demander si la petite taille du matériel est, elle aussi, corrélée avec le jeune âge du défunt. Mais comment modéliser l'adjectif « petit » ? Nous sommes bien au-delà de la modélisation naïve que nous préconisions au début de cette section. Les résultats de cette expérimentation nous encouragent à offrir aux experts qui le souhaitent une formation avancée sur la description de documents. Aussi, un certain nombre de séminaires ont été mis en place dans nos plateformes d'expérimentation (notamment au CRATA, équipe de recherche travaillant sur l'antiquité classique à l'Université Toulouse le Mirail).

5. Conclusion

Nous avons tout d'abord rappelé que l'organisation des bibliothèques (comme celle des documents) est en général basée sur un modèle arborescent, probablement en raison de l'influence de la philosophie d'Aristote sur notre manière de voir la connaissance. De manière à sortir de ce schéma simpliste, nous avons étudié certaines alternatives. Cependant, du fait qu'elles s'appuient toujours sur un modèle du domaine considéré comme fixe et extérieur, ces alternatives nous ont semblé telles quelles inapplicables à notre approche : celle d'une modélisation dynamique effectuée par les experts eux-mêmes et non par des tiers.

Nous avons ensuite proposé un modèle appelé « réseau de description » permettant à chaque expert de superposer au corpus sa propre structure, son propre point de vue. Nous avons vu comment, grâce à un mécanisme de filtre inductif, on pouvait assister l'expert à arpenter l'espace documentaire conjointement à travers plusieurs points de vue.

Enfin, nous avons exposé une étude de cas portant sur une lecture historiographique de la publication d'une fouille de nécropole. Nous avons considéré les structures suivantes : la structure bibliographique du rapport (en colonne), la structure par sépulture et la typologie des vestiges (tombes, mobilier, ossements). Cette expérimentation nous a encouragé d'une part à revoir notre gestion des contextes de lecture (pour afficher par exemple une carte de répartition du matériel archéologique) et, d'autre part, à rédiger un « guide des bons usages » à l'intention des experts souhaitant créer des modèles complexes.

Si la modélisation à base de points de vue semble prometteuse, il semble cependant nécessaire de l'accompagner d'autres assistants. Un premier consisterait en la création d'un espace intersubjectif permettant d'expliciter les relations entre points de vue et de faire ressortir ainsi les zones d'achoppement. Un second correspondrait à la constitution d'un espace diachronique offrant la possibilité de visualiser la dynamique des points de vue. Ces deux aspects sont actuellement à l'étude dans le cas de la manipulation par l'archéologue du document d'architecture et du temps archéologique.

6. Références bibliographiques

[Aristote-300] Aristote, *Organon : I. Catégories ; II. De l'interprétation* (Trad. J. Tricot), Paris : Vrin, 1959, 153 p.

[Bruneau76] Bruneau Ph., « Quatre propos sur l'archéologie nouvelle » [en ligne], In : *Bulletin de Correspondance Hellénique*, n°100, Athènes : Ecole française d'Athènes, 1976, pp.103-130. Disponible sur Internet : <http://cefael.efa.gr/horde/rayge/detail.php?site_id=1&actionID=page&series_id=BCH&volume_number=100&issue_number=1&startpos=105>

[BurrowEtEklund94] Burrow A., Eklund P.W., "Visual structure representations and conceptual graphs" [en ligne], In: *Proceedings of the fourth international*

workshop on Peirce: A conceptual graph workbench, Maryland, August 19, 1994, pp. 4-10. Disponible sur Internet : <http://citeseer.nj.nec.com/ellis94 proceedings.html>

[Bush45] Bush V., "As we may think" [en ligne], *The Atlantic Monthly*, #176, July 1945, pp. 101-108. Disponible sur Internet : <http://www.theatlantic.com/unbound/ flashbks/computer/bushf.htm>

[Chalendar97] Chalendar G.(de), *Abstractions de schémas à partir de situations agrégées* [en ligne], DEA de sciences cognitives, Universités Paris-Sud XI, 1997. Disponible sur Internet : <http://www.limsi.fr/Individu/gael/MemoireDEA>

[Dubois99] Dubois F., *Archéologie et mode de formation de la nécropole : une approche cognitive*, Rapport de projet de fin d'études, EFA, 1999, 8 p. + Annexes.

[EklundEtCole02] Eklund P., Cole R., "Structured ontology and information retrieval for email search and discovery" [en ligne], In: *Proceedings of the thirteenth International Symposium on Methodologies for Intelligent Systems* [ISMIS'2002], Lyon, June 26-29, 2002, Lecture Notes in Artificial Intelligence #2366, Berlin : Springer-Verlag, pp. 75-84. Disponible sur Internet : <http://link.springer.de/link/service/series/0558/papers/2366/23660075.pdf>

[Genest00] Genest D., *Extension du modèle des graphes conceptuels pour la recherche d'informations*, Thèse de doctorat en Informatique, Université Montpellier II, 2000, 181 p.

[Iacovella97] Iacovella A., « Etudes des proximités dans l'espace funéraire : Le cas de la nécropole occidentale de Mégara Hyblaea », *Archeologia e Calculatori*, 8, 1997, pp. 67-102.

[Martin96] Martin P., *Exploitation de graphes conceptuels et de documents structurés et hypertextes pour l'acquisition de connaissances et la recherche d'informations* [en ligne], Thèse en informatique, Université de Nice - Sophia Antipolis, 1996. Disponible sur Internet : <ftp://ftp.inria.fr/INRIA/publication/Theses/TU-0431>

[MechkourEtAl95] Mechkour M., Berrut C., Chiaramella Y., "Using a conceptual graph framework for image retrieval", In: *The International Conference on Multi-Media Modeling* MMM'95, Nov. 14-17, 1995, pp. 127-142.

[OrsiEtCavallari1892] Orsi P., Cavallari F.S., "Megara Hyblaea", In : *Monumenti Antichi dei Linci*, 1, 1892, Colonnes 799-818 (extrait).

[PredigerEtWille99] Prediger S., Wille R., "The lattice of concept graphs of a relationnaly (sic) scaled context" [en ligne], In: *Seventh International Conference on Conceptual Structures*, LNCS #1640, Berlin : Springer-Verlag, 1999, pp.401-414. Disponible sur Internet : <http://wwwbib.mathematik.tu-darmstadt.de/ Math-Net/Preprints/Listen/files/2033.ps.gz>

D’un corpus d’images à une base d’images :

Une plateforme combinant syntaxe et sémantique et une méthodologie de prototypage

L. Besson, A. Da Costa, E. Leclercq, M.N. Terrasse

Laboratoire LE2I – Université de Bourgogne - France
prenom.nom@u.bourgogne.fr

Résumé :

Avec la multiplication des domaines qui utilisent l'image comme support d'information, le besoin de créer des bases d'images s'amplifie. Les techniques de recherche d'images par le contenu deviennent cruciales dans un tel contexte. Cependant, si les techniques proposées sont de plus en plus efficaces, elles sont aussi de plus en plus variées et spécialisées. Par conséquent, la tâche du concepteur de bases d'image -qui doit effectuer un choix difficile et lourd de conséquences- est ardue. L'objectif de cet article est de présenter une plate-forme qui permet d'appliquer aux bases d'images les méthodes de développement des systèmes d'information.

Cette plate-forme repose sur un modèle générique permettant de représenter les images de façon globale ou locale, en utilisant des attributs syntaxiques ou sémantiques, et avec des relations spatiales ou sémantiques entre objets de l'image. Le modèle générique doit être instancié pour une application donnée. L'instanciation du modèle est propagée vers le corpus d'images pour définir l'interface d'acquisition : choix des algorithmes d'extraction, stratégie de combinaison des extracteurs puis utilisation d'extracteurs pré-existants afin de produire les résumés des images. De même l'instanciation du modèle est propagée vers les utilisateurs pour définir la ou les interface(s) de recherche : indexation et classification des images. Une telle plate-forme doit permettre d'une part de séparer les phases de conception et d'implémentation de la base d'image, d'autre part de construire rapidement un prototype qui pourra être validé avant la mise en exploitation.

Mots-clés : Ingénierie des bases d'image - Recherche par le contenu Représentation de la syntaxe et de la sémantique.

Abstract:

With the increasing number of domains that use images as an information media, the need for establishing image databases grows. Content-based image retrieval is becoming crucial in such a context. However, as the proposed search techniques are increasingly efficient, they are also increasingly diverse and specialized. Thus, an image database designer faces difficult choices. Our objective is to present a framework which first enables image database designers to use information system engineering methodologies. Furthermore, our framework allows designers to combine syntactical and semantic descriptions of images.

Our framework is based on a generic model for representation of images. By using this model, it is possible to represent images either globally or locally (as a hierarchy of objects/zones), with syntactical and semantic attributes, with spatial and semantic relations between objects/zones. Such a generic model needs to be instantiated for a given application. All the choices made at the model instantiation are propagated to the lower level of our framework in order to generate an extraction interface. These choices are also propagated to the upper level of our framework in order to generate search interfaces (indexing and classification of images). Such a framework enables image database engineers to design, prototype and tests and then implement and optimize their databases.

Keywords : Image database engineering - Content-based image retrieval - Syntax and semantics representation.

1. Introduction

Qu'elle appréhende directement la réalité (comme une radiographie) ou qu'elle ait un rôle symbolique (comme les peintures rupestres), l'image est un important vecteur de transmission d'information. Avec le développement de nouvelles techniques de production, de diffusion et de traitement d'images, les techniques de recherche dans des bases d'images sont un des composants (et des problèmes) majeurs des technologies pour l'image.

La première difficulté (intrinsèque) des mécanismes de recherche d'images dans une base est liée à un double problème de volume : volume de la base en terme du nombre d'images qu'elle contient et volume de chaque image qui est une somme complexe d'informations. Les interfaces des bases de données images proposent donc deux approches pour traiter les problèmes de volume. La première approche consiste à diminuer virtuellement le volume de la base d'images en ne considérant plus comme unité d'accès une image mais un groupe d'images considérées comme similaires : c'est l'approche **classification**. La seconde approche consiste à diminuer virtuellement le volume d'une image en réduisant chaque image à un résumé : c'est

l'approche **indexation**. Dans les deux cas il est essentiel de disposer d'un mécanisme permettant d'extraire de chaque image une partie significative « minimale et suffisante » afin de construire le résumé de l'image et, éventuellement, de décider si deux images appartiennent ou non à la même classe. Le coeur des deux approches est donc bien le même : extraire d'une image une sous-description représentative (le résumé de l'image) et être capable d'induire une similarité entre images à partir d'une similarité entre résumés. Les mécanismes permettant d'extraire un résumé à partir d'une image sont généralement classés en deux catégories : les mécanismes syntaxiques et les mécanismes sémantiques (dont les noms ont varié d'un auteur à l'autre : objets de l'image et objets du domaine [GWJ91], représentations numériques ou symboliques, notions d'image et image virtuelle [PSTT01], etc.). Dans la suite nous parlerons de zones dans le cadre d'une approche purement syntaxique et d'objets le reste du temps. L'approche syntaxique consiste à réduire l'image (ou une zone de l'image) à un vecteur de paramètres physiques : les paramètres choisis ainsi que leur domaine peuvent varier considérablement d'une approche à l'autre. Certains systèmes utilisent plutôt la couleur [WJ97], d'autres plutôt les contours ou les textures [FFN+93}, d'autres des caractéristiques propres au signal [LTMD01]. Cette approche syntaxique présente l'avantage d'être entièrement automatisable via des algorithmes de traitement d'images. Les approches sémantiques consistent à réduire l'image à un ensemble d'objets sémantiques (identifiés par leur signification dans le monde réel) reliés éventuellement par des relations sémantiques (qui correspondent à une interprétation de l'image comme reproduction d'une partie du monde réel). Cette approche nécessite l'intervention d'un expert et l'existence d'un modèle permettant de représenter la complexité du monde réel qui fait référence.

La seconde difficulté – extrinsèque celle-là – relève de la prise en compte précise des besoins des utilisateurs. Ce problème a été étudié par Denos [DBM97]. Denos formalise la différence d'évaluation de la pertinence d'une réponse à une question par un humain (qui va travailler en fonction des données disponibles et du contexte de la recherche : c'est la notion de pertinence utilisateur) et par un système informatique (qui sera limité aux données disponibles : c'est la notion de pertinence système). Dans le cadre d'un système fermé (toutes les images font référence au même champ sémantique, tous les utilisateurs ont les mêmes objectifs et la même expertise par rapport aux images), la différence entre la pertinence utilisateur et la pertinence système peut être contrôlée. Par contre, pour un système ouvert (en particulier dans le cadre d'applications réparties telles que le web) il est impossible de contrôler la différence entre la pertinence utilisateur et celle (figée) du système. Par exemple, en mode syntaxique, on peut espérer trouver des paramètres qui correspondent aux critères d'analyse des experts du domaine ; ces critères risquent fort d'être totalement indécryptables par des utilisateurs non experts. De même en mode sémantique, les annotations des experts d'un domaine peuvent ne pas être discriminantes pour les experts d'un autre domaine (ou pour des utilisateurs quelconques). Després [DL00] en montre des exemples dans le cadre des systèmes bibliographiques, les ontologies en sont aussi des exemples flagrants [Gua95]. Santini *& al.* vont même plus loin en considérant qu'une partie de la sémantique

d'une base de données provient de l'interaction des utilisateurs avec cette base [SGJ01]

La mise en oeuvre d'une base d'images se réduit donc plus ou moins à un choix – difficile – entre un mécanisme stable mais de qualité moyenne (approche syntaxique) et un mécanisme instable avec de bons résultats ponctuels et un risque de totale inadéquation (approche sémantique). Quelques propositions ont cependant été faites pour tenter de diminuer l'écart entre ces deux extrêmes tels que le bouclage de pertinence ou l'introduction de thesaurus et de systèmes de déduction [Fou02]. En pratique, si on fait l'hypothèse que les techniques de recherche par le contenu sont assez efficaces pour satisfaire la plupart des besoins, le concepteur d'une base d'images reste néanmoins aux prises avec d'une part un corpus d'images et d'autre part les besoins des utilisateurs. Les caractéristiques du corpus d'images et du domaine d'application doivent être exploitées au mieux afin de satisfaire l'ensemble des utilisateurs. Par exemple, Chabot [OS95] utilise des propriétés sémantiques (annotations) et des connaissances afin de piloter l'analyse des images ; le système KMeD [TCC+94] utilise une architecture multi-couches basée sur les contours, les formes et les objets, spécialisée pour le domaine médical et qui permet une identification quasi automatique des objets ; le système PICTION [SB94] combine les annotations textuelles et les techniques de traitement d'images pour proposer un module de traitement du langage naturel permettant l'évaluation des requêtes utilisateur.

Il est donc difficile de trouver le point d'équilibre entre toutes les contraintes pesant sur une base d'images tout en exploitant au mieux l'information potentiellement disponible. Nous pensons que les méthodologies mises en oeuvre dans d'autres domaines des systèmes d'information peuvent être, avec grand profit, appliquées à la construction des bases d'images. A cet effet, nous proposons une plate-forme de prototypage qui permet de passer – à moindre coût – d'un corpus d'images à une base d'images correctement profilée pour ses utilisateurs potentiels. Cette plate-forme a pour coeur un **modèle générique** qui doit permettre de représenter les images de la plupart des applications. Ce modèle permet de décrire les images soit *globalement* (l'image est vue comme un tout, une unité de description) soit *localement* (l'image est constituée d'objets/zones individuellement identifiables). Les attributs associés aux images et aux objets/zones de l'image peuvent être soit *syntaxiques* (e.g., couleur, intensité, forme) soit *sémantiques* (e.g., rivière, bâtiment, route). Dans le cas des descriptions locales, *plusieurs types de relations entre objets/zones* d'une même image peuvent être mis en place : composition (de l'image à partir des objets/zones), relations spatiales (e.g., au nord de, au sud-est de) et topologiques (e.g., voisinage, intersection), relations sémantiques (e.g., la porte fortifiée du camp romain). Dans le cadre d'une application donnée, ce modèle est instancié : on choisit les attributs (syntaxiques, sémantiques ou mixtes) et les relations qui vont constituer les résumés des images. En utilisant la **couche basse de la plate-forme**, l'instanciation du modèle est propagée pour définir l'interface d'acquisition des images. En utilisant la **couche haute de la plate-forme**, l'instanciation du modèle est propagée pour définir l'interface de recherche.

2. Description de la plate-forme et première approche de la méthodologie associée

L'architecture de notre plate-forme est illustrée en figure 1. Son noyau est constitué d'un couple **modèle générique et bibliothèque de mesures de similarité**. Le modèle générique permet – dans un formalisme unique – une description des images qui est adaptée à la fois à un traitement purement basé sur la syntaxe (comme par exemple les transformées en ondelettes) et à un traitement à forte composante sémantique. Notre modèle générique doit être instancié pour un corpus d'images donné et un jeu de fonctionnalités demandées par les utilisateurs potentiels, c'est-à-dire en fonction des informations disponibles, visibles dans les images d'une part et en fonction d'autre part des informations nécessaires aux traitements qui devront être effectués. La description d'une image dans ce modèle instancié constitue le résumé de l'image.

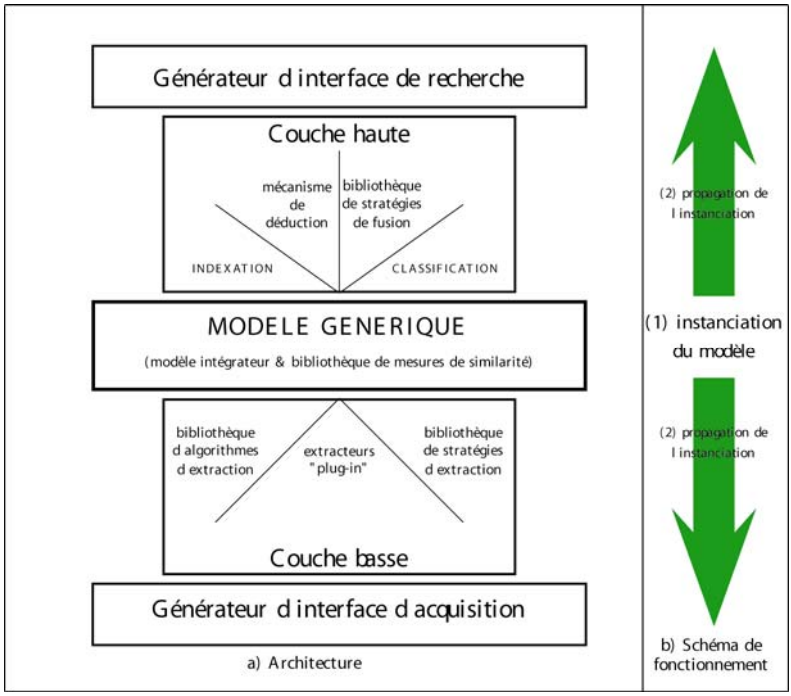


Figure 1 : Présentation de notre plate-forme

L'instanciation du modèle peut alors être propagée vers le corpus d'images (en utilisant les ressources de la **couche basse** de notre plate-forme) pour mettre en place l'**interface d'acquisition** des images. Le concepteur de la base d'image utilise la bibliothèque d'extracteurs et de stratégies d'extraction pour définir puis implémenter les outils d'extraction qui vont construire les résumés des images du corpus. Le choix des algorithmes d'extraction, la stratégie de combinaison des extracteurs sont déterminés dans une phase d'analyse. La phase d'implémentation « combine » des extracteurs pré-existants (fournis avec la plate-forme) ou ad hoc (développés spécialement pour l'application concernée) pour mettre en place des outils d'extraction qui vont construire les résumés des images du corpus.

De même, l'instanciation du modèle est propagée vers les utilisateurs via la **couche haute** de notre plate-forme – qui offre des mécanismes d'indexation et de classification, des mécanismes de déduction, une bibliothèque de stratégies de fusion syntaxique/sémantique – afin de mettre en place l'**interface de recherche** de la base d'images (ou éventuellement les interfaces de recherche appropriée aux différents types d'utilisateurs).

Dans la suite de cette section, nous donnons quelques détails sur la couche basse de la plate-forme et sur la méthode de construction de l'interface d'acquisition.

La couche basse de notre plate-forme

Ainsi que nous l'avons brièvement décrit ci-dessus, le modèle instancié qui a été choisi pour une application donnée détermine le contenu des résumés des images (en termes d'attributs et de relations entre objets). Il est donc nécessaire de mettre en oeuvre divers extracteurs d'information pour construire les résumés à partir des images. Nous distinguons trois types d'extracteurs qui peuvent être automatiques, semi-automatiques ou manuels (e.g., la plupart des annotateurs sémantiques). Le premier type, dit des *extracteurs syntaxiques*, calcule des paramètres physiques tels que couleur moyenne, énergie, etc. Le second type, dit des *extracteurs géométriques*, isole dans l'image des formes géométriques simples (cercles, triangles, rectangles, etc.) à partir d'un contour ou d'une texture. Le troisième type, dit des *annotateurs sémantiques*, permet – par intervention d'un expert du domaine sur lequel porte la base d'images – d'ajouter des méta-données à l'image. Notre bibliothèque d'extracteurs est organisée comme décrit en figure 2. Le point d'accès privilégié est une bibliothèque d'algorithmes d'extraction (e.g., algorithme de Roth & Levine). Pour chacun de ces algorithmes sont données : la liste des paramètres qu'il extrait d'une image (e.g., couleur moyenne, histogramme d'intensité, relation de voisinage, datation, nature archéologique), la liste des paramètres dont il a besoin en entrée, la liste des modules de code implémentant cet algorithme. Les seuls paramètres considérés ici sont ceux qui doivent être extraits de l'image. Un dictionnaire de paramètres sert de référence afin de faciliter l'utilisation de cette bibliothèque. Cette bibliothèque est bien entendu évolutive : elle sera étendue au fur et à mesure des besoins et des expériences menées.

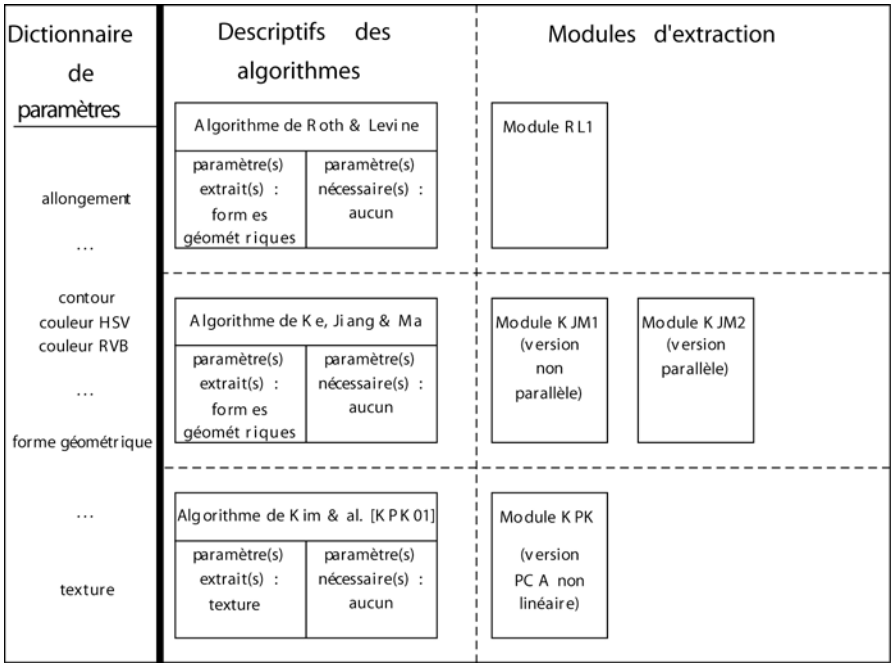


Figure 2 : Structure de la bibliothèque d'algorithmes d'extraction

Construction de l'interface d'acquisition

A partir de la liste des paramètres à extraire de l'image, le concepteur de la base d'image lance une construction incrémentale (automatique ou semi-automatique) de la liste des algorithmes d'extraction nécessaires. Ce processus consiste à supprimer à chaque étape au moins un paramètre de la liste, en ajoutant éventuellement de nouveaux paramètres à cette liste. Le processus est poursuivi jusqu'à une situation de point fixe pour laquelle il n'y a plus de paramètre à extraire qui ne soit pas couvert par au moins un des algorithmes retenus. A partir du moment où l'extraction de tous les paramètres est assurée, il reste à résoudre un ensemble de dépendances. Le concepteur de la base d'image lance alors un tri topologique [GT96]pour déterminer les ordres possibles pour l'enchaînement des algorithmes d'extraction. La dernière étape est le choix d'un des ordres possibles et la construction (par combinaison de composants) d'une interface d'acquisition à partir des modules d'implémentation des algorithmes sélectionnés et ordonnés. Au delà de la phase de prototypage, cette première version de l'interface sera optimisée. L'optimisation devra porter sur l'élimination des calculs redondants (à la charge de l'équipe d'informaticiens) ainsi que sur l'optimisation des modules d'extractions proprement dit (sous contrôle d'un expert en segmentation).

3. Le modèle de description des images

Ainsi que nous l'avons signalé, notre modèle décrit les images selon deux points de vue orthogonaux. Un premier point de vue traite de la granularité [AKDGB,ZAL01] de description des images : globale ou locale. Un second point de vue traite de la nature des attributs et relations utilisés pour la description : sémantiques ou syntaxiques. Le modèle que nous proposons est capable de répondre à toutes les combinaisons des niveaux de granularité avec la nature des informations représentées dans les attributs. Par exemple, une description **locale syntaxique** incluant la couleur est utilisée dans QBIC [GWJ91] ainsi que dans le projet IKONA [GB].

Nous avons testé un tel type de description avec des transformées en ondelettes appliquées aux images d'une base paléontologique. Chaque imagette produite par la transformée en ondelettes est considérée comme une « zone » de l'image. Les informations attachées à chaque zone sont calculées de façon automatique à partir des valeurs des paramètres de la transformée en ondelettes. De même, une description **globale syntaxique** peut être par exemple un histogramme de couleurs ou d'intensités comme dans ImageGrep [WJ97]. Un attribut syntaxique global est le plus souvent soit la valeur moyenne soit la valeur la plus significative d'un paramètre physique de l'image. Nous avons utilisé une description **locale sémantique** sur une base de vues aériennes de sites archéologiques. Dans une telle description les zones identifiées dans l'image sont des objets (ayant une interprétation dans le monde réel). La composition de l'image à partir des objets est hiérarchique. Une description **globale sémantique** est encore souvent utilisée pour décrire par des mots-clés les images d'une base mise à disposition d'experts du domaine d'application (e.g., base d'images d'une agence photographique, base d'images médicales) : chaque image étant représentée par des attributs date, lieu, sujet, etc. Les descriptions sémantiques globales sont en général créées par des experts du domaine qui annotent les images.

Le modèle que nous présentons est générique et multi-granulaire. La généricité repose sur la possibilité d'instancier plusieurs éléments de la description d'une image en fonction du domaine d'application et des fonctionnalités prévues. Le choix des attributs décrivant une image, par exemple, est fortement lié au domaine d'application. Pour une application comme la base de vues archéologiques, la couleur n'est pas utilisable pour décrire les objets. En effet cette base étant composée de vues aériennes prises avec différentes techniques (photographie couleur, noir et blanc ou infrarouge, etc.) ainsi qu'à différentes saisons, un même objet peut avoir des couleurs différentes d'une image à l'autre. En revanche, pour cette application, la forme est un paramètre essentiel. Nous appelons *paramètre discriminant* un paramètre qui est « syntaxiquement » détectable dans les images du corpus à traiter et *paramètre significatif* un paramètre qui a un sens dans le contexte de l'application.

La multi-granularité repose sur une organisation hiérarchique de la représentation d'une image. L'image est (éventuellement) décomposée en objets (ou zones) qui peuvent eux-mêmes être décomposés. Dans la suite nous parlerons d'objets (pour objets/zones) afin de ne pas alourdir inutilement le texte. A chaque objet sont associés des attributs syntaxiques et sémantiques. Deux objets peuvent être associés entre eux par des relations syntaxiques, spatiales et sémantiques. Dans la suite de cette section, nous présentons la formalisation de notre modèle générique. Nous décrivons ensuite ce qu'est la phase d'instanciation du modèle que nous illustrons par un exemple.

Un aperçu du modèle

La description des images est basée sur la description d'objets (*objets simples* détectables dans l'image, *objets complexes* qui sont composés à partir des objets simples, l'image elle-même qui est traitée comme un objet). Nous notons par O l'ensemble des objets apparaissant dans la description d'une image. Cet ensemble O est l'union de deux sous-ensembles O_s et O_c qui contiennent respectivement les objets simples et complexes. Dans le cas d'une description globale des images, l'ensemble O est réduit à un seul objet – considéré comme simple – qui est l'image elle-même.

Chaque objet de O est décrit par un tuple comprenant : un identificateur d'objet (noté *idf*), une géométrie (notée *geo*) qui doit permettre de retrouver l'ensemble des pixels de l'image associés à cet objet (cette géométrie est optionnelle), des tuples d'attributs physiques et sémantiques (respectivement notés *AttP* et *AttS*). Nous notons *Descr* l'ensemble des tuples décrivant les images et nous mettons ainsi en place une fonction de description d'objets notée d :

$$\begin{aligned} d : \quad & O \quad \rightarrow \quad Descr \\ & O \quad \rightarrow \quad < idf(o),\ geo(o),\ AttP(o),\ AttS(o) > \end{aligned}$$

Entre les objets de l'ensemble O nous pouvons définir plusieurs relations : relations de composition qui déterminent une vision hiérarchique de l'image, relations spatiales (distance, direction, relation topologique), relations sémantiques. Il peut y avoir plusieurs relations de composition dans la description d'une image ; ces relations correspondent alors à plusieurs points de vue sur la façon dont l'image est construite. Nous imposons qu'il existe pour chaque image une relation de décomposition privilégiée prenant en compte tous les objets simples et l'image elle-même. Les relations entre objets de O – quelque soit leur nature – sont représentées par des graphes étiquetés. Nous utilisons les définitions suivantes :

- Un graphe étiqueté R_i est un tuple $< ETIQ_i, A_i, \square_i >$ tel que :

  1) L'ensemble des étiquettes est noté $ETIQ_i$. Dans certains cas, cet ensemble sera réduit à une constante : par exemple pour certaines relations de composition pour lesquelles il n'y a pas différents cas de mise en relation de deux objets.
  2) L'ensemble des arcs A_i est un sous-ensemble de $O \times O$ ($A_i \subset O \times O$). En fonction de la nature de la relation, diverses contraintes peuvent être mises en place sur cet ensemble d'arcs A_i.
  3) La fonction $\square$ d'étiquetage $\square_i$ associe à chaque arc une étiquette : $\square$

- L'ensemble R des graphes étiquetés est noté $\{< ETIQ_i, A_i, \square_i >\}_{i=1..n}$. Pour faciliter les traitements ultérieurs, nous imposons que la numérotation des graphes étiquetés constitutifs de R permette de regrouper les relations de même nature (composition, spatiale, sémantique).

Phase d'instanciation du modèle

La phase d'instanciation du modèle a pour objectif de déterminer, pour une application donnée, la structure de description qui sera commune à toutes les images et à tous les objets des images ainsi que les relations entre objets qui vont être prises en compte. Pour cela nous proposons les étapes suivantes :

- Détermination (par un spécialiste de segmentation par exemple) des *paramètres discriminables* pour le corpus d'images. Il s'agit des paramètres qui sont calculables (si on doit extraire automatiquement les paramètres) ou visibles (si on travaille de façon manuelle ou semi-automatique).
- Détermination (par les spécialistes du domaine) des *paramètres significatifs* c'est-à-dire des paramètres qui sont sémantiquement utilisables. Il s'agit ici

d'éliminer les paramètres qui n'ont pas de sens pour le corpus d'images (par exemple parce que leurs valeurs ne sont pas cohérentes d'une image à l'autre).

- Détermination (par les analystes et concepteurs de la base) des *paramètres nécessaires* à la réalisation des fonctionnalités demandées.
- Choix du jeu de paramètres du modèle à partir des paramètres discriminants, significatifs et nécessaires. Cette étape consiste à trouver un équilibre entre ce qui est nécessaire et ce qui est disponible. Trois cas principaux peuvent se produire. Dans le cas le plus favorable, toutes les informations nécessaires sont disponibles. Dans le cas intermédiaire, certaines informations nécessaires sont couvertes par des paramètres discriminants mais pas significatifs. Il faut dans ce cas effectuer une analyse précise des méta-données associées aux images pour tenter de rendre significatifs les paramètres manquants. Dans le cas le plus défavorable, certaines informations nécessaires sont couvertes par des paramètres significatifs mais non discriminants. Dans ce cas, il faut soit améliorer la qualité du corpus d'images soit réduire le jeu de fonctionnalités qui sera proposé.

  Les paramètres ainsi choisis sont ensuite répartis dans les catégories syntaxique ou sémantique et les tuples correspondants sont définis. Pour chaque paramètre il reste à définir son ensemble de valeurs possibles.
- Si le modèle est local, les relations entre objets (autres que la relation de composition privilégiée) sont définies dans les trois catégories : composition, spatiales et sémantiques. Pour chacune des relations on définit l'ensemble des étiquettes qui sera associé à son graphe.

Exemple d'instanciation du modèle

Dans cette partie nous allons présenter un exemple d'instanciation du modèle pour une application d'archéologie aérienne. Cet exemple d'instanciation est tiré d'un projet consistant à passer d'un corpus de vues aériennes de sites de la région Bourgogne (sous forme d'images et de fiches techniques sur papier) à une base d'images numérisées et annotées[1].

Le corpus contient plus de 60 000 images et a été produit sur une période de trente ans environ dans l'objectif d'inventorier des vestiges archéologiques. La base d'images est très hétérogène en terme de qualité des images car elle comprend des vues prises sous différents angles (ce qui provoque des déformations sur les images), avec différentes techniques (photographie traditionnelle, photographie infrarouge etc.). A chaque image sont associées des informations sémantiques provenant de fiches techniques. Ces informations sont soit externes à l'image (telles que la date de prise de vue, le nom de l'opérateur, la localisation), soit liées à la nature archéologique du contenu ou

[1] Cependant, l'image utilisée pour illustrer cette instanciation provient d'une autre base d'image d'archéologie aérienne [Ua].

du contexte de la photo (les mots clés utilisés dans cette partie font référence à un thésaurus spécialisé).

Certaines des informations sémantiques sont elles aussi de qualité variable : lors de la prise de vue, toutes les images n'ont pas été localisées de manière exacte, pour certaines photos il n'y a aucune indication d'orientation, etc. L'exploitation de la base d'images passe par plusieurs types de requêtes tels que : la recherche d'images par des critères sémantiques globaux (on peut par exemple rechercher les images « *qui ont été prises pendant le mois d'octobre 1978* » ou bien rechercher les images « *qui ont été prises dans une rayon de 3 km autour du site romain des sources de la Seine* »), la recherche d'images par des critères sémantiques locaux (on peut par exemple rechercher les images « *qui contiennent une trace de motte féodal*e »), la recherche d'images en fonction de leur contenu (on peut ainsi vouloir rechercher toutes les images « *qui contiennent une zone avec une texture de pavage en galets* »).

Lors de la modélisation, l'image est décomposée en objets. Ces objets correspondent à des formes géométriques simples : bâtiments visibles en surface, infrastructures (telles que segments de routes, rivières, ponts)ou traces de bâtiments anciens. Pour ces derniers, les contours peuvent n'être que partiellement détectables sur la vue aérienne. Une fois les objets extraits, les relations spatiales (distance, direction, adjacence...) entre ces objets sont calculées puis modélisées.

Dans le cadre de cette application, nous avons utilisé une interface d'extraction manuelle [Ler03] afin de simuler la détection des contours simples (géométriques ou courbes simples). Un expert (spécialiste d'archéologie) a indiqué quels sont les contours importants et quels sont leurs attributs sémantiques. Dans cette instanciation de notre modèle, les objets sont décrits par leur géométrie et des attributs sémantiques mais ils n'ont pas d'attributs physiques.

Pour chaque objet, le tuple d'attributs sémantiques $AttS = < loc, tso, dat >$ comprend :

- Une localisation globale (e.g., Massala), notée *loc*, qui n'a de sens que sur l'image elle-même. Cette localisation est facultative ;
- Un type principal d'objet, noté *tpo*, qui indique la nature – archéologique ou non – de l'objet (e.g., contrevallation, rivière). Ce type principal est obligatoire ;
- Un type secondaire d'objet, noté *tso*, qui précise le type principal (e.g., fragment de contrevallation, segment amont de rivière). Ce type secondaire est facultatif ;
- Une datation de l'objet (e.g., 73, contemporain) qui est optionnelle.

Nous détaillons ci-dessous l'instanciation du modèle pour cette application et nous donnons un exemple d'image (figure 3.a) avec les contours conservés par l'expert (figure 3.b) et la représentation de l'image dans le modèle instancié (figure 3.c). Dans la figure 3.c, les liens entre objets représentent la composition privilégiée, les cadres des objets simples sont en gras.

Dans la figure 4.a, nous donnons les directions de placement de l'un des objets (l'objet *AB*) par rapport à tous les autres objets.

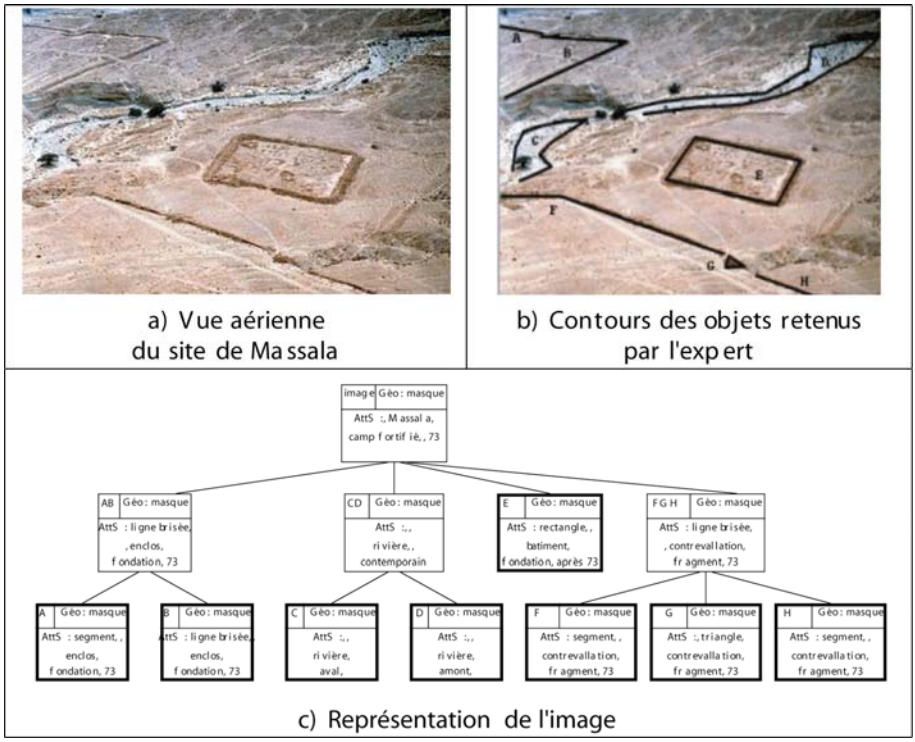


Figure 3 : Le camp de Massala [Ua]}

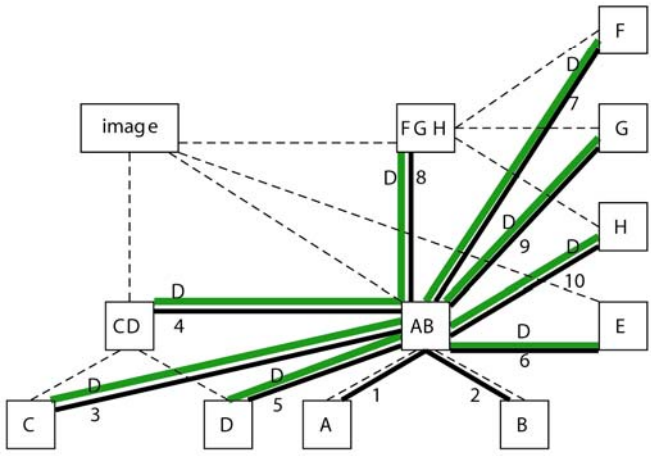
a) vue aérienne originale et b) avec les contours des objets retenus par l'expert (en utilisant une interface couplée avec une base de données pour le stockage des objets ainsi définis que nous avons développée [Ler03] sur la base d'un éditeur graphique en licence GNU [BLP]), c) représentation de l'image (objets et relation de composition privilégiée).

En figure 4.b, nous illustrons les relations (relation topologiques en trait gras clair et relations de direction en traits gras sombres) entre cet objet *AB* et les autres objets. La seule étiquette utilisée pour la relation topologique est D pour

« Disconnected ». Les étiquettes utilisées pour la relation de direction font référence à la table donnée en partie a) de la figure. La combinaison des deux structures (l'arbre qui décrit la composition des objets et les graphes qui décrivent les autres relations entre objets) modélise tous les aspects de l'image.

Placement de l objet AB par rapport à l objet	selon le vecteur <N O, N, NE, O, E, SO, S, SE>	Label associé au vecteur
A	< 1, 0, 0, 0, 0, 0, 0, 0>	1
B	<0, 0, 0, 0, 0, 0, 0.205128, 0>	2
C	<0, 0, 0, 0, 0, 0.266667, 0.733333, 0>	3
CD	< 0, 0, 0, 0, 0, 0, 0, 1>	4
D	<0, 0, 0, 0, 0.631952, 0, 0, 0.368048>	5
E	< 0, 0, 0, 0, 0, 0, 0, 1>	6
F	<0, 0, 0, 0, 0, 0, 0.228261, 0.771739>	7
FG H	< 0, 0, 0, 0, 0, 0, 0, 1>	8
G	< 0, 0, 0, 0, 0, 0, 0, 1>	9
H	< 0, 0, 0, 0, 0, 0, 0, 1>	10

a) Table des relations de direction concernant l' objet AB



b) Les arcs étiquetés des relations portant sur l' objet AB

Figure 4 : Le camp de Massala [Ua] : relations de l'objet AB avec les autres objets

L'ensemble des graphes étiquetés utilisés dans cette instanciation est :

$$\{< ETIQ_i, A_i, \square_i >\}_{i=1..3}$$

avec :

- Pour le graphe d'indice 1, qui est le graphe de la relation de composition privilégiée, un ensemble d'étiquettes réduit à une constante (cette étiquette constante n'est pas visualisée dans la figure 4.a).
- Pour le graphe d'indice 2, qui est le graphe de la relation de direction, l'ensemble des étiquettes est l'ensemble des vecteurs de direction de la forme *<N, S, E, O, NE, NO, SE, SO>* (pour respectivement Nord, Sud, Est, Ouest, Nord-Est, Nord-Ouest, Sud-Est et Sud-Ouest). La relation de direction que nous utilisons est une relation à base de logique floue que nous avons développée pour cette application archéologique [BTY01]. Les calculs des vecteurs de direction sont effectués dans la base de données liée à l'interface d'acquisition des objets.
- Pour le graphe d'indice 3, qui est le graphe de la relation topologique, nous utilisons certaines des relations topologiques issues de la méthode des neufs intersections d'Egenhofer [EF91].

4. Conclusion

Nous avons présenté les grandes lignes d'une plate-forme pour la création de bases d'images. Cette plate-forme permet – par combinaison de différents composants autonomes mais intégrés via un modèle et une mesure de similarité génériques – d'utiliser dans la description des images des informations syntaxiques ou sémantiques qui peuvent être locales ou globales et d'accéder à la base en mode recherche (via un mécanisme d'indexation) ou en mode exploratoire (via un mécanisme de classification). La phase d'instanciation permet de cibler le système sur le type d'application visé par le choix des composants les plus adaptés au type d'images de la base (si la base n'est pas trop hétérogène). Le paramétrage de la mesure de similarité ainsi que certaines des relations spatiales proposées permettent d'adapter le système non seulement à l'application mais aussi aux mécanismes de perception humains.

Notre objectif à court terme est de définir les autres composants et particulièrement d'intégrer dans la bibliothèque d'extracteurs des modules qui doivent être fournis par d'autres équipes de notre laboratoire.

A moyen terme, nous prévoyons de mettre en place le mécanisme de déduction permettant de garantir la cohérence des informations sémantiques associées – à différents niveaux de détail – à une même image. Ce mécanisme de déduction doit aussi permettre d'utiliser alternativement des informations syntaxiques et

sémantiques pour une même recherche. Ceci devrait nous permettre de passer à la phase d'utilisation de notre plate-forme pour définir et tester des stratégies de fusion syntaxique-sémantique. Nous prévoyons de définir un protocole de validation des stratégies (et de construire une base d'images pouvant servir de « benchmark » pour comparer ces stratégies).

5. Bibliographie

[AKDGB] W. Al-Khatib, Y.F. Day, A. Ghafoor, and P.B. Berra. Semantic Modeling and Knowledge Representation in Multimedia Databases. *Transactions on Knowledge and Data Engineering*, 11(1), 1999.

[BLP] A. Basset, C. Lignier, and P. Pascal. Editeur graphique vectoriel. GNU General Public License, Available at URL http://editeurgraphique.free.fr/.

[BTY01] L. Besson, M.N. Terrasse, and K. Yétongnon. A fuzzy approach to direction relations. In *CBMI 2001, Content-Base Multimedia Indexing*, pp. 169-176, Brescia, Italy, September 2001.

[DBM97] N. Denos, C. Berrut, and M. Mechkour. An image retrieval system based on the visualization of system relevance via documents. In *Proc. of the Int. Conf. on Database and Expert Systems Applications*, DEXA'97, pp. 214-224, 1997.

[DL00] M. Després-Lonnet. Thésaurus iconographiques et modèles culturels. *Document Numérique*, 4(1-2):153-165, 2000.

[EF91] M.J. Egenhofer and R. Franzosa. Point-Set Topological Spatial Relations. *International Journal of Geographical Information Systems*, 5(2):161-174, 1991.

[FFN+93] C. Faloutsos, M. Flickner, W. Niblack, D. Petrovic, W. Equitz, and R. Barber. Efficient and Effective Querying by Image Content. Technical report, Researsh Report, IBM Almaden Research Center, 1993.

[Fou02] J. Fournier. *Indexation d'images par le contenu et recherche interactive dans les bases généralistes*. Phd-thesis, Université de Cergy-Pontoise, 2002.

[GB] V. Gouet and N. Boujemaa. Object-based Queries Using Color Points of Interest. Imedia Project, Inria Rocquencourt, Available at URL citeseer.nj.nec.com/ 577185.html.

[GT96] W.K. Grassmann and J.-P. Tremblay. *Logic and Discrete mathematics, A Computer Sciences Perspective*. Prentice Hall, 1996. ISBN 0-13-501206-6.

[Gua95] N. Guarino. Formal Ontology, Conceptual Analysis and Knowledge Representation. *Int. Journal of Human and Computer Studies, Special Issue on Formal Ontology, Conceptual Analysis and Knowledge Representation*, 1995.

[GWJ91] A. Gupta, T.E. Weymouth, and R. Jain. Semantic Queries with Pictures: The VIMSYS model. In G.M. Lohman, A. Sernadas, and R. Camps, editors, *17th International Conference on Very Large Data Bases*, September 3-6, 1991, Barcelona, Catalonia, Spain, Proceedings, pp. 69-79. Morgan Kaufmann, 1991.

[KPK01] K.I. Kim, S.H. Park, and H.J. KIM. Kernel Principal Component Analysis for Texture Classification. *Signal Processing Letters*, 8(2), 2001.

[Ler03] B. Leroux. Combinaison des caractéristiques physiques et sémantiques pour la recherche dans les bases de données images. Technical Report, 2003.

[LTMD01] J. Landré F. Truchetet, S. Montuire, and B. David. Content-based Multiresolution Indexing and Retrieval of Paleontology Images. *SPIE Proc. of Storage and Retrieval for Media Databases* - San Jose - CA - USA, 4315:482-489, 2001.

[OS95] V. Ogle and M. Stonebraker. Chabot: Retrieval from a Relational Database of Images. *IEEE Computer*, 28(9):40-48, September 1995.

[PSTT01] G. Petraglia, M. Sebillo, M. Tucci, and G. Tortora. Virtual Images for Similarity Retrieval in Image Databases. *Transactions on Knowledge and Data Engineering*, 13(6), November/December 2001.

[SB94] R.K. Srihari and D.T. Burhans. Visual Semantics: Extracting Visual Information from Text Accompanying Pictures. In *Proc. of AAAI '94*, Seattle, WA, USA, 1994.

[SGJ01] S. Santini, A. Gupta, and R. Jain . Emergent Semantics Trough Interaction in Image Databases. *Transactions on Knowledge and Data Engineering*, 13(3), 2001.

[TCC+94] R.K. Taira, A.F. Cardenas, W.W. Chu, C.M. Breant, J.D.N. Dionisio, C.C. Hsu, and I.T. Ieong. An object-oriented data model for skeletal development. In *Proc. of the SPIE*, 1994.

[Ua] URL=http://www.archeologie-aerienne.culture.gouv.fr. L'archéologie aérienne dans la France du nord.

[WJ97] D.A. White and R. Jain. Imagegrep: Fast visual pattern matching in image databases. In *Proc. SPIE: Storage and Retrieval for Image and Video Databases*, volume 3022, pp. 96-107, 1997.

[ZAL01] X.M. Zhou, C.H. Ang, and T.W. Ling. Image Retrieval Based on Object's Orientation Spatial Relationship. *Pattern Recognition Letters*, 22, 2001.

CEDERILIC : constitution d'un livre et d'un index numérique*

Jean Charlet[1], Touria Aït el Mekki[2], Didier Bourigault[3], Adeline Nazarenko[2], Régine Teulier[4], Baruk Toledano[5]

[1]*STIM/DSI/AP–HP & INSERM ERM 202, Université Paris VI - France*
jc@biomath.jussieu.fr

[2]*LIPN, CNRS - Université Paris-Nord - France*
{taem,nazarenko}@lipn.univ-paris13.fr

[3]*ERSS, CNRS-Université Toulouse Le Mirail - France*
didier.bourigault@univ-tlse2.fr

[4]*CRG - France*
teulier@poly-polytechnique.fr

[5]*LIP6, Université Paris VI - France*
baruk.toledano@lips.fr

Résumé :

Nous décrivons une expérience en grandeur réelle de constitution d'un index thématique pour un ouvrage scientifique. Cet ouvrage est constitué d'une sélection de vingt-et-un articles de trois éditions des journées Ingénierie des connaissances (1999-2001). Ce corpus a été traité par l'analyseur SYNTEX puis par le système INDDOC, logiciel dédié à la constitution d'index. Ce travail a été réalisé dans un contexte entièrement numérique, c'est-à-dire à partir de fichiers numériques et pour constituer la collection des articles de l'ouvrage en un ensemble de fichiers HTML au sein duquel l'utilisateur navigue via un navigateur. Nous présentons les principaux problèmes rencontrés et les solutions adoptées.

Mots-clés : livre numérique, indexation, ingénierie des connaissances, acquisition de connaissances à partir de textes, structuration de terminologie, condensation de l'information, XML, DTD DocBook.

---

* CEDERILIC pour « Cédérom pour indexer le livre IC » est un projet soutenu par France Télécom. En dehors de la forte activité de recherche suscitée par le projet, le soutien a principalement consisté dans le financement du stage de DESS de Baruk Toledano qui a réalisé les programmes de transformation et d'enrichissement des fichiers.

Abstract:

We describe a real experiment in order to build a thematic index of a scientific book. This book is a compilation of 21 articles from the French Knowledge Engineering conferences (1999-2001). The corpus has been analysed by SYNTEX then by INDDOC, software dedicated to index formation. This work has been realized in a full digital context, with digital HTML articles and HTML index. The user uses a browser for exploring the articles through the index. We describe the work, the main problems and the chosen solutions.

Keywords: digital book, indexation, knowledge engineering, knowledge acquisition from texts, terminology structuration, XML, DTD DocBook.

1. Introduction

Nous décrivons une expérience en grandeur réelle de constitution d'un index thématique pour un ouvrage scientifique. Cet ouvrage [TEUL 04] est constitué d'une sélection de vingt-et-un articles de trois éditions des journées Ingénierie des connaissances (1999-2001). Ce travail a été réalisé dans un contexte entièrement numérique, c'est-à-dire à partir de fichiers numériques et pour constituer la collection des articles de l'ouvrage en un ensemble de fichiers HTML que l'utilisateur peut consulter via un navigateur. Ce travail tire parti des expériences acquises par les auteurs :

- La constitution d'un index pour le livre sur l'Ingénierie des connaissances [CHAR 00], regroupant 35 articles des années 1995-1998 effectuée par D. Bourigault et J. Charlet [BOUR 99]. Plutôt que de faire appel aux auteurs des articles, ce travail innovait en exploitant les résultats fournis par un outil de traitement automatique des langues, l'analyseur syntaxique de corpus LEXTER (prédécesseur de SYNTEX [BOUR 00]) à partir de l'analyse automatique du corpus électronique constitué des trente-cinq articles sélectionnés. Si le repérage des candidats termes (cf. § 3) nous a rapidement paru n'être qu'un point de départ, l'expérience nous a permis de repérer un certain nombre de difficultés et d'y apporter des solutions [BOUR 99].
- Le développement d'un système de constitution d'index par T. Aït El Mekki et A. Nazarenko, INDDOC [AITE 02]. Tirant parti des enseignements du travail précédent, l'équipe du LIPN a proposé une réflexion et une nouvelle architecture de constitution d'index. Cette architecture considère un index comme une ressource, constituée à partir d'un corpus, que des outils permettent d'ébaucher (index ébauche), que l'utilisateur complète (index source) et qui peut être visualisé.

Ainsi, que ce soit la première expérience, vis-à-vis de la complexité de l'index construit, ou la seconde, vis-à-vis de la complexité des fonctions attendues, tout concourait au développement d'un index numérique permettant de naviguer dans

une collection d'articles numériques. Nous avons donc décidé de monter un projet qui visait à associer à un livre « papier », un cédérom proposant les articles indexés et permettant d'y accéder via l'index. Le projet comporte cinq étapes principales (cf. figure 1) :

1. La transformation des articles du format d'origine (RTF) dans un format XML,
2. L'enrichissement de ce format selon plusieurs contraintes (visualisation, indexation),
3. Le traitement du corpus ainsi constitué par SYNTEX pour obtenir les candidats termes nécessaires à l'étape suivante,
4. La constitution de l'index grâce à INDDOC,
5. La réalisation finale des fichiers à visualiser et de l'interface de navigation.

Nous décrivons, section 2, les tenants et aboutissants de notre approche, section 3, la constitution des ressources XML, section 4, le repérage des candidats termes par SYNTEX, section 5, le mode de constitution de l'index par INDDOC, et, section 6, la constitution de la ressource HTML sur laquelle navigue l'utilisateur. La section 7 décrit l'expérimentation d'un point de vue qualitatif et quantitatif. Enfin, nous essayons de tirer des conclusions et de proposer des perspectives à ce travail dans la section 8.

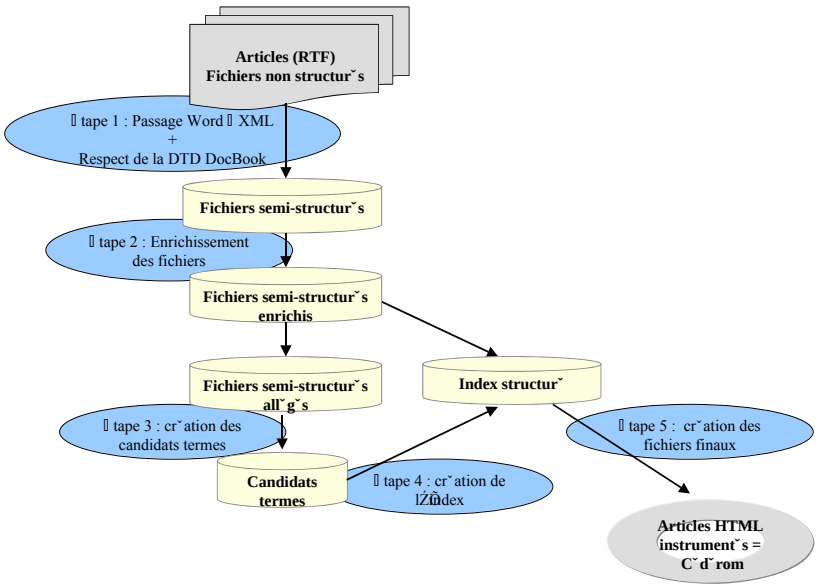


Figure 1 : Étapes principales de réalisation du cédérom

2. Approche

En nous appuyant sur la démarche proposée dans [AITE 02], nous concevons la construction d'un index comme un processus en deux étapes suivies d'une visualisation :

1. L'acquisition de l'index source qui est une base de connaissances qui contient le réseau des descripteurs accompagnés des renvois au texte.
2. La génération qui produit un ou plusieurs index dérivés (vues) à partir de l'index source en fonction des contraintes éditoriales. La génération exploite ainsi l'ordre de pertinence établi en 1 pour ne sélectionner qu'un certain nombre de descripteurs et de renvois par descripteur. Elle peut également ne conserver qu'un sous-ensemble de relations sémantiques. Cette génération se fait sur la base de feuilles de styles produites par l'indexeur ou de modèles prédéfinis.
3. La visualisation finale est produite. Même si on peut représenter un index sous différentes formes dans un contexte numérique, dans le cadre de cette expérimentation, nous avons choisi de ne mettre en œuvre qu'un index textuel pour éviter, dans un premier temps, les problèmes de visualisation graphique.

On distingue deux grandes catégories de logiciels d'aide à la création d'index de fin de livres :

1. Les logiciels de gestion d'index que l'on trouve aujourd'hui dans les traitements de texte grand public et qui demandent à l'indexeur de saisir l'intégralité des entrées d'index et des renvois aux textes. Celui-ci le fait généralement à partir d'une lecture sur épreuve du document ou, plus rarement, à l'écran. Ces logiciels prennent donc en charge le tri alphabétique (qui est parfois complexe lorsqu'il est croisé avec les niveaux d'index) et la mise en forme du document « index » (format de sortie et feuille de style).
2. Les logiciels d'acquisition d'index qui proposent un premier jeu d'entrées d'index et de renvois au texte. Ces logiciels vont plus loin dans l'accompagnement de l'utilisateur. Pour ce faire, ils s'appuient sur la structure du document (cf. HTML INDEXER[1], IXGEN[2]). Ils reposent parfois sur une analyse linguistique pour l'extraction de groupes nominaux (f. SYNTACTICA[3], INDEXING ONLINE[4]) mais la recherche d'occurrences pour le

[1] http://www.html-indexer.com/
[2] http://www.fsatools.com/
[3] http://www.syntactica.com/login/login1.htm
[4] http://www.indexingonline.com/index.php

calcul des renvois se limite à une recherche de chaînes de caractères et ne tient pas compte de critères linguistiques.

En dehors des distinctions de casse, ces logiciels ne prennent pas en compte la variation: deux formes fléchies différentes rattachées à un même lemme sont proposées comme deux entrées différentes. Ils n'aident en rien l'indexeur à sélectionner ce qui doit figurer dans l'index. Enfin, ils limitent les entrées d'index aux seuls syntagmes nominaux. Dans ce travail, nous proposons une approche globale tenant compte de ces déficiences et nous proposons des évolutions.

3. De RTF au document semi-structuré instrumenté

Le travail consiste donc, dans un premier temps, à créer une chaîne permettant la transformation de documents RTF en document HTML avec un travail d'enrichissement sur un format intermédiaire, dit « pivot ». Deux choix s'imposaient d'eux-mêmes : *(a)* Puisque nous devions construire une ressource générique, suffisamment structurée, destinée à s'enrichir et à être analysée par les outils du projet, XML était le candidat idéal d'autant plus que de très nombreux programmes prêts à interpréter et transformer des fichiers XML existent ; *(b)* la DTD DocBook[5], standard de l'édition numérique, était le parfait support des enrichissements prévus[6]. À ce stade de notre travail, nous n'avons pas exploité les TOPIC MAPS, norme destinée, entre autres, à représenter des index et fondée sur XML. Il semble cependant que ce soit un bon candidat pour la représentation des index numériques et nous envisageons de l'adopter à l'avenir.

Le premier défi consiste à tirer des informations de structure d'un document RTF qui ne possède aucune méta-information. La solution retenue exploite « les styles » typographiques de Word. Dans un document HTML, les balises indiquent le statut de l'élément auquel elle se rapporte. Des règles éditoriales peuvent être attachées à chaque type de balise. De même qu'une balise HTML définit certains cas l'apparence de l'élément auquel elle se rapporte (i.e. l'utilisation d'une balise <H1> implique que le contenu sera en gras à l'écran avec une police bien plus grande que le reste du texte). Dans notre cas, c'est le style du texte qui va donner du sens au texte (i.e. un titre de niveau 1 – en général, appelé Titre 1 –, nous permet de délimiter une frontière entre deux sections de niveau 1). Pour faire cela, après étude des propriétés de différents logiciels, nous avons choisi la version libre d'un logiciel

[5] http://www.oasis-open.org/docbook/documentation/reference/html/docbook.html

[6] Une autre DTD sert de standard pour les documents numériques, c'est la DTD TEI de la *Text Electronic Initiative* mais elle correspond à des textes plus littéraires, au contraire de la DTD DocBook utilisée pour des documentations techniques (LINUX) et comportant d'origine des éléments de description des index.

de transformation de RTF, UPCAST[7], qui construit des fichiers respectant la DTD UPCAST. Ces mêmes fichiers XML ont une structure proche de ceux respectant la DTD DOCBOOK et peuvent donc être traduit pour respecter cette nouvelle grammaire via un programme XSL univoque[8].

La suite du travail a consisté à enrichir la représentation ainsi créée pour permettre de construire des fichiers HTML finaux instrumentés (tables de matières par fichiers, liens entre le texte et la bibliographie, repérage des auteurs, institutions, courriel, repérage des figures comme fichiers externes, etc.) à l'aide de modules logiciels écrits en PERL ou XSL. En parallèle, des fichiers « allégés » sont créés[9] pour être fournis en entrée de SYNTEX (cf. figure 2).

4. Créer des candidats termes avec SYNTEX

SYNTEX [BOUR 00] est un analyseur syntaxique de corpus. Il existe actuellement deux versions, pour le français et pour l'anglais, qui ont été utilisées dans plusieurs projets [BOUR 04]. Le résultat de l'analyse effectuée par SYNTEX est un réseau de mots et de syntagmes : un syntagme verbal (resp. nominal, adjectival) est un groupe de mots dont la tête syntaxique est un verbe (resp. nom, adjectif). Par exemple, *révéler une lésion osseuse* est un syntagme verbal dont la tête syntaxique est le verbe *révéler* et l'expansion le syntagme nominal *lésion osseuse*. Dans le domaine du livre sur l'Ingénierie des connaissances, *modèle conceptuel de l'application* a pour tête syntaxique *modèle conceptuel* et pour expansion le nom *application*. Dans le réseau construit, dit « réseau terminologique », chaque syntagme est relié d'une part à sa tête (lien T) et d'autre part à ses expansions (lien E – cf. figure 3). Les éléments du réseau (mots et syntagmes) sont appelés « candidats termes ». À chaque candidat terme est associé un certain nombre d'informations numériques, sur lesquelles l'utilisateur peut se baser pour organiser son dépouillement :

---

[7] http://www.

[8] L'efficacité de cette transformation suppose que les auteurs aient respecté au mieux les indications éditoriales de styles et la syntaxe de tous les éléments pertinents pour le traitement comme, par exemple, les appels des références bibliographiques. Ce problème sera rediscuté en conclusion (Cf. § 8).

[9] Il s'agit de texte brut avec seulement un identifiant unique associé à chaque paragraphe permettant un repérage univoque de chaque partie du texte.

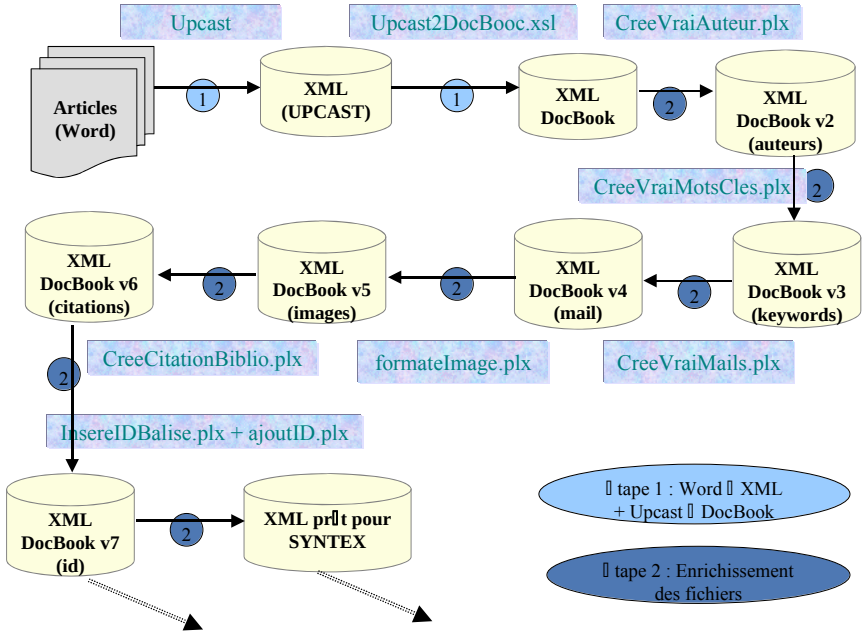


Figure 2 : De Word/RTF à des fichiers XML enrichis – Étapes 2 et 3

- **Fréquence** : c'est le nombre d'occurrences du candidat terme détectées par le logiciel dans le corpus. L'interface d'analyse des résultats permet à l'analyste d'accéder à l'ensemble des contextes d'apparition du candidat terme dans le corpus. Dans le travail exposé ici, cet accès, crucial, est assuré par le système INDDOC.
- **Productivité en tête** (resp. expansion) : c'est le nombre de « descendants en tête » (resp. « descendants en expansion ») du candidat terme, c'est-à-dire le nombre de candidats termes plus complexes qui ont le candidat terme en position tête (resp. expansion). À partir de ces informations, l'analyste peut visualiser des listes paradigmatiques de candidats termes partageant la même tête ou la même expansion (cf. figure 3), ce qui le guide vers la constitution de taxinomies locales. La difficulté essentielle pour l'utilisateur vient de la masse des résultats issus de l'extraction. Même s'il existe de nombreux travaux fort intéressants sur le filtrage statistique de candidats termes extraits automatiquement de corpus [DAIL 94;

MAYN 01; NAKA 01], l'expérience montre qu'aucune mesure statistique ne peut suppléer l'expertise de l'analyste, en particulier parce qu'il y a toujours des candidats termes de fréquence 1 dont l'analyse est intéressante. De façon générale, sachant qu'il ne pourra analyser tous les candidats termes extraits du corpus, l'analyste doit adopter une stratégie optimale qui, étant donné le temps qu'il consacre à l'analyse terminologique et le type de la ressource à construire, lui garantit que, parmi les candidats qui auront échappé à son analyse, la proportion de ceux qui auraient pu être pertinents est faible. Ce constat demande, dans le cas de notre projet, que l'étape suivante, prise en charge par INDDOC, tienne compte de ces particularités (cf. § 5).

modèle conceptuel
Tête de :
modèle conceptuel de l'application
modèle conceptuel des données
modèle conceptuel des traitements
Expansion de :
construction d'un modèle conceptuel
validation d'un modèle conceptuel

Figure 3 : Extrait du réseau terminologique construit par SYNTEX autour du candidat terme modèle conceptuel

5. Créer un index avec INDDOC

On élabore donc l'index à partir de la liste des candidats termes et du réseau syntaxique produits par SYNTEX. Pour cela, il faut :

1. Structurer cette liste en réseau sémantique : nous nous appuyons sur les liens tête-expansion produits par SYNTEX, mais il faut interpréter et enrichir ce réseau, *a minima* en introduisant des relations hiérarchiques entre les futures entrées et sous-entrées et des liens d'équivalence sémantique pour établir des références croisées d'une entrée à l'autre ;
2. Calculer les renvois au texte pour que l'index permette effectivement d'accéder aux passages pertinents, qui sont de taille variable ;
3. Introduire une mesure de pertinence pour permettre le filtrage et le tri des informations.

L'acquisition de l'index source procède en deux étapes :

- La création automatique d'une ébauche d'index. Cette étape part du texte et produit un index source. Elle permet de construire le contenu de l'index (la liste structurée des descripteurs et les renvois au texte). Elle repose sur des techniques de structuration de terminologie pour construire le réseau, sur des techniques de segmentation de texte pour établir les renvois au texte et sur des mesures de pertinence pour trier et sélectionner l'information dans l'index. Nous détaillons les différentes étapes de ce processus ci-après
- La validation de l'ébauche d'index. Cette étape interactive repose sur une interface qui permet à l'auteur de l'index de visualiser l'ébauche, de la modifier et de l'enrichir. L'index peut-être visualisé sous différentes formes. L'interface de validation permet d'ajouter, de supprimer une fiche, ou de modifier un descripteur ou un renvoi. . En cas de doute, l'indexeur peut, à tout instant, consulter les segments de textes associés à une entrée. Le travail de validation reste coûteux mais l'interface permet de l'organiser et d'en assurer la cohérence.

Dans ce qui suit, nous ne présentons pas l'interface de validation [AITEL02] mais nous décrivons le processus qui permet de construire automatiquement un index aussi « bon » que possible. La construction de l'ébauche d'index consiste notamment à filtrer les candidats termes, à les organiser en réseau, à calculer leurs occurrences et à mesurer le poids des termes et de leurs occurrences (cf. figure 4).

5.1 Le filtrage

Dans un premier temps, la liste des candidats termes produite par SYNTEX est filtrée. Seuls les termes nominaux sont conservés comme entrées potentielles (les termes verbaux peuvent être pris en compte mais seulement comme variantes d'un terme nominal). L'application d'un antidictionnaire permet d'éliminer une partie des termes non pertinents pour former des entrées d'index. L'application de règles approximatives de racinisation permet de réduire encore la liste initiale en regroupant certains termes.

5.2 La structuration

INDDOC tente de repérer différents types de relations entre les termes : variation morphosyntaxique, synonymie, hyperonymie notamment. Il s'appuie pour ce faire sur les résultats obtenus en structuration de terminologie (FASTER [JACQ 96], PROMETHE [MORI 99], SYNOTERM [HAMO 01]), la principale difficulté vient de ce que les outils existants ont été conçus pour acquérir un type particulier de relation et du fait qu'il semble impossible d'avoir une méthode unique de détection de relations.

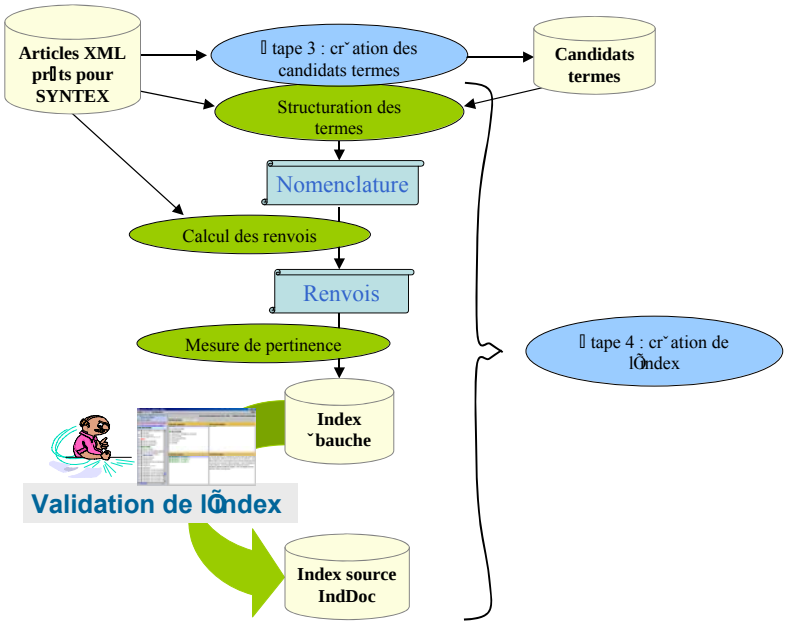


Figure 4 : Génération de l'index source – Étapes 3 et 4

En pratique, l'intégration ne doit pas se faire en fonction du type de relation visée mais selon la méthode utilisée. INDDOC comporte ainsi 2 sous-modules de structuration [AITE 03] :

- Le premier repose sur la structure interne des termes et exploite directement les résultats de FASTER et SYNOTERM.
- Le second module exploite l'information contextuelle. Pour cela, nous avons développé un module de recherche de relations à base de patrons.

Nous avons écrit une base de patrons pour l'hyperonymie et la méronymie qui s'appliquent à tout type de corpus. INDDOC exploite ainsi à la fois les relations tête-expansion et des patrons génériques[10] (ex. « SN1 être SN2 », « SN1 Verbe-

[10] SN pour syntagme nominal.

Composition SN2 »...) pour établir les relations hiérarchiques[11]. Nous intégrons l'ensemble des liens produits par les différents modules dans un réseau commun. Les résultats sont parfois redondants et nous conservons l'union des ensembles de relations produits par chaque outil.

5.3 Le calcul des renvois

Le calcul des renvois consiste, pour chaque entrée et chaque sous-entrée, à établir la liste de ses occurrences dans le texte, la difficulté étant de sélectionner les plus pertinentes et de définir la taille de l'empan de texte auquel il est pertinent de renvoyer[12]… Pour identifier les segments de renvoi, nous partons d'une segmentation absolue qui ne dépend que du document. À ce stade, on segmente le document en unités documentaires minimales[13] (UDMs), puis on élargit ces UDMs en unités documentaires élargies (UDEs) en fonction des marqueurs linguistiques et typographiques tout en respectant la structure logique du document (une UDE ne peut pas être à cheval sur deux sections de document). Cette segmentation permet donc de découper le document en UDEs linguistiquement ou typographiquement homogènes. À l'issue de cette phase, le document est représenté comme une liste d'UDEs. Ensuite, nous procédons à une segmentation relative qui dépend d'un descripteur donné. Cette phase est nécessaire pour établir la liste des segments de renvoi (liste des renvois) d'un descripteur. Elle comporte trois étapes :

(1) Identification des segments de renvoi (les UDEs qui contiennent le descripteur ou une de ses variantes) ;
(2) Regroupement des segments de l'étape 1 qui sont adjacents dans le texte du document, ce qui permet d'obtenir une liste simplifiée de segments de renvoi ;
(3) Généralisation de la séquence des segments d'une même section et sous-section en un unique renvoi à la section, lorsqu'une partie suffisamment grande de la section figure dans la liste des segments établie à l'étape 2.

Une partie de ce regroupement est également laissée à la charge de l'auteur qui a la possibilité dans une interface de validation de regrouper plusieurs occurrences qui lui semblent proches sous un seul renvoi. Il peut ainsi compléter et corriger le regroupement qui est basé sur des heuristiques génériques.

---

[11] À l'avenir, on pourrait envisager de laisser la possibilité à l'auteur de l'index d'introduire des patrons spécifiques.

[12] Précisons que les pages ne sont pas les bonnes unités pour un index dédié à un ouvrage numérique.

[13] Un extrait de texte de type paragraphe, phrase ou même de mots. Le type de l'extrait dépend de la tâche à réaliser (indexation, résumé automatique etc.). Dans notre cas, c'est le paragraphe

5.4 La mesure de pertinence

Une fois identifiés les renvois associés à une entrée, il reste à les classer par ordre de pertinence. Nous nous inspirons de l'approche TF/IDF pour l'évaluation de la pertinence de différentes clefs d'indexation dans une base documentaire mais notre mesure [AITE 04] permet de prendre en compte, outre le poids d'un mot dans l'ensemble du document et sa fréquence dans le segment de renvoi, le poids d'une occurrence particulière (qui peut être mise en valeur typographiquement, par exemple) et le poids des segments où il cité (lequel dépend en retour du poids des termes qu'il comporte). Notre critère de pertinence tient ainsi compte des paramètres qui sont traditionnellement exploités par les indexeurs [NANC 93 ; THEC 03] : la typographie, la présence d'une occurrence dans un titre, une mise en relief discursive... Notre mesure repose ainsi sur différents marqueurs ce qui la rend plus robuste aux variations de genre, de domaine et de style.

6. Instrumenter un fonds documentaire indexé

À ce stade du processus, l'index source est créé au format INDDOC (DTD). Il faut le traduire dans un format respectant la DTD DOCBOOK et faire un certain nombre de traitements pour permettre sa traduction en HTML, en particulier sur la forme des termes dans les articles et des renvois dans l'index devant respecter la syntaxe des ancres HTML. C'est durant cette traduction qu'a lieu la génération de l'index dérivé en fonction des contraintes éditoriales que l'on s'est données : affichage de l'index hiérarchique sur 2 niveaux, conservation du typage de certains liens (« voir aussi », hyperonymie), abandon des autres (projetés dans « voir aussi »). La fin de l'instrumentation est la génération de l'index HTML, la création des articles HTML et la mise en place de l'interface de navigation qui comporte, dans cette première version, 3 « frames », (1) table des matières des articles, (2) articles et (3) index. C'est à ce niveau qu'est visualisé l'empan des renvois. Cette visualisation pose problème : si nous visualisons les renvois d'index dans les textes avec de la couleur, ce que nous faisons, nous pouvons utiliser pour cela des fonctions JAVASCRIPT mais avec un défaut majeur : la non portabilité de ce langage d'un navigateur à l'autre. La solution choisie est donc de générer autant de fichiers HTML coloriés qu'il y a d'entrées d'index pour un article. Cela augmente le volume de l'ensemble des fichiers mais rend l'interface totalement portable (cf. figure 5).

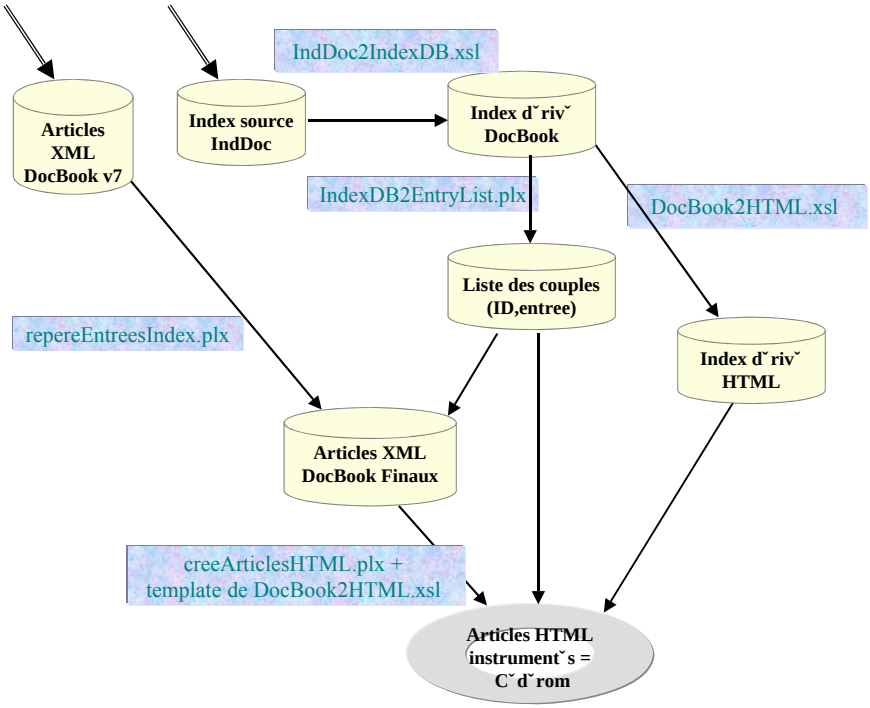


Figure 5 : Génération des fichiers HTML instrumentés – Étape 5

7. Retour d'expérience

7.1 Résultats de INDDOC

Filtrage

Le corpus comporte environ 177 000 mots. SYNTEX fournit une liste de 32 334 candidats termes. De cette liste, nous retenons 17 521 candidats termes nominaux. Le filtrage à base d'antidictionnaire et de racinisation ramène cette liste à quelques 2 800 candidats termes qui sont autant d'entrées potentielles de l'index.

Structuration

Pour la structuration, nous prenons en compte toutes les variantes de termes. Nous travaillons donc sur la liste complète de 10 008 candidats termes fournie par SYNTEX. À partir de cette liste, le module de structuration de INDDOC

calcule 4 440 relations sémantiques. Comme les termes ne sont pas validés au préalable, certaines relations relient en fait des termes non pertinents (par ex. *différents membres : partie différente, haut niveau : niveau supérieur*). Cette proportion de relations s'avère relativement faible si on la compare avec des monographies [AITE 04]. Cela peut être dû à la nature de l'ouvrage qui est un recueil d'articles d'auteurs différents, à forte diversité stylistique.

Calcul des renvois

Nous avons constaté que la segmentation réduit effectivement le nombre de renvois (cf. tableau 1), même si là encore cette réduction est moins forte que sur d'autres corpus. On observe en effet que le facteur de réduction de la segmentation dépend de la nature du document (monographie vs collection) et du style de la rédaction (le style littéraire emploie plus de marqueurs linguistiques ; ce qui augmente le facteur de réduction dans le passage des UDMs aux UDEs). L'intérêt global de cette étape apparaît quand on compare le nombre de renvois obtenus (2 997) est avec le nombre d'occurrences des différentes entrées de l'index (28 342)[14] .

Par comparaison avec d'autres corpus étudiés, on observe aussi (cf. tableau 1) que les étapes de simplification et de généralisation sont moins marquées dans notre corpus dont les articles sont généralement fortement structurés.

Nb d'UDMs	1 085	Segmentation absolue
Nb d'UDEs	907	
Nb de seg. De renvois	3 097	Segmentation relative
Nb de seg. après simplification	3 008	
Nb de seg. après génération	2 997	
Nb de UDMs occurrences	28 342	

Tableau 1 : Nombre d'UD et de segments aux différentes étapes

Mesure de pertinence

Nous avons appliqué la mesure de pertinence de INDDOC sur notre corpus. Il est difficile d'évaluer cette mesure en tant que telle et globalement le tri obtenu mais on peut observer sur des exemples le bon comportement de notre mesure de pertinence.

[14] Il s'agit en réalité du nombre d'UDMs occurrences, *i.e.* de la somme, pour chaque entrée, du nombre de paragraphes dans lesquels elle figure (un paragraphe donné peut donc figurer deux fois, associés à deux entrées différentes), résultat qu'on obtiendrait par un calcul naïf des renvois.

Considérons le descripteur « Modélisation » qui a 4 renvois : le premier dans l'ordre du corpus, S1, apparaît dans une introduction. Le deuxième segment S2 regroupe une sous-section qui traite de la « modélisation ». Dans le troisième segment S3, le descripteur apparaît en début de segment mais le segment lui-même est inclus dans une conclusion. Le troisième segment S4 correspond à une section qui traite de la « modélisation ».

Le système a ordonné les renvois en privilégiant S4 pour la quantité de l'information apportée, devant S2 dont l'apport d'information est plus faible. Le renvoi S1 est placé en dernière position parce qu'il s'agit d'un paragraphe de l'introduction. S3 apparaît dans une conclusion mais il apporte davantage d'information et le descripteur apparaît au début du segment, ce qui lui confère plus d'importance.

7.2 Validation de l'index

La validation de l'index vient de s'effectuer avec le système décrit. À partir du corpus disponible, l'interface INDDOC nous a proposé 2 700 termes comme possibles entrées d'index. Ces termes étant à valider, structurer en niveaux et à rattacher au corpus. Nous avons retenu un peu moins de 1 000 termes comme entrées d'index et environ le double de liens vers les textes. Avant de discuter plus avant des critiques et questions en suspens, on peut noter que :

- À l'inverse des constitutions d'index sur papier, et cette remarque est aussi valable pour le travail précédent, le choix des entrées d'index se fait par suppression au sein d'une liste « large », à l'inverse d'un travail standard, repérant les entrées d'index dans n texte en partant de zéro. Cela amène à la constitution d'un index très riche.
- Vis-à-vis de cet index très riche, l'usage via une interface Web doit être observé et évalué, peu d'expériences ayant été faites dans ce domaine, sauf pour des documentations techniques. On peut penser, mais ce doit être validé, que la richesse de l'index est compensée par la facilité de navigation.

8. Conclusion et perspectives

Revenons sur les points critiques de l'expérience précédente en 1999 [BOUR 99] :

⇨ **L'empan d'un renvoi**. C'est l'un des aspects originaux ; la question de la prise en charge est prise en compte dès le départ dans le système et l'interface et aide l'indexeur à choisir cet empan. Par ailleurs, les mesures de pertinence du système INDDOC proposent les renvois dans un ordre, justement pertinent, pour prendre en compte cette question.

- ⇨ **La difficulté de choisir ce qui est une bonne entrée d'index**. La méthodologie et le système INDDOC en particulier sont une réponse partielle à cette difficulté mais cette question recevra toujours une réponse en termes de choix humain.
- ⇨ **La structuration de l'index**. Le système INDDOC et la méthodologie mise en œuvre ici est beaucoup plus riche que précédemment : le système et l'interface permettent en plus de typer les relations entre les entrées de l'index, beaucoup plus même que ce que nous avons choisi de faire (cf. infra).
- ⇨ **La validation**. Rien n'a changé. Puisque le choix des entrées d'index est un choix de l'indexeur alors que les articles sont écrits par d'autres, on pourrait faire valider ce choix avec les auteurs des articles. Cette procédure, testée précédemment, n'a pas été mise en œuvre ici. En revanche, comme dans l'expérience précédente, l'indexation se fait à 2 pour éviter l'idiosyncrasie d'un travail solitaire.

Pour le reste, deux enseignements peuvent déjà être tirés de ce travail :

- Construire des fichiers XML semi-structurés à partir de Word est un travail sans fin : venant d'un éditeur WYSIWYG utilisant de nombreux caractères spéciaux pour la mise en page ou la possibilité de générer des figures en interne, nous avons été obligés à des corrections incessantes. Ainsi, exemple parmi tant d'autres, un caractère qui semble être le « caractère blanc » peut être un « blanc collant » ou un « blanc italique ». Pour les programmes qui enrichissent les fichiers XML, cela veut dire l'obligation de prendre en charge de nouvelles contraintes d'expression à chaque nouvel article ou nouveau morceau d'article. Sur la question des figures, nous avons été obligés de redéfinir (ce qui a amené à refaire des figures mais c'est un choix qui avait été fait de toute façon au départ) toutes les figures comme des fichiers externes, Sinon il était impossible de faire des fichiers XML cohérents avec les articles « papier » et affichant correctement les figures. Enfin, un éditeur comme celui précité, ne pousse pas les auteurs à respecter les styles que nous leur demandions d'utiliser très précisément pour pouvoir créer des fichiers semi-structurés avec UPCAST (cf. § 3). Il en aurait été autrement à partir d'un formateur de texte de type LATEX.
- Les travaux des auteurs de INDDOC permettaient de générer des index avec un accès graphique beaucoup plus riche que l'index textuel que nous avons décidé de construire de prime abord. On pouvait par exemple. envisager trois vues différentes qui correspondent à trois stratégies de recherche différentes : (1) la recherche par descripteur qui privilégie l'exhaustivité et la précision de l'information en permettant de visualiser l'ensemble des informations qui se rattachent à un descripteur sous la forme d'une étoile ; (2) la recherche par réseau qui donne une vue globale sur la nomenclature de l'index ; et (3) la recherche thématique qui permet d'accéder à un terme

puis à l'ensemble des descripteurs qui relèvent de ce thème. Ce mode de structuration de l'information se rapproche des approches thésauriques et des pratiques lexicographiques anglosaxones.

Par rapport à cette dernière recherche thématique, il faut noter que, comme nous le remarquons dans [BOUR 99], nous avons fait le choix de construire un index qui rend compte des usages des auteurs, à l'inverse des index thématiques. Ces derniers correspondent à des ressources beaucoup plus normalisées, qui peuvent aller jusqu'à des ontologies [CHAR 02]. Enfin, en construisant l'index source, nous nous sommes réservés la possibilité de construire un index dérivé plus riche en répertoriant plus de relations que celles utilisées pour le présent travail.

9. Références bibliographiques

[AITE 02] Ait El Mekki T., Nazarenko A., « Comment aider un auteur à construire l'index d'ouvrage ? », *Colloque International sur la Fouille de Texte*, Tunis, 2002, pp. 141- 157.

[AITE 03] Ait El Mekki T., Nazarenko A., « Le réseau terminologique, un élément central pour les index de fin de livre », *Actes des cinquièmes rencontres Terminologie et Intelligence*, Strasbourg, 2003, pp. 1-10.

[AITE 04] Ait El Mekki T., Nazarenko A., « Une mesure de pertinence pour le tri de l'information dans un index de fin de livre », TALN04, Fès, 2004 (Soumis)

[NANC 93] Nancy C. Mulvany, « *Indexing Books* », The University of Chicago Press, 1993.

[BOUR 99] Bourigault D., Charlet J., Construction d'un index thématique de l'Ingénierie des connaissances, Actes de la conférence IC'99, Massy-Palaiseau/ Polytechnique, 1999.

[BOUR 00] Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », Cahiers de Grammaires. Univ. Toulouse - Le Mirail, n° 25, 2000, pp. 131-151.

[BOUR 04] Bourigault D., Aussenac-Gilles N., Charlet J., Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas, *Revue d'Intelligence Artificielle*, 2004.

[CHAR 00] Charlet J., Zacklad M., Kassel G. & Bourigault D. [2000] (éd.), Ingénierie des connaissances. Évolutions récentes et nouveaux défis, « Coll. technique et scientifique des télécommunications », Eyrolles, Paris, 632 p.

[CHAR 02] Charlet J., *L'ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*, Mémoire d'habilitation à diriger des recherches. Paris VI, décembre 2002.

[DAIL 94] Daille B., *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*, thèse en Informatique Fondamentale, Univ. de Paris 7, Paris, 1994.

[JACQ 96] Jacquemin, C., « A symbolic and surgical acquisition of terms through variation. » , in S. Wermter, E. Riloff, and G. Scheler, ed., Connectionist,

Statistical and Symbolic Approaches to Learning for Natural Language Processing, Springer, Heidelberg, 1996.

[HAMO 01] Hamon T., Nazarenko A., « Detection of synonymy links between terms: experiment and results », in Bourigault D, L'Homme M.-C., Jacquemin C. (éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, 2001, pp. 185-208.

[MAYN 01] Maynard D., Ananiadou S., Term extraction using similarity-based approach, in Bourigault D, L'Homme M.-C., Jacquemin C. (éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, Amsterdam, 2001, pp. 261-78.

[MORI 99] Morin E., « Des patrons lexico-syntaxiques pour aider au dépouillement terminologique », *Traitement Automatique des Langues*, 1999, 40(1): 143-166.

[NAKA 01] Nakagawa H., Experimental evaluation of ranking and selection methods in term extraction, in Bourigault D, L'Homme M.-C., Jacquemin C. (éd.), *Recent advances in computational terminology*, John Benjamins Publishing Company, 2001, pp. 303-26.

[TEUL 04] Teulier R.,Charlet J., Tchounikine P., Ingénierie des connaissances, L'harmattan, Paris, 2004. À paraître.

[THEC 03] *The Chicago Manuel of Style*, chapter 18, fifteenth Edition, The University of Chicago Press Staff, 1993.

Un mode de mise en scène théâtrale directement inspiré de la fouille interactive de données numériques

Alain Bonardi[1], **Francis Rousseaux**[2]

[1]*Equipe Intelligence Artificielle et Robotique Mobile*
Université Paris 8 - 2 rue de La Liberté - 93526 Saint-Denis Cedex 02 - France

alain.bonardi@wanadoo.fr

[2] *Equipe projet SemanticHIFI de l'IRCAM et CReSTIC de l'Université de Reims*
1, place Igor-Stravinsky, 75004 Paris - France

francis.rousseaux@ircam.fr

Résumé :

La mise en scène de théâtre traditionnelle repose sur une approche formelle de la similarité s'appuyant sur des ontologies dramaturgiques et des variations d'instanciation. Inspirés par la fouille de données numériques interactive, qui suggère des approches différentes, nous rendons compte de recherches théâtrales utilisant l'ordinateur comme partenaire de l'acteur pour échapper à la spécification *a priori* des rôles.

Mots-clés : similarité, instanciation, ontologies, fouille de données interactive, théâtre inter-média.

Abstract:

In this paper, we show to what extent traditional theatre staging is based on a formal approach of similarity using a dramaturgical ontology and instanciation variations. Drawing our inspiration from the opposite approach through interactive data mining, we hereby account for theatre researches using computers as actor partners to escape *a priori* specification.

Keywords: similarity, instanciation, ontologies, interactive data mining, inter-media theatre.

1. Introduction

Dans cet article, nous réfléchissons à la mise en scène de théâtre et ses évolutions dans le contexte du dialogue acteurs-ordinateurs, en mobilisant les catégories de l'informatique. Nous avons déjà abordé l'approche inverse en montrant comment des progiciels tels que Powerpoint™ sont fondés sur des conceptions théâtrales et plus particulièrement scénographiques, liées aux notions d'avant-plan et d'arrière-plan [ROUSSEAUX & BONARDI02].

Distinguant deux approches de la similarité en informatique – inspirées par les ontologies *vs* par la fouille de données interactive, nous les utilisons pour comprendre la mise en scène au sens traditionnel et ses nouvelles modalités liées à l'utilisation d'ordinateurs comme partenaires des acteurs. Notre recherche s'appuie sur un exemple de réalisation, la pièce de théâtre inter-média *La traversée de la nuit* de Geneviève de Gaulle.

2. L'approche traditionnelle de la mise en scène dans une perspective informatique

Face à un texte de théâtre, chaque metteur en scène souhaite proposer sa lecture/interprétation. Rappelons en effet que l'interprétation n'est pas immanente au texte, malgré les indications parfois nombreuses (préface, didascalies) de l'auteur. Un texte ne peut exister sur scène sans exégèse du metteur en scène.

2.1 Ontologies de la dramaturgie et variations d'instanciation

Cette lecture/interprétation est toujours un effort pour créer des formes. Essayons d'en rendre compte dans une perspective informatique. La démarche du metteur en scène commence par l'établissement d'une ontologie synthétique de la dramaturgie : on y décrit les personnages sous forme de types (ce en quoi le théâtre de boulevard par exemple excelle ave son trio, mari, femme et amant !) et d'instanciations, en indiquant le nom du personnage, sa situation au début de la pièce, son costume. Le déroulement de la pièce propose des variations d'instanciation[1] : le spectateur découvre que tel ou tel personnage est différent de ce

[1] *instanciation* est un anglicisme couramment utilisé par les informaticiens, qui renvoie au mot *instance* signifiant *exemple, cas ;* l'instanciation généralise en quelque sorte l'opération, utilisée par les mathématiciens, d'affectation d'une valeur numérique à une variable : pour parler du réel, les informaticiens instancient des classes abstraites, décrétant ainsi que telle ou telle entité est un cas particulier d'une classe, elle-même reliée à d'autres classes par des hiérarchies de généralité et/ou des propriétés formelles, l'ensemble du dispositif [PERROT94] constituant ce qu'on appelle parfois une *ontologie* (les ontologies prétendent ainsi décrire des pans de connaissances mondaines très utilisées en intelligence artificielle), parfois une *conception à objets* (une conception à objets est constituée de

qu'il imaginait au départ. Ces variations d'instanciation peuvent parfois aboutir à des révisions d'ontologie. C'est par exemple l'enjeu, autant métaphysique que théâtral, de la pièce *El burlador de Sevilla* du dramaturge espagnol Tirso de Molina (1630) qui inaugure le mythe de Don Juan : ce personnage peut-il être sauvé en reconnaissant ses fautes *in extremis* avant sa mort ? L'ontologie du personnage peut-elle être radicalement modifiée à la toute fin de la pièce ?

2.2 Une approche formelle de la similarité à base d'ontologies

Dans cette approche traditionnelle du théâtre, la notion de similarité par les ontologies est centrale. Le metteur en scène règle chaque scène ou passage faisant unité en le considérant comme un exemple dans un ensemble de cas fournis par la littérature théâtrale. Expliquer un personnage à l'acteur qui le joue revient à le pointer dans l'ontologie proposée et à relier cette ontologie à celle d'autres pièces ou d'autres lectures de la même pièce par d'autres metteurs en scène, pour donner à comprendre par un exemple dit « similaire ».

En généralisant, il s'agit d'une *approche formelle*, dans laquelle on représente l'exemple comme une instance d'une structure générale embrassant tous les cas, et on cherche les similarités en faisant varier l'instanciation. Cette approche présente l'avantage de fournir une explication du caractère « similaire à l'exemple » de la proposition, voire une mesure de distance: c'est par ce biais qu'un concept récapitulatif en *intension* peut être créé. Les ontologies permettent de rechercher les similarités à un exemple en demeurant dans l'enceinte du concept, quitte à passer au concept immédiatement plus général quand la quête est infructueuse. Ceci s'applique à bien d'autres activités que le théâtre, par exemple l'organisation de la vente de disques compacts (CD) dans un grand magasin de disques [ROUSSEAUX & BONARDI04b]. En effet, la pratique de l'achat de CD prescrit subrepticement nos activités musicales et le rangement *a priori* dans des bacs de vente, à évolution lente, est structuré par l'acquisition marchande et la notion de genre.

3. L'approche de la similarité par la fouille de données interactive

Au niveau informatique, il existe une autre manière d'aborder la question de la similarité : c'est **l'approche fouille de données interactive**, dans laquelle on représente l'exemple comme une spécialisation de l'ensemble des cas, et on cherche d'autres spécialisations voisines, mais sans disposer par avance d'une ontologie. L'utilisateur accepte de la façonner à sa main avec l'aide interactive de la machine, de manière *ad hoc*. Il s'agit d'une approche ***en extension*** : façonner une similarité

---

graphes d'héritage conçus pour donner lieu à des programmes informatiques par simple instanciation de paramètres clés).

revient à façonner une liste de contenus de forme similaire par des opérations rectificatives successives mobilisant le calcul numérique en interaction interprétative avec des actions rectificatrices sur les contenus et leur forme (du côté de l'utilisateur, provoqué par les propositions de la machine).

L'activité de *music-ripping* illustre cette approche [ROUSSEAUX & BONARDI04]. Elle consiste en la manipulation créative de contenus audio-numériques passant par des gestes de modification, aboutement, suppression, etc., associés aux interfaces informatiques. Lorsque l'activité pratiquée est une écoute signée [DONIN04], une écoute/composition/production, son objet devient le grain élémentaire d'écoute/composition/production, un *échantillon*, constamment modifié, réorganisé, re-mixé et renommé [PACHET03] par l'utilisateur.

Remarquons que nous touchons là à une des différences fondamentales entre les mathématiques et l'intelligence artificielle. En effet, les mathématiques posent l'équivalence entre ***l'intension*** et ***l'extension***. La notion de classe d'équivalence est par exemple présente sur les deux versants : d'un côté, on peut relier deux éléments individuels entre eux en vérifiant qu'ils appartiennent ou pas à une même classe ; de l'autre, on recouvre les ensembles (par exemple l'ensemble des entiers relatifs) avec un nombre minimal de classes. Les noyaux d'endomorphismes jouent le même rôle en algèbre vectorielle : on peut vérifier que deux vecteurs individuels appartiennent ou non au même noyau et on peut dans le cas d'endomorphismes diagonalisables recouvrir un espace vectoriel par un nombre fini de noyaux. En revanche, dans le domaine de l'intelligence artificielle, il n'y a pas d'équivalence entre **intension** et **extension**, sans doute parce que l'équivalence entre un spécimen (comme instance particulière d'une catégorie) et une singularité (perçue comme traversant le réel) n'est tout au plus acceptable que comme mauvaise heuristique (car niant la notion même de situation).

4. Retour au théâtre : la *Traversée de la nuit*

Que peut donner cette approche informatique de fouille de données interactive au niveau de la mise en scène de théâtre ? Elle suppose l'introduction de l'ordinateur selon un mode de dialogue entre acteurs et machines. Une mise en scène peut-elle ne plus se conformer à une ontologie préexistante mais échapper à la spécification *a priori* en s'appuyant sur des interactions multi-modales ? C'est le sens de la recherche que nous avons menée dans le spectacle de théâtre inter-média *La traversée de la nuit*[2], sur le texte de Geneviève de Gaulle-Anthonioz [DE GAULLE98], évoquant son emprisonnement au cachot du camp de Ravensbrück à la fin de la Deuxième Guerre Mondiale.

[2] Pièce de théâtre donnée les 21, 22 et 23 novembre 2003 au Centre des Arts d'Enghien-les-Bains (95). Mise en scène : Christine Zeppenfeld ; comédiennes : Valérie Le Louédec et Magali Bruneau ; conception multimédia : Alain Bonardi et Nathalie Dazin ; musique : Stéphane Grémaud ; lumières : Thierry Fratissier.

4.1 Les interactions multi-modales dans la Traversée de la nuit

La mise en scène de *La traversée de la nuit*, repose sur un système homme-machine « autarcique » : une comédienne, Valérie Le Louédec, disant l'intégralité du texte, une danseuse, Magali Bruneau, accomplissant un certain nombre de gestes inspirés du théâtre Nô et un ordinateur multimédia, acteur artificiel. L'ordinateur se manifeste sous forme d'images projetées sur un écran de fond de scène de très vastes dimensions (la comédienne et la danseuse en voient toujours au moins une partie sans se retourner), provoquant la réaction des deux comédiennes, notamment de la danseuse adaptant la réalisation de sa gestuelle aux mouvements et qualités de l'image. Or, les deux actrices sur scène constituent les deux versants du même personnage – conscient et inconscient, selon les traditions du *shite* et du *waki* du théâtre Nô. Entraînée dans ses déplacements par la danseuse, la comédienne adapte elle aussi sa déclamation, sans compter les moments où elle regarde aussi l'écran. Pour boucler la boucle, l'ordinateur capte les états émotionnels de la voix de la comédienne.

Figure 1 : Exemple de génération d'images sur l'écran de fond de scène dans La traversée de la nuit *(Valérie Le Louédec à gauche, Magali Bruneau à droite; photographie : Julien Piedpremier)*

4.2 Description informatique du système homme-machine

L'implémentation informatique du système homme-machine est fondée sur un réseau de neurones d'analyse de la voix en entrée et un système multi-agents générateur d'images en sortie.

Le système informatique multimédia temps réel mis en œuvre est constitué en entrée d'un réseau de neurones destiné à reconnaître des états émotionnels dans la voix de la comédienne et en sortie d'un système multi-agents générateur d'images projetées sur l'écran. L'ensemble a été codé en utilisant la plateforme de développement graphique temps réel Max/MSP/Jitter.

Le réseau de neurones a été entraîné en mode supervisé pendant plusieurs mois par rapport à une liste d'états émotionnels que s'imposait la comédienne durant la lecture du texte complet. La voix en entrée est traitée phrase par phrase, chacune donnant lieu au calcul d'un vecteur de douze composantes : quatre d'entre elles concernent la prononciation des voyelles (formants), quatre d'entre elles représentent le caractère bruité de la voix et donc la prononciation des consonnes; les quatre derniers paramètres s'attachent à la prosodie (courbe d'amplitude de la voix dans la phrase). Pour chaque vecteur présenté en entrée, le réseau de neurones fournit un état émotionnel « reconnu ».

Le système multi-agents permet la génération temps réel d'images projetées en fond de scène. Les agents sont comme des « colleurs d'affiches » dynamiques qui construiraient ensemble des images toujours renouvelées.

- Chaque agent possède un petit modèle psychologique de sensibilité (positive ou négative), qui réagit selon les séquences de texte, aux états émotionnels du réseau de neurones. Il en résulte, en fonction de ce qu'indique le réseau de neurones, et en fonction des poids de sensibilité, une humeur qui conditionne leur « volonté » d'accomplir les tâches à mener.
- Les agents coopèrent à un but qui est l'optimisation d'une fonction d'utilité de l'image (une différente par séquence de texte).
- Les agents se coordonnent dans l'exécution de ce buts commun par rapport à l'état émotionnel reconnu par le réseau de neurones, par un mécanisme de compensation d'humeurs : ceux qui sont « d'excellente humeur » (grande valeur positive) concèdent un peu de leur ardeur à ceux qui ont une humeur très négative.
- Les agents communiquent entre eux deux à deux à période fixe en se transmettant leurs humeurs respectives.
- L'environnement des agents est constitué des états émotionnels reconnu par le réseau de neurones, du repère d'événement indiquant à quel endroit on se trouve dans la pièce et de valeurs propres à la séquence du texte correspondante et des indications d'un agent-observateur indiquant les qualités de l'image globale construite.

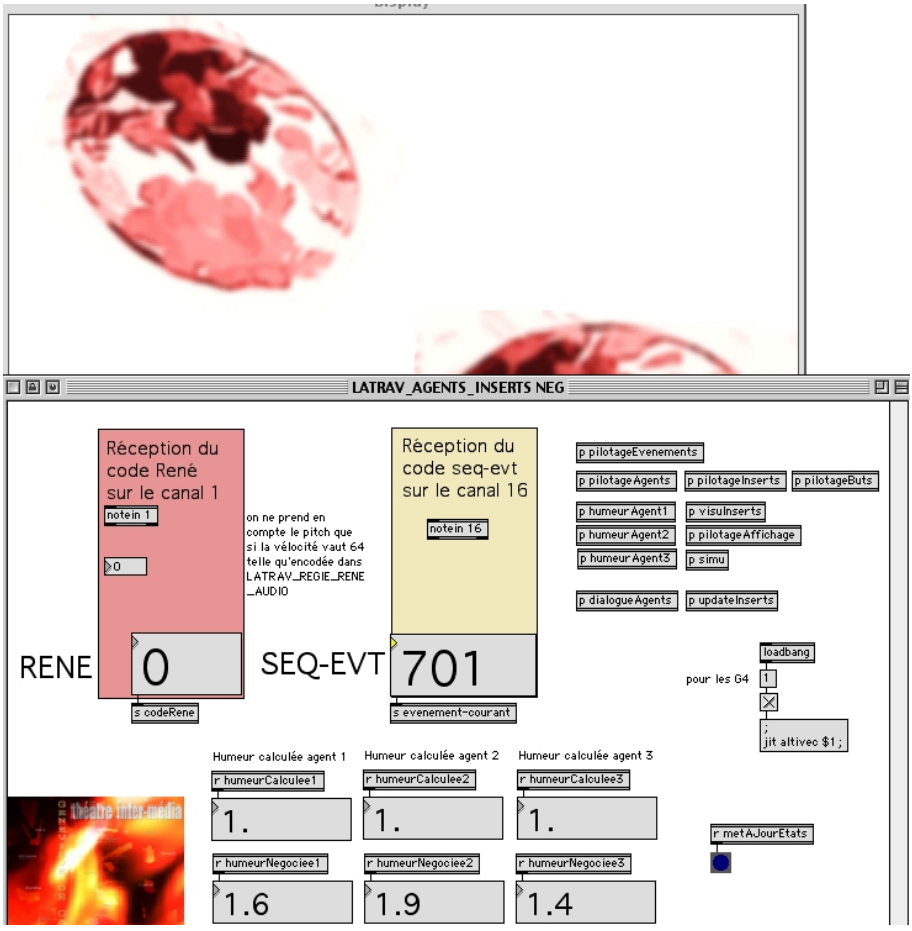


Figure 2 : Exemples de patchs Max/MSP/Jitter. En arrière, deux agents autonomes portant des fragments d'images dans La traversée de la nuit ; en avant, une partie de l'écran de pilotage (source : Alain Bonardi)

5. Conclusion

Nous avons montré comment l'approche informatique de la similarité fondée sur la fouille de données interactive peut inspirer de nouvelles modalités de mise en scène de théâtre associant acteurs et ordinateurs. Sortir de la mise en scène traditionnelle fondée sur les ontologies conduit irréversiblement à un affadissement de l'instanciation au profit de la manipulation active de contenus numériques passant par des transformations de données souvent irréversibles ; en cela, mixer une compilation ou une séquence musicale dans un logiciel *ad hoc* ressemble en profondeur à l'établissement *live* face au public d'une continuité dramatique, lorsque les ordinateurs, devenus acteurs, et les comédiens, se provoquent mutuellement. Dans les deux situations, la machine est engagée dans un fonctionnement heuristique.

Remarquons enfin que ces nouvelles approches de théâtre inter-média ouvrent des possibilités de simulation homme-machine : le metteur en scène peut utiliser ces systèmes pour la simulation dynamique de ses idées.

6. Références bibliographiques

[DE GAULLE98] de Gaulle, Geneviève, *La traversée de la nuit*, Éditions du Seuil, 1998.

[DONIN04] Donin, Nicolas, *Towards Organised Listening: Some Aspects of the Signed Listening Project at Ircam*, Organised Sound, Cambridge University Press, 2004.

[PACHET03] Pachet, François, *Nom de fichiers : Le nom*, actes du séminaire STP de la MSH Paris, 2003.

[PERROT94] Perrot, Jean-François, *Des objets aux connaissances*, Journée Méthodes objets et Intelligence Artificielle : Frontières, Ponts et Synergies, Paris RIA, juin 1994.

[ROUSSEAUX & BONARDI04a] Rousseaux, Francis; Bonardi, Alain, *Music-ripping : des pratiques qui provoquent la musicologie*, Musicae Scientiae, 2004.

[ROUSSEAUX & BONARDI04b] Rousseaux, Francis; Bonardi, Alain, *Reconcile Art and Culture on the Web*, 1st Workshop on Philosophy and Informatics, Cologne, 2004.

[ROUSSEAUX & BONARDI02] Rousseaux, Francis; Bonardi, Alain, « Vagabonds, pédants ou philistins : choisir en beauté », in *Art lyrique et art numérique*, Observatoire Musical Français, Série Conférences et séminaires n °13, 2002.

Session 4

Sémantique linguistique et applications documentaires

Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet

Mathieu Valette

Centre de Recherche en Ingénierie Multilingue (CRIM)
Institut National des Langues et Civilisations Orientales (INaLCO)
2 rue de Lille, 75343 Paris cedex 07 - France

mathieu.valette@inalco.fr

Résumé :

La demande pressante des institutions en matière de protection des usagers contre les contenus illicites ou préjudiciables sur Internet (racisme, xénophobie, pédophilie) invite à dépasser les systèmes de filtrage automatique conventionnels basés sur des listes de mots-clés ou des annuaires d'adresses préétablies, peu efficaces et exigeant de fréquentes mises à jour. *PRINCIP*, la plate-forme multilingue de détection de pages Web racistes dont nous présentons quelques aspects, met en jeu une analyse sémantique globale, multicritère, et différentielle des documents. Elle repose à la fois sur les propositions théoriques de la linguistique interprétative et les possibilités offertes par l'implémentation dans un système multi-agents, tout en se démarquant des approches ontologiques classiques.

Mots-clés : Internet, détection automatique, sémantique interprétative.

Abstract:

The authorities pressing needs regarding Web-users protection against illegal or abusive content on the Net – racism, xenophobia, paedophilia – have implied setting aside conventional key-word-based filtering systems as well as black lists, given their lack of efficiency and the need for frequent updating. PRINCIP, the multilingual platform for filtering racist pages on the Web is based on a global,

multi-criteria differential semantic analysis of Web pages based on the breakthroughs of interpretative semantics as well as the opportunities arising from implementation in a Multi-Agent System, in contrast to conventional ontological approaches.

Keywords: Internet, automatic detection, interpretative semantics.

1. Problématique

Le projet PRINCIP (Plate-forme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet, http://www.princip.net) est un système de détection automatique des pages Web racistes, xénophobes développé conjointement par plusieurs laboratoires de recherche européens[1]. Il repose sur une critique des systèmes de filtrage actuels, et notamment sur ceux qui recourent à de simples listes de mots-clés (CybertSitter, CyberPatrol). Ceux-ci témoignent en effet d'une approche naïve du texte raciste, suggérant qu'il y a des mots racistes et des mots qui ne le sont pas, sans considération pour leur mise en texte. Autrement dit, ces systèmes reposent sur un préjugé ontologique discutable, comme si le racisme était une langue de spécialité avec une terminologie stable et univoque : il y aurait des concepts racistes et des mots leur correspondant.

Pourtant, l'analyse des textes racistes montre une toute autre réalité : d'une part, en tant qu'expression d'une opinion, le racisme n'est pas un discours référentiel, mais relève davantage de la rhétorique[2] ; d'autre part – et en conséquence – sa caractérisation et sa détection impliquent la prise en compte de l'intertextualité inhérente au Web, manifestée, dans le cas présent, par la présence de sites *sur* le racisme, c'est-à-dire antiracistes, qui partagent avec les textes racistes une part non négligeable de leur vocabulaire. En bref, l'idée de *mots*-clés racistes s'avère peu pertinente : les traits sémantiques caractéristiques du texte raciste se situent en deçà, ou au-delà de ces mots-clés, privilégiés par l'approche ontologique.

Sans prétendre décrire de manière exhaustive l'ensemble des stratégies mise en œuvre pour détecter le racisme dans le cadre de PRINCIP, nous exposerons dans le présent article différents aspects de l'approche sémantique non ontologique que nous

1 Financé intégralement par la Commission Européenne, dans le cadre du *Safer Internet Action Plan*, le consortium PRINCIP comprend notamment le Centre de Recherche en Ingénierie Multilingue de l'Institut National des Langues et Civilisations Orientales de Paris, le Laboratoire d'Informatique de l'Université Paris 6 - Pierre et Marie Curie, l'Institut für Germanistik de l'université Otto-von-Guericke à Magdebourg, la School of Applied Language and Intercultural Studies de la Dublin City University.

2 Les fondements culturels de cette rhétorique ont fort bien été étudiés par Denis Blondin [BLON95] : en bref, elle repose sur l'opposition Nous *vs.* les Autres.

privilégions. Nous aborderons le problème de l'intertextualité et du choix théorique qui en découle (§ 2) ; puis nous présenterons les différents critères sémantiques retenus pour la caractérisation des documents racistes sur Internet selon les trois paliers de description du texte proposés par la sémantique interprétative de François Rastier (§ 3) ; enfin, nous décrirons brièvement les options retenues en termes d'implémentation dans un système multi-agents (§ 4).

2. L'intertextualité de l'Internet

2.1 Racisme et antiracisme, frontières et recouvrements

De précédent travaux d'analyse du discours, notamment ceux de Simone Bonnafous [BONN89, 91] et Pierre-André Taguieff [TAGU88], ont mis en évidence la dialectique qui oppose et lie tout à la fois les auteurs antiracistes aux auteurs racistes. Pour ces raisons, détecter les pages Web racistes n'est envisageable qu'à la condition de prendre en compte l'*intertextualité* avec d'autant plus d'attention qu'elle est massive et généralisée sur Internet, où les contenus ne sont pas qualifiés ni hiérarchisés par les moteurs de recherche.

Globalement, les modalités de l'intertextualité des documents racistes et antiracistes relèvent de la citation (les antiracistes citent les textes racistes) et de l'appropriation (les racistes s'approprient le vocabulaire antiraciste).

La rhétorique antiraciste consiste en effet à déconstruire l'argumentation des textes racistes, de sorte qu'une large place est faite aux citations, celles-ci pouvant aller du simple mot au paragraphe, voire davantage. Par conséquent, les lexies les plus stables et les plus ancrées dans le vocabulaire des auteurs racistes, c'est-à-dire celles qui feraient de bons candidats *a priori* à la constitution d'une liste de mots-clés, sont celles dont les auteurs antiracistes vont faire un usage critique privilégié. En somme, le discours rapporté fausse sensiblement les statistiques sur corpus. Par exemple, les lexies « *Race blanche* », ou « *bougnoule* », réputée racistes, sont en fait, dans environ deux tiers des cas, actualisées dans des textes antiracistes. Le phénomène est le même pour le vocabulaire xénophobe d'extrême droite (« *immigrationisme* », « *immigration-invasion* », « *complot judéo-maçonnique* », etc.).

Parallèlement, les auteurs racistes s'approprient certaines lexies antiracistes notoires. Par exemple « *pote* », emblème lexical de l'association SOS-Racisme (*La marche des potes* en 1983), s'il n'est plus guère utilisé par celle-ci que dans des lexies composées figées (par exemple, les associations de quartiers « *les maisons des potes* »), est remotivé par les auteurs racistes, qui l'emploient à des fins euphémiques. Il en est de même pour le verlan « *beur* » (ou « *beurette* »), également popularisé par la lutte contre le racisme du début des années 80 : dans notre corpus d'analyse, 77,22 % des occurrences relèvent en fait des textes racistes.

Cette intertextualité trouve d'autres formes de manifestations plus problématiques encore, parce qu'elles ressortissent à une rhétorique de la page Web. Ainsi, tel site raciste reproduira *in extenso* un article de la presse non raciste (*L'Express, Le Monde*) s'il traite d'un fait de société qui intéresse son propos xénophobe (par exemple, les tournantes, viols collectifs commis dans les quartiers défavorisés, thème alors associé à la purification ethnique). Dans ce cas, l'euphémisation est maximale, car le Webmestre n'a pas à ajouter le moindre commentaire : le péritexte (sommaire, liens connexes) suffit aux lecteurs pour mesurer son intention.

La prise en compte de l'intertextualité impose donc de dépasser l'idée qu'il existe des concepts racistes et antiracistes (ou non racistes) actualisés de part et d'autre d'une frontière idéologique. Le matériau lexical raciste s'avère un point d'accès à la problématique, mais ne suffit pas, loin de là, à sa détection. Le racisme est l'expression d'une opinion, non la description d'un univers conceptuel.

2.2 Une approche différentielle des textes

Si, comme nous l'a enseigné Saussure et à sa suite, la sémantique structurale, la valeur linguistique est définie par des oppositions, il apparaît légitime d'adopter une approche différentielle des textes racistes et antiracistes. La sémantique différentielle apporte la solution théorique adéquate à ce cas de figure, en décrivant les éléments signifiants de la langue dans des systèmes d'oppositions et non sur un mode référentiel.

Nous présenterons dans cet article une interprétation et une mise en application de quelques unes des propositions théoriques de François Rastier émises dans le cadre de la sémantique interprétative (cf. [RAST94, 01]). À la différence d'autres travaux récents auxquels on pourra légitimement comparer notre approche ([BEUS98], [TANG97], [THLI98]), nous faisons un usage opportuniste de la théorie, n'en retenant que certains aspects jugés particulièrement adéquats à la problématique de la détection du racisme. L'objectif de PRINCIP en effet est de détecter convenablement les documents racistes sur Internet, non d'évaluer l'applicabilité de la sémantique interprétative.

En l'occurrence, nous présenterons dans cet article une interprétation et une exploitation de l'opposition *fond sémantique* vs. *forme sémantique*. Dans la sémantique interprétative, le fond sémantique est assimilé à une certaine catégorie d'unités textuelles : les *isotopies*, organisées en faisceaux (une isotopie est l'effet de récurrence d'un même sème) tandis que les formes sémantiques correspondent à une autre catégorie d'unités textuelles que sont les *molécules sémiques* (groupe stable de sèmes non nécessairement lexicalisé).

L'hypothèse principale qui préside à PRINCIP est que les textes racistes et antiracistes partagent un même fond commun mais qu'ils se distinguent par la *saillance* de formes sémantiques soit racistes, soit antiracistes. Ce sont donc les notions générales de fond et de formes sémantiques que nous retiendrons, plutôt que les unités sémantiques auxquelles elles correspondent théoriquement.

Nous aborderons cette question aux trois niveaux d'analyse du texte définis par François Rastier :

1. Le niveau microsémantique, où nous étudierons les règles de constitution des lexies racistes ou antiracistes ;
2. Le niveau mésosémantique, où seront abordées les unités textuelles non lexicalisées, ou n'ayant pas de lexicalisation privilégiée : isotopies sémantiques, molécules sémiques) ;
3. Le niveau macrosémantique, celui des discours et des genres textuels déterminés par un ensemble hétérogène d'indices d'expression.

3. Les critères sémantiques pour la caractérisation

3.1 Niveau macrosémantique : le global et le local

Alors que le filtrage par mots-clés repose sur un seul palier de la description linguistique, la détection multicritère mise en place par PRINCIP s'appuie sur plusieurs paliers de complexité textuelle : lexie, période ou section, et texte, ce dernier jugé primordial dans le cadre de la sémantique interprétative dans la mesure où il détermine le sens des unités de paliers inférieurs (cf. [RAST94, 01]).

La thèse défendue par François Rastier dans sa sémantique interprétative, selon laquelle le global (le texte) détermine le local (le signe) apparaît en effet particulièrement adaptée au filtrage automatique des textes d'opinion, même à un niveau d'analyse relativement rudimentaire. Les données locales, dans les textes racistes, relèvent des lexies susceptibles d'être citées par les antiracistes. Les données quantitatives non spécifiquement lexicales, conditionnées par le genre textuel, seront assimilées à des données globales.

Nous avons distingué deux types de données globales :

1. Celles, proprement textuelles, relevants des genres et des discours dans lesquels sont actualisés les textes racistes (ou antiracistes) ;
2. Celles, infratextuelles, qui ressortissent à une sémiotique plus générale des documents Web (images, polices de caractères, code couleurs, etc.).

Si mettre au même niveau deux types de données *a priori* fort différents peut surprendre, le rôle interprétatif des données de structuration du document HTML apparaît pourtant, comme nous allons le voir, déterminant, au même titre que les enluminures médiévales. La page Web structurée en HTML, quel qu'en soit le contenu, est soumise à des contraintes intertextuelles fortes qui déterminent la forme du document et les formes du texte. Autrement dit, une page Web, même « vide », présente déjà un fond structurel commun à toutes les pages du sites, que ce soit au niveau des étiquettes HTML elles-mêmes (structuration de la page, métadonnées) ou

du matériel lexical affiché à l'écran (par exemple le *péritexte* : sommaire, rubrique, etc.). L'ensemble constitue ce que nous appellerons la *signature sémiotique* du site.

Ainsi, sur un corpus comprenant la totalité des pages d'un site raciste donné[3], nous avons mesuré que sur le texte seul, 24,75 % des occurrences de formes appartenaient à ces informations péritextuelles communes à toutes les pages du site. Sur la source HTML, étiquettes et péritexte confondus, ce pourcentage atteint 47,45 % – c'est dire le poids de ces données souvent oubliées en linguistique textuelle. Lié à la question des genres du Web, leur statut herméneutique reste à approfondir.

3.1.1 Données globales et genres textuels

La catégorisation manuelle des corpus d'apprentissage a permis de dresser l'inventaire des genres et des discours dans lesquels s'inscrivent la plupart des textes racistes. On a relevé principalement des discours littéraires (textes de chansons, récits, témoignage), politiques (tract, discours, programme) et journalistiques ou idéologiques (article, pamphlet, opinion, faits-divers).

Mais l'un des genres privilégiés des auteurs racistes est le pamphlet ou le libelle. Cela se manifeste par des informations textuelles caractérisant la diatribe et la polémique : points d'exclamation, adverbes de négation ou d'évaluation dénotant un style outré ou hyperbolique (« *jamais* », « *rien* », etc.), pronom et désinence de la deuxième personne du pluriel, morphèmes dépréciatifs (« *-âtre-* ») ou vulgaires (« *foutr-* »), etc.

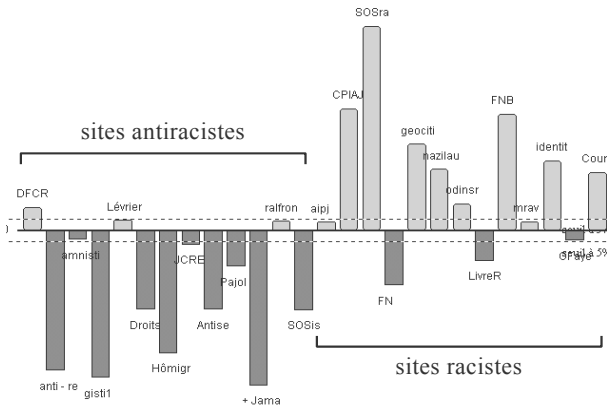


Figure 1 : Fréquences relatives d'un ensemble de formes caractérisant la diatribe dans un corpus constitué de 13 sites antiracistes (à gauche) et 13 sites racistes (à droite)

[3] SOS-racaille.org qui a été interdit en 2003, mais dont il existe quelques avatars aujourd'hui.

Comme les textes antiracistes sont rarement pamphlétaires, ces critères d'expression sont sensiblement plus fréquents dans les documents racistes qu'antiracistes, comme l'illustre la figure 1 où sont présentées les fréquences relatives cumulées d'une sélection de formes participant au genre pamphlétaire (« *jamais* », « *rien* », « *peu* », « *tout* », « *trop* », et le point d'exclamation) dans un ensemble de 13 sites antiracistes et 13 sites racistes. Les colonnes sombres marquent un déficit, les colonnes claires un excédent par rapport à une fréquence théorique indiquée par la ligne médiane. De cet écart quantitatif, la plate-forme PRINCIP infère une différence sémantique suffisante pour catégoriser adéquatement les documents.

Ainsi, lorsque PRINCIP aura à traiter un document comprenant, par exemple, une occurrence de la lexie raciste « *immigration-invasion* »[4], elle évaluera l'opinion de l'auteur à partir des critères d'expression dits de *bas niveau* (i.e. hors lexies racistes) présentes dans le texte et calculera le « taux » de racisme et le « taux » d'antiracisme du document. En d'autres termes, les données globales (comme le genre, qui conditionne les critères d'expression de bas niveau) ont une incidence sur les données locales (lexies) dans la mesure où elles leur donnent un *sens* raciste ou antiraciste.

Parce qu'ils sont *a priori* sans lien sémantique avec les critères lexicaux de haut niveau, et que, par conséquent, ils demeurent en deçà de la conscience des auteurs et des lecteurs des textes, ces critères de bas niveau relèvent d'un *implicite inconscient* (déterminé par le choix du genre par l'auteur)[5]. Ce défaut de conscience s'avère crucial dans la mesure où il assure la pérennité du système : si les lexies se périment, les genres, eux, s'avèrent beaucoup plus stables dans le temps.

Les textes antiracistes présentent eux aussi des critères d'expression de bas niveau. Légalistes, ils comportent les indices du genre ; par exemple, certaines entités nommées et dates anciennes y sont fréquentes et témoignent, par-delà la problématique des genres, d'une mémoire des événements (textes de loi, action historique contre le racisme, etc.) complètement absente des textes racistes.

3.1.2 Données globales infratextuelles

Les critères globaux retenus par PRINCIP ne sont pas seulement linguistiques, et relèvent aussi de la structuration du document numérique. Le code HTML fournit de précieux critères d'expression.

Il apparaît que globalement, les auteurs antiracistes ont davantage recours aux étiquettes de mise en forme que les antiracistes, qui se distinguent quant à eux par un usage plus poussé des possibilités multimodales du HTML. En d'autres termes, les antiracistes mettent en ligne des textes quand les racistes produisent des documents Internet. Quelques exemples en donneront une bonne illustration.

---

[4] Sur la collecte des lexies racistes, cf. [VALE04], pp. 1109-1110.

[5] Nous savons gré à François Rastier d'avoir porté notre attention sur ce point.

L'organisation et la hiérarchisation des parties d'un texte en titres et sous-titres (indiquées par les balises <H1>, <H2>, <H3>, etc.) est banale dans les textes antiracistes mais rarissime dans les textes racistes : 45,14 % des textes antiracistes y recourent contre seulement 1,67 % des textes racistes. Les taux de précision sont respectivement de 96,42 % et 3,58%[6]. Les listes structurées (balises <UL>, <OL>, <LI>) sont également une spécificité antiraciste (42,89 % des documents antiracistes en contiennent contre 14,15 % des documents racistes, et le taux de précision est de 75,19 % au bénéfice des antiracistes).

Dans la mesure où, comme nous l'avons vu, les textes antiracistes donnent une large place aux citations, les balises qui leur sont spécialement dédiées (<CITE>, <BLOCKQUOTE>) y apparaissent sensiblement plus fréquemment (dans 18,45 % des cas, et seulement dans 4,84 % des textes racistes), et ce avec une précision de 79,22 %.

Les polices de caractère peuvent également être très discriminantes. Dans notre corpus contrasté, la police Verdana apparaît très spécifique aux pages racistes (le taux de précision raciste est de 92 % et le taux de rappel antiraciste de seulement 3 %). Les balises emphatiques (<EM>, <STRONG>), qui tendent à se substituer aux classiques italiques et gras, sont privilégiées par les antiracistes (dans 56,85 % des cas, avec une précision de 73,87 %), et ignorée des racistes dont la préférence semble aller vers le souligné (<U>) : 38,36 % des documents racistes en contiennent contre 12,96 % des documents antiracistes ; pour une précision raciste de 74,73 %.

Si les antiracistes sont des gens de l'écrit, comme semble l'attester leur sens de la composition, les racistes se sont appropriés le potentiel multimodal du Web avec davantage d'adresse. Dans ces travaux de description sémantique des images racistes sur Internet réalisés pour PRINCIP, Monica Nincinski [NINC04] a montré que les codes couleur ne sont pas les mêmes chez les antiracistes et chez les racistes. Chez ces derniers, ils reposent en partie sur un contraste clair-obscur qui fait écho à la rhétorique que l'on rencontre dans les textes (nous vs. les autres)

En effet, l'opposition entre le rouge et noir, sans évidemment être une exclusivité du racisme, apparaît très caractérisante pour une approche différentielle racisme/antiracisme. Le rouge, notamment domine dans les sites racistes : nous avons en effet mesuré que dans un corpus constitué de pages HTML racistes et antiracistes de 20 millions de caractères, les deux tiers des occurrences des principales étiquettes correspondant à cette région de la palette chromatique (rouge primaire : #FF0000; rouge profond : #990000, rouge sang : #CC0000) se trouvent dans les pages racistes, avec un pic à 92 % pour le rouge sang.

De même, 80,28 % des images JPEG de notre corpus proviennent des sites racistes. Elles sont par ailleurs présentes dans 44,5 % des pages racistes et dans seulement 10,97 % des pages antiracistes.

---

[6] Le *rappel* est le rapport du nombre de documents pertinents sélectionnés au nombre total de documents du sous-corpus considéré. La *précision*, dans l'acception qui est la nôtre, est le rapport du nombre de documents sélectionnés pertinents au nombre total de documents sélectionnés dans l'ensemble du corpus. Le cumul des précisions équivaut donc à 100 %.

Si racistes et antiracistes font un usage semblable (statistiquement parlant) des image au format GIF, lorsqu'une image, quel que soit son format, est placée en arrière-plan (balise <BODY BACKGROUND=>), il s'agit dans 77,7 % des cas d'une page raciste. La présence de bannières est également un critère discriminant. Sur notre corpus de test, la totalité des étiquettes <BANNERS> se trouve dans les pages racistes.

Enfin, si l'hypertextualité (liens internes au site) semblent bien maîtrisée par les deux parties, les racistes, là encore prennent un léger avantage en ce qui concerne la connectivité (liens externe) : 71,32 % des documents racistes contiennent au moins un lien, et parmi ces documents, la moyenne dépasse les trois liens par page (3,1 liens/page) ; tandis qu'une moitié seulement des documents antiracistes (50,62 %) s'ouvre vers la Toile, en proposant, en moyenne, à peine plus d'un lien (1,07 liens/page). Enfin, les Webmestres racistes offrent plus volontiers la possibilité d'un contact par courrier électronique que leur détracteurs : 76,73 % des occurrences de l'étiquette correspondante sont le fait des textes racistes.

3.2 Niveau mésosémantique : les unités textuelles

Si PRINCIP relativise l'importance des concepts et des mots-clés qui y sont associés dans son approche du texte raciste, cela ne signifie pas pour autant que les *unités* textuelles y sont négligées, bien au contraire ; il s'agit de privilégier d'autres unités textuelles, qui du point de vue du traitement, correspondent à des *cooccurrences* de morphèmes ou de mots[7], et dans une perspective strictement sémantique, de traits récurrents et de groupes de traits. À la différence des mots isolés, les unités textuelles peuvent donc être discontinues. Leur actualisation ne dépend pas de la présence de la totalité des items qui la composent, elle est graduelle, de sorte qu'il est possible de faire évoluer le seuil de présence des items à partir duquel une unité est considérée comme actualisée. Il peut être relativement bas si le document contient déjà beaucoup d'indices. En bref, d'un point de vue quantitatif, une unité textuelle n'est pas soit présente, soit absente, elle est plus ou moins présente. Parmi ces unités, les *thèmes* (ou *molécules sémiques*) ont été particulièrement étudiés, dans une perspective que nous allons présenter maintenant. D'autres pistes de recherches ont été également explorées, notamment, les *isotopies sémantiques*. En raison de contraintes éditoriales, nous n'en traiterons pas ici, sinon, ultérieurement, dans le paragraphe 4 : « Analyse multicritère et système multi-agents », à des fins illustratives.

Le thème sémantique (ou *molécule sémique*) consiste en un « groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier » ([RAST94], p. 223). Comme nous n'avons pas envisagé de constituer des

[7] Dans l'acception qui est la nôtre, deux items sont en cooccurrence lorsqu'ils sont actualisés dans une même fenêtre prédéterminée (paragraphe, alinéa, etc.). Nous opposons ainsi la cooccurrence à la *collocation*, où les items sont immédiatement voisins.

dictionnaires sémiques, nous avons déterminé plusieurs façons d'exploiter la notion de thèmes telle que la conçoit la sémantique interprétative. La plus productive consiste à isoler les cooccurrents d'une lexie relevant du fond sémantique dans des contextes racistes puis antiracistes, de manière à en identifier les spécificités. Ces cooccurrents sont rebaptisés des *corrélats* lorsqu'ils sont jugés qualifiants sémantiquement et qu'ils sont *saillants* (pour une discussion, lire [RAST01], pp. 211-213). Ils relèvent alors des formes sémantiques (racistes ou antiracistes). La seconde façon consiste à distinguer les différents éléments lexicaux d'un micro-récit fréquent dans les textes racistes ou antiracistes, comme par exemple les viols collectifs (assimilés à la purification ethnique) chez les racistes ou l'organisation de manifestations (ressortissant à la pratique de la vie associative) chez les antiracistes.

Pour isoler les corrélats (formes saillantes) d'une lexie appartenant au fond sémantique, nous avons utilisé les sorties du logiciel de lexicométrie Hyperbase (Étienne Brunet, université de Nice, http://ancilla.unice.fr/), avant de développer notre propre chaîne de traitement. La procédure est la suivante :

1. Relevé de tous les contextes de la lexie étudiée (ou mot-pôle, *mp*) dans le sous-corpus d'apprentissage raciste (*scr*), puis dans le sous-corpus antiraciste (*sca*),
2. Raboutage de l'ensemble de ces contextes de manière à constituer deux nouveaux textes, un raciste (*ntr*) et un antiraciste (*nta*),
3. Mesure à l'aide d'un test d'écart-réduit (loi normale) des spécificités de ntr et nta par rapport à *scr* et *sca*,
4. Sélection des corrélats dans la liste des spécificités (cooccurrents) obtenue.

Le tableau de la figure 2 représente une sélection réalisée sur deux sous-corpus, l'un composé de textes racistes (à gauche), l'autre de textes antiracistes (à droite), à partir du mot-pôle « *immigration* ».

Il s'ensuit la neutralisation des cooccurrents peu ou non pertinents du mot-pôle (par exemple, dans le cas présent : « *clandestine* », « *insécurité* ») et sa qualification *sémantique* par delà sa signification propre (le concept d'immigration).

Ainsi, les corrélats racistes d'« *immigration* » participent de lexies composées telles que « *immigration incontrôlée* », « *immigration croissante* », « *immigration-invasion* », « *immigration-colonisation* », « *immigration de peuplement* », etc., tandis que les corrélats antiracistes suggèrent une problématique des « *flux migratoires* » et de la « *fermeture des frontières* ».

Les noms mentionnés par les racistes sont des personnalités politiques (Le Pen, Chirac) ou, plus incidemment, des idéologues de la lutte contre l'immigration, tandis que les « entités nommées » antiracistes relèvent de documents législatifs (rapport Weil, lois Debré, traité d'Amsterdam, etc.).

Mot-pôle : « immigration » (fond sémantique)			
Environnement raciste du mot-pôle (forme sémantique raciste) :		*Environnement antiraciste du mot-pôle (forme sémantique antiraciste) :*	
écart-réduit	*forme*	*écart-réduit*	*forme*
32.14	incontrôlée	43.10	clandestine
26.16	clandestine	25.11	politique
25.78	insécurité	23.33	flux
21.97	massive	19.61	frontières
21.27	intégration	17.87	zéro
18.58	invasion	16.83	migratoires
18.16	colonisation	15.71	Weil
18.10	ratée	14.78	insécurité
16.77	peuplement	13.34	intégration
16.05	chômage	13.19	fermeture
15.21	extra	12.93	chômage
14.81	problèmes	12.18	maîtrise
14.39	population	12.17	Amsterdam
14.01	regroupement	11.89	émigration
13.44	démographique	11.58	asile
13.07	musulman	11.58	question

Figure 2 : thème sémantique d'« immigration » (extrait)

Enfin, les populations sont qualifiées de façon continentale, géographique, ethnique ou confessionnelle dans les textes racistes (« *extra-européens* », « *afro-arabes* », « *afro-maghrébins* », « *musulmans* ») et par leurs origines nationales par les antiracistes (« *turque* », « *italiens* », « *portugais* », « *algérienne* », etc.). Si cette procédure de détection n'est pas, à l'heure où nous écrivons ces lignes, complètement stabilisée, on observe pour le moment un gain en termes de précision de l'ordre de 30 % (en moyenne) par rapport aux valeurs du mot-pôle.

3.3 Niveau microsémantique : la composition des lexies

L'opposition fond/formes sémantiques appliquée au niveau lexical nous a été inspirée par la très grande créativité lexicale des auteurs racistes. Nous avons constaté que nombres des lexies racistes étaient composées d'un ou quelques morphèmes du fond sémantique et d'un ou quelques morphèmes de la forme sémantique racistes.

Une illustration exemplaire peut-être donnée par le couple d'antonymes « *judéophobie* » et « *judéophilie* » : tous deux partagent un morphème du fond sémantique (« *judéo-* ») et s'opposent par leurs suffixes : « *-phobie* » relève de la forme antiraciste tandis que « *-philie* » est une forme raciste. En collaboration avec

Anne-Laure Jousse (INaLCO), nous avons ainsi constitué plusieurs dictionnaires morphémiques et étudié les principales règles de constitution des lexies.

La constitution du dictionnaire morphémique est déterminée par le taux de rappel et de précision de différents morphèmes (« *euro-* », « *franc-* », « *démocr-* », etc., ou des entités nommées telles que « *LICRA* », etc.). La précision d'un morphème du fond sémantique doit tendre vers 50 % raciste, 50 % antiraciste et 0 % neutre, quand son rappel, sans être déterminant, doit être le plus élevé possible.

Le lexème « *démocr* », par exemple, a été retenu pour le fond sémantique parce que ses valeurs sont, de ce point de vue, excellentes (cf. figure 3). Bien que moins exemplaire, l'entité nommée « *LICRA* » présente cependant des valeurs intéressantes (cf. figure 4) : son taux de rappel est relativement bas, mais il s'agit d'un fond « parfait » dans la mesure où, de par sa signification très restreinte (i.e. *Ligue Internationale contre le racisme et l'Antisémitisme*), elle n'est en pratique pas actualisée dans des textes dit neutres (i.e. ne relevant ni du racisme ni de l'antiracisme).

démocr-	*rappel*	*précision*
Corpus raciste	32,92	49,96
Corpus antiraciste	29,67	45,03
Corpus neutre	3,29	5,01

Figure 3 : Mesures (rappel et précision) du morphème démocr- (fond sémantique)

LICRA	*rappel*	*Précision*
Corpus raciste	2,19	49,45
Corpus antiraciste	2,24	50,55
Corpus neutre	0	0

Figure 4 : Mesures (rappel et précision) du morphème LICRA (fond sémantique)

	rappel	*précision*
-mafi(a)- (*mafia*, *mafieux*, etc.)	5,61	61,46
-ouill- (magouille, fripouille, etc.)	6,09	70,68
-man- (*israëlomane*, etc.)	23,65	68,37
-crass- (*crasseux*)	1,96	76,72

Figure 5 : Mesures de quelques morphèmes de la forme sémantique raciste[8]

[8] NB : Pour faciliter la lecture, nous ne donnons pas les mesures de rappel et de précision de ces morphèmes pour les autres sous-corpus. Ils sont évidemment très inférieurs. Par exemple, le taux de rappel antiraciste de *« -ouille- »* est de 1,24 % et sa précision de 14,45 % ; le taux de rappel neutre est de 1,28 % et sa précision de 14,86 %.

	rappel	*précision*
-phob- (xénophobe, islamophobie, etc.)	22,44	72,48
circul- (circuler, circulation)	22,19	67,77
universit- (universitaire, université, etc.)	12,46	58,17
résid- (résider, etc.)	19,45	57,31

Figure 6 : Mesures de quelques morphèmes de la forme sémantique antiraciste

Pour les formes sémantiques, on a choisi des morphèmes ayant un taux de précision élevé, puisque c'est la précision qui permet de différencier les deux formes sémantiques. Le taux de rappel, à cet égard, est moins déterminant. Les exemples proposés ci-dessous constituent de ce point de vue de bons spécimens. La figure 5 présente les taux de rappel et de précision de morphèmes racistes, la figure 6, ceux de quelques morphèmes antiracistes.

L'objectif de cette approche morphémique est d'anticiper sur la néologie des racistes. Ainsi, pour s'en tenir aux exemples présentés ici, si les lexies « *licrasse* » « *licrasseux* » (Fond « LICRA » + forme raciste « *crass* ») ont été repérées lors de tests sur Internet en avril 2004, ce n'est pas le cas de « *licrassouille* » qui pourtant, pourrait fort bien être actualisée un jour, d'autant plus que les lexies « *démocrasseux* » et « *démocrassouille* », par exemple, qui repose sur le même principe de composition , sont attestées dans nos corpus avec une précision raciste de 100 %.

4. Analyse multicritère et système multi-agents

La plate-forme de détection est implémentée au moyen d'un système multi-agents développé au sein de l'équipe OASIS (Objets et Agents pour Systèmes d'information et de Simulation) du Laboratoire d'informatique de Paris 6 (cf. par exemple [SLOD03]).

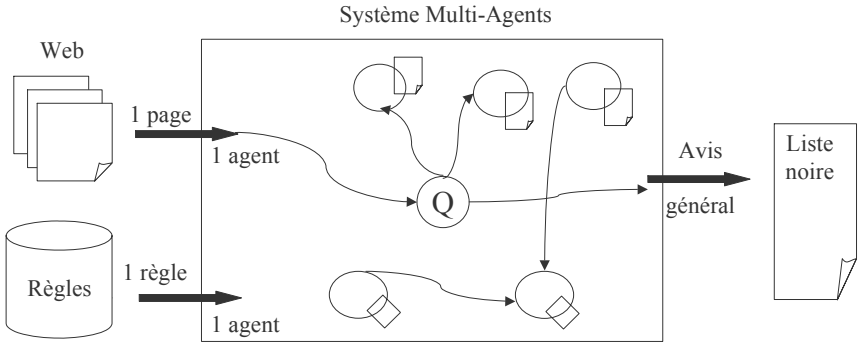


Figure 7 : Plate-forme multi-agents pour l'implémentation de PRINCIP

Les règles linguistiques permettant de détecter le racisme seront environ 300 par langue. Chacune de ces règles a la capacité d'exprimer une opinion sur les documents qu'on lui présente. C'est la somme de plusieurs opinions, parfois contradictoires qui permet de donner un avis général sur chaque document. En entrée du système multi-agents, chaque page Web est associée à un agent et se voit allouée un temps de traitement. Par ailleurs, chaque règle linguistique est associée à un agent.

Un agent de requête (Q, sur la figure 7) introduit l'ensemble des documents à analyser dans le système, ce qui a pour effet de générer autant d'agents-document, lesquels décident des agents-règles à appliquer à leur document, suivant un ensemble de critères complexes (vitesse d'exécution de la règle, nature des informations linguistiques qu'elle contient, fiabilité et adéquation de son jugement au document en présence, etc.). Chaque agent-règle exprime alors un vote sur le document. Lorsque le temps de traitement alloué est atteint, un « dépouillement » des votes est effectué. Si le résultat est jugé satisfaisant, l'agent de requête exprime un avis « raciste », « antiraciste », « ni l'un ni l'autre » ou n'en exprime aucun si le décompte des votes ne le permet pas. Dans ce dernier cas, un complément de temps de traitement peut être dispensé si des indices laissent néanmoins supposer que le document pourrait être raciste. Une catégorie particulière est réservée pour les documents qui apparaissent à la fois racistes et antiracistes.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<rule language="FR" ID="FR_dis:outgroup_myth_destruction_113">
        <active value="true"/>
        <procedure value="java:RulePatternMatch"/>
        <technique value="PatternMatcher"/>
        <clue>désagr[éè]g</clue>
        <clue>désorganis</cl
        <clue>alt[èé]r</clue>
        <clue>d[ée]labr</clue>
        <description/>
        <function value="dis:outgroup_myth_destruction"/>
        <input value="text"/>
        <level value="substring:root"/>
        <output value="racism_evidence" weight="3" precision="85.88"/>
        <recall value="4.15"/>
        <speed value="232"/>
        <threshold value="0"/>
</rule>
```

(1) Isotopie sémantique composée d'un ensemble de lexèmes

(2) Information sémantique : rhétorique de la destruction

(3) Information sur les performances de la règle : poids, précision, rappel, vitesse d'exécution, seuil

*Figure 8 : Un exemple de règle sémantique : fragment de l'isotopie /*destruction*/*

La figure 8 présente l'exemple minimal et simplifié d'une règle, en l'occurrence, une règle isotopique. Nous avons mis en évidence trois de ces caractéristiques principales. En premier lieu, un ensemble d'éléments (ici, des

lexèmes) qui constituent l'isotopie ou une partie de l'isotopie sémantique retenue (1) : ces différents lexèmes partagent un trait sémique /*altérité*/. L'isotopie relève d'une catégorie préalablement identifiée comme appartenant au mythe de la destruction. Cette qualification est précisée dans un champ prévu à cet effet (2). Elle est utilisée par l'agent de requête lors du choix des règles à appliquer sur un document. Enfin, la règle a été « mesurée » selon un ensemble de critères (3) : précision, rappel, poids, vitesse d'exécution, etc. Tous ces critères, et d'autres non détaillés ici, sont utilisés dans les stratégies coopératives du système multi-agents.

La plate-forme PRINCIP sera opérationnelle en juillet 2004. Ses performances sont actuellement évaluées par la Ligue Belge des Droits de l'Homme.

5. Conclusion

Dominée par l'approche ontologique (le web dit « sémantique ») et essentiellement cantonnée à la constitution de terminologie et à la veille, la problématique de la détection et de la catégorisation automatique est appelée à connaître quelques inflexions théoriques. Les textes susceptibles d'être traités automatiquement ne sont plus seulement ceux, univoques, des sciences et techniques. Internet, notamment, et la masse considérable de documents qui y circulent, et celle incommensurable de ceux qui y circuleront demain, créent de nouvelles demandes en termes de catégorisation, de classification et de filtrage : ce ne sont plus seulement des outils de *recherche* dont l'utilisateur a besoin, mais des outils d'*interprétation*. Ce sont donc de nouvelles méthodologies d'analyse des textes que les théoriciens doivent proposer aux ingénieurs du traitement automatique du langage.

La plate-forme PRINCIP, théoriquement fondée sur la sémantique interprétative, s'inscrit dans cette évolution. En se posant la question primordiale des genres textuels et de l'intertextualité sur le net, en travaillant avec des outils théoriques proprement sémantiques (et non seulement ontologiques), c'est-à-dire sur des morphèmes, des unités sémantiques non lexicalisées, voire sur des étiquettes HTML en tant qu'elles participent à la structuration du texte, plutôt que sur des concepts, des mots isolés et des phrases, l'équipe du projet PRINCIP entend participer à ce débat.

NB : Les travaux de synthèse présentés ici ont pour cadre un projet collectif. Je suis redevable aux membres du consortium PRINCIP de l'ensemble de ces réflexions, et notamment à l'équipe du Centre de Recherche en Ingénierie Multilingue de l'INaLCO. J'ai en outre plaisir à remercier Évelyne Bourion, Aurélien Slodzian et François Rastier pour les remarques, suggestions et critiques qu'ils ont formulées afin d'améliorer cet article. Je tiens également à signifier ma reconnaissance à Emmanuel Cohen, Alexander Estacio Moreno et Anne-Laure Jousse pour leur participation à ces recherches.

6. Bibliographie

[BEUS] Beust, P., *Contribution a un modèle interactionniste du sens. Amorce d'une compétence interprétative pour les machines*, Thèse de doctorat, Caen, 1998.

[BLON95] Blondin, D., *Les deux espèces humaines. Autopsie du racisme ordinaire*, Paris, L'Harmattan, 1995.

[BONN89] Bonnafous et P.A. Taguieff (éd.), *Racisme et antiracisme : frontières et recouvrements*, Mots 18, 1989.

[BONN91] Bonnafous S., *L'immigration prise aux mots*, Kimé, 1991.

[NINC04] Nicinski, M., « Typologie et description sémantique des images utilisées dans les sites Internet racistes », *Caractérisation des contenus de l'Internet : au-delà du lexique, l'approche sémantique*, journée ATALA organisée par F. Rastier, N. Grabar, T. Beauvisage, 31 janvier 2004, Paris.

[RAST94] Rastier, F., Cavazza, M., Abeillé, A., *Sémantique pour l'analyse: de la linguistique à l'informatique*, Paris, Masson, 1994.

[RAST01] Rastier, F., *Arts et sciences du texte*, Paris, PUF, 2001.

[SLOD03] Slodzian, A., Aknine, S., « Intelligent Agents for Tracking Racist Documents on the Internet », *Workshop on Intelligent Techniques for Web Personalization (ITWP '03)*, Acapulco (Mexique) 2003.

[TAGU88] Taguieff, P.-A., *La force du préjugé. Essai sur le racisme et ses doubles*, Paris, La découverte, 1988.

[TANG97] Tanguy, L., *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative*, Thèse de Doctorat, Rennes 1, 1997.

[THLI98] Thlivitis, T., *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*, Thèse de doctorat, Rennes 1, 1998.

[VALE04] Valette, M., Grabar, N., « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? L'exemple du projet PRINCIP », *Le poids des mots, Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT)*, 10-12 mars 2004, Louvain-la-Neuve (Belgique), G. Purnelle, C. Fairon, A. Dister, éds., UCL-Presses Universitaires de Louvain, 2004, pp. 1106-1116.

Modèle sémantique et interactions pour l'analyse de documents

Vincent Perlerin, Stéphane Ferrari

GREYC – UMR 6072 CNRS – Université de Caen
14032 Caen cedex - France

{Vincent.Perlerin,Stephane.Ferrari}@info.unicaen.fr

Résumé :

Cet article présente différents cas d'utilisation d'un modèle de Traitement Automatique des Langues (TAL), centré sur l'utilisateur, où l'interaction joue un rôle central. Nous présentons différents outils de visualisation de documents et de navigation dans des collections de documents afin de déterminer les différents facteurs et leur importance respective quant aux modalités d'interaction à proposer dans les phases d'utilisation de ce modèle.

Mots-clés : TAL, modèle centré utilisateur, navigation documentaire, visualisation d'informations textuelles.

Abstract :

This paper describes an user-centred NLP (Natural Language Processing) model mainly based on Human/Computer interaction. We present different tools for document visualization and browsing. We analyze each step of interaction with the applications based on this model in order to determine the deciding factors of such interactions.

Keywords : NLP, user-centred model, document browsing, textual information visualization.

1. Introduction

Dans cet article, nous proposons d'analyser les interactions avec l'utilisateur dans des applications informatiques traitant du document numérique. Pour étayer nos propos, nous présentons différents outils de visualisation de documents et de navigation dans des collections de documents. Tous sont développés autour d'un même modèle de Traitement Automatique des Langues (TAL) où l'utilisateur joue un rôle central. Notre objectif est de déterminer quels sont les différents facteurs et leur importance respective quant aux modalités d'interaction à proposer dans les phases d'utilisation de ce modèle.

Dans une première section (2.1), nous présentons LUCIA (*Located User-Centred Interpretative Analyser*), le modèle sur lequel nos travaux se fondent. LUCIA est à la fois un modèle de représentation lexicale et un modèle pour l'analyse interprétative de textes. C'est l'utilisateur qui, assisté de la machine, fournit les ressources du système et exprime par là même un point de vue particulier sur le lexique des domaines de son intérêt. Dans la section 2.2, en regard des premiers besoins d'interaction ayant émergé de la présentation du modèle, nous ferons un rapide état de l'art quant aux propositions du domaine de la visualisation et de la navigation documentaire. Nous présentons ensuite un premier logiciel (LUCIABuilder en 3) développé pour créer et gérer les ressources décrites selon le modèle LUCIA. Cette étape est le point de départ de toute application du modèle. Nous montrons ici comment LUCIABuilder met en œuvre des principes de visualisation des ressources permettant selon nous à l'utilisateur de se familiariser avec les notions linguistiques sous-jacentes indépendamment des formats et des langages informatiques manipulés. Dans la troisième partie (4), nous présentons deux applications du modèle : un projet de traitement des métaphores ayant fait l'objet d'une validation sur corpus (projet IsoMeta) et un projet d'assistance à la recherche documentaire (RD). Ces deux projets, dont les tenants et les aboutissants restent très différents, permettent de distinguer parmi les besoins d'interactions qu'ils ont créés des aspects génériques et d'autres dont nous analysons la spécificité vis-à-vis du modèle, de la tâche, du type d'utilisateur et des corpus manipulés. Nous y décrivons les interfaces que nous avons développées en comparant nos approches à des travaux relatifs. Nous concluons enfin en rappelant le rôle essentiel de l'utilisateur dans l'ensemble des interactions observées.

2. Cadre et objectifs

Dans cet article, nous présentons plusieurs applications construites ou fondées autour d'un même modèle de TAL. Dans cette partie, nous décrivons succinctement ce modèle pour introduire les besoins d'interaction inhérents à son utilisation, notamment la visualisation de documents et la navigation dans des ensembles documentaires. Nous présentons ensuite un rapide état de l'art des travaux de ce domaine.

2.1 Un modèle pour le TAL

LUCIA est à la fois un modèle de représentation lexicale et un modèle pour l'analyse interprétative de textes. Il s'inspire entre autres de la Linguistique Structurale européenne contemporaine. En tant que modèle de représentation lexicale, il permet à un utilisateur d'exprimer des connaissances et un point de vue sur le lexique d'un domaine en organisant des entrées lexicales (lexies) selon deux critères principaux :

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

Il est par exemple possible de regrouper dans un premier temps les lexies *bus* et *train* pour rendre compte du fait qu'elles décrivent toutes deux, dans le contexte de la tâche de l'utilisateur, des véhicules, puis de les différencier pour préciser si elles se rapportent plutôt à une portée intra-urbaine ou extra-urbaine. Ces deux critères organisationnels s'expriment à l'aide de jeux d'oppositions appelés des couples attributs/valeurs. L'exemple précédent peut ainsi s'exprimer par l'utilisation de l'attribut binaire « Portée du transport » opposant les valeurs « intra-urbain » et « extra-urbain » – on note cet attribut [Portée du transport : intra-urbain vs. extra-urbain] – *bus* actualisant la valeur « intra-urbain » et *train* la valeur « extra-urbain ». Ce regroupement forme une catégorie que l'on peut nommer en l'occurrence « Véhicules ». On la représente alors à l'aide d'une table comme sur le modèle suivant (figure 1).

VEHICULES	**Portée du transport**
bus	*intra-urbain*
train	*extra-urbain*

Figure 1 : Table LUCIA à un attribut

Plusieurs attributs peuvent être combinés pour décrire les lexies au sein d'une même catégorie. Les lexies placées dans les lignes de la table ainsi obtenue correspondent à l'actualisation spécifique des valeurs des attributs mis en jeu. Dans l'exemple de la figure 2, deux attributs binaires sont utilisés pour caractériser la catégorie des « Véhicules » : [Portée du transport : intra-urbain *vs*. extra-urbain] et [Mode de transport : routier *vs*. ferroviaire].

VEHICULES	Portée du transport	Mode de transport
bus	*intra-urbain*	*routier*
métro, tramway	*intra-urbain*	*ferroviaire*
	extra-urbain	*routier*
train	*extra-urbain*	*ferroviaire*

Figure 2 : Table LUCIA à deux attributs

Comme dans la figure 2, certaines lignes de tables LUCIA peuvent apparaître vides ou supporter plusieurs lexies. Dans le premier cas, cela signifie que l'utilisateur n'a pas trouvé, jusqu'à cette étape, de représentants lexicaux correspondants à la combinaison des valeurs d'attributs. Dans le second cas, cela signifie que l'utilisateur ne considère pas intéressant pour sa tâche de distinguer les lexies de cette ligne. Les ressources LUCIA sont dynamiques : elles peuvent être modifiées à tout moment, en particulier en cas de changement de point de vue de l'utilisateur sur les données ou en fonction des résultats obtenus de ces dernières (cf. infra). En particulier, une ligne de table peut faire l'objet d'une sous-catégorisation pour ajouter une précision aux ressources comme dans la figure 3.

VÉHICULES	Portée du transport	Mode de transport
bus	*intra-urbain*	*routier*
	intra-urbain	*ferroviaire*
car, autocar	*extra-urbain*	*routier*
train	*extra-urbain*	*ferroviaire*

VÉHICULES FERROVIAIRES INTRA-URBAINS	Localisation des voies
tramway, métro aérien	*air-libre*
métro	*sous-terrain*

Figure 3 : Deux tables LUCIA reliés par un lien d'héritage sémique

Dans la figure 3, le lien de la ligne de la table « Véhicules » correspondant à [Portée du transport : intra-urbain] et [Mode de Transport : ferroviaire] vers la table « Véhicules ferroviaires intra-urbains » est un lien d'héritage sémique. Il permet par exemple à la lexie *tramway* d'hériter, en plus de [Localisation des voies : air-libre], des deux attributs/valeurs de la ligne de laquelle part le lien. Plusieurs tables peuvent ainsi être utilisées (reliées entre elles ou non) pour décrire le lexique d'un domaine. Un tel ensemble s'appelle un *dispositif*. Toutes ces notions sont décrites plus en détail dans d'autres publications [Beust, 1998, Nicolle *et al.*, 2002, Perlerin *et al.*,

2003] et n'ont pas un rôle central dans cet article. Nous tenons simplement à montrer la complexité relative du modèle utilisé afin d'expliquer et de justifier la multiplicité des points de vues ou représentations visuelles présentées à l'utilisateur, acteur principal de cette approche interactionniste de la sémantique computationnelle.

En tant que modèle pour l'analyse interprétative de textes, LUCIA s'inspire de la notion d'isotopie, qui est ici envisagée comme la récurrence d'un même attribut au sein d'une même entité textuelle (corpus, texte, paragraphe…). Ainsi, dans la phrase (1) et en fonction de la table présentée en figure 1 la récurrence de l'attribut [Pression] constitue une isotopie.

> (1) *« S'il s'agit d'un anticyclone éphémère entre 2 passages de dépressions (dorsale), l'air est en général un peu plus frais et porteur d'une instabilité un peu plus marquée. »*[1]

On note que cet attribut est ici présent avec deux actualisations différentes, ce qui distingue le modèle LUCIA d'autres approches de l'isotopie. L'utilisateur du modèle ayant entière liberté dans son choix d'attributs pour décrire un domaine particulier, LUCIA permet de projeter, en analyse sur un texte, son point de vue particulier, en repérant les isotopies issues de ses propres représentations. Nous renvoyons à [Perlerin *et al.*, 2003] (*op. cit.*) pour plus de détails sur ce modèle. Hormis la phase de construction des ressources, les interactions nécessaires à l'utilisation du modèle LUCIA concernent essentiellement la visualisation de documents et la navigation dans des ensembles documentaires. Nous présentons maintenant un rapide état de l'art de ce domaine.

2.2 Des travaux proches de notre problématique

Le plus souvent, lorsque les travaux de TAL s'intéressent aux modalités d'interaction avec les utilisateurs, il s'agit de proposer des moyens de visualisation de résultats calculés sur un document ou un ensemble de documents. Au niveau du document ou du texte, les solutions sont nombreuses pour par exemple permettre la visualisation de relations de similarité. Dans [Salon *et al.*, 1995], on propose ainsi de projeter la représentation 2D d'un texte sur le périmètre d'un cercle et d'en relier au moyen de segments les passages calculés comme similaires. Jacquemin et Jardino [Jacquemin *et al.*, 2002] proposent 3D-XV un logiciel de visualisation de documents volumineux par une technique de projection et de coloration permettant la mise en valeur de passages relevant d'une thématique particulière. Minel [2001] dans le cadre des résumés automatique de textes, propose quant à lui un module à la plate-forme CONTEXTO permettant le coloriage et la représentation schématique en rectangles colorés de textes pour en accélérer la lecture. Au niveau des ensembles documentaires, l'exemple le plus connu reste les *TilesBars* de Hearst [1995]. A la suite d'une requête à système de recherche documentaire (RD), un ensemble de rectangles correspondant chacun à un document jugé pertinent par le système est

[1] http://www.portalpes.com/meteo_alpes/page_droite/aNWA.htm

soumis à l'utilisateur. 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. Il existe d'autres techniques de présentation d'ensemble de documents : par exemple sous-forme d'arbres de hiérarchie en 3D (le *Cone Tree* [Robertson *et al.*, 1991]), ou d'arbres hyperboliques (l'*Hyperbolic Tree* [Lamping, 1995]). Certaines méthodes permettent aussi d'apprécier la proximité entre documents sous la forme d'inclusions de rectangles (le *Tree-Map* [Johnson *et al.*, 1991]) ou à l'aide de représentations en perspectives (le *Document Lens* [Robertson *et al.*, 1993]). L'intérêt grandissant pour les travaux sur corpus a fait croître les besoins de techniques de visualisations spécifiques en TAL. Les nombreux logiciels dédiés à l'analyse de données textuelles, comme par exemple les logiciels HYPERBASE d'Etienne Brunet[2], LEXICO3 de l'équipe CLA2T de Paris 3[3] ou encore LEXICA de la société Le Sphinx[4] proposent ainsi tous des biais de visualisation de résultats d'analyses sur corpus dépassant les simples listes textuelles avec des projections en Analyse de Composantes Principales (ACP), des courbes, des graphiques en secteurs, etc. Certains de ces logiciels utilisent également la couleur pour mettre en évidence des termes dans les documents pour faciliter l'étude de leurs distributions (c'est le cas par exemple de TROPES de la société Acetic[5] ou ALCEST de la société Image[6]). 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…). Les solutions proposées ont longtemps été très gourmandes en ressources logicielles et très onéreuses. L'intérêt de ces techniques n'est d'ailleurs pas toujours probant. Leur évaluation est difficile et s'effectue principalement par l'expérimentation sur des groupes d'utilisateurs comme dans [Cribbin *et al.*, 2001] pour une comparaison de techniques de visualisation multidimensionnelles ou encore dans [Cockburn *et al.*, 2002] pour une étude sur l'intérêt de la 2D et de la 3D dans les systèmes de gestion de documents. Pour ces systèmes, l'élaboration des ressources utilisées pour les calculs fait rarement l'objet d'une réflexion poussée en termes d'interaction et ceci pour deux raisons principales : soit les ressources sont fournies *a priori* (à partir de processus automatiques ou de données partagées – type ontologies ou thésaurus), soit la constitution des ressources est réservée à des spécialistes ayant des compétences en informatique ou dans un langage formel donné.

Nous avons vu en 2.1 que LUCIA place l'utilisateur au cœur même du système. Dans toutes les phases d'utilisation du modèle, i.e. aussi bien dans la phase de constitution des ressources que dans les phases d'exploitation des résultats obtenus du calcul de la machine, l'utilisateur est l'acteur principal. La machine a donc à charge de l'assister le plus efficacement possible et également de s'adapter à

2 http://ancilla.unice.fr/~brunet/pub/hyperbase.html
3 http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicowww/lexico3.htm
4 http://www.lesphinx-developpement.fr/
5 http://www.acetic.fr/tropesfr.htm
6 http://www.image.cict.fr/Index_Alceste.htm

son niveau de compétence et à sa tâche. Les modalités d'interaction sont donc ici d'autant plus importantes qu'elles vont conditionner l'efficacité du système et la capacité de l'utilisateur à en comprendre le fonctionnement. Dans la suite, nous proposons deux cadres d'utilisations différents pour LUCIA. S'ils s'adressent à deux types d'utilisateurs distincts, ils présentent cependant une partie centrale commune : la constitution des ressources.

3. Gestion interactive des ressources

La première application liée à un modèle de représentation lexicale est la gestion des ressources qu'il permet de manipuler : création, révision, *etc*. Nous présentons dans cette partie LUCIABuilder, logiciel d'étude *open source*[7] qui a été développé pour ce besoin. Il intègre notamment des fonctionnalités classiques de gestion de données :

- ⇨ Création et modification de structures :
  - Précision des attributs à utiliser
  - Regroupement de ces attributs en tables et dispositifs
  - Précision des liens d'héritage de lignes vers tables
- ⇨ Ajout et modification des données (lexicales) :
  - Positionnement de formes canoniques dans les tables (lemmes)
  - Choix des formes graphiques associées à chaque forme canonique (flexions)

Bien que toutes ces données soient représentées en machine dans un format standard (XML), nous avons ressenti le besoin de proposer à l'utilisateur une application entièrement dédiée plutôt qu'un éditeur standard adapté au langage de représentation informatique sous-jacent. Notre objectif est ici de permettre à l'utilisateur de se familiariser avec les concepts qu'il manipule pour l'aider à mieux comprendre le comportement du modèle de TAL, sans devoir manipuler des modèles informatiques. De ce fait, LUCIABuilder offre, en plus des fonctionnalités énoncées, de nombreuses vues différentes sur les structures et les données.

Avant d'utiliser LUCIABuilder, il faut connaître, au moins partiellement, le lexique du ou des domaines considérés pour la tâche à réaliser. Pour cela, s'il s'agit de prendre connaissance de ces domaines à l'aide d'une collection de textes, il est par exemple possible d'utiliser MemLabor [Perlerin, 2002], un outil annexe fournissant une liste des graphies obtenues d'une analyse distributionnelle. Quelles que soient la tâche et la méthode utilisées, LUCIABuilder permet principalement la

[7] LUCIABuilder est disponible sur notre site : http://www.info.unicaen.fr/~perlerin/recherche/

structuration de lexique(s) en dispositifs LUCIA à travers une interface graphique[8] composée de 5 panels principaux d'interaction. Le premier panel (figure 4) permet de créer les attributs que l'utilisateur juge adéquats pour sa représentation. Les attributs créés peuvent être stockés dans différents ensembles, selon qu'ils sont jugés caractéristiques d'un domaine ou non.

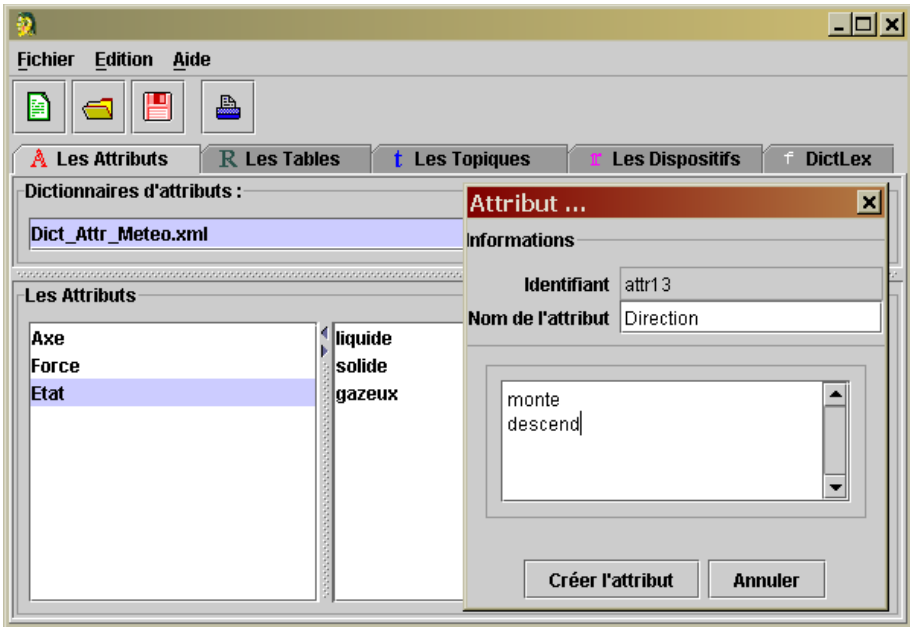


Figure 4 : LUCIABuilder : construction des attributs

Le second panel (figure 5) permet le regroupement d'attributs pour composer les tables décrivant des lexies d'une même catégorie. L'utilisateur choisit les attributs et la machine calcule automatiquement la combinatoire de leurs valeurs. L'ensemble est aussitôt présenté sous la forme d'une table, pour habituer l'utilisateur à cette notion, et lui permettre la saisie des données. Les entrées lexicales qu'il décide d'y faire figurer pourront être mises en relation avec des formes graphiques à l'aide du dernier panel (voir plus loin). Les tables sont stockées dans des fichiers correspondant aux dispositifs LUCIA (et implicitement aux domaines décrits). L'interface force l'utilisateur à exploiter cette notion en présentant la liste des noms des tables d'un même dispositif dans la partie gauche de l'interface.

[8] LUCIABuilder est réalisé en Java (JDK 1.4) et exploite une interface SWING.

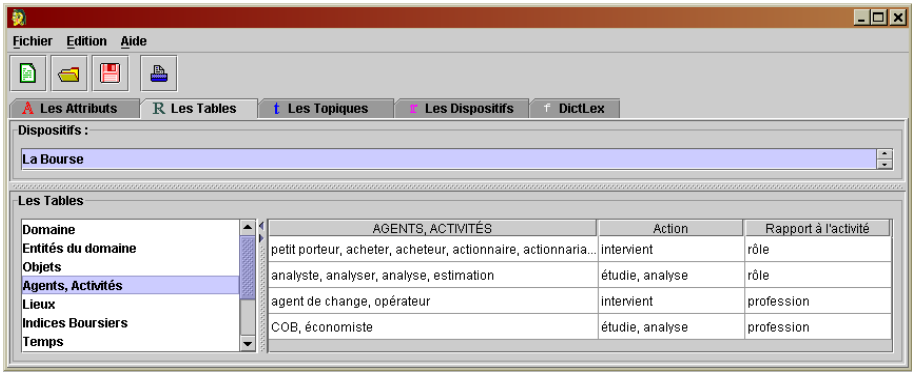


Figure 5 :LUCIABuilder : construction des Tables

Dans le dispositif « La Bourse », une table intitulée « Agents, Activités » regroupe des lexies telles que « petit porteur » et « analyser », partageant ici les attributs [Action] et [Rapport à l'activité]

Le troisième panel (figure 6) présente un point de vue différent sur les tables : une représentation dite en *topiques* (inspirée des travaux de Jacques Coursil [Coursil, 1992]). Chaque ligne d'une table y est représentée par un rectangle déplaçable à l'aide de la souris. Il contient le premier mot de la ligne ou, si elle est vide, les valeurs des attributs qui lui correspondent. Les rectangles constituent les sommets d'un graphe dont chaque arête représente, et permet d'apprécier, les différences mises en jeu entre les lexies des deux lignes associées aux sommets reliés. Les arêtes portent les noms des attributs différenciant les sommets. Le nombre maximal de valeurs d'attributs qui diffèrent entre deux lignes d'une même table est le nombre d'attributs utilisés pour construire la table, et il n'est pas limité. LUCIABuilder offre de ce fait la possibilité à l'utilisateur de préciser le nombre maximal ou le nombre exact de valeurs d'attributs différenciant les sommets à relier. Ceci permet de ne pas surcharger cette représentation en topiques. Les topiques permettent d'obtenir un point de vue différent sur les informations supportées par les tables, elles montrent en particulier plus nettement le nombre et la nature des différences proposées entre les instances d'une même catégorie.

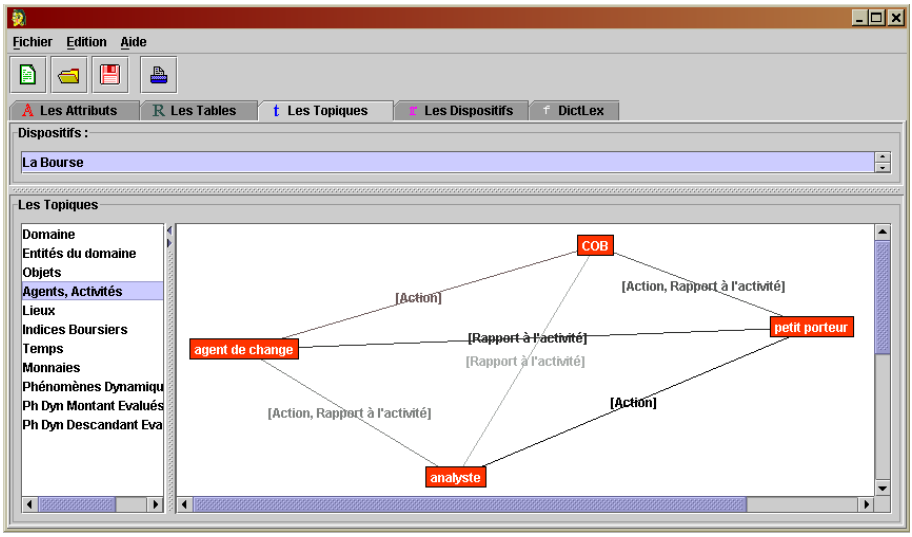


Figure 6 : LUCIABuilder : visualisation en topique

Pour la table « Agents, Activités » de « La Bourse », l'utilisateur a différencié, par exemple, « analyste » et « petit porteur » par une actualisation différente de l'attribut [Action].

Le quatrième Panel (figure 7) permet de créer les liens d'héritage dans un même dispositif, via une vue interactive simplifiée. Les tables y sont représentées par des rectangles déplaçables contenant uniquement leur nom. La première lexie d'une ligne servant de point de départ à un lien d'héritage est affichée sur l'arête correspondante, la flèche pointe vers la table héritant.

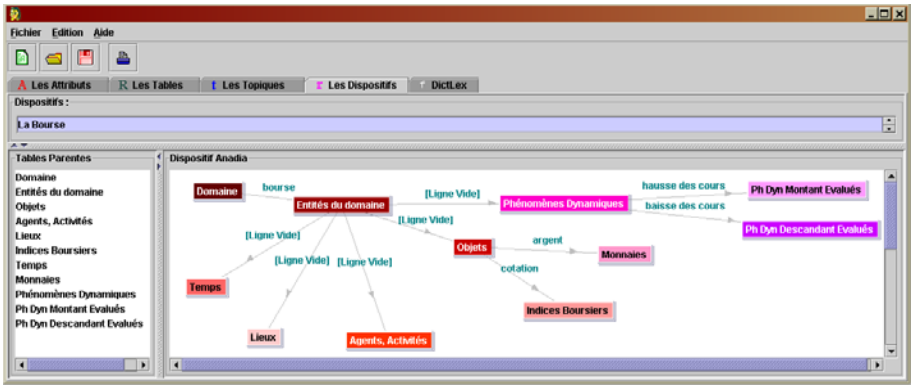


Figure 7 : LUCIABuilder : visualisation d'un dispositif

Exemple de lien d'héritage : la ligne de la table « Objets » contenant le mot « argent » est ici reliée à la table « Monnaies »

Dans les panels 3 et 4, l'utilisateur peut affecter une couleur à chaque table d'un dispositif. Ces couleurs sont utilisées dans les traitements ultérieurs, lors des différents affichages de résultats d'analyse. Elles doivent aider l'utilisateur à interpréter les résultats en les situant par rapport à sa propre structuration lexicale. La section suivante illustre différentes utilisations de cette particularité du modèle. D'autres vues exploitant ces couleurs sont proposées dans l'interface. Il s'agit cette fois de familiariser l'utilisateur avec ses propres descriptions plus qu'avec le modèle. En particulier, nous avons exploité l'interactivité du langage SVG pour produire une vue d'un dispositif (figure 8). Cette vue permet à l'utilisateur d'apprécier la totalité des informations du dispositif, des entrées lexicales figurant sur chaque ligne de chaque table aux liens d'héritages. Les afficheurs SVG possèdent une fonction zoom, qui permet d'obtenir la vue d'ensemble présentée ici autant que la lecture des entrées lexicales des lignes de chaque table.

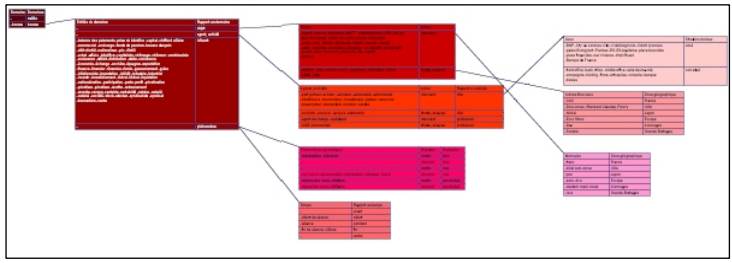


Figure 8 : Vue d'ensemble d'un dispositif complet au format SVG
Pour un même domaine, des variations d'une même couleur ont été affectées aux tables

Le cinquième et dernier panel permet de mettre en relation les lexies saisies dans les lignes des tables avec leurs flexions possibles, ou plus précisément dans l'implémentation actuelle, avec une liste de graphies. Pour cela, plutôt que demander à l'utilisateur de décrire ces graphies, nous cherchons à lui en proposer automatiquement la liste. Actuellement, cette liste est calculée à partir d'une base de données lexicales issue de Mahtlex, développée à l'IRIT, Toulouse. L'utilisateur peut valider l'ensemble des graphies associées à une entrée (ex : les graphies du verbe « pouvoir »), créer ses propres graphies (ex : ajouter « K7 » aux graphies du nom « cassette »). Il est prévu d'inclure dans les fonctionnalités de ce panel un générateur de flexions pour pallier les lacunes des lexiques statiques. Les aspects flexionnels n'étant au cœur du modèle utilisé, nous recherchons ici à accélérer cette phase, de manière à ne pas focaliser l'utilisateur sur un problème annexe à celui de la structuration lexico-sémantique proprement dite.

Comme il est montré dans [Beust, 1998] (*op. cit.*), une application telle LUCIABuilder permet en définitive à l'utilisateur de partager avec la machine un certain point de vue sémantique sur un domaine donné. Nous présentons dans la partie suivante deux utilisations distinctes de ressources ainsi créées, de manière à analyser cette fois comment la machine peut présenter ses résultats d'analyse en exploitant ce savoir sémantique partagé, et quelles interactions en découlent.

4. Applications

Nous présentons dans cette partie deux applications distinctes du modèle LUCIA : l'une pour l'étude d'une métaphore conceptuelle et l'autre pour une aide personnalisée à la RD. Nous cherchons à montrer, sur des fonctionnalités communes aux deux applications, en quoi les interactions et les représentations visuelles peuvent être différentes. Nous analysons entre autres en quoi ces différences dépendent du caractère centré sur l'utilisateur du modèle LUCIA.

4.1 Présentation succincte des applications

Le projet IsoMeta, dédié à l'étude d'une métaphore conceptuelle, s'adresse à des spécialistes de la langue. Il a pour objectif d'analyser en quoi LUCIA permet de détecter et de fournir une aide à l'interprétation des métaphores. Nous renvoyons à [Beust *et al.*, 2003, Perlerin *et al.*, 2002] pour une présentation plus approfondie. Les propositions faites dans ce projet ont été validées par l'étude d'une métaphore conceptuelle particulière : la *météorologie boursière*. Certains corpus traitant de la bourse contiennent en effet de nombreux emplois métaphoriques de termes météorologiques : « *tempête sur la bourse* », « *le Dow Jones : thermomètre de Wall Street* »... La validation a donc été menée sur un corpus thématique, constitué d'environ 600 articles du journal *Le Monde* traitant d'économie. Le modèle permet ici de décrire les lexiques des domaines source et cible de la métaphore (resp. la

météorologie et la bourse) par la création de deux dispositifs. Pour observer les résultats d'analyse, les besoins d'interaction sont centrés autour de :

- La navigation dans la collection d'articles pour repérer ceux pouvant contenir des emplois de la métaphore ;
- L'identification et l'analyse détaillée des zones de textes pertinentes dans ces articles.

Pour l'aide personnalisée à la RD, le modèle offre à l'utilisateur le moyen de préciser son point de vue sur le lexique d'un ou plusieurs domaines, tant dans son choix de lexies que dans la manière dont il les organise. Ceci peut constituer une aide efficace pour rechercher *a posteriori* des documents relatifs au(x) domaine(s) décrit(s). Mais l'efficacité de cette aide ne se limite en aucun cas à la qualité des résultats de la recherche. En effet, la qualité de l'interaction entre l'utilisateur et ces résultats y contribue aussi en grande partie. Dans cette optique, les besoins d'interaction vont une nouvelle fois dans le sens d'une aide à la navigation parmi une collection de documents (ici des réponses à une recherche donnée), avec toujours un principe de double situation :

- Présentation graphique de la collection pour une identification rapide des documents les plus pertinents ;
- Au sein d'un de ces documents, identification et analyse des parties de textes les plus intéressantes.

En faisant le parallèle entre ces deux applications, nous montrons dans la suite de cette section les principes communs et les spécificités de ces deux phases d'interaction.

4.2 Navigation dans un ensemble documentaire

Dans les deux applications, une collection de documents existe, et l'utilisateur doit naviguer dans cette collection pour y repérer les documents qui l'intéressent le plus. Cet aspect apporte donc une dimension générique aux interfaces proposées. Les travaux du même champ d'application cités précédemment (2.2) permettraient d'améliorer l'interface proposée notamment en prenant en compte une relation d'ordre – cet aspect constitue l'une des perspectives de nos travaux. La présentation de tous les documents est une fonctionnalité générique, mais dans cette collection, la représentation de chaque document y diffère selon l'application.

Pour étudier la métaphore conventionnelle *météorologie boursière*, deux dispositifs ont été construits : l'un correspondant à la bourse, l'autre à la météo. Chacune des lexies de ces dispositifs est repérée automatiquement au sein des documents du corpus. Pour étudier le phénomène, la première tâche est de naviguer dans le corpus et d'y repérer les documents susceptibles de receler une telle métaphore, i.e. des documents où au moins une lexie en rapport avec la météo apparaît. Pour faciliter ce repérage, nous utilisons une interface (figure 9) regroupant l'ensemble des documents traités par les modules d'analyse. Chaque document y est

représenté sous la forme d'un graphique en histogrammes. Chaque barre d'un histogramme correspond à une table d'un dispositif. Elle hérite de la couleur associée à la table et sa hauteur est proportionnelle au nombre de lexies du document qui appartiennent à cette table. Le passage de la souris sur l'une des barres permet en outre l'affichage du nom de la table et du nombre de lexies repérées lui appartenant. Un lien hypertexte permet enfin d'accéder au document représenté par l'ensemble de l'histogramme.

Figure 9 : Visualisation du corpus analysé pour le projet IsoMeta

L'ensemble des histogrammes est actuellement représenté au sein d'une même page HTML qui offre ainsi une vision d'ensemble du corpus traité. Cette présentation reste très simpliste : les documents sont ordonnés selon leur place dans la collection initiale. Ce qui nous intéresse ici est surtout la manière dont la représentation d'un document dans cette collection est adaptée à l'utilisateur et à la tâche. Les informations présentes dans les histogrammes sont très précises au regard du modèle de structuration lexicale utilisé, et nécessitent une bonne appréhension de ce modèle de la part de l'utilisateur. Le choix de couleurs dominantes opposées pour les deux domaines facilitent le repérage rapide des articles intéressants (la météo apparaît ici en vert, la bourse en rouge). Un tel choix n'est pas inhérent au modèle mais conseillé lorsqu'il y a plusieurs dispositifs. Notons que la taille des barres n'est pas relativisée, car la collection est constituée de textes de tailles comparables. Cet aspect introduit une dépendance de nos représentations au matériau traité, cette fois indépendante du modèle et de l'utilisateur.

Dans le cadre de la RD, nous proposons aux utilisateurs de construire des dispositifs pour filtrer et réordonner des résultats provenant de systèmes classiques tels que les moteurs de recherche sur l'Internet. Ce filtrage et ce classement se fondent sur la notion d'isotopie introduite en 2.1. Au niveau des représentations proposées et par rapport à la tâche précédente, des différences majeures apparaissent :

- Les tailles des documents à traiter ne sont plus nécessairement comparables ;
- Les descriptions sémantiques peuvent être moins complexes que précédemment ;
- L'intérêt n'est plus simplement de repérer un thème dans un document mais de pouvoir rapidement en évaluer l'importance relative.

Pour faciliter la navigation dans les listes de résultats, nous proposons en conséquence une représentation schématique du document, conservant son aspect visuel global, et intégrant une coloration des parties de texte correspondant aux thèmes attendus par l'utilisateur. Techniquement, à l'heure actuelle, nous ne proposons cette représentation que pour des pages HTML. Elle exploite le format SVG (figure 10) pour permettre à l'avenir plus d'interaction, comme l'insertion de liens hypertextes directement vers les parties intéressantes des documents.

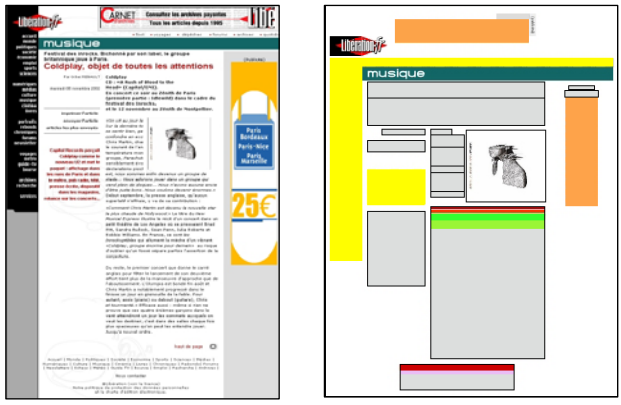


Figure 10 : Visualisation d'un document pour la RD :

A gauche un article du journal Libération (format HTML), à droite la représentation schématique SVG produite automatiquement avec coloriage de certaines parties du texte

Cette représentation schématique ayant pour but de conserver l'aspect visuel du document, elle intègre les images d'origine. Nous avons en outre exploité les spécificités des documents Web en repérant des zones particulières et en leur attribuant une représentation graphique mise en couleur en fonction de leur nature : index (parties de la page HTML regroupant beaucoup d'hyperliens) en jaune dans l'exemple, publicités (selon les URL d'accès aux documents et leur format) en

orange. Les autres parties de texte sont par défaut en gris, mais peuvent faire l'objet d'une coloration thématique supplémentaire. Elles sont jugées intéressantes lorsqu'elles sont supports d'isotopies relevant des tables construites par l'utilisateur. Ceci est représenté par la superposition d'un rectangle de la couleur associée à la table, et dont la taille est proportionnelle au nombre d'isotopies trouvées dans la partie. À l'aide de ce schéma, nous proposons de visualiser rapidement la proportion des thèmes relativement aux parties des documents et non au document dans son ensemble.

Cette interface, comme par exemple le *Document Lens* [Mackinlay *et al.*, 1993], présente une vision macroscopique des documents permettant d'ignorer dans une première étape le texte dans ses détails et de repérer les invariants aspectuels propres à certains media. Ici, le contexte d'apparition des documents n'est pas directement accessible : ils sont obtenus automatiquement d'un système de RD et ne présentent donc pas nécessairement de rapports physiques (hyperliens) ni de similarité ni de genre. Nous pouvons en revanche envisager une technique de visualisation dépendante d'une hiérarchie se fondant sur les liens d'héritage au sein d'un dispositif. Dans [Card, 1999], l'auteur décrit deux manières de présenter des données hiérarchiques : par connexion - *connection* (Cone Tree [Robertson *et al.*, 1991] (*op.cit.*) ou Hyperbolic Tree [Lamping, 1995] (*op. cit.*)) ou par inclusion - *enclosure* (TreeMaps [Johnson *et al.*, 1991] (*op. cit.*)). Les connexions utilisent des liens visuels entre les parents et les enfants de la structure alors que les inclusions tendent à présenter les enfants comme partie intégrante des parents. Comme nous avons pu le voir dans la description du logiciel LUCIABuilder, la présentation des dispositifs se fait selon le principe des connexions (figure 3). Cette technique utilisée pour les ensembles documentaires nous semble d'autant plus appropriée qu'elle pourrait être totalement analogue dans sa présentation aux dispositifs construits par les utilisateurs. La hiérarchie pourrait ainsi représenter les documents classés en fonction des proportions des lexies des tables des dispositifs et reliés entre eux selon les liens d'héritages. Il faut cependant noter que le choix d'une telle technique de visualisation ne dépend pas uniquement du modèle (existence d'une hiérarchie), mais aussi de l'utilisateur et de sa manière d'envisager la tâche de RD. En effet, pour une recherche peu précise, un utilisateur novice peut avoir tendance à se contenter d'un dispositif peu structuré. L'idée nous semble cependant intéressante pour des experts réalisant une veille technologique, désireux de structurer leur domaine.

Pour les deux applications présentées, la tâche de navigation au sein d'une collection de documents peut donc être réalisée via différentes représentations tant de la collection que des documents eux-mêmes. Si des choix s'imposent pour d'autres modèles, la grande liberté qu'offre le modèle LUCIA ne permet pas de conclure quant aux représentations les mieux adaptées. Au contraire, elle fait apparaître leur dépendance vis-à-vis de l'utilisateur, s'exprimant dès la construction des ressources, et de la tâche. Nous proposons de continuer à analyser ces dépendances sur la tâche de représentation d'un document unique pour observation de résultats d'analyse.

4.3 Identification des zones pertinentes des documents

Dans les deux tâches proposées, une fois un document pressenti comme pertinent, l'utilisateur a besoin d'accéder à l'information attendue. Cette information n'est pas nécessairement le document dans sa totalité. La représentation d'un document pour observation des résultats d'analyse par l'utilisateur doit donc aussi permettre de situer rapidement les parties intéressantes. Dans le cadre d'IsoMeta, c'est la présence ou non d'une métaphore qui permet de décider si une partie est intéressante. Dans le cadre de la RD, ce sont les thèmes de chaque partie qui servent de base pour juger de l'intérêt d'un document.

Techniquement, dans le cadre d'IsoMeta, les documents analysés sont codés en HTML. Leur affichage dans un navigateur représente la deuxième partie de notre interface pour l'analyse des métaphores (figure 11). Chacune des lexies appartenant aux dispositifs y est mise en valeur par une coloration en arrière plan avec la couleur de la table correspondante. La coloration des mots est maintenant couramment utilisée pour l'exploration textuelle. Par exemple, le moteur de recherche Google propose cette facilité grâce aux liens « en cache » proposés avec chaque URL de la liste des résultats d'une requête au moteur. Cette technique de coloration est pratique pour repérer rapidement l'information pertinente à l'écran, nous envisageons donc aussi son utilisation pour la RD. Mais dans ce cadre, les documents peuvent être trop volumineux pour que cette technique seule suffise. Elle reste donc à compléter, par exemple par l'utilisation en amont des représentations SVG interactives permettant un accès direct aux parties pertinentes.

Figure 11 : Visualisation d'un document colorié pour IsoMeta

Une fois une partie de texte pertinente repérée, il reste éventuellement à présenter plus d'informations quant aux analyses effectuées. Dans le cadre d'IsoMeta, le passage de la souris sur une lexie mise en valeur permet l'affichage des représentations issues des dispositifs (le nom de la table, et les attributs/valeurs actualisés – figure 11). Cette technique permet une vision microscopique de la sémantique des documents, utile à l'utilisateur expert pour l'analyse de la dynamique des attributs/valeurs (sur le modèle de la dynamique sémique) mise en jeu dans les emplois métaphoriques. Une telle quantité d'informations n'est cependant pas nécessaire dans le cadre de la RD pour la plupart des utilisateurs. Les remarques faites plus haut s'appliquent ici encore : un novice se contentant d'un dispositif peu structuré n'aura pas besoin de plus d'informations une fois les zones de textes pertinentes repérées, tandis qu'un expert réalisant une veille technologique peut désirer accéder à une vision plus fine de la dynamique locale.

De manière formelle, l'interface de lecture proposée dans la figure 11 pour l'observation des résultats d'analyse de documents répond aux exigences classiques de ce type d'application :

- Elle est interactive (les couleurs et les groupes thématiques peuvent être modifiés par l'utilisateur, le passage de la souris sur certaines zones permet l'affichage d'informations ciblées…) ;
- Elle propose l'abstraction de certaines données pour rendre leur contenu plus explicite par l'utilisation des graphiques en histogrammes ou une représentation schématique en SVG ;
- Elle propose la mise en valeur de données pertinentes pour l'étude par un moyen graphique facilement repérable et issu des propositions de l'utilisateur ;
- Elle permet enfin un accès immédiat aux documents initiaux par l'utilisation de liens hypertextes.

Mais l'ensemble de ces critères ne permettent pas de conclure quant à son adéquation à une situation donnée. C'est selon nous parce que le modèle LUCIA propose à l'utilisateur d'interagir en partageant sa connaissance dès la phase de structuration lexicale qu'il est difficile, sinon inapproprié, de fournir une interface de lecture de résultats unique. Il ne nous paraît pas impossible d'exploiter la même interface dans les deux cadres applicatifs. C'est donc l'utilisateur qui nous semble peser ici dans le choix de la meilleure interface de lecture. Ce poids ne s'exprime probablement pas que par la manière dont il structure son lexique, mais aussi par la façon dont il désire aborder la tâche d'analyse des résultats.

5. Conclusion et perspectives

Dans cet article, nous avons cherché à analyser les interactions mises en jeu dans les applications du modèle LUCIA, un modèle de TAL fondé sur l'exploitation de lexiques sémantiques structurés. Après une présentation de ce modèle et de différents travaux relatifs à la visualisation de résultats d'analyse de documents, nous avons commencé par montrer en quoi la constitution des ressources constituait le premier lieu d'interaction pour ce modèle. L'interface LUCIABuilder offre à l'utilisateur la possibilité de structurer lui-même ses ressources lexicales. Afin de familiariser l'utilisateur avec le modèle de TAL sous-jacent, l'interface propose de nombreuses vues différentes des mêmes objets, et ne fait surtout pas intervenir de connaissances relatives aux modèles informatiques utilisés pour leur représentation (XML) ou leur manipulation (XSL, classes JAVA). Elle prépare en outre déjà à la phase d'observation des résultats par l'attribution de qualités graphiques aux objets créés, notamment des couleurs associées à des ensembles de lexies d'une même catégorie, d'un même domaine.

Nous avons ensuite étudié plus particulièrement deux applications du modèle : l'une pour l'analyse d'une métaphore conceptuelle (projet IsoMeta), l'autre pour une aide à la recherche documentaire. Ces applications permettent d'envisager des fonctionnalités génériques telles la navigation au sein d'une collection de documents analysés pour y sélectionner les plus pertinents puis la lecture plus précise d'un de ces documents pour observer le résultat local des analyses. Bien que les interactions nécessaires à la réalisation de ces tâches soient comparables, nous avons vu que les représentations visuelles des résultats d'analyse ne le sont pas nécessairement. Cependant, il semble que c'est essentiellement l'utilisateur qui influe sur l'adéquation de ces représentations, et dans une moindre mesure, la tâche et le type de document traité.

Il est difficile de conclure quoi que ce soit à partir d'une simple étude de cas. Mais il est intéressant de constater le rôle prédominant de l'utilisateur dans les interactions. Avec un modèle fixé, lorsque l'utilisateur a le choix des représentations lexicales sémantiques, c'est avant tout lui qui oriente l'ensemble des interactions à venir en précisant en quelque sorte dès le départ les bribes de connaissances qu'il veut bien partager avec la machine.

Remerciements

Le projet IsoMeta est mené conjointement avec Pierre Beust du GREYC. La base lexicale BDLex a été transformée au format XML avec l'aide de Pierre-Sylvain Luquet de GREYC. Nous remercions les relecteurs de cet article pour leurs commentaires.

6. Références bibliographiques

Beust, P., Ferrari S. et Perlerin, V. (2003*). NLP model and tools for detecting and interpreting metaphors in domain-specific corpora*. In Proceedings of Corpus Linguistics, 2003. Main Conference and Interdisciplinary Workshop on Corpus-Based Approaches to Figurative Language. UK, Lancaster.

Beust, P. (1998). *Contribution à un modèle interactionniste du sens*. Thèse en vue de l'obtention du grade de Docteur en Informatique de l'université de Caen.

Card, S.K., *et al.* (1999). *Information Visualization: Using Vision to Think*. San Francisco, California. 1-34, Morgan Kaufmann Publishers.

Cockburn, A. et McKenzie, B. (2002). *Evaluating the Effectiveness of Spatial Memory in 2D and 3D Physical and Virtual Environments*. In Actes de CHI 2002. Mineapolis, MD, USA.

Coursil, J. (1992). *Grammaire analytique du français contemporain - Essai d'intelligence artificielle et de linguistique générale*. Thèse en vue de l'obtention du grade de Docteur en Informatique de l'université de Caen.

Cribbin, T. et Chen, C. (2001). *Visual-Spatial Exploration of Thematic Spaces: A Comparative Study of Three Visualisation Models*. In Actes de Electronic Imaging 2001: Visual Data Exploration and Analysis VIII. San Francisco, CA, USA.

Hearst, M.A. (1995). *TileBars: Visualization of Term Distribution Information in Full Text Information Access*. Proceedings of the Conference on Human Factors in Computing Systems CHI'95. [ACM]

Jacquemin, C. et Jardino, M. (2002). *Une interface 3D multi-échelle pour la visualisation*. In Actes de IHM'2002. Poitiers.

Johnson, B. et Schneiderman B. (1991). *Tree-maps: A space-filling approach to the visualization of hierarchical information structures*. In Proceedings of IEEE Visualization '91. New York. pp. 284-291.

Lamping, J. (1995). *A focus+context technique based on hyperbolic geometry for viewing large hierarchies*. Proceedings of the Conference on Human Factors in Computing Systems CHI'95. [ACM Press]

Mackinlay, J.D. et Robertson, G.G. (1993). *The Document Lens*. Proceedings of the ACM User Interface and Software Technology conference (UIST'93). pp. 101-108.

Minel, J.L. (2001). *Filtrage sémantique (du résumé automatique à la fouille de textes)*. Hermès Sciences Publication, Lavoisier. Paris.

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

Perlerin, V. (2002). *Memlabor, un environnement de création, de gestion et de manipulation de corpus de textes*. RECITAL 2002. Tome 1. Nancy. pp. 507-516.

Perlerin, V., Beust, P. et Ferrari, S. (2002). *Métaphores et dynamique sémique*. 2ème Journée de Linguistique de Corpus. Lorient.

Perlerin, V. et Pierre, B. (2003). *Pour une instrumentation informatique du sens*. Variation, construction et instrumentation du sens. Hermès Science Publications, Lavoisier. Paris. [8], pp. 197-228.

Robertson, G.G. et Mackinlay Jock D. (1993). *The Document Lens*. In Actes de UIST'93 (ACM Symposium on User Interface Software and Technology). Atlanta, USA. pp 101-107.

Robertson, G.G., Mackinlay Jock D. et Stuart K, C. (1991). *Cone Trees: Animated 3D Visualizations of Hierarchical Information*. In Proceedings ACM Conf.Human Factors in Computing Systems, CHI '91. pp. 189-194., ACM Press.

Salton, G., Allan, J., Buckley, C. et Singhal, A. (1995). *Automatic analysis , theme generation and summarization of machine readale text*. Science. VL264. [3], pp. 1421-1426.

Quelques contenus généraux au service des documents

Dominique Dutoit[1,2], Patrick de Torcy[2], Yann Picand[2]

[1]*CRISCO, Université de Caen, 14032 Caen cedex - France*
[2]*MEMODATA, 17 rue Dumont d'Urville, 14000 Caen - France*

d.dutoit@crisco.unicaen.fr

http://www.memodata.com

Résumé :

Nous débutons par une réflexion générale sur *forme et contenu*. Il s'ensuit quelques remarques de précaution concernant ces termes, ces remarques conduisant à une redéfinition du signe comme contenu cognitif. Comme nos travaux portent sur des contenus relativement petits par rapport au document entier, nous nous interrogeons sur les contributions linguistiques à la gestion documentaire. Nous présentons enfin certains de nos outils et les assemblons brièvement dans deux applications documentaires concrètes.

Mots-clés : Forme, contenu, signe, document, langue, cinématique d'application GED, architecture d'outils de gestion de textes.

1. Introduction

Des traitements plus sémantiques des documents électroniques seraient dorénavant envisageables alors même qu'avant ils l'étaient moins. Cette assertion que l'on trouve finalement dans l'appel à contribution de CIDE 7 est suffisamment générale pour, appliquée improprement, donner lieu à des interprétations aporétiques de n'importe quel résultat –ou absence de résultat– que quiconque soumettrait. De ce fait, dans une première partie, nous explorons quelques attendus possibles de cette assertion. Une deuxième partie est consacrée à la description de quelques contributions concrètes de nos travaux au traitement « sémantique » des documents électroniques. Nous espérons que grâce à la première partie, le lecteur pourra juger de l'orientation, des résultats, etc., présentés dans la seconde partie avec cette

objectivité naïve de celui qui regarde ce que quelque chose fait plutôt que ce que quelque chose est.

2. Contenu textuel et contenu sémantique

Nous sommes engagés depuis 1989 dans l'exploration des opérations possibles qui pourraient être exécutées par un système ayant des notions des instructions associées à des éléments que l'on croit constitués des parties d'un énoncé, d'un acte de langage, d'un document etc. Finalement, l'espace de notre travail est limité comme suit :

- Un système (artificiel) effectue des opérations. Ces opérations appartiennent à l'intersection entre les opérations souhaitables (pour accéder à d'autres opérations) et les opérations programmables (selon les opérations disponibles).
- Ces opérations correspondent à l'exécution d'instructions que l'on croit potentiellement associées à une partie du texte, chacune de ces parties constituant précisément des touts construits par lesdites opérations : cela constitue un tout parce que précisément telles ou telles opérations ont pu s'appliquer. A ce point, évidemment, rien n'indique que ce tout appartient à d'autres parties du texte, c'est à dire que ce tout soit perçu comme tel par quelque autre opération dans le système. De plus, il n'y a aucune raison de penser que ce tout ait une quelconque existence propre en dehors du système.

2.1 Contenu textuel – les taches d'encre noire

Incapable de lire le chinois, à l'instar d'une machine, je n'ai pas la moindre idée de ce qui constituerait ne serait-ce qu'une partie (donc un tout particulier) d'un texte chinois. Face au français, aucune machine ne pourra jamais déterminer sans notre éclairage si « i » en tant que bitmap constitue deux symboles distincts, autrement dit deux parties – la barre d'un côté, le point de l'autre – ou est un tout constitué de deux parties. Les codifications ASCII, ANSI, etc., ont résolu le problème comme suit :

- En déclarant sans vergogne le « i » en tant que tout dans la réalité de certaines langues écrites : ce « i » est doté d'un numéro.
- En associant ce « i » à diverses représentations graphiques acceptables.

Il n'y a pas de place ici pour la moindre émergence. Il y a incorporation du discernement du « i » dans un corps unique qui ne peut être qu'un « i » idéalisé. Quand bien même nous serions face à des millions de traces de i, nous ne pourrions avoir au mieux qu'un « i » statistique, qu'un tout potentiel. Le « i » constitué comme un tout à prendre globalement suppose – mille expressions diront la même chose – une existence propre du « i », un « i » en tant que « i », un « i » indépendant d'une

quelconque occurrence de « i », un « i » dont on affirme l'existence quelque part dans notre pensée, nulle part ailleurs, un « i » improuvable, un « i » intersubjectif, c'est à dire un « i » subjectif. A ce point, où le « i » ne peut que rester non fondé, ce « i » pourra appartenir à tel ou tel système graphique, ce système ne pouvant pas plus émerger du fait de son emploi : il est donné et accepté comme tel si on veut l'utiliser.

Comment interpréter la proposition suivante de CIDE 7 : « la mise en avant du *sens* a longtemps été regardée avec beaucoup de scepticisme au profit de traitements dits de *surface*, s'attachant à la *forme* par opposition au *contenu* ».

Ce scepticisme se fonde sur plusieurs raisons : nature du sémantique, complexité, nécessité d'une expérience du monde, subjectivité du sens etc. Réfléchissons un moment sur la possibilité d'une étude des formes.

Une forme peut-elle être discernée en tant que telle ? La définition de l'intérieur, de la frontière et souvent d'un environnement bien plus large est nécessaire. Or cette définition n'appartient pas à la forme elle-même : c'est nous qui la donnons. Si nous appelons *contenu* cette définition, l'existence même de la forme suppose, nécessite préalablement l'existence du contenu. Ce contenu est intérieur à nous, toujours supposé, jamais réalisé, la forme elle, c'est à dire une partie distinguée d'une chose quelconque, n'existe que selon ce contenu. Hors ce contenu, la forme n'a pas de limite. La forme n'ayant pas de limite en tant que telle, l'univers ne contient qu'une seule forme : lui-même. Le reste n'est que discernement, c'est-à-dire résultat d'une chose intangible, inobjective, sans partie, strictement opératrice. Plusieurs mots sont ici possibles : *contenu* (de la forme), *discernement* de la forme, *définition* de la forme, mais aussi, naïvement : *sens* de la forme.

De tout cela, il résulte que l'objectif d'un traitement de surface qui s'attacherait exclusivement à la forme hors toute définition de contenu est radicalement inaccessible : je peux à la limite, par la pensée, me rendre en un bord de l'univers, ou en son centre, mais aucunement, par la pensée, me représenter cet objectif de l'étude d'une forme en tant que telle. Au contraire, il me faudra toujours ajouter quelque contenu (la formation du « i », la formation d'un mot particulier), éventuellement son nom (le « i », ledit mot), etc. De même que l'attribut « i » ne préexiste pas aux ensembles de pixels, aux longueurs d'ondes acoustiques (etc.) constituant le « i », l'attribut « mot » ne préexiste pas aux ensembles de sons, symboles graphiques (etc.) constituant ce mot. Et, puisqu'il s'agit d'une chose reconnue, il est également vrai que l'ensemble particulier de pixels, la liste particulière de caractères formant un mot (il faut aussi considérer l'orientation de la lecture, etc.) n'existent possiblement en tant que chose à considérer globalement, tout, partie d'une construction plus large, qu'après que leurs intériorités et leurs frontières soient données. Il s'agit ici de dire comme Pascal que tout et parties se connaissent simultanément, mais aussi, que tout et parties n'existent que comme objet de notre pensée. Ainsi, les formes seules n'ayant pas d'existence propre, la subjectivité s'impose – radicale – dans toute (dé)composition. Ce qui rend la forme présente à notre esprit n'est pas la forme, mais notre esprit qui conçoit cette composition-ci ou cette composition-là comme quelque chose à prendre ensemble, à *comprendre*.

Quand nous disons : *voici*, *telle forme a telle propriété*, nous utilisons une formule courante mais à considérer systématiquement comme suit : *voici*, *je discerne telle forme au moyen des opérations lambda*, *au moyen de tel contenu*. Ainsi, si nous définissons un document comme un objet composé de plusieurs parties, ces parties sont définies par des contenus qui préexistent.

2.2 Contenu sémantique – des formes sans trace

A ce point, il faut distinguer clairement (au moins) deux types de formes sur lesquelles s'applique notre esprit. Le premier de ces objets est la réalité : *la semaine dernière, j'ai été chargé par un sanglier*. A l'instar de n'importe quelle forme, le sanglier n'est sanglier qu'en tant que distingué comme tel de son environnement. Cette distinction faisant le sanglier est parfois nommée : *sanglier*.

Que j'aie une perception ou non de cette réalité, si je ne m'étais pas écarté, j'aurais été blessé. Sans organe de perception, j'aurais été frappé par un objet sans forme. Avec cette perception, j'aurais été frappé par un objet doté d'une forme.

Le deuxième de ces objets est la trace de la réalité, c'est à dire toute chose qui n'est pas la forme perçue de l'objet réel concerné mais la perception d'une forme autre renvoyant à l'évocation de l'objet concerné : la trace dans le sol ou le mot *sanglier*. Avec le langage, aveugle ou non, je suis frappé par un sanglier.

Bien évidemment, une trace n'existe pas comme telle. Une trace est d'abord une forme et comme toute forme, la trace n'existe que dans l'esprit qui la discerne. Mais la trace existe doublement dans cet esprit : en tant que forme discernée par ses intérieurs, frontières et extérieurs mais aussi en tant que tout construisant un objet qui existe en dehors des limites formelles de la trace, un objet strictement pensé, une conception sans forme, un idéal.

Reprenons une dernière fois notre « i ». Nous avons dit de lui que :

- **Il n'existe pas d'occurrence de « i » en soi**. Il n'existe que des systèmes capables de repérer, de distinguer un « i » dans une occurrence de i. Il n'est pas nécessaire d'avoir une connaissance abstraite du « i » pour être capable de le reconnaître. Par exemple, un protozoaire qui reconnaît dans son environnement une particule qu'il peut absorber, n'a pas besoin d'une connaissance abstraite de cette particule. Il a juste besoin d'avoir une fonction de discernement de la particule. Cette fonction peut par exemple être que les parties du protozoaire épousent les parties de la particule, mais ce peut être (et c'est sûrement) une toute autre fonction.
- **Le contenu préexiste à la forme**. Le contenu préexiste à toutes les occurrences de la forme. Ainsi, nous espérons que notre protozoaire pourra reconnaître la plupart des corpuscules dont il a besoin. Dans une certaine mesure, ce contenu définit un idéal de la forme : il vaut toutes les occurrences de cette forme, passées, présentes et à venir. Il n'est pas pour autant un quelconque prototype, une sélection particulière parmi toutes ces formes.

- **Le contenu incorpore toutes les formes possibles**. Or ces formes peuvent apparaître sous des apparences multiples. Le contenu ne peut alors qu'être une opération, c'est à dire une définition en intension.
- **Le contenu ne renferme pas nécessairement tous les détails de la forme**. Ainsi, pour notre protozoaire, seuls quelques signaux chimiques pourraient suffire, étant entendu maintenant qu'une erreur est toujours possible.
- **L'intension du contenu peut s'activer pour une partie incomplète de la forme**. Notre protozoaire ne discerne pas toutes les parties de la particule qu'il peut tester. Je n'ai jamais vu, pas plus qu'un physicien ou un mathématicien, et je ne verrai jamais, un cube complet. J'infère le cube depuis la partie que j'aperçois.
- **Une partie d'un contenu est un contenu, un contenu est composé de plusieurs parties-contenus**. Le rapport entre partie-contenu et contenu-tout est abductif (*hypothèse du tout possible, le tout réel n'existant pas nécessairement*).
- **Les parties d'un contenu appartiennent à d'autres contenus**. Le cube se voit de dessus comme un carré. Il y a recouvrement des parties.

Nous renvoyons pour une synthèse des conséquences logiques de la relation partie-tout sur des parties homogènes au tout à [SMITH98].

A notre sens, il ne faut pas chercher une quelconque gnoséologie dans ce qui a été dit ci-dessus ni même dans ce qui va être dit. S'agissant de faits pouvant même être vus à travers l'activité d'une amibe, le propos ne porte ni sur un néo-kantisme ni, pour la suite, sur un rapport entre le langage et le monde des objets à la manière d'E. Cassirer. Au moment d'énoncer qu'un simple i ne peut aucunement se définir dans une « doctrine où tout se tien(drait), et qui paraît(rait) s'imposer par sa rigueur » [BERGSON11], qu'aucun i ne tiendra jamais dans une géométrie raisonnable (nous essaierons ailleurs de nous expliquer là-dessus), nous rappelons seulement le mot du philosophe : « mais qui ne voit que ses spéculations (du théoricien) sont alors purement abstraites et qu'elles portent, non pas sur les choses mêmes, mais sur l'idée trop simple qu'il se fait d'elles avant de les avoir étudiées empiriquement ».

Reprenons.

Doté d'un contenu sans forme, mais qui a pour destination le discernement d'une forme particulière, laquelle ne prend le titre de forme que du fait de son action, il m'est possible de nommer le contenu qui définit cette forme.

- **Un contenu peut avoir un nom**. J'appelle « i » le contenu qui me permet de discerner un i. J'appelle *sanglier* le contenu qui discerne un sanglier.
- **Le même nom pourra être utilisé par des contenus d'espaces différents**. J'utilise « i » pour le contenu sonore i et « i » pour le contenu graphique de « i ». J'appelle « table » des contenus très différents.

- **Ce nom peut servir à relier ces contenus d'espaces différents en un contenu abstrait, c'est à dire en un contenu qui n'aura jamais la moindre forme**. Je n'ai jamais vu un « i » sonore, ni plus qu'entendu un « i » graphique. Face à une occurrence de « i », bien que ce mot soit monosémique d'une certaine manière, certains contextes me font sélectionner un « i » plutôt qu'un autre. Soit « i » dans « ami ». Ce « i » est un i graphique. Mais il est aussi le « i » prononçable. Regardons maintenant certaines bordures d'écran de Minitel. Certaines sont faites du symbole graphique « i ». Mais sont-ce bien des « i » ? Soit encore un animal qui pousserait un magnifique « i ». Pouvez-vous défendre que ce « i » est vraiment un « i » ? Ainsi, comme depuis un cube vu en partie, nous devons reconstruire un cube entier, nous ne constatons jamais qu'une partie d'un « i ». Mais cette partie doit effectivement être une partie : la reconstruction, l'abduction d'une totalité du « i » ne doit pas conduire à une aberration (*nécessité d'un tout possible*). Ces réflexions font plus que limiter l'intérêt d'une ontologie locale à la localité pour laquelle elle a été conçue (l'alphabet seul définit mal le « i »). Le rapport entre partie-contenu et contenu-tout est abductif (*hypothèse du tout imaginaire possible, ce tout ne pouvant plus apparaître dans une même phase de la réalité*).
- **Ce nom peut servir à introduire chaque contenu-espace dans une ontologie de cet espace**. Pour un i graphique, il peut s'agir d'un mot écrit, d'une ontologie des lettres, d'un système de numérotation, d'une bordure de minitel, etc.
- **Ce nom peut servir à introduire tous les contenus-espaces dans une ontologie multi-spatiales**. Pour le « i » abstrait, l'alphabet et le système phonématique.
- Au final, **nous nous retrouvons par le simple fait d'avoir ajouté un nom à un contenu** discernant des formes considérées comme extérieures à nous, **avec un nouveau type de forme, des formes intérieures à nous qui n'existent qu'en tant que nous les discernons**.

L'étude du discernement de ces formes abstraites, c'est à dire l'étude des contenus marqués par des signifiants est l'objet de la sémantique. Cet objet ne se résume pas au couple saussurien signifiant/signifié. Effectivement, un hypergraphe pouvant se ramener à un point [LARSEN98], si signifiant et signifié étaient indépendants, il resterait possible de ramener chacun d'eux à des points et de les relier par une barre horizontale. Mais signifiant et signifié d'un signe sont tous les deux des constructions méréologiques dont les parties entretiennent entre elles et avec les parties de l'autre bord des liens à tous niveaux. Insécable, le signe ne peut se ramener à deux parties reliées par une simple barre.

3. La pragmatique abstraite

Le signe ne peut se ramener à deux parties reliées par une simple barre : il en est ainsi des signes utilisés pour désigner des catégories linguistiques (§ 3.1). Au lieu de dire que telle chose a telle catégorie il nous faut alors dire telle chose a telle catégorie selon tel constructeur, tel contenu nommé par la catégorie (§ 3.2). Mais l'implémentation d'une telle relativité suppose des outils nouveaux (§ 3.3). Alors que l'on voudrait encore fonder nos théories sur des origines, un nouveau cadre pourrait nous affranchir de ces origines (§ 3.4). Une théorie sans fondement prédicatif doit voir le jour (§ 3.5). Elle ne se justifie que par et pour l'ordinateur (§ 3.6).

3.1 A propos des catégories

A bien des égards, les catégories telles qu'on les entend ordinairement reflètent la théorie des ensembles. En tant qu'accumulation de catégories d'horizons variés, notre ressource linguistique appelée *Le Dictionnaire Intégral* (LDI) la reflète aussi. Mais pour cette raison même qu'il est une accumulation, il perd un peu de cette cohérence que la théorie semble apporter. Cela motive parfois l'anathème final sanctionnant notre pratique : elle ne respecte pas les propriétés élémentaires d'un arbre de Porphyre. 2 387 années après la naissance du grand philosophe, une bonne représentation devrait se limiter à la présentation d'un genre et d'une différence spécifique pour coder l'espèce, tout le reste n'étant que figure de l'esprit. Mais s'agissant précisément d'esprit avec le langage humain, notre discipline (théorie de l'information) devrait changer de conception.

La linguistique a, elle aussi, bien des difficultés à s'affranchir des catégories. C'est un peu ce que note [Culioli81, p35] :

> *Etant donné la multiplicité des situations, selon les langues, on peut parfois éprouver quelque gêne à employer certains termes habituels ; il arrive alors que l'on recourt à des termes plus généraux, mais peut-être aussi plus vagues : au lieu de nom, on parlera de nominal, au lieu de verbe, on parlera de verbal ou verboïde qui signifient qu'il ne s'agit pas tout à fait ou exclusivement de noms ou de verbes, mais cependant laissent entendre qu'on estime essentielle la distinction nom-verbe*.

Depuis quelques années, concernant les catégories, une conviction se répand : quand nous observons tout ce qu'on appelle « nom » dans les ouvrages, et que nous nous demandons quelles peuvent (ou quelle peut) bien être la(es) propriété(s) commune(s) qui les unisse(nt) tous, nous ne trouvons que l'ensemble vide. Il n'est pas question de faire ici un procès à la catégorie *nom* plus qu'à une autre catégorie. Nous essayons simplement de souligner qu'à chaque fois que nous rencontrons la catégorie *nom*, il faudrait avoir comme précision en quel(s) sens, pour quelle(s) propriété(s) intensionnelle(s), ce mot est utilisé. Il nous faut lire la liste suivante de sept noms et en dégager les propriétés communes : *acabit, aboutissement, deux, guerre, lundi, maison et pour*. Pour autant, il serait assurément inconséquent de

retirer ces mots de la catégorie *nom*. *Nom* est comme *i* : il ne se réalise que dans la pensée. Dans des mesures différentes, le *i* graphique dans un mot, le *i* sonore dans une voix humaine, le *i* de l'écran minitel, le *i* crié par un oiseau sont tous bien des *i*. Et il nous faut définir ces mesures qui consistent en des moments de la pensée (une évolution d'un système abstrait). Dans la campagne actuelle d'évaluation des analyseurs syntaxiques, des acteurs sont en lice pour déterminer si *manger* dans *manger un bonbon plairait à Lise* est un nominal (tag="nom") ou un verbe (tag="v"). La bonne réponse est la suivante : il est un nom selon tel et tel contenu légitime définissant la catégorie *nom* qui comme la catégorie *i* ne peut pas par principe se réaliser en entier dans une occurrence et il est un verbe selon tel et tel contenu légitime définissant la catégorie *verbe* qui comme la catégorie *i* ne peut pas se réaliser par principe en entier dans une occurrence. De toutes les façons que l'on prenne le problème, l'évaluation devient bien plus fausse que les systèmes évalués : en effet, les systèmes évalués ont plus de mobilité que notre conception de l'évaluation.

3.2 Relativité de l'espace de représentation

Ainsi toute caractérisation n'est vraie que dans un lieu, un espace qui lui est propre : par exemple dans le dispositif abstrait correspondant à l'instant de la pensée de celui qui les a conçues. Il n'est pas d'assertion générale valable. Une proposition comme :

$$\forall l\_\forall x\_P(x)$$

avec l un domaine, un lieu quelconque, *peut toujours* s'avérer fausse. Il n'importe pas de savoir si *2+2=4* est une proposition généralement vraie ; ce qui importe est de savoir dans quel lieu cette proposition est vraie. Pour de nombreux auteurs, la méréologie est utile pour mieux rendre compte de la réalité : il n'est pas facile de rendre compte d'une table avec un ensemble de points. Elle est le plus souvent adaptée tel quel (pour les quelques auteurs qui ont pris conscience de son intérêt, la plupart des autres essayant par exemple de construire des espaces vectoriels) dans les travaux de traitement automatique des langues ou de représentation des connaissances. Elle peut prendre alors le nom de méréotopologie et revendique parfois son utilisation pour la mise au point d'une physique naïve. Or il n'est pas question non plus de cela ici, mais seulement de la prescription élémentaire suivante : on s'interdira dorénavant de dire *ceci est cela*, mais l'on acceptera davantage quelque chose comme : *ceci est cela dans tel lieu*. Korzyski et Lienewsky le dirent en leur temps d'une façon différente. Le simple ajout de la notion de lieu de pertinence d'une assertion pourrait avoir une portée considérable pour une discipline comme la sémiologie : en effet, une sémiologie endogénéisant ce lieu **aurait un domaine sans bord, indéfiniment extensible**.

3.3 Graphe de graphes, hypergraphe

Soit l'affirmation suivante [CULIOLI81, p134] : *Là encore, le formalisme des treilles est plus adéquat que la représentation linéaire parenthésée. Seule une présentation formelle dans le langage algébrique des treilles permettrait de montrer comment se constitue un énoncé (plus généralement une famille structurée d'énoncés*). Or il advient que ces treillis que nous utilisons depuis quinze ans rencontrent également trois limites qui rendent impossible de montrer comment se constitue une famille structurée d'énoncés. Deux de ces limites appartiennent à la topologie générale. La troisième de ces limites concerne particulièrement le langage.

Considérons un même objet présent dans plusieurs espaces. C'est le cas de la plupart des objets. Cet objet peut être un objet abstrait comme notre idée du « i » (se dit [i], s'écrit *i*, est une lettre, etc., selon notre horloge psychique), notre idée de « maison » (abrite, lieu, est un nom, etc., selon la même horloge), ou un objet concret (cette tasse remplie de café à tel moment du temps physique).

La première limite est que les différents espaces de validité de chaque assertion ne peuvent pas être donnés dans un treillis. Un treillis reste fondamentalement homogène (sauf à placer dans des types de relation des choses qui ne devraient pas y être). Une solution courante à ce problème consiste à constituer différents treillis ou différentes bases de données. Mais alors nous rencontrons d'importantes difficultés à établir des liens entre ces espaces.

La deuxième de ces limites est qu'un treillis ne peut pas enregistrer des objets composites dans un même espace, des objets qui présenteraient plusieurs parties organisées et visibles dans un tout particulier de cet espace. Par exemple, le treillis ne peut pas disposer la position des pieds d'une table par rapport à son plateau ni plus que décrire un syntagme particulier comme formant un tout. Nous trouvons par exemple une difficulté importante, largement étudiée sous le nom de couplage syntaxe-sémantique. Pour une entreprise comme la nôtre qui dispose depuis l'origine d'un axe paradigmatique massif cette question théorique a été inquiétante.

La troisième de ces limites est que dans un treillis un même objet ne peut pas être à la fois dans le sommet gauche d'une relation orientée (côté signifiant par exemple) et dans la partie droite de cette relation. Par exemple, dans un treillis les expressions suivantes *le mot nom est un nom* ou bien *le mot cheval est un nom* sont totalement inconsistantes, ne construisent aucune forme méréologique intéressante. Par ce fait, un moyen formel insuffisant comme le treillis impose la constitution d'un métalangage alors même qu'il n'est certainement pas raisonnable de vouloir traiter d'un système qui s'engendre lui-même (*La langue engendre la langue* [CULIOLI81, p. 24]) par le fait primitif qu'il incorpore sa propre métalangue (*La métalangue est dans la langue* [*Ibid* p. 29]). Si la *langue sert, en effet, de système sémiotique interprétant de tout autre système sémiotique signifiant*, [*Ibid* p. 28], si de plus *une langue naturelle est un système sémiotique ayant quelques caractéristiques qui en font un système sémiotique fondamental à la base de toute classification*, on ne peut rien tenter de moins que la fabrication d'un système artificiel qui serait la théorie et approcherait la fonctionnalité. [*Ibid* p. 33] : *Vouloir*

se passer entièrement de toute langue naturelle pour décrire une langue naturelle, cela est illusoire et ce fait relève des particularités spécifiques des langues. Il est peu probable que la gestion d'hypergraphes (graphe de graphes) ayant pour sommet tantôt des points, tantôt des graphes de points et tantôt des formes méréologiques (composées de plusieurs parties dotées de parties plus complexes que celles fournies par la liste ou la liste de listes) nous éloigne davantage de la fonctionnalité nécessaire que la gestion de taxinomies ou de treillis. Donnons quelques exemples des possibilités offertes.

3.4 De la possibilité d'abandonner la recherche des origines

En linguistique, les origines sont nombreuses. Elles peuvent concerner l'organisation de toute la langue, l'organisation des sens d'un mot, la signification d'une occurrence particulière dans un énoncé par rapport à un glossaire de signification, la signification d'une occurrence d'un mot dans un énoncé par rapport à une théorie. Notre propos ici n'est pas de critiquer ces différentes tentatives car il faut savoir profiter des éclairages qu'elles ont apportés, et apporteront encore, mais de contribuer modestement à limiter l'adhésion sous la forme de croyance plutôt que sous une forme utilitariste. Après tout, si la langue était un couteau suisse, nous admettrions bien que chaque outil a besoin d'une présentation spécifique.

⇨ **Primitive structurante pour l'ensemble de la langue**

- Primitive sémique : par exemple, quel terme trait sémique doit-on choisir entre les traits portés par les mots *élément, ensemble, partie et tout* ?
- Primitive lexicale : même exemple, et quelle subordination entre *couleur* et *rouge* d'une part, entre *bois* et *forêt* d'autre part.

⇨ **L'organisation des sens d'un mot**

On trouve évidemment l'organisation *sens propre*, *sens figuré*, etc., mais le lecteur pourra trouver une littérature critique abondante sur le sujet. Plus intéressante est l'idée de *forme schématique* qui constitue un point important de l'œuvre de Culioli. On notera ici le nombre singulier de l'expression. Il ne s'agit nullement pour un mot comme *lit* de pouvoir apparaître d'emblée dans l'espace théorique traitant de la langue avec plusieurs formes schématiques. Soit un arbre dérivant au fil de l'eau : il faudrait au contraire que ce qui émerge constitue un seul morceau, ceci rassurant sur l'unicité de l'arbre tout entier, partie immergée et cinématique de l'arbre dans l'eau inclus. [FRAENKEL91] note à ce propos : *Cette question conduit à reformuler la question sous une forme plus brutale : le mot lit a-t-il un sens ? Ces observations tendent à montrer que la réponse est négative dès lors que l'on appréhende ce sens comme décomposable en sèmes ou en paramètres sémantiques*. Ainsi, si l'on présente *lit* comme émergeant à travers différents sèmes (indépendants), *lit n'aurait plus de*

sens unique. Pour éviter cela, il faudrait que les sens et tous les emplois de *lit* provinssent d'une souche visible en son entier. Quelle serait cette souche pour *i* ?

⇨ **Un sens répertorié identifiable dans une occurrence**

C'est la croyance aujourd'hui la plus répandue en traitement automatique de la sémantique lexicale et dans certaines manifestations de la linguistique. Cette croyance voudrait que pour être scientifique :

- Un lexique présente un découpage des sens d'un mot tel que ce découpage entretienne des relations bijectives avec les significations en contexte. Cela a fait dire et permit de voir écrit que le dictionnaire Le Robert est assez mal construit du fait que des promotions d'étudiants n'ont pas pu se mettre d'accord sur le marquage en terme de numéro de signification du Robert de la plupart des occurrences des unités lexicales présente dans un corpus sur lequel chaque étudiant avait comme mission d'effectuer ce marquage.
- Un programme informatique soit capable, comme dans la conférence américaine SENSEVAL, de retrouver ce marquage humain.

Au plan linguistique, cette attitude peut être rapprochée de celle qui conduit à insérer dans un système de catégories grammaticales des éléments comme *personne* et marquant avec ce trait d'une façon semblable le mot samouraï dans *le samouraï mangeait* et dans *le samouraï aimait à combattre*. Là encore, ce qui importe est de savoir dans quelle mesure précise samouraï est une personne dans ces énoncés.

Nous posâmes avec un groupe d'amis la question suivante à un orateur linguiste :

Le groupe : *Combien avez-vous d'enfants ?*

L'orateur : *2,5 enfants, peut-être m'en donnerez-vous davantage ou un peu moins.*

Le groupe : *Comment avez-vous 2,5 enfants*.

L'orateur : *C'est le nombre moyen d'enfants qu'un autre groupe m'a accordé quand j'ai raconté mon histoire*.

Cet orateur avait comme candidat enfant un fils biologique non connu jusqu'à l'âge de 9 ans, non reconnu, non élevé jusqu'à cet âge, une fille élevée, reconnue mais non biologique, une autre fille, demi-soeur biologique de la première, reconnue et élevée depuis la naissance et une dernière fille, non biologique, mais sœur du fils biologique, et élevée avec lui depuis l'âge de la rencontre de sa famille perdue, 5 années auparavant. Comme l'orateur aimait fort ces enfants, par vote, notre groupe lui consentit 3 et ⅓ enfants.

Comme pour *i*, pour *père de*, pour *enfant de* il nous revient toujours de calculer ce que l'on veut dire par là, et pour le moins quel *sème* est attaché (l'état-civil, la

biologie, l'éducation, l'amour). Quel est pour vous le sens primaire de *père de* ? Ou bien, s'il n'a pas de sens primaire mais seulement un complexe de significations, quel trait allez-vous choisir dans le Robert pour chacun de ces enfants si d'aventure vous n'en deviez choisir qu'un seul ? Et puis, en quoi « père de » serait-il toujours une personne ? Qui de Chronos ou du Verbe est le père des signes ?

3.5 La théorie du blanc et la pragmatique abstraite

[CULIOLI81 p. 55] note à propos des explications d'un linguiste concernant la formation de certains énoncés en Mùuré : *Par des substitutions, des élisions et des effacements tous injustifiés, comme nous venons de le voir, il est clair que l'on pourrait non seulement dériver n'importe quelle langue, mais aussi passer d'une langue à l'autre. Il est douteux cependant que cela soit d'un quelconque intérêt*.

Ce que nous appelons blanc s'oppose très précisément au noir. Les théories noires étudient les textes, les taches d'encre sur le papier. Ces théories s'adressent à l'homme et ont pour résultat d'expliquer un texte par un autre texte. C'est en cela que nous les appelons théories noires : le nouveau texte, ensemble de taches d'encre, reste bien évidemment à expliquer.

A l'inverse, le blanc s'intéresse à tout ce qui dans le texte écrit est rempli par l'espace, le vide ; tout ce qui n'aura jamais de phénomène. Dans l'espace idéel, il s'agit des dispositions informationnelles, des représentations que l'on se fait d'un texte. Comme depuis l'invention des catégories grammaticales, ces catégories appartiennent à certaines de ces représentations, les catégories appartiennent au blanc en tant que résultat particulier d'un certain type de discernement. Si ce discernement particulier présente une certaine inter-subjectivité alors il doit être retenu. Si d'une façon générale, les catégories linguistiques ont à voir avec les autres types de représentation (d'une façon ou d'une autre, ces catégories délimitent des régions), il n'en reste pas moins vrai que ces catégories dans leur extension actuelle ou même future ne fournissent ni fondement ni finalité particulier à la théorie du blanc. Après tout, peu importe une analyse fine de *quelle est la couleur du cheval blanc d'Henri IV* si, l'effectuant, l'élève et l'ordinateur, tous les deux savants en grammaire, ne savent pas répondre. Mais ce qui est vrai pour la linguistique, l'est aussi pour les statistiques : une bonne réponse au problème posé ci-dessus fondée sur des critères statistiques, pour toute utile qu'elle soit, n'en serait pas davantage une bonne réponse pour la théorie du blanc. Car cette dernière s'intéresse principalement aux opérations informationnelles que chaque occurrence de mots convoque.

Tout le début de cet article a été focalisé sur le point suivant : quand nous rencontrons un signe lambda, par exemple *i* ou bien *nom*, ou bien *lit*, ou bien *père de*, etc., il nous faut toujours expliquer pour quelle partie ces énonciations de *i*, de *nom*, de *lit*, de *père de*, etc., sont pertinentes. Or il advient que cette explication est le résultat d'opérations particulières qui, puisqu'elles sont rarement données dans les articles, entretiennent des disputes sur la nature des choses plutôt que sur les opérations associées à ces choses.

Nous ne souhaitons pas cependant conclure cette longue section en laissant la possibilité de croire que si effectivement dans les textes les opérations cognitives ne sont pas données, au moins sommes-nous assurés que dans les textes l'ensemble des marqueurs de ces opérations nous serait donné sous la forme des mots que nous trouvons. En effet, la plupart des mots qui permettraient de retrouver une représentation informationnelle sont absents du discours. Prenons trois exemples pour montrer comment la théorie du blanc devra autant s'intéresser aux « mots » manquants, aux blancs de l'énonciation qu'aux mots présents.

⇨ **Exemple 1 : « remplissage » ou « effacement » imposé par la grammaire**

Soit la première phrase des règles officielles du jeu d'échec en anglais et sa traduction, officielle également, en français :

- *The game of chess is played by two opponents*.
- *Le jeu d'échecs se joue à deux*.

⇨ **Exemple 2 : « Effacement » permis par la compréhension du co-texte**

Soit : *Pierre boit-il du vin ?*

Il nous faut comparer en terme d'aspect les réponses suivantes :

- *Oui ; Dominique boit du vin*.
- *Oui ; Dominique boit un verre de vin*.

⇨ **Exemple 3 : « Effacement » permis par la compréhension des mots**

Il faut se demander pourquoi *quelle est la couleur du cheval blanc d'Henri IV* a normalement pour réponse *blanc*. Est-ce ici question d'intelligence, de connaissances socioculturelles inaccessibles ou bien est-ce moins magiquement que l'on aurait oublié de modéliser convenablement ce qu'un dictionnaire de langue nous indique ? Une question est de savoir si l'énoncé sert véritablement à établir un lien direct entre cheval et blanc ou effectue fondamentalement autre chose. Cette autre chose pourrait être l'évocation par deux points des éléments d'un réseau systématique compris dans l'organisation de la langue, parfois même du langage.

Dans *cheval blanc*, ce n'est pas le cheval qui est blanc. Analyse :

- Qu'est blanc à cette position ?

  Blanc est une valeur.

- Une valeur a-t-elle une existence propre ?

  Une valeur suppose une fonction, un caractère, une propriété etc. qui retourne cette valeur.

- Quel caractère correspond à blanc ?

  Le caractère *couleur*, mais de nombreux caractères n'ont pas de signifiants.

- De quoi couleur est-il de la façon la moins métaphorique le caractère ?

  D'une surface particulière. Une surface particulière correspond à une extraction d'un corps opaque lors d'une observation. Cette surface est une valeur.

- De quoi une surface est-elle une valeur ? etc. Mais ce etc. s'avérera inutile.

La présentation lapidaire par rapport au véritable cadre qu'il faut mettre en place, aboutit finalement à la reformulation suivante :

Quelle est la valeur de la couleur de la surface appelée cheval d'Henry IV qui a une couleur de valeur blanche ?

Ce genre de dispositif présente l'avantage de n'être pas fondé sur une connaissance d'un cheval en trois dimensions. Le texte précédent n'affirme ni la troisième dimension ni l'animal. Alors la résolution du problème n'emploie pas ces connaissances. Le cheval d'Henry IV pourra rester une plaisanterie scolaire.

⇨ **Exemple n : généralisation**

Il y a quelque temps, nous avons eu la charge d'essayer de produire toutes les façons de demander la valeur du cours de Bourse de la société Toto :

Je veux le cours de Toto. Puis-je avoir, etc.

Pour la partie sémantique je veux + ∅, nous avons trouvé plus d'un 1 300 000 énoncés pragmatiquement équivalents, sans épuiser la question.

Pour la partie sémantique *cours de toto*, nous avons trouvé plus d'un 90 000 énoncés, sans épuiser la question. Tous ces *énoncés avaient des blancs, ici ou là. Qui ne voit que l'énoncé vouloir* un chat est impensable tel quel, sans ajout de quelque chose. Ce quelque chose est nécessairement une prédication souhaitée impliquant le chat. C'est par exemple : *Je veux avoir un chat*. [VICTORRI99] aurait écrit « *Je roule pour un chat* » mais cela est nettement plus compliqué.

Nous pourrons retenir en dernier ressort l'ordre de complexité dans lequel il faut maintenant replacer nos exemples d'effacement. L'exemple le plus simple est fourni par *cheval blanc*. La construction à donner y est aisée et nous l'avons grossièrement esquissée. Evidemment, la façon dont *verre* produit un effet imperfectif est bien plus difficile à transmettre à un ordinateur et cela suppose des représentations intermédiaires. Enfin, le premier des exemples présente une difficulté inhérente à tout ce qui est général : il reste à fournir un contour assez précis du phénomène.

L'expression *pragmatique abstraite* est née des faits que :

1. Dans chaque occurrence d'un signe il faille saisir une compréhension particulière et que,
2. La compréhension particulière à saisir peut procéder d'un signe absent de l'énoncé hors toute situation hors texte.

> Avançant dorénavant sans filet puisqu'il faut ajouter et retrancher à l'envie, il convient que l'ordinateur détermine lui-même une cible. Et il semblerait bien que la seule cible disponible soit d'emblée une forme élaborée de modélisation et de représentation. Finalement l'oxymore « pragmatique abstraite » prend pour sens ce qu'Austin a dit des actes de discours, mais dans l'espace abstrait de la représentation. Ici, tous les candidats-contenus sont bienvenus.

4. Au service des documents

Réalisant dès le départ pour la partie « sémantique » un treillis, le problème que nous avons eu à résoudre n'a pas tellement été le problème de la détermination du numéro de sens d'une occurrence dans un texte. En fait, nous nous sommes à un moment donné, entre 1991 et 1993, posés ce genre de questions mais il est vite apparu que ce que l'on appellerait ici « sens » nous apporterait d'avantage en terme de reconnaissance sociale à court terme que de solutions permettant d'avancer dans la résolution des *problèmes généraux*. Concernant la sémantique lexicale, nous en sommes venus à nous poser la question suivante : quelle lieu abstrait, quel domaine une occurrence d'un mot active-t-elle (sachant un contexte) ? Etant capable de découvrir une partie de ces régions abstraites, sous la forme de points grossiers, quelle utilité pouvons-nous en tirer. A côté, nous nous sommes égarés longtemps en matière de grammaire, où sous la pression d'un ensemble d'étiquettes ou de fonctions présentés comme des standards, nous avons durablement cherché à trouver en ces étiquettes des fins en soi. Aujourd'hui nos systèmes peuvent « réapprendre » à tagger selon n'importe quel jeu d'étiquettes en quelques jours de calcul. Finalement, nous parlerons dans cette dernière partie davantage de sémantique lexicale que d'autre chose. Mais nous devons en premier lieu présenter la batterie d'outils qui permettent d'aborder quelques traitements de sémantique lexicale. Nous terminons notre présentation par un exemple concret d'application documentaire.

4.1 Des outils pour travailler

Les outils sont des contenus capables d'effectuer des opérations (en particulier de discernement) sur des formes présentes dans le document ou sur des formes absentes de ce dernier (mais inférables). Nous appelons traitement de surface tout traitement qui fabrique un token avec sa langue, sa catégorie grammaticale, sa fréquence, son paradigme flexionnel, son contexte dans la phrase. L'ensemble des outils appartient à la boîte à outils Sémiographe™ (java). La boîte à outils Sémiographe™ correspond elle-même, pour l'essentiel, à une compilation des données du Dictionnaire Intégral™ (LDI). LDI est géré par Lexidiom™. Lexidiom est rédigé en java et est motorisé par Firebird. Les données alphanumériques sont toutes UNICODE FSS.

4.1.1 Les outils de la surface

La boîte à outils actuelle est constituée de la liste suivante :

1. Correction et recherche phonétique.

2. Reconnaisseur/correcteur d'erreurs clavier (inversion, répétition, absence, etc.).

3. Reconnaisseur de langue.

4. Reconnaisseur de mots simples et mots composés : reconnaissance, flexion, lemmatisation, catégorie grammaticale, fréquence etc.

5. Reconnaisseurs Chiffres, formules chimiques, équations, ordinaux, dates, adresses, mail, http, ftp, symboles de liste numérotée etc.

6. Reconnaissance entités nommées.

4.1.2 Les outils lexico-sémantiques

Ces outils concernent des calculs monolingues et interlingues. Ils ont un fondement génératif (fonctions lexicales), sémasiologique, onomasiologique et componentiel.

Fonctions monolingues

7. Fonctions lexicales monolingues

Elles reflètent une interprétation relativement libre des fonctions lexicales prévues dans la théorie sens→texte d'Igor Mel'çuk. Plusieurs fonctions de sens→texte sont incluses dans le treillis du Dictionnaire Intégral et donc ne sont accessibles que par une opération à effectuer sur ledit treillis. Au total, les 43 fonctions différentes du Dictionnaire ont 110 000 occurrences.

8. Les consignes d'emploi

• 'niveau de langue' (ex : litt.), 'fréquence' (ex. rare), 'origine' (ex. de l'anglais), 'statut officiel' (ex. recommandé), 'domaine' (ex. maritime) sont donnés.

9. Informations encyclopédiques prenant la forme de fonction lexicale

• des lieux inclus et des lieux incluants

Lieux inclus (Calvados)={Caen(ville), Bayeux(ville), ... Orne (rivière), ... }

Environ 90 000 relations de ce type existent dans le Dictionnaire Intégral.

Fonctions inter-langues

Les fonctions inter-langues sont organisées autour de la notion de Synset (ensemble de synonymes), à la manière de WordNet [FELLBAUM98] et EuroWordNet [VOSSEN98]. Ces fonctions ont 350 000 occurrences actuellement.

Structure componentielle

Avec le fait (récent et encore peu pris en compte dans nos algorithmes) qu'une relation quelconque dans le Dictionnaire Intégral est toujours validée selon un lieu particulier (une langue, une ontologie locale etc., dans un hypergraphe), la structure componentielle est l'une des exclusivités de notre ressource. Cette structure est aussi la plus ancienne couche du dictionnaire : épine dorsale et infrastructure, elle existe depuis le début de nos travaux (1989). La question que nous posions alors était la suivante : considérant un mot tel que *yen*, ce mot doit-il être classé dans les monnaies, en tant qu'hyponyme, ou selon un Etat particulier : le Japon. Les deux classements étant pertinents selon deux points de vue différents, des espaces abstraits différents, nous avons conçu au départ un outil gérant des graphes orientés plutôt que des arbres. La démarche que nous avons tentée de suivre était la suivante :

- Découpage des mots en concepts autonomes (ici, \monnaie et \Japon[1])
- Faire vivre les concepts créés (remplir \monnaie avec les monnaies des différents pays, \Japon avec les choses définies par le Japon)
- Rechercher dans les mots de ces concepts un ou plusieurs éventuels hyperonymes (ex : monnaie)
- Définir les concepts créés par d'autres concepts plus élémentaires, récursivement. Faire vivre ces concepts, c'est-à-dire les remplir. (pour \monnaie, \unité et \échange monétaire par ex.), rechercher les hyperonymes, etc.
- Essayer, pour les termes polysémiques, de dégager un signifié de puissance sous la forme de plusieurs de ces concepts.

Le résultat est un énorme treillis dans lequel chaque nœud (mot ou concept) a en moyenne 2,1 pères. Dans ce treillis déjà largement décrit [DUTOIT00], il est possible de retrouver des éléments qui appartiennent à ce que l'on pourrait appeler des ontologies locales, des isotopies sémantiques ou des champs référentiels selon ce que l'on obtient, la terminologie linguistique utilisée et le regard porté sur le résultat.

Les fonctions de base de la structure componentielle sont les suivantes :

10. Extraction des sèmes partagés et des sèmes de différenciation entre un lexème et un ou plusieurs autres lexèmes. Ex : de *yen* vers *Japon*, le sème commun est \Japon, la différence est \unité monétaire c à d \unité+\échange monétaire c à d,etc.

Cette fonction permet par exemple de tester la saturation d'un mot par un texte : pour *yen*, si le texte comporte des éléments renvoyant \monnaie (ex : acheter) et des éléments renvoyant \Japon (ex : saké), le mot *yen* est dit saturé.

[1] Le signe \ note des éléments considérés comme non linguistiques. Ces informations sont à l'usage du lexicologue.

11. Extraction rapide des sèmes partagés : fournit la partie commune, rapidement, mais avec moins de contrôle sur les chemins que la fonction précédente. Cette fonction ne permet pas de tester la saturation des traits. Elle est néanmoins utile quand on veut comparer rapidement un mot avec un texte (par exemple).

L'ensemble des résultats obtenus est associé à une métrique. Cette métrique peut varier sensiblement selon l'application. L'ensemble (obtention des résultats par la définition de chemins et métrique) est défini d'une façon extérieure au programme par la personne chargée de développer un service.

4.2 Deux réalisations

Nous décrivons dans cette section deux services utilisant le Sémiographe. Elles sont le fait de clients. L'un est une importante société de GED. L'autre est une multinationale ayant des activités en GED.

⇨ **Gestion d'un fond documentaire**

Les spécifications ont été :

- Faciliter la maintenance du thésaurus métier en organisant une veille terminologique depuis un flux de documents métier,
- Faciliter l'accès au thésaurus par les internautes (11000 termes dans le thésaurus),
- Permettre l'accès à un document en "langue naturelle".

➢ *Gestion du flux et veille terminologique*

La SSII cliente disposait déjà d'un outil de séquençage (repérage des séquences répétées) et d'un algorithme de clustering (présentation des séquences trouvées dans l'environnement des termes du corpus). Ce client a choisi d'enchaîner :

→3[2] 'Reconnaisseur de langue'→4 'Mots simples' →6 'Reconnaisseurs' →7'Etiqueteur'→

➢ *Accès thesaurus*

Une projection de ce thésaurus métier est effectuée sur le Dictionnaire Intégral :

→4 'Mots simples' →5 'Mots composés' →12 'Extraction des sèmes partagés et des sèmes de différenciation'→

L'outil 12 un peu lent est ici adapté pour repérer les parties de l'architecture du thésaurus qui se retrouvent dans LDI.

[2] Numéro du paragraphe décrivant l'outil. La flèche donne une idée de la cinématique.

➢ *Permettre l'accès à un document en « langue naturelle »*

L'implémentation retenue favorisa le recours aux fonctions lexicales.

⇨ **Gestion documentaire multilingue**

Le progiciel concerné est une plate-forme de gestion documentaire multilingue.

La satisfaction du besoin est passée par la production d'une ressource multilingue obtenue par projection du Dictionnaire Intégral sur des vues particulières :

→9 'Fonctions lexicales monolingues' →12 'Fonction lexicale interlingue' →13 'Extraction des sèmes partagés'→

Ce genre d'application est suffisamment demandé pour :

- Envisager un développement parallèle en d'autres langues,
- Augmenter les utilisations de la ressource,
- Entretenir des liens étroits avec les autres « ontologies » générales comme WordNet [FELLBAUM98], EuroWordNet [VOSSEN98], Balkanet [STAMOU02] et maintenant SUMO [FLATER03].

5. Conclusion

Bien évidemment, les liens qui existent entre contenus des documents et documents sont multiples et variés. Mais il est possible d'extraire certains contenus dans la perspective d'une application plutôt que d'un domaine : les domaines n'étant jamais, par définition, suffisamment précis, l'affûtage de quelques outils de traitement de la langue est une alternative responsable.

Il y a quelques années, cela a été presque une surprise de constater que notre travail analogique, essentiellement mais seulement associationniste, pouvait rendre service à l'industrie. Aujourd'hui, nous le regardons aussi comme une analyse de l'existant, quelque chose sur lequel on pourra fonder de nouveaux contenus. Ces nouveaux contenus seront peut-être de notre vivant ceux de la *pragmatique abstraite* si la communauté est prête à supporter le défi. En attendant, nous pourrons toujours suivre le développement des ontologies locales du Web Sémantique.

Remerciements

Ils vont à Charles Hunt, mathématicien, pour son questionnement insistant et sa réactivité finale quand il a été question d'orienter la recherche d'outils plus larges. Ils vont à Marianne Dabbadie pour ses relecture et critique. Ils vont à Nadine Lucas qui aida à conserver un ancrage dans la pratique. Ils vont enfin au comité CIDE 2004 qui a montré patience et tolérance.

6. Références bibliographiques

[BERGSON11] Bergson, Henri, 1911, « La conscience et la vie »", *réédition Magnard*, 1986

[CASSIRER33] Cassirer, Ernst, 1933, « Le langage et la construction du monde des objets », in *Psychologie du langage*, Paris 1933

[CULIOLI81] Culioli, A., Desclés, J.P., « Systèmes de représentations linguistiques et métalinguistiques : les catégories grammaticales et le problème de la description de langues peu étudiées », *Collection ERA 642*, numéro spécial, n° ISSN 0223-548X, 1981.

[DESCLES91] Desclés Jean-Pierre, « Au sujet des catégories grammaticales », *La théorie d'Antoine Culioli, Ouverture et incidence*, (Ophrys, 1991).

[DUTOIT00] Dutoit, Dominique, « Quelques opérations sens→texte et texte→sens ... », *PhD thesis*, Univ. de Caen, 2000, pp. 132-140.

[FELLBAUM98]Fellbaum, Christian, "WordNet : An electronic lexical database", Cambridge, *MIT press*, 1998.

[FLATER03] Flater, D,. 2003. "SUMO2LOOM Documentation. ", *National Institutes of Standards and Technology Report*, January 9, WERB, 2003.

[FRAENKEL91] Fraenkel, J.J. et Lebaud, D., « Lexique et opérations », *La théorie d'Antoine Culioli, Ouverture et incidence* (Ophrys, 1991).

[LARSEN98] Larsen, Niels C., Molz, Martin K., Normark, Kurt, "Graph abstractions as the basis of an extensible Editing Tool", Denmark, *Aalboorg University*, 1998.

[STAMOU02] Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufis D., Koeva S., Totkov G., Dutoit D., Grigoriadou M., "Balkanet: A multilingual Semantic Network for Balkan Languages", *First International WordNet Conference*, Mysore India.

[SMITH96] Smith, Barry, "Mereotopology: a theory of parts and boundaries", *Data and Knowledge engineering*, 20 (1996), pp. 287-303.

[VOSSEN98] Vossen, Piek, "EuroWordNet : a multilingual database with lexical semantic networks", *Dordrecht, Kluwer Academic Publisher*, 1998.

[VICTORRI99] Victorri, Bernard, « Le sens grammatical », *Revue Langue, Larousse*, 1999.

Les documents auto-explicatifs : une voie pour offrir l'accès au sens aux lecteurs

Hervé Blanchon, Christian Boitet

Laboratoire CLIPS-GET
BP 53, 38041 Grenoble cedex 9 - France

{herve.blanchon,christian.boitet}@imag.fr

Résumé :

Dans le cadre du projet LIDIA, nous avons montré que, dans de nombreuses situations de traduction multicible, la Traduction Automatique Fondée sur le Dialogue (TAFD) peut être une très bonne alternative à des outils classiques d'aide au traducteur ou de traduction automatique, même pour des langages contrôlés. Nos premières expériences ont montré le besoin de conserver une mémoire des intentions de l'auteur au moyen d'annotations de désambiguïsation qui transforment le document source original en un document auto-explicatif (DAE). Dans cet article, nous présentons un moyen d'intégrer ces annotations dans un document XML et de les exploiter afin de permettre à des lecteurs de comprendre les intentions de l'auteur. Nous montrerons aussi qu'une fois traduit dans une langue cible, un DAE peut être transformé de façon automatique en un DAE dans cette même langue cible.

Mots-clés : désambiguïsation interactive, traduction automatique fondée sur le dialogue, modèle de document, document actif.

Abstract:

In the framework of the LIDIA project, we have shown that, in the context of multi-target translation, Dialogue-Based Machine Translation (DBMT) may be a good alternative to classical professional translator tools or machine translation tools, even in the context of controlled languages. Our first experiments have shown the need to keep a memory of the author's intentions using disambiguating annotations which transform the original source document into a self-explaining document (SED). In this paper we present a mean to integrate those annotations within an XML document and exploit then in order to allows readers to understand the author's intent. We will also

show that, once translated into a target language, a SED may be automatically transformed into a SED in this target language.

Keywords: interactive disambiguation, dialogue-based machine translation, document model, active document.

1. Introduction

On rencontre beaucoup de travaux sur le document électronique, que ceux-ci concernent les aspects multimédia ou l'annotation du contenu. Mais nous n'en connaissons aucun qui vise à permettre d'annoter par le sens désiré par l'auteur en cas d'ambiguïté. Si nous sommes, pour l'instant, les seuls à travailler sur cette idée, évidemment porteuse de nombreuses applications, c'est sans doute qu'elle n'a pu être concrètement envisagée qu'après avoir mis en place et expérimenté le nouveau paradigme de la Traduction Automatisée Fondée sur le Dialogue (TAFD).

La TAFD a été proposée dans le cadre du projet LIDIA [Boitet, C., 1990]. Il s'agit de permettre à un auteur monolingue de produire des traductions de qualité des documents qu'il rédige. Dès qu'il rencontre des ambiguïtés qu'il n'est pas capable de résoudre automatiquement, le système pose des questions à l'auteur afin qu'il désambiguïse interactivement son document. Ce type de système doit permettre la traduction de haute qualité de documents qui ne peuvent être traduits faute de temps, de traducteurs ou de solution automatisée satisfaisante.

Au cours du développement de la maquette LIDIA-1 [Boitet, C., *et al.*, 1994, Boitet, C., *et al.*, 1995a], nous avons naturellement été conduits à l'idée que les informations obtenues par le système lors de la phase de désambiguïsation interactive pourraient être conservées afin d'enrichir le document avec le sens qu'il véhicule. Un tel document enrichi serait alors un Document Auto-Explicatif (DAE). Un visualiseur de DAE pourrait montrer au lecteur où se trouvent les ambiguïtés et, à sa demande, préciser le sens choisi par l'auteur.

Dans cet article, nous présentons d'abord la notion d'ambiguïté en langue naturelle et en proposons une définition formelle. Nous décrivons ensuite le projet LIDIA et notre premier démonstrateur, en expliquant aussi comment produire un DAE en utilisant les informations collectées au cours de la phase de désambiguïsation interactive. Nous décrivons ensuite un nouveau démonstrateur qui permet de construire un DAE au cours de l'étape d'analyse, ainsi qu'une première réalisation d'un visualiseur de DAE. Nous concluons brièvement après avoir présenté les perspectives à court et moyen terme de ce travail.

2. Vers une définition formelle de l'ambiguïté

Pour « traiter informatiquement l'ambiguïté », il faut pouvoir définir une ambiguïté comme un objet. Or les définitions usuelles sont seulement du type : « un énoncé est ambigu s'il a, au moins, deux interprétations différentes ». Nous avons besoin d'une notion plus précise permettant en particulier de classer, de rechercher, et de visualiser les ambiguïtés, puis de les traiter pour construire des dialogues de désambiguïsation.

Nous rappelons d'abord ce qu'est l'ambiguïté en linguistique. Nous montrons ensuite les implications de l'ambiguïté pour le Traitement Automatique des Langues Naturelles. Nous proposons finalement une définition formelle utile pour le TALN.

2.1 L'ambiguïté en linguistique

Pour Catherine Fuchs [Fuchs, C., 1996], « pour qu'il y ait **ambiguïté linguistique**, il faut que les différentes significations en jeu soient prédictibles en langue, c'est-à-dire que l'analyse linguistique puisse en rendre compte : tous les **dictionnaires** du français doivent consigner le fait que la suite graphique *bière* correspond à deux unités lexicales, désignant respectivement la « boisson » et le « cercueil » ; toute **grammaire** du français doit prendre en compte le fait que la séquence « N1 faire V-infinitif N2 à N3 »[1] recouvre deux types de relations sous-jacentes correspondant respectivement à « N1 faire que X V N2 à N3 »[2] et à « N1 faire que N3 V N2 »[3]. »

L'ambiguïté peut être **virtuelle** ou **effective** : elle est virtuelle lorsque le contexte linguistique sélectionne l'une des significations ; elle est effective lorsque le contexte linguistique autorise plusieurs interprétations. Dans les exemples suivants, l'ambiguïté de bière reste virtuelle : [exemple 1] *la soif les poussa à commander deux bières au bar* (bière=boisson) et [exemple 2] *les croque-morts descendirent la bière dans la fosse* (bière=cercueil.) Par contre, dans la phrase *comme il faisait très chaud le jour de l'enterrement, on sortit la bière* (bière= ?), l'ambiguïté est effective.

2.2 L'ambiguïté en analyse automatique

Un analyseur automatique fournit deux classes de réponses vis-à-vis de l'ambiguïté : il peut être incapable de repérer certaines ambiguïtés effectives en langue (par exemple s'il connaît uniquement l'unité lexicale de bière qui correspond à boisson) ; il peut aussi considérer certaines ambiguïtés virtuelles comme des ambiguïtés effectives (par exemple il n'a pas les connaissances nécessaires pour choisir le sens boisson pour bière dans l'exemple 1). Dans le premier cas,

---

1 Exemple : *j'ai fait porter des fleurs à Lucie*.
2 Soit : *j'ai fait que quelqu'un porte des fleurs à Lucie*, pour notre exemple.
3 Soit : *j'ai fait que Lucie porte des fleurs*, pour notre exemple.

l'analyseur manifeste un défaut de couverture de la langue. Dans le second, l'analyseur manifeste un défaut d'« intelligence ».

2.3 L'ambiguïté comme un objet formel

Jusqu'à présent, nous avons dit qu'une phrase est ambiguë. En fait, on peut presque toujours réduire la localisation d'une ambiguïté à une partie de la phrase. Nous allons formaliser cela [Boitet, C., *et al.*, 1995b].

Prenons par exemple la phrase suivante :

(1) Do you know where the international telephone services are located?

Le fragment souligné contient une ambiguïté d'attachement qui peut être représentée par deux squelettes [Black, E., *et al.*, 1993] :

[international telephone] services / international [telephone services]

Cependant, il n'est pas suffisant de considérer cette séquence isolément. Prenons comme exemple la phrase suivante :

(2) The international telephone services many countries.

L'ambiguïté a disparu ! Dans la pratique, il est très fréquent que l'ambiguïté relative à un fragment apparaisse, disparaisse puis réapparaisse lorsque l'on augmente le contexte du fragment. Ainsi, pour définir proprement une ambiguïté, il faut considérer le fragment à l'intérieur d'une phrase et clarifier l'idée que le fragment utile est le plus court fragment sur lequel l'ambiguïté puisse être observée.

De manière formelle, on peut donc dire qu'un fragment **F** présente une ambiguïté de degré **n (n≥2)** dans une phrase **U** s'il possède **n** représentations différentes qui peuvent être utilisées pour produire une représentation complète de **U**. Pour être le support de l'ambiguïté, **F** doit être minimal relativement à l'ambiguïté considérée. Cela signifie que **F**, et les **n** représentations qui lui sont associées, ne peut être réduit à un fragment strictement plus petit **F'**, et ses **n** sous-représentations associées, sans perdre la première propriété.

Dans l'exemple (1), le fragment "the international telephone services" associé à ses deux représentations the [international telephone] services / the international [telephone services], n'est pas minimal car il peut être réduit au fragment "international telephone services", associé à ses deux représentations [international telephone] services / international [telephone services], qui est minimal.

Nous proposons la définition formelle suivante [Boitet, C., 1994] :

Une ambiguïté **A** de degré **n** (n≥2) relative à un système de représentation **R**, peut être formellement définie comme :

A = (U, F, <S_1, S_2, …, S_m>, <s_1, s_2, … , s_n), m≥n où :

- **U** est une phrase complète, appelée le contexte de l'ambiguïté.
- **F** est un fragment de **U**, habituellement, mais non nécessairement connexe, le ***support*** de l'ambiguïté.
- les **S_1, S_2, …, S_m** sont les représentations distinctes de **U** dans **R**, et les **s_1, s_2, … , s_n** leurs sous-parties représentant **F** telles que ∀i,j **; $s_i \neq s_j$**.
- Condition de minimalité :
  Soit **F'** un fragment de **U** strictement contenu dans **F**, et **s'$_1$**, **s'$_2$**, … , **s'$_n$** les parties respectives de **s_1, s_2, … , s_n** correspondant à **F'**. Il existe alors au moins une paire **s'$_i$, s'$_j$** (i≠j) telle que **s'$_i$=s'$_j$.**

Le type de l'ambiguïté **A** dépend de la différence qui caractérise les **s_i**. Il doit être défini relativement à chaque **R** particulier.

Figure 1 : Définition formelle d'une ambiguïté

3. LIDIA-1 : vers les DAE

3.1 LIDIA : un projet de TA Fondée sur le Dialogue

Les efforts passés visant à améliorer la qualité des traductions produites par des systèmes de TA ont montré que la TA de haute qualité est possible, mais seulement pour des typologies de textes (domaine, style) très contraintes. On peut citer, par exemple, les bulletins météorologiques (METEO, TAUM, anglais→français), les bulletins boursiers (ALT/Flash, NTT, japonais→anglais), ou les documents techniques (BV/aéro/FE pour les manuels de maintenance d'avions, Systran pour des documents XEROX en anglais contrôlé).

La Traduction Automatique Fondée sur le Dialogue de haute qualité est un nouveau paradigme pour des situations traductionnelles pour lesquelles les autres approches – fondées sur la langue, fondées sur la connaissance – ne sont pas appropriées [Boitet, C., *et al.*, 1995a]. En TAFD, bien que les sources de connaissances linguistiques soient encore cruciales, et que des connaissances extra-linguistiques puissent être utilisées si elles sont disponibles, l'emphase est mise sur la pré-édition indirecte au moyen d'un dialogue de désambiguïsation avec l'auteur afin d'obtenir des traductions de haute qualité sans révision.

La première situation que nous avons considérée est la production de documents multilingues sous la forme de documents HyperCard. HyperCard est un environnement de production de documents hypertextes dont les pages sont des

« cartes ». Les cartes contiennent différents types d'objets, dont des champs textuels. Du point de vue linguistique, nous utilisons une approche fondée sur un transfert multiniveau avec des acceptions, propriétés, et relations interlingues. Notre première maquette, LIDIA-1, démontre l'idée avec un document HyperCard qui présente, en contexte, des phrases ambiguës en français. Ce document peut être traduit vers l'anglais, l'allemand et le russe. Bien que cette maquette soit réduite du point de vue de sa couverture linguistique, elle montre bien le potentiel de l'approche.

3.2 LIDIA-1 : un premier démonstrateur

L'utilisateur peut activer les traitements LIDIA les plus fréquents grâce à une palette d'outils. La première ligne d'outils (figure 2), considérée de gauche à droite, permet de traduire l'objet sélectionné, et de voir la progression des traitements, les annotations, et la rétrotraduction en français. La seconde ligne permet de se déplacer parmi les cartes qui composent le document.

Après l'analyse, un bouton (**? !!** - figure 3) apparaît sur de l'objet à traduire si son contenu est ambigu et nécessite donc une désambiguïsation interactive.

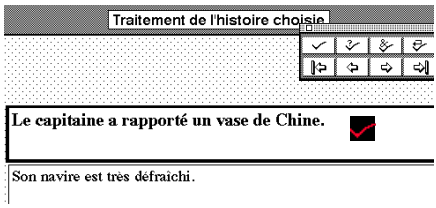


Figure 2 : sélection d'un champ textuel à traduire

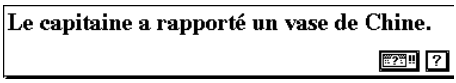


Figure 3 : Signalement de questions de désambiguïsation en suspens

Lorsqu'il décide de résoudre les ambiguïtés concernant un objet particulier, l'utilisateur clique sur ce bouton et les questions sont proposées comme ci-dessous, à l'aide de « rephrasages » simples.

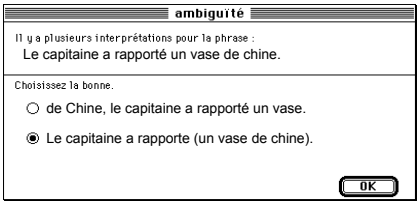


Figure 4 : Désambiguïsation structurale

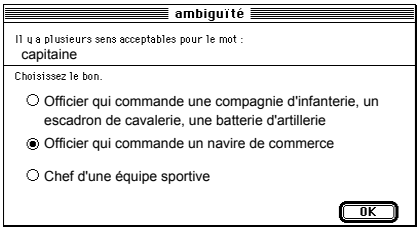


Figure 5 : Désambiguïsation de polysémie

La figure 6 montre la traduction de la phrase « le capitaine a rapporté un vase de Chine » dans deux contextes différents.

Figure 6 : Traduction en allemand d'une phrase dans deux contextes différents

3.3 Production d'un DAE dans le contexte de la TAFD

Le concept de DAE a été proposé et motivé dans [Boitet, C., 1994]. Nous donnons ici un bref aperçu (figure 8) des étapes de traitement mises en œuvre et des structures de données produites dans le cadre de notre architecture pour LIDIA-1. Nous montrerons aussi comment la production de DAE en langues source et cible s'y intègre.

Chaque phrase du texte en langue source est d'abord analysée pour produire une structure *mmc-source* (multisolution, multiniveau[4], concrète[5]). Cette structure *mmc* est alors utilisée pour construire un arbre des questions qui seront posées à l'auteur. À l'issue de l'étape de désambiguïsation interactive, le système obtient la structure *umc-source* (unisolution, multiniveau, concrète) non ambiguë choisie par l'auteur. Cette structure *umc* est ensuite abstraite en une structure *uma-source* (unisolution, multiniveau, abstraite[5]).

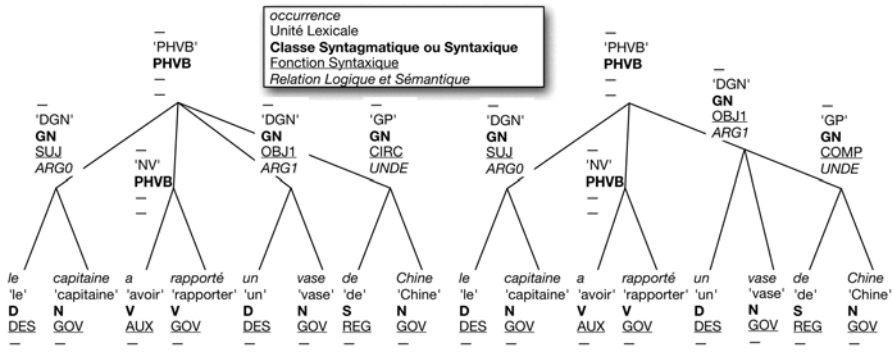


Figure 7 : Exemple de structure mmc

[4] La structure contient trois niveaux d'interprétation linguistique : le niveau des classes syntaxiques et syntagmatiques, le niveau des fonctions syntaxiques, et le niveau des relations logiques et sémantiques.

[5] Une représentation « concrète » d'un texte est telle qu'on retrouve le texte représenté grâce à un parcours canonique de la structure (mot des feuilles pour un arbre de constituants, parcours infixé pour un arbre de dépendances). Sinon, la structure est dite « abstraite ».

Un composant de transfert lexical et structural produit maintenant une structure *gma-cible* (génératrice, multiniveau, abstraite[5]). Une structure *gma* est plus générale et génératrice qu'une structure *uma* car les niveaux de surface (fonctions syntaxiques, catégories syntagmatiques…) peuvent ne pas être renseignés. Dans ce cas, ce sont des préférences du transfert qui les instancieront.

L'étape de sélection de paraphrase produit une structure *uma-cible* qui est homogène à la structure qui serait produite en analysant puis en désambiguïsant interactivement le texte cible qui va être généré. Le processus de traduction se termine avec les générations syntaxique et morphologique.

Lors des étapes de la traduction, ou de l'analyse uniquement, les informations nécessaires à la construction d'un DAE sont conservées. La figure 8 montre un diagramme fonctionnel des processus que nous venons de décrire.

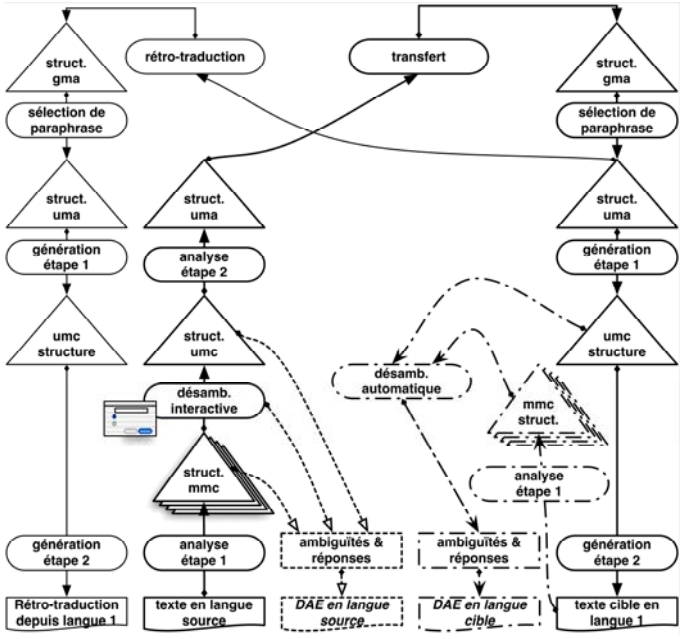


Figure 8 : Organisation linguicielle en TAFD et production de DAE

4. Production d'un DAE avec la maquette LIDIA-2

LIDIA-2 est une version en Java de l'environnement d'accès aux services de TADF. Dans cette version, nous avons utilisé un désambiguïseur de l'anglais [Blanchon, H., 1995] directement utilisable dans la nouvelle architecture d'intégration des composants que nous avons mise en œuvre.

4.1 Exemple de session

L'auteur personnalise d'abord son environnement. Il peut alors créer un nouveau document ou ouvrir un document existant. La fenêtre du document est divisée en deux sections : la partie supérieure est la fenêtre d'édition, la partie inférieure affiche des informations relatives à l'état du traitement du document.

Après que l'auteur a demandé l'analyse du document (figure 9), les phrases ambiguës sont colorées en brun et les phrases non ambiguës (comme la première phrase de la figure 9) en vert. On peut lire dans la figure 9 que le texte contient sept phrases ambiguës et une phrase qui ne l'est pas.

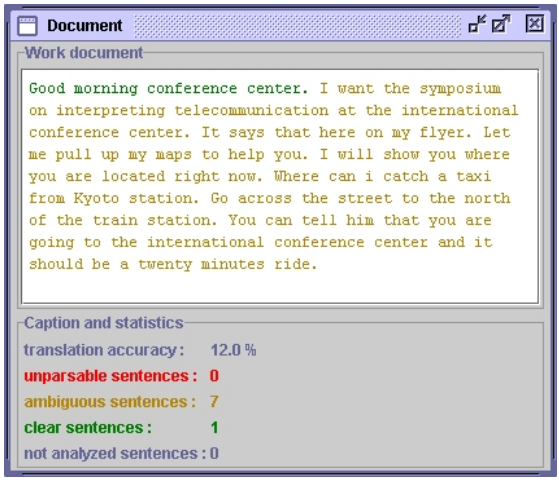


Figure 9 : Fenêtre de document LIDIA-2 (après analyse)

Lorsque l'auteur clique deux fois sur une phrase ambiguë, le dialogue de désambiguïsation relatif à cette phrase est activé. L'ordre des questions correspond à un parcours depuis la racine jusqu'à une feuille dans l'arbre des questions.

Par exemple, lorsque l'auteur choisit la phrase "I want the symposium on interpreting telecommunications at the international conference center", une première question (figure 10) lui est proposée.

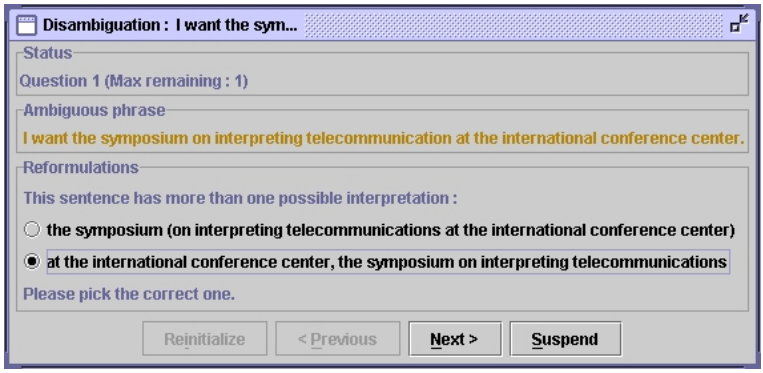


Figure 10 : Première question de désambiguïsation de la phase exemple[6]

La partie inférieure de la fenêtre lui montre qu'il répond à la première question et qu'il devra ensuite répondre à au plus une autre question. À tout moment, il peut aussi arrêter la session de désambiguïsation (bouton Suspend). Lorsqu'il a répondu à une question, il passe à la question suivante avec le bouton Next.

En cours de session (figure 11), l'auteur peut aussi revenir à la question précédente (bouton *Previous*), ou alors recommencer la session (bouton *Reinitialize*). Lorsque qu'il a répondu à toutes les questions (figure 12), l'utilisateur peut clore la session (bouton *Close*). On doit répondre à toutes les questions en une seule fois.

[6] La phrase concernée par le dialogue de désambiguïsation peut être comprise de deux façons : « je veux le symposium sur l'interprétation de télécommunications au centre de conférences international(es) » [un objet direct] ou « je veux le symposium sur l'interprétation de télécommunications *qui se déroule* **au centre de conférences international(es)** » [un objet direct et **un circonstant**].

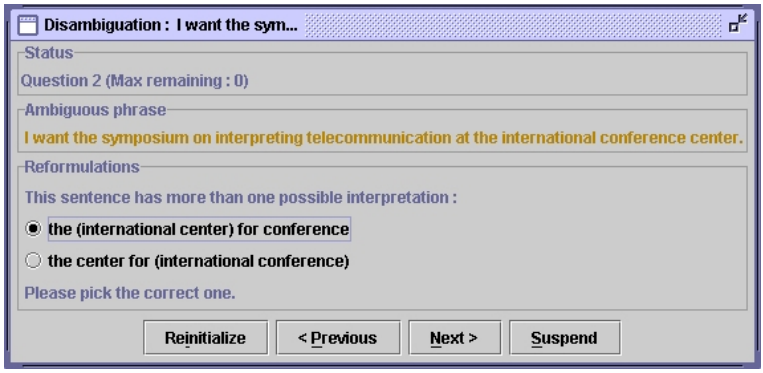


Figure 11 : Seconde question de désambiguïsation de la phrase exemple[7]

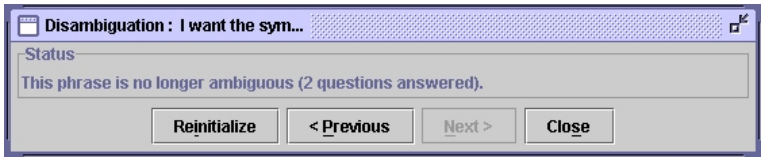


Figure 12 : Fin de la session de désambiguïsation de la phrase exemple

4.2 Document LIDIA-2

Le document XML produit par LIDIA-2 est manipulé par l'API DOM. Lorsque l'auteur ouvre un document existant, sa syntaxe est vérifiée avec l'API SAX.

Ce document contient une entête (**<description>**, figure 13) et un contenu (**<content>**, figure 14) — le texte — sous forme de paragraphes (**<paragraphe>**) et de phrases (**<phrase>**), entrées par l'utilisateur, qui sont enrichies par les informations collectées lors des différentes étape de traitement.

Pour chaque phrase du texte, le contenu comprend la langue source, le texte de la phrase, l'arbre des questions, ainsi que la ou les traductions obtenues dans les langues cibles choisies. L'arbre des questions est une représentation *à la Lisp* de l'arbre des questions produit par le module de désambiguïsation, enrichie par une trace du chemin suivi par l'utilisateur lors de la désambiguïsation effective.

[7] S'agit-il d' « un centre international de conférence(s) » ou d' « un centre de conférence(s) internationale(s) » ?

```
<description>
   <title><![CDATA[A trip to Tokyo]]></title>
   <language><![CDATA[ENG]]></language>
   <auteur>
     <firstname><![CDATA[herve]]></firstname>
     <lastname><![CDATA[blanchon]]></lastname>
   </auteur>
 </description>
```

Figure 13 : Descripteur d'un document LIDIA-2

```
<phrase source="ENG" stamp="51054803544695">
  <original><![CDATA[ I will show you where you are located right
now.]]></original>
  <question>
    <reformulation choix="NON"><![CDATA[I will show you (where you are
located right now).]]>
      <analyse><![CDATA[…]]></analyse>
    </reformulation>
    <reformulation choix="OUI"><![CDATA[right now, I will show you where you
are located.]]>
      <analyse><![CDATA[…]]></analyse>
    </reformulation>
  </question>
  <traduction cible="FRA"><![CDATA[Je vais tout de suite vous montrer où vous
êtes.]]></traduction>
</phrase>
```

Figure 14 : Extrait d'un fichier LIDIA-2 pour une phrase

4.3 Filtrage vers un DAE

Pour produire le DAE associé au document LIDIA-2 en cours, celui-ci est filtré. Le DAE conserve l'entête du document LIDIA-2. On conserve du contenu son organisation en paragraphes et phrases. Pour chaque phrase, on retient le texte d'origine et la trace du parcours de l'auteur dans l'arbre des questions.

5. Visualisation d'un DAE

Nous concevons un DAE comme un document autonome et « portable » qui doit pouvoir être diffusé sur PC et PDA.

5.1 Objectif et contraintes

Afin de permettre la lecture d'un DAE, nous proposons donc un visualiseur sous la forme d'une application indépendante.

Un tel visualiseur doit permettre à un lecteur de lire le contenu du document et d'appréhender le « sens exact » de que ce l'auteur a voulu dire. À cette fin, le lecteur doit être prévenu que certains segments peuvent avoir plusieurs interprétations (« sens »). Le visualiseur doit alors être capable de révéler, à la demande, le « sens » choisi par l'auteur lors de la phase de désambiguïsation interactive.

Nous avons implémenté, en Java, une première version d'un tel visualiseur.

5.2 Lecture active à l'aide du visualiseur

Le visualiseur permet au lecteur d'ouvrir un DAE dont le contenu textuel est alors affiché (figure 15).

À ce point, les segments ambigus ne sont pas surlignés (cf. section 6.1). Pour obtenir les informations relatives aux différentes interprétations possibles d'une phrase, l'utilisateur doit cliquer deux fois sur son texte. Une boîte de dialogue apparaît alors. Elle permet au lecteur de naviguer dans l'arbre des questions en voyant les rephrasages sélectionnés par l'auteur (**=> … <=**) lors de la désambiguïsation, comme le montre la figure 16.

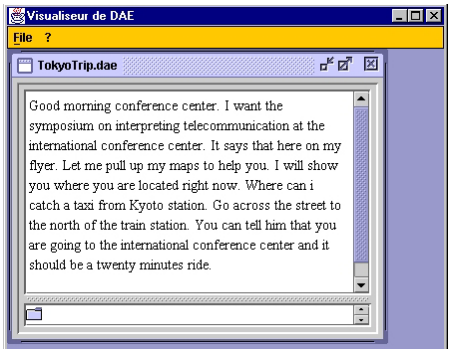


Figure 15 : Interface du visualiseur de DAE

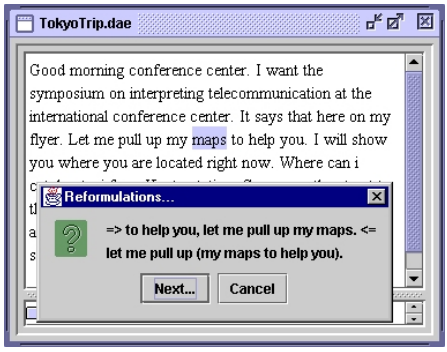


Figure 16 : Affichage de la lecture à retenir pour la phrase sélectionnée[8]

[8] Celui qui parle veut-il dire : « pour vous aider, laissez-moi vous présenter mes cartes » ou « laissez-moi vous présenter mes cartes conçues pour vous aider ».

5.3 Perspectives à court terme

Afin d'améliorer l'implémentation de LIDIA-2, nous avons plusieurs objectifs à court terme. Les perspectives à long terme sont détaillées dans la section 6.

Intégrer les linguiciels et le désambiguïseur du français

Notre premier objectif à court terme est d'intégrer dans la nouvelle architecture les modules d'analyse, de désambiguïsation interactive [Boitet, C., *et al.*, 1994, Boitet, C., *et al.*, 1995a], de transfert et de génération (vers l'anglais, l'allemand et le russe) développés pour la maquette LIDIA-1. Cela nous permettrait d'avoir une plate-forme d'expérimentation plus riche.

Rendre la désambiguïsation modifiable

Dans certains cas, il peut être intéressant de refaire la désambiguïsation interactive, soit pour corriger un résultat de traduction (la désambiguïsation interactive aurait dans ce cas été mal faite), soit pour produire une nouvelle traduction afin de montrer l'intérêt de la désambiguïsation.

Toutes les informations nécessaires sont déjà disponibles dans un document LIDIA-2. Ainsi, une nouvelle désambiguïsation pourrait être effectuée de manière autonome (hors ligne). Si le nouveau parcours de désambiguïsation est le même que le précédent, les bonnes traductions auront déjà été calculées (si l'on vise plusieurs langues cibles). Si le parcours est différent les traductions antérieures devront être écartées et de nouvelles traductions devront être produites (en ligne).

Créer automatiquement des corpus multilingues auto-explicatifs alignés

Comme dit plus haut, LIDIA-2 peut accepter des demandes de traduction vers plusieurs langues cibles. Les traductions sont conservées dans le document LIDIA-2.

Il pourrait donc être intéressant d'exporter un document multilingue déjà aligné au niveau de sa structure. On pourrait même envisager de conserver dans le document LIDIA-2 toutes les structures intermédiaires produites lors du processus de traduction pour calculer automatiquement différents alignements pour chaque phrase (au niveau des mots, des segments, des syntagmes).

Cela pourrait être utile, par exemple, en apprentissage des langues, et aussi pour l'étude contrastive des ambiguïtés. On sait également que les besoins de corpus alignés croissent avec le développement de moteurs statistiques pour le traitement de la langue naturelle, notamment en traduction.

6. Perspectives à long terme

Nos objectifs à long terme sont ceux qui ont un impact sur les modules de la chaîne de traduction (analyse, transfert, génération), et sur le module de désambiguïsation interactive.

6.1 Présenter le support de l'ambiguïté dans un DAE

Afin d'améliorer la présentation d'un DAE, il convient de localiser précisément les ambiguïtés en utilisant leur support (voir la définition au 2.3. Nous présentons brièvement la façon dont les ambiguïtés sont actuellement détectées.

Détection d'ambiguïtés dans les modules de DI actuels

Dans les modules actuels de préparation des arbres de questions, un type d'ambiguïté est décrit comme la cooccurrence, parmi les différentes solutions présentes dans la structure mmc, de différents schémas d'arbre — appelés aussi patrons — contenant des variables de nœud (**X, Y, Z, U, V**) ou de forêt (**p0, p1, p2, p3**). Ces différents schémas sont regroupés dans un faisceau comme dans la figure 17.

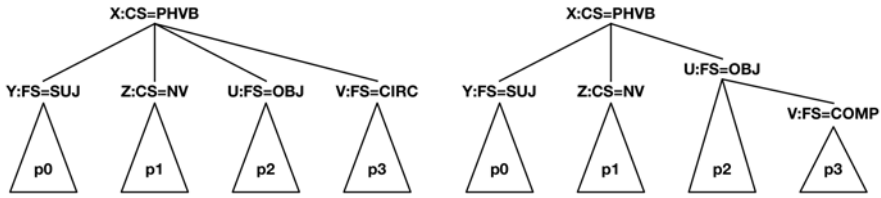


Figure 17 : Faisceau de description d'un type d'ambiguïté

Un item de question est produit pour chacun des patrons qui composent le faisceau sous forme de « rephrasage ». Chaque « rephrasage » est produit en faisant un certain nombre d'opérations sur les variables de forêt apparaissant dans le patron. Ainsi, pour le faisceau de la figure 17, les méthodes de rephrasage associées au deux patrons sont : Texte(**p2**)Virgule() Texte(**p0**) Texte(**p1**) Texte(**p2**) pour le patron de gauche et Texte(**p0**) Texte(**p1**) Parenthèse(Texte(**p2**), Texte(**p3**)) pour le patron de droite (cf. figure 4 pour une réalisation de ces méthodes de rephrasage).

Vers un processus de désambiguïsation interactive utilisant le support

Les patrons que nous utilisons actuellement n'utilisent pas la notion de support de l'ambiguïté. En effet, les patrons capturent très souvent un segment plus

grand que le support afin de permettre un rephrasage plus compréhensible de l'ambiguïté.

Pour qu'un DAE soit vraiment utile, il faudrait que les supports des ambiguïtés soient indiqués afin que le système puisse indiquer clairement les segments qui posent un problème d'interprétation. Il faut donc que le processus de désambiguïsation interactive fournisse cette information. Deux voies sont possibles.

Dans une première approche, on peut choisir d'attacher la description du support de l'ambiguïté à chaque faisceau. Le support peut, en effet, être défini à partir des variables utilisées dans les patrons.

La seconde approche, proposée dans [Boitet, C., *et al.*, 1995a], oblige à changer la description des ambiguïtés, en utilisant pour ce faire uniquement leur support. Cependant, pour faire des rephrasages du même type[9] que ceux que nous produisons actuellement, il faudrait décrire les informations supplémentaires à utiliser lors de la fabrication de ceux-ci.

Une première étude a été conduite dans ce sens. On propose de construire automatiquement les patrons sur le support à partir d'un corpus d'analyses multiples. Les informations manquant pour le rephrasage seraient retrouvées au moyen d'heuristiques [Irgadian, F., 2002].

6.2 Autoriser une désambiguïsation interactive incomplète

Dans le contexte de vraies applications, l'analyseur rencontrera un grand nombre d'ambiguïtés. Il est donc possible que, pour certaines phases, l'arbre des questions ait une telle profondeur que l'auteur n'accepte de répondre qu'aux questions cruciales.

Supposons, par exemple, qu'une phrase de longueur **N** possède $\mathbf{k^N}$ interprétations et que les descripteurs d'ambiguïté sont constitués en moyenne de **p** patrons. **(k/b).N** questions désambiguïseraient complètement cette phrase. Si par exemple **(k/b)=1/2**, il y aurait alors 120 questions pour une page de 240 mots. Bien qu'il ne faille pas plus de 10 minutes pour répondre à toutes ces questions[10], si chaque réponse prend 5 secondes, l'auteur peut vouloir consacrer moins de temps à la désambiguïsation interactive.

En d'autres termes, étant donnée une structure *mmc*, quelques réponses à des questions de désambiguïsation et, éventuellement, des préférences utilisateur, le système doit être capable de faire des choix et de produire une traduction unique ou alors une représentation factorisée explicitant les différentes traductions possibles en

[9] Nous avons montré [Blanchon, H., *et al.*, 1996, Blanchon, H., *et al.*, 1997] que les rephrasages actuellement produits permettent aux utilisateurs de distinguer clairement les différentes interprétations et de choisir la bonne.

[10] Ce temps doit être comparé aux mesures de temps de travail fournies par les traducteurs professionnels. Pour une page et pour chaque langue cible, il faut compter 1 heure pour produire une première traduction et 20 minutes de postédition.

langue cible. Afin d'implémenter une telle stratégie, il est nécessaire que les modules utilisés puissent mettre en œuvre des techniques heuristiques de désambiguïsation automatique ou soient capables de manipuler des structures ambiguës.

6.3 Certifier le sens

À partir du degré de complétion de la désambiguïsation d'une phrase, et en prenant en compte la « crucialité pour la traduction » des ambiguïtés non résolues, il est sans doute possible de calculer un « niveau de certification du sens » associé à la traduction, et de le calculer au niveau des paragraphes, des sections, etc., jusqu'au document lui-même.

6.4 Créer des DAE en langues cibles

Montrons maintenant comment l'architecture que nous proposons permet aussi de produire des DAE en langue cible. Comme l'étape de génération produit une structure intermédiaire équivalente à une structure d'analyse désambiguïsée (*umc*), il suffit de faire une analyse multiple (*mmc*) des phrases effectivement générées, puis de construire un arbre des questions concernant ces phrases.

Sachant que l'on connaît la structure *umc* à retenir, on peut calculer automatiquement les réponses aux questions de désambiguïsation : pondre à la place d'un lecteur en langue cible. On pourra donc produire un DAE en langue cible sans intervention humaine.

Atteindre cet objectif est cependant difficile en pratique puisqu'il faut disposer d'un analyseur multiple dans chacune des langues traitées (source et cibles). Nous espérons construire un prototype complet implémentant cette idée grâce à des coopérations internationales.

7. Conclusion

Nous avons montré une première implémentation du concept de document auto-explicatif. Cette idée se situe dans le champ de la recherche sur les documents actifs [Quint, V., *et al.*, 1994]. Nous travaillons sur un environnement LIDIA intégré à un éditeur de documents XML *à la Thot* (http://opera.inrialpes.fr/Thot.en.html).

Notre structure de DAE est assez simple car toute l'information contenue dans un tel document n'est pas encore au format XML. Par exemple, la structure *mmc* et l'arbre des questions sont représentés dans un formalisme *à la Lisp*, ce qui nécessite des modules de gestion spécifiques, alors qu'un traitement avec DOM serait plus efficace et portable.

Cependant, ces deux premières étapes (le nouvel environnement LIDIA-2 et le visualiseur de DAE) représentent des résultats originaux, et les perspectives de ce travail sont variées, tant au plan pratique qu'au plan théorique.

8. Références bibliographiques

[Black, E., *et al.*, 1993] Black, E., Garside, R. & Leech, G. (1993). *Statistically-Driven Grammars of English: the IBM/Lancaster Approach*. Rodopi. Amsterdam. 248 p.

[Blanchon, H., 1995] Blanchon, H. (1995). *An Interactive Disambiguation Module for English Natural Language Utterances*. Proc. NLPRS'95. Seoul, Korea. Dec 4-7, 1995. vol. 2/2: pp. 550-555.

[Blanchon, H., *et al.*, 1996] Blanchon, H. & Fais, L. (1996). *How to ask Users About What they Mean: Two Experiments & Results*. Proc. MIDDIM'96. Le col de porte, Isère, France. 12-14 Août 1996. vol. 1/1: pp. 238-259.

[Blanchon, H., *et al.*, 1997] Blanchon, H. & Fais, L. (1997). *Asking Users About What They Mean: Two Experiments & Results*. Proc. HCI'97. San Francisco, California. August 24-29, 1997. vol. 2/2: pp. 609-912.

[Boitet, C., 1990] Boitet, C. (1990). *Towards Personal MT : general design, dialogue structure, potential role of speech*. Proc. Coling-90. Helsinki. 20-25 Août 1990. vol. 3/3: pp. 30-35.

[Boitet, C., 1994] Boitet, C. (1994). *Dialogue-Based MT and self explaining documents as an alternative to MAHT and MT of controlled languages*. Proc. Machine Translation Ten Years On. Cranfield, England. Oct. 12-14, 1994. 7 p.

[Boitet, C., *et al.*, 1994] Boitet, C. & Blanchon, H. (1994). *Promesse et problèmes de la "TAO pour tous" après LIDIA-1, une première maquette*. in Langages, "Le traducteur et l'ordinateur"l. vol. 116, décembre 1994: pp. 20-47.

[Boitet, C., *et al.*, 1995a] Boitet, C. & Tomokiyo, M. (1995b). *Ambiguities & ambiguity labelling: towards ambiguity databases*. Proc. RANLP'95 (Recent Advances in NLP). Tzigov Chark, Bulgaria. 14-16 September, 1995. vol. 1/1: pp. 13-26.

[Fuchs, C., 1996] Fuchs, C. (1996). *Les ambiguïtés du français*. Ophris. Paris. 184 p.

[Irgadian, F., 2002] Irgadian, F. (2002). *Traduction Interactive Fondée sur le Dialogue. Rap. Université Stendhal*. Rapport de stage de Maîtrise en Industries de la Langue. 20 juin 2002. 109 p.

[Quint, V., *et al.*, 1994] Quint, V. & Vatton, I. (1994). *Making structured documents active*. In Electronic Publishing Origination, Dissemination, and Design. vol. 7(2): pp. 55-74.

Session 5

Indexation et recherche d'information

Analyses sémantiques pour la navigation textuelle

Eric Crestan[1,2], Claude de Loupy[1,3], Luc Manigot[1]

[1]*Sinequa, 51-54 rue Ledru Rollin, 92400 Ivry-sur-Seine - France*

{crestan,loupy,manigot}@sinequa.com

[2]*Laboratoire Informatique d'Avignon, BP 1228, Agroparc,*
339 chemin des Meinajaries, 84911 Avignon cedex 9 – France

[3]*Laboratoire MoDyCo – UMR 7114, Université de Paris 10,*
Bâtiment L, 200 avenue de la République, 92001 Nanterre cedex - France

Résumé :

Il devient de plus en plus clair que les performances pures des outils de recherche documentaire stagnent [HARM00]. Afin de passer au-delà de ce seuil, il faut replacer l'utilisateur au centre du processus de recherche. Les techniques présentées dans cet article abordent la question de l'aide à la navigation et l'interaction avec l'utilisateur. L'extraction et la mise en surbrillance des entités nommées dans les documents permettent une navigation intuitive et une lecture plus rapide des documents. L'extraction de concepts liés à une requête permet de désambiguïsation ou de palier une sous-spécification de celle-ci. La classification d'articles sur plusieurs ensembles de classes permet, par croisement de classes, un filtrage pertinent des documents réponses. L'apport pour l'utilisateur de telles techniques est clair mais il est très difficile de le chiffrer car les environnements d'évaluation type TREC, CLEF ou Amaryllis ne permettent pas d'évaluer le temps de recherche.

Mots-clés : Moteur de Recherche Sémantique, Concepts, Entités Nommées, Classification, Aide à la Navigation.

Abstract:

It is now clear that document retrieval systems have come to a stagnation point [HARM00]. In order to overcome this limitation, we need to put the user in the center of the search process. In this paper, we propose several techniques to tackle this problem through navigation help. Named entities extraction and highlighting enable an intuitive navigation and faster document reading. Users can filter returned documents thanks to automatically extracted concepts. A document classification also allows intuitive navigation and document filtering. The improvements of such techniques seem evident, but none of the proposed evaluation campaigns (TREC, CLEF or Amaryllis) takes into account the time a user needs to succeed in her/his research.

Keywords : Semantic Search Engine, Concepts, Named Entities, Classification, Navigation Help.

1. Introduction

Il devient de plus en plus clair que les performances pures des outils de recherche documentaire stagnent [HARM00]. Les travaux sur les heuristiques de rapprochement requête/documents ainsi que sur l'apport d'analyses linguistiques ne montrent pas de progrès révolutionnaire depuis déjà un certain temps. Pourtant, les utilisateurs ne sont toujours pas satisfaits des performances des outils actuels. Il faut donc les faire intervenir de manière plus active dans le processus de recherche.

Deux moyens peuvent être utilisés pour cela : l'interaction et l'aide à la navigation (ce dernier faisant plus ou moins intervenir le premier). Dans cet article, nous nous intéressons plus précisément à l'aide à la navigation. L'outil que nous présentons est basé sur des analyses syntaxiques et sémantiques permettant à l'utilisateur d'appréhender plus rapidement les documents répondant à sa requête. Plutôt que de parcourir la liste des centaines de documents qui répondent, il peut commencer à filtrer en indiquant des « concepts » ou des entités nommées (lieux, personnes, etc.) qui l'intéressent et qui sont extraits de manière automatique par le système. Ensuite, l'utilisateur est aidé dans sa compréhension rapide des documents par un surlignage des entités jugées importantes (personnes, lieux, entreprises, dates, etc.).

Cet article s'articule autour de 6 sections. L'introduction est donnée dans cette section. Le moteur d'indexation et de recherche sémantique *Intuition* de Sinequa est décrit dans la section 2. Dans la section 3, les entités nommées sont utilisées pour augmenter les possibilités de navigation et pour la visualisation des articles de presse. Une approche par extraction de concepts est présentée en section 4. Puis, l'intérêt de la classification a priori de document pour la navigation est présenté en section 5. Enfin nous conclurons en section 6.

2. Moteur de recherche sémantique

Les systèmes de recherche documentaire, actuellement les plus performants, utilisent tous des statistiques sur la fréquence d'occurrence des mots dans les textes. Cela s'explique par le fait qu'ils doivent non seulement être capables de traiter une quantité importante de documents, mais aussi de pouvoir traiter des textes provenant de différents domaines. Pour autant, les moteurs de recherche ne sont pas tous identiques. Il existe différentes méthodes pour aborder les documents, en faisant notamment intervenir des traitements linguistiques dans le processus.

Le moteur de recherche *Intuition* de Sinequa aborde les documents sur différents points de vue. Il combine à la fois une recherche classique sur les mots, ainsi qu'une recherche basée sur un niveau plus sémantique.

2.1 Recherche sur les mots

Plusieurs éléments sont à prendre en compte en vue d'indexer des documents. L'un des aspects les plus importants concerne la langue qui sera traitée par le moteur. En effet, les langues, bien que se ressemblant parfois, peuvent demander des traitements bien différents. Aucun moteur ne peut se prévaloir d'être totalement indépendant de la langue, car des langues comme le thaï requièrent un minimum de travail au niveau de la segmentation (l'écriture thaï est non-flexionnelle mais ne comporte pas d'espace entre les mots). Il existe néanmoins trois grands types de traitements possibles sur les mots. Le premier est en fait un « non-traitement ». C'est entre autre la méthode utilisée par le moteur de recherche Internet *Google*, qui indexe les formes fléchies des mots. Ainsi, un utilisateur cherchant « *automobile* », ne trouvera pas de documents contenant seulement le mot *automobiles* au pluriel. Cela peut provoquer du silence sur le questionnement. Le second traitement, qui a été largement utilisé dans les dix dernières années, est l'utilisation d'un « raciniseur » (stemmer). Ce processus consiste à tronquer les mots pour les ramener à une forme dite racine. Ainsi, les mots *déménageur*, *déménagera* et *déménagement* seront tronqués à la forme racine *déménag*. Cela permet de résoudre le problème de silence cité plus haut, mais fait apparaître, dans certains cas, un nouveau problème lié au bruit. Par exemple, le substantif *portes* et le verbe *portera* serons ramenés à la même racine *port*, ce qui peut causer des problèmes de bruit lorsqu'un utilisateur recherche des documents parlant de *portes en bois*. La dernière technique consiste à ramener tous les mots à leur forme de base en effectuant une lemmatisation. C'est notamment le traitement utilisé par le moteur d'indexation et de recherche *Intuition*. Dans l'exemple précédent, le substantif *portes* sera ramené à sa forme de base *porte*, alors que le verbe *portera* sera ramené à sa forme infinitive *porter*. En outre, dans cet exemple, une levée d'ambiguïté est effectuée grâce à une analyse syntaxique du contexte permettant de déterminer que le mot *portes* n'est pas une forme conjuguée du verbe *porter*. Toutefois, bien que le problème de bruit soit diminué, nous nous retrouvons à nouveau avec un problème de silence dans le cas de *déménageur*,

déménagera et *déménagement*. Pour palier cette difficulté, nous utilisons, en plus des traitements linguistiques, des dictionnaires de dérivation et de synonymie.

Lors de l'indexation par *Intuition*, les documents sont transformés en un vecteur de mots (formes de base) dans lequel le poids attribué à chaque mot dépend de sa fréquence d'occurrence dans le document (TF) et le nombre de documents dans lesquels il apparaît (IDF). Le rapprochement d'un document avec une requête est fait en calculant l'angle entre le vecteur du document et le vecteur de la requête. Une des méthodes de calcul d'angle les plus connues à l'heure actuelle est celle du *Cosinus* proposée par [SALT83]. *Intuition* utilise une autre mesure du fait des faibles performances de *Cosine* pour des requêtes courtes [WILK95].

Malgré tout cela, une indexation par les mots seuls est parfois insuffisante. Par exemple, pour une requête « *crise économique en France* », plusieurs documents traitant du sujet peuvent être oubliés car ils ne citent pas les mots de la requête ou un de leurs synonymes. C'est le cas pour le terme *crise économique* qui est une notion qui doit être traitée sur un niveau sémantique plutôt qu'au niveau des mots.

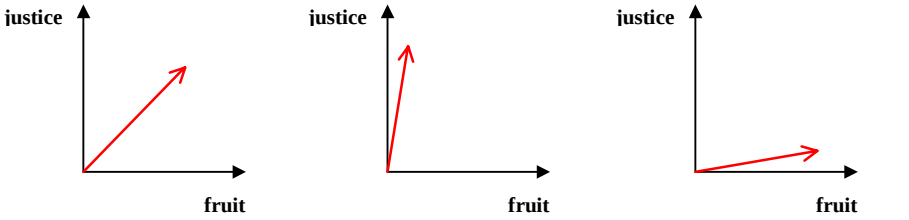
2.2 Recherche sémantique

Outre la recherche classique par les mots, le moteur de recherche *Intuition* utilise en parallèle un second mode de recherche basé sur l'analyse des sens. Le but de cette recherche est, dans un premier temps, de favoriser des documents dont la sémantique globale est proche de celle de la requête, mais aussi de palier le problème de silence.

A la fin des années 90, Sinequa a développé un dictionnaire basé sur une approche thématique de l'univers [MANI97]. L'idée consiste à répartir « *l'univers des mots* » sur un espace avec un nombre de dimension fixe. Environ 800 dimensions (ou sacs de mots) ont été utilisées pour classer des mots (ou termes) de la langue française. Le choix de ces 800 dimensions a été réalisé pour assurer une proximité sémantique moyenne dans le cadre d'une recherche d'information. Des milliers de dimensions distinctes n'aurait pas apportés une souplesse suffisante, et le choix de quelques dizaines aurait induit des approximations trop brutales. Ces 800 dimensions ne reposent donc pas sur une ontologie standard, mais représentent une couverture suffisante des langues traitées pour le besoin exprimé. Dans ce dictionnaire, un mot peut appartenir à plusieurs dimensions à la fois. Par exemple, le mot avocat appartiendra à la dimension *justice/juridique*, mais également à la dimension *fuit/aliment*. En parallèle de l'indexation sur les mots des documents, chaque document est converti en un vecteur sémantique à 800 composantes. Le poids attribué à chaque dimension dépend principalement du nombre de termes trouvés dans le document lui appartenant. Une désambiguïsation locale est effectuée pour renforcer le poids attribué à une dimension lorsqu'un mot appartient à plusieurs d'entre elles.

Comme pour la recherche sur les mots, les requêtes sont également converties sous forme vectorielle, puis l'angle est calculé entre le vecteur sémantique de la requête et celui des documents. L'avantage de ce changement d'espace est que la

cooccurrence, dans la requête, de mots appartenant à une même thématique, renforce le poids de la dimension correspondant à cette thématique. Ainsi, dans l'exemple présenté en figure 1, il n'est pas possible dans le cas de la requête 1 de lever l'ambiguïté sur le mot *avocat*. Par contre, la cooccurrence des mots *avocat* et *Cour* dans la requête 2, permet de lever l'ambiguïté sur le premier. De même pour la requête 3, les mots *récolte* et *avocat* ont un descripteur en commun qui est fruit, ce qui permet de lever l'ambiguïté sémantique de la requête et de favoriser les documents parlant de fruits.



Requête 1 : *avocat* Requête 2 : *avocat à la Cour* Requête 3 : *récolte des avocats*

Figure 1: Modèle vectoriel sémantique

Toutefois, il est à noter que dans le cas de la requête 3, l'utilisateur aurait pu vouloir des documents parlent de la « récolte des honneurs des avocats », mais nous conviendrons que la requête serait dès lors mal posée, car impossible à désambiguïser même par un être humain.

Le score de pertinence globale pour un document, par rapport à une requête, est calculé par combinaison linéaire entre le score sur les mots et le score sur les sens.

3. Entités Nommées

L'approche sémantique des documents a plusieurs facettes. Les entités nommées sont une autre manière de voir la sémantique (documentaire). En effet, le fait de savoir que, dans un contexte particulier, *Charles de Gaulle* est un nom propre de personne et non pas un porte-avions ou un aéroport, permet une restriction de la recherche. De plus, la visualisation par surlignage des différents types d'entité détectés permet une lecture plus rapide et conviviale des textes. Dans le cadre de

l'application Diva-Press[1], une extraction d'entités nommées a été appliquée à des publications quotidiennes généralistes ou financières.

3.1 Extraction des entités nommées

Les systèmes d'extraction d'entités nommées sont principalement basés sur deux approches. La plus ancienne consiste à manuellement définir des règles linguistiques pour la détection de chaque type d'entités. Nous pouvons notamment citer les systèmes UNO [IWAN95] et LOLITA [MORG95] qui ont été utilisés dans le cadre de la campagne d'évaluation MUC-6 [GRIS95]. Cette approche est coûteuse en temps de développement, mais donne globalement de très bons résultats. La seconde approche repose sur l'apprentissage de modèle et les statistiques. Les techniques employées sont entre autres les Modèles de Markov Cachés comme pour le système NYMBLE [BIKE97] ou l'apprentissage automatique de règles comme c'est le cas du système ALEMBIC [ABER95]. L'inconvénient d'une telle méthode est qu'il faut disposer d'un corpus d'apprentissage pour entraîner les modèles.

Un système à base de règles linguistiques, manuellement définies, a été implémenté dans le cadre de ce travail. Il est basé sur une cascade de transducteurs s'appuyant sur des lexiques spécifiques. La définition des règles avait déjà été faite lors de notre participation à la campagne d'évaluation des moteurs de questions-réponses TREC-11 [VOOR00], mais celle-ci portait sur l'anglais. De nouveaux transducteurs ont été réalisés pour la détection des entités des types suivants:

- Noms de personnes (*Jacques Chirac, George W. Bush, Messier…*),
- Sociétés/Organisations (*Canal +, Organisation des Nations Unies, Dupont Corp…*),
- Lieux (*Paris, Allemagne, Rhône-Alpes…*),
- Temporel (*12 décembre 99, samedi soir, 1997…*),
- Chiffres (*12 %, 30 K€, 30 milliards de dollars…*).

Deux modes de détection ont été mis en place pour la reconnaissance des sociétés. Tout d'abord, un premier transducteur détecte les sociétés uniquement sur lexique. Cela permet d'avoir une réussite quasi parfaite sur celles-ci. Puis, un second transducteur détecte les sociétés restantes par des règles contextuelles. Cela permet de proposer de nouvelles sociétés à ajouter au lexique.

Un traitement complémentaire, dit de « normalisation », est appliqué lors de la reconnaissance des noms de personnes. Il est destiné à trouver les occurrences des patronymes qui auraient pu échapper à des règles linguistiques. Pour chaque nom détecté par le transducteur, le patronyme est extrait du nom et ensuite recherché dans le reste du document.

[1] Diva-Press est un spécialiste de la diffusion électronique de la presse économique et financière française. http://www.diva-press.com/

L'exemple[2] de la figure 2 montre un cas de normalisation de noms de personnes. Les noms *Bush* et *Blair*, dans le titre, ne peuvent être détectés grâce au contexte. En effet, le verbe *déstabiliser* ne prend pas toujours un humain en temps qu'objet direct. Le fait d'avoir détecté les occurrences de *George W. Bush* et de *Tony Blair* dans le corps de l'article, permet d'utiliser cette connaissance pour détecter les occurrences des deux patronymes dans le titre.

Cette méthode est également utilisée pour différencier un nom de société et de personne qui seraient homographes. Par exemple le groupe *Lagardère* sera sans doute détecté en temps que société grâce au contexte, mais le simple mot *Lagardère* sera identifié en tant que nom de personne grâce à la présence à d'autres endroits de l'article du nom complet *Jean-Luc Lagardère*. Dans certains cas, l'ambiguïté ne peut être levée et des choix sont à effectuer.

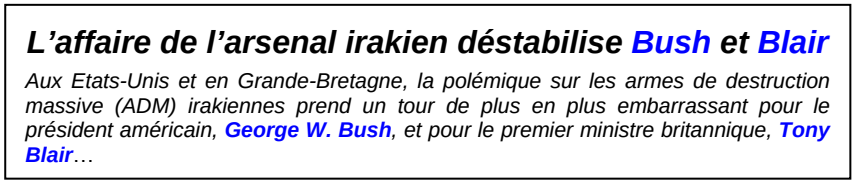


Figure 2 : Exemple de normalisation pour les noms de personne

3.2 Visualisation

La mise en surbrillance des entités nommées ne date pas d'hier. Il a plus 2 000 ans de cela, en Egypte, les titres des hauts dignitaires étaient écrits avec de l'encre rouge alors que le reste du texte était en noir. D'autre part, dans le livre « *De Geographia* », écrit aux environs de 1300, le nom des pays et régions était également écrit en rouge. Au-delà de l'aspect esthétique de la chose, le surlignage des entités nommées permet un survol rapide des documents et facilite le repérage des paragraphes et des sujets traités.

Ainsi, le surlignage des sociétés permet de voir d'un coup d'œil que le document (figure 3) parle de la compagnie *Shell*, et cite notamment les compagnies *Exxon Mobil* et *BP*. De plus, il apparaît immédiatement que *Philip Watts* est la seule personne mentionnée dans cet article.

[2] Echantillon d'un article du journal Le Monde en date du 07.02.2004

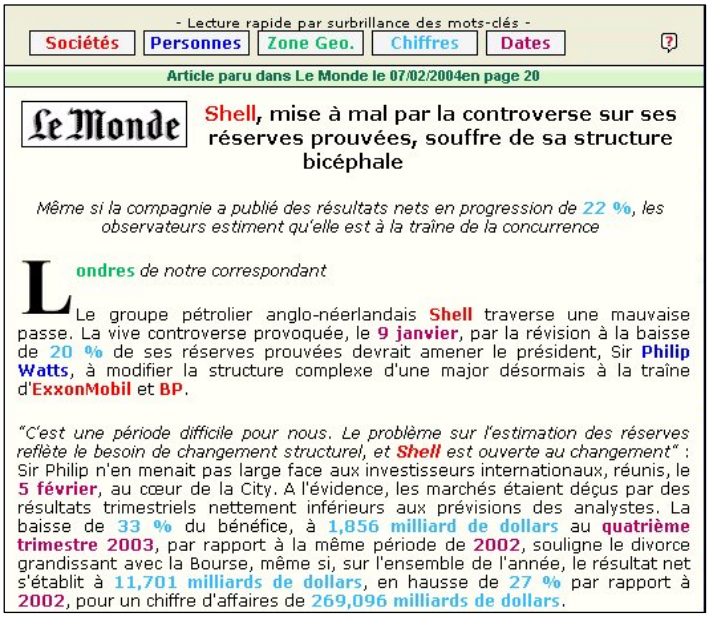


Figure 3: Article du journal Le Monde *issu de Diva-Press*

3.3 Navigation

Les entités détectées dans les documents peuvent également être utilisées pour guider l'utilisateur et accélérer sa recherche d'information. A chaque document correspond une liste d'entités de différents types. Lors de l'indexation, les entités sont déversées dans des « colonnes » (ou champs) spécifiques de la base *Intuition*. Ainsi, dans le cas de Diva-Press, une colonne existe pour chaque type d'entités. Ces colonnes, une fois remplies, peuvent être utilisées pour effectuer des statistiques par rapport à une requête. Pour une requête donnée, les entités nommées de type *personne* et *société* sont sommées sur la liste des réponses retournées par le moteur. Les deux listes d'entités présentées dans les tableaux 1a et 1b, ont été extraites sur la requête « *licenciement* ».

Score	Sociétés
7 %	Medef
3 %	SNCF
3 %	Boeing
2 %	Air Littoral
2 %	Alcatel
2 %	Danone

Score	Personnalités
5%	François Fillon
3%	Michel de Virville
3%	Jean-Pierre Raffarin
1%	Igor Landau
1%	Jean Marimbert
1%	Gerhard Schröder

Tableau 1: Sociétés (1a) et Personnalités (1b) extraites sur la requête « licenciement *»*

Les scores correspondent au taux de représentativité d'une entité par rapport à l'ensemble des entités de même type. Ainsi, le *Medef* représente 7 % des entités de type *société* et *François Fillon* représente 5 % des entités de type *personne* parmi les réponses a cette requête.

L'intérêt de cette approche est évident : elle permet à l'utilisateur de spécifier sa requête en restreignant la liste de documents réponses à un sous-ensemble, défini par la présence de l'entité choisie. Cela permet également à une personne novice du domaine, de rapidement acquérir des connaissances sur celui-ci. En effet, de par son poste de Ministre des Affaires sociales, *François Fillon* est étroitement lié à une thématique de travail dans laquelle se trouve le substantif *licenciement*.

4. Conceptualisation

L'utilisation des entités nommées peut n'être d'aucune utilité dans certains cas, ou ne pas correspondre aux attentes de l'utilisateur. Pour remédier à ce problème, une extraction de « *concepts* » clés a été mise en place. Le terme *concept* désigne en fait un groupe nominal minimal, composé de 1 à 3 mots, qui peut être assimilé dans la plupart des cas à des unités de sens. Les groupes nominaux ne contenant qu'un seul mot sont en fait des abréviations ou des noms propres. Ils sont détectés lors de l'indexation des documents par un automate tel celui présenté en figure 4.

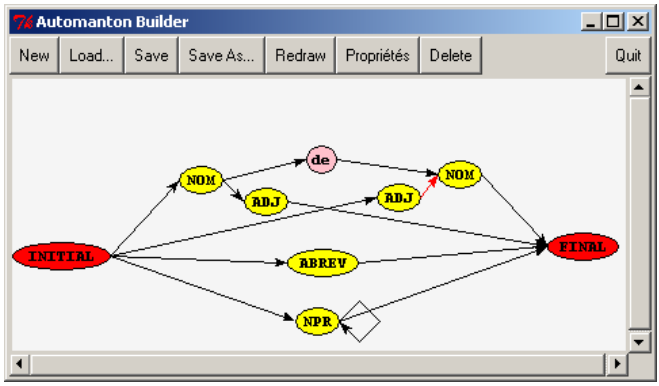


Figure 4 : Automate minimal pour la détection des concepts

Seul les séquences simples, de type *NOM ADJ* ou *NOM de NOM*, sont prises en compte par l'automate. Cela permet, d'une part, d'accroître leur fréquence d'occurrence dans le corpus et d'autre part, d'avoir des séquences assez longues pour lever l'ambiguïté sémantique sur les mots qui les composent. Les seuls mots simples qui sont acceptés en tant que concepts sont les abréviations et les noms

propres qui ne sont généralement pas ambigus. Les mots inconnus, donc non présents dans le dictionnaire, sont également considérés comme des noms propres. Cela peut parfois conduire à des erreurs, mais le mode de calcul de pertinence des concepts (voir plus loin) permet de limiter ce problème.

Les concepts sont ensuite stockés, pour chaque document, dans une colonne dédiée. Le calcul de pertinence de chaque concept, par rapport à une requête, se fait en comparant sa fréquence, dans la liste des 1 000 premiers documents rapportés par le moteur, avec sa fréquence dans l'ensemble de la base. Nous utilisons pour cela une méthode de calcul de spécificité en considérant les concepts comme des termes.

Concepts
licenciement économique
procédures de licenciement
plans de licenciements
plans sociaux
contrats de travail
prud'hommes
motif économique
indemnités de licenciement
CDD
lettre de licenciement

Tableau 2 : Concepts liés à la requête licenciement

La liste de concepts présentée dans le tableau 2 est triée par pertinence. Elle est donnée pour la requête *licenciement* sur la base de Diva-Press. Le mot *licenciement* revient souvent parmi les concepts, ce qui est normal par rapport à la requête. Toutefois, d'autres concepts ne contiennent pas ce substantif, mais sont tout de même représentés de par leur importance dans le sous-corpus défini par la requête. Ils sont tous sémantiquement liés à la requête et non ambigus. Si les termes simples avaient également été sélectionnés, le nom *plan* aurait probablement été détecté comme pertinent par rapport à la requête. Cependant, ce nom est très polysémique lorsqu'il est pris seul. La connaissance de la requête qui a généré ce concept, aurait permis de lever l'ambiguïté sur ce mot, mais dans certains cas, la requête est elle-même ambiguë et ne permet pas de lever cette ambiguïté. Prenons un autre exemple de requête avec le substantif *barrage*[3]. Ce mot comporte plusieurs ambiguïtés sémantiques qui ont pu échapper à l'utilisateur alors qu'il saisissait sa requête. Le tableau 3 donne les concepts liés à cette requête.

[3] Exemple classique d'ambiguïté sémantique en français, notamment utilisé dans les travaux de Veronis [VERO03].

Nous pouvons clairement distinguer que ces concepts ne sont pas tous liés à une même thématique. Ainsi, les concepts *tirs de barrage*, *matches de barrages* et *barrages retour* sont explicitement liés à une thématique sportive. Alors que, *barrage routier* et *construction de barrages* sont quant à eux liés aux domaines du transport pour le premier et des constructions hydrauliques pour le second. Cela peut être assimilé à une méthode de désambiguïsation sémantique en contexte et guider l'utilisateur dans sa recherche.

Concepts
tirs de barrage
matches de barrages
barrage routier
Golbey Épinal
Stade Clermontois
barrage militaire
Rueil
construction de barrages
Maurienne
barrages retour

Tableau 3 : Concepts liés à la requête barrage

5. Classification pour la navigation thématique

L'homme a depuis toujours essayé de classer les éléments du monde qui l'entoure. L'un des précurseurs du domaine fut Aristote qui proposa, dès l'antiquité, un modèle de classification pour les sciences. Pour ce qui est de la classification automatique de textes, les premières approches remontent aux années 60 et étaient principalement basées sur l'ingénierie des connaissances. Depuis le début des années 90, les approches basées sur l'apprentissage sont devenues les plus populaires au sein de la communauté. Un état de l'art du domaine est disponible dans [SEBA02].

La classification de documents, selon des classes préalablement définies, permet un accès plus rapide aux documents escomptés et limite le risque de bruit. La sélection des classes est toutefois un exercice à la fois difficile et primordial pour cette tâche. Dans le cadre de Diva-Press, les documents sont classés en même temps selon 4 ensembles de classes qui sont :

- **Thèmes** :

Ce sont des sujets de société d'ordre général (ex : *décentralisation, consommation, places financières, droit des sociétés…*),

- **Secteurs** :
  Ce sont les secteurs d'activité abordés dans les articles (ex : *secteur automobile, secteur de la santé, assurances, gestion d'entreprise…*),
- **Sociétés** :
  Ce sont des sociétés préalablement définies dans un thésaurus (ex : *Air Liquide, Peugeot SA, Vivendi Environnement…*),
- **Zones géographiques** :
  Comme pour les sociétés, les zones géographiques ou politiques sont définies dans un thésaurus (ex : *Amérique centrale, France, Union Européenne, G7…*).

Nous observons immédiatement que ces ensembles ne demandent pas le même traitement. En effet, la classification par sociétés et zones géographiques ne revient quasiment qu'à détecter la présence de l'entité du type attendu dans les documents, alors que, dans le cas de la classification par thèmes et secteurs, la classification se fait grâce à un vocabulaire beaucoup plus important. Nous nous attacherons à présenter les deux techniques employées pour la classification automatique dans les sections suivantes.

5.1 Classification par sociétés et zones géographiques

La classification par sociétés et zones géographiques ne consiste pas seulement à classer les documents selon l'apparition des noms de sociétés ou de pays. Lorsqu'une société est citée une seule fois dans un document, il n'est pas forcément pertinent d'indexer ce document sur cette société. C'est le cas notamment de la société *Club Méditerranée* dans cet extrait d'article.

…

*Ces actions terroristes, qui n'avaient pas fait de victimes mais occasionné d'importants dégâts matériels, avaient été revendiquées par les clandestins du Canal historique du Front de libération nationale de la Corse (**FLNC**), dirigé dans le nord de l'île par Charles Pieri. « Gilbert Trigano – **l'ancien PDG du Club Méditerranée** – m'avait donné le conseil de ne pas trop me développer en Corse », s'est souvenu, devant le juge, le 20 janvier, M. Maillot.. A l'époque, la revendication du **FLNC** ne l'avait guère surpris : « Je n'ai pas été étonné de ces revendications, a-t-il indiqué à Philippe Courroye. Après ces attentats, j'ai été contacté par le **Sporting Club de Bastia** ».*

…

Figure 5 : Extrait d'un article du journal Le Monde

Pour répondre à cette problématique, une heuristique a été mise en place prenant en compte plusieurs facteurs comme la fréquence d'occurrence des entités et leur position dans le document. Une société dont le nom apparaît dans un titre ou en début d'article a plus d'importance qu'une société n'apparaissant qu'une fois en fin d'article. Une société ne sera retenue que lorsque son score dépasse un certain seuil, qui est proportionnel à la taille du document. Le taux de variabilité des noms de sociétés est un facteur important à prendre en compte. Pour résoudre ce problème, il est nécessaire de gérer les variantes de nom comme dans le cas de *Peugeot* et *Peugeot SA*. Quant aux zones géographiques, il est nécessaire d'utiliser un thésaurus pour rattacher les villes à leur(s) zones respective(s). Ainsi, si un document contient les villes *Paris*, *Berlin* et *Rome*, le document sera indexé sur *Union Européenne*.

L'avantage de cette technique de classification est qu'elle ne requiert pas de corpus d'apprentissage, mais seulement un corpus de développement pour mettre au point le système des pondérations. De plus, elle profite pleinement de la détection d'entités effectuée lors de l'indexation des documents.

5.2 Classification par thèmes et secteurs

La classification automatique par thèmes et secteurs ne peux pas employer la même technique que pour les sociétés et les zones géographiques. Un thème ne peut être détecté par la seule présence de son libellé dans un document. Comme c'est le cas dans la collection *Reuters* [LEWI97], l'application Diva-Press demande de classer les documents selon plusieurs centaines de thèmes et de secteurs. De plus, un document peut appartenir à un nombre de thèmes et de secteurs allant de 0 à N. La différence majeure dans cette application, c'est qu'elle n'est qu'une aide à la classification manuelle. Une fois les documents catégorisés, des annotateurs vérifient les propositions du système et les modifient le cas échéant. Cela a pour avantage d'accroître la taille du corpus d'apprentissage et d'avoir un retour sur les performances du système.

L'approche utilisée pour cette application est celles des N Plus Proches Voisins (NPPV) [COVE67]. Bien que n'étant pas la meilleure méthode recensée dans l'état de l'art, elle a l'avantage d'être directement exploitable par le biais du moteur *Intuition*. Cette technique consiste à trouver les K documents, qui ont déjà été catégorisés, les plus proches d'un document d que l'on veut classifier. Pour chaque nouveau document de la base, un calcul de score de proximité est utilisé pour ramener les 100 documents les plus similaires. Cette méthode est apparentée à celle utilisée pour le rapprochement des requêtes avec les documents réponses. Elle est basée sur un calcul d'angle entre les vecteurs du nouveau document et ceux des autres documents de la base déjà catégorisés. L'approche sémantique présentée en section 2.2 est également employée pour augmenter le score des documents liés sémantiquement. Le score pour chaque thème (ou secteur) v est calculé en sommant, sur les N premiers documents rapportés, le rapport du score de chaque document ϑ avec son rang j (voir équation 5.1).

$$score(v) = \frac{\delta_v}{\delta} \times 100 \qquad (5.1)$$

où $\delta_v = \sum_{j=1}^{N} \sum_{i=1}^{M_j} \frac{\vartheta^{\alpha}}{j^{\beta}} \times \sigma(v)$ et $v_j^i = v \rightarrow \sigma(v) = 1$; $v_j^i \neq v \rightarrow \sigma(v) = 0$

La puissance α joue donc sur l'importance donnée au score de pertinence du document, alors que la puissance β influe sur le score relatif au rang du document. La variable M_j correspond au nombre de classes proposées pour le document situé au rang j. Etant donné que les documents proches sont les mêmes pour les deux classifications, le temps de traitement est quasiment nul pour la seconde classification.

Les tests effectués lors des phases préliminaires nous ont permis d'obtenir des scores avoisinant les 70 % de précision pour les thèmes et les secteurs. Néanmoins, la grande quantité de classes, ainsi que le fait d'avoir une quadruple classification, oblige les annotateurs à vérifier la totalité des documents car la qualité est primordiale pour leur besoin.

La classification des articles selon plusieurs ensembles de classes permet d'effectuer des requêtes complexe combinant requête plein-texte avec un filtrage par thèmes, secteurs, sociétés et/ou zones géographiques. De plus, la diversité des classes permet une grande couverture, que ce soit au niveau des entités nommées ou au niveau de thèmes et secteurs d'activité. Les articles classés peuvent également être utilisés pour extraire les concepts liés à une classe particulière. Par exemple, lorsque l'on sélectionne uniquement les documents appartenant au *secteur automobile*, on obtient la liste de concepts du tableau 4.

Concepts liés
constructeurs automobiles
Denis Fainsilber
GM
marché automobile
Volkswagen
TOYOTA
Ford
Nissan
Chrysler
Renault

Tableau 4 : Concepts liés au secteur automobile

Nous retrouvons dans cette liste les noms des principaux constructeurs automobiles, ainsi que le nom de *Denis Fainsilber*, qui est le spécialiste des transports au journal *Les Echos*.

6. Conclusion

Les possibilités d'amélioration des performances pures des moteurs de recherche semblent réduites depuis la mise en œuvre de techniques d'indexation combinant les analyses statistiques, syntaxiques et sémantiques des documents. Toutefois, le repositionnement des utilisateurs au centre du processus de recherche semble ouvrir de nouvelles perspectives. La détection et visualisation des entités nommées permettent une lecture plus rapide et intuitive des articles de journaux. Les statistiques sont, dès lors, d'une grande utilité pour extraire la distribution de chaque type d'entités par rapport à une requête. Grâce à la liste des concepts, une sous-spécification dans une requête ne nécessite plus forcement une re-formulation de celle-ci. Les concepts peuvent être également utilisés comme moyen de désambiguïsation sémantique des requêtes. Enfin, l'utilisation d'une classification a priori des documents, autorise une meilleure appréhension du contenu des documents avant même de consulter leur corps. Utilisées comme éléments de filtrage, les catégories donnent le choix à l'utilisateur de restreindre la portée de sa requête.

Les évaluations à la TREC [HARM93] ne sont pas adaptées pour juger des performances de systèmes utilisant des fonctionnalités avancées de navigation. Le problème majeur est que l'utilisateur devrait être mis au centre du processus d'évaluation, et qu'il ne faudrait plus juger les moteurs par les seules mesures de précision/rappel, mais aussi par le temps passé à chercher une réponse adéquate.

7. Références bibliographiques

[ABER95] Aberdeen J., Burger J., Day D., Hirschman L., Robinson P. and Vilain M., "MITRE: Description of the ALEMBIC system used for MUC-6". In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, pp. 141–155, Columbia, MD, 1995. NIST, Morgan-Kaufmann Publishers.

[BIKE97] Bikel D., Miller S., Schwartz R., and Weischedel R. "NYMBLE: a high-performance learning name-finder". In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, pp. 194–201,Washington, D.C., 1997. ACL.

[COVE67] Cover T. and Hart P., *Nearest neighbor pattern classification*. IEEE Trans. Inform. Theory, IT13, pp. 21-27, 1967.

[GRIS95] Grishman R. and Sundheim B., "Design of the MUC-6 evaluation". In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, 1995. NIST, Morgan-Kaufmann Publishers.

[HARM00] Harman D., "What We Have Learned, and not learned, from TREC", *Proceedings of the 22nd Annual Colloquium on IR Research*, Sidney Sussex College, Cambridge, Angleterre, pp. 2–20, April, 2000.

[HARM93] Harman D., Overview of the First Text Retrieval Conference, *National Institute of Standards and Technology*, Special Publication 500-207, 1993.

[IWAN95] Iwanska L., Croll M., Yoon T., and Adams M., "Wayne state university: Description of the UNO natural language processing system as used for MUC-6". In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, 1995. NIST, Morgan-Kaufmann Publishers.

[LEWI97] Lewis, D. D., Reuters-21578 text categorization test collection, http://www.research.att.com/lewis, 1997.

[MANI97] Manigot L., Pelletier B. : « Intuition, une approche mathématique et sémantique du traitement d'informations textuelles ». *Actes du colloque International Fractal 1997*, pp. 287-291. BULAG, Université de Franche Comté, Revue Annuelle 1996-1997.

[MORG95] Morgan R., Garigliano R., Callaghan P., Poria S., Smith M., Urbanowicz A., Collingham R., Costantino M., Cooper C., and the LOLITA Group, "University of Durham: Description of the LOLITA system as used for MUC-6". In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, Columbia, MD, 1995. NIST, Morgan-Kaufmann Publishers.

[SALT83] Salton G., McGill M.J., *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.

[SEBA02] Sebastiani F., *Machine learning in automated text categorization*. ACM Computing Surveys, 34(1):1-47, 2002.

[VERO03] Véronis J., « Hyperlex : cartographie lexicale pour la recherche d'informations ». *Actes de la Conférence Traitement Automatique des Langues (TALN'2003)*, pp. 265-274. Batz-sur-mer (France): ATALA, 2003.

[VOOR00] Voorhees E., Tice D., "Building a Question Answering Test Collection", *Proceedings of SIGIR-2000*, pp. 200-207, July, 2000.

[WILK95] Wilkinson R., Zobel J., Sacks-Davis R, "Similarity measures for short queries". In *Fourth Text Retrieval Conference (TREC-4)*, Gaithersburg, MD, pp. 277-285, 1995.

Sélection des traits et catégorisation thématique dans un corpus de pages personnelles Web

Martine Hurault-Plantet

LIMSI-CNRS, BP 133, 91403 Orsay cedex - France

Martine.Hurault-Plantet@limsi.fr

Résumé :

Nous présentons dans cet article une méthode de sélection des traits sémantiques dans un corpus de pages personnelles Web. Les traits sémantiques sont ceux qui représentent le sens de ce qui est présenté par l'auteur de la page Web, et s'opposent aux autres traits caractérisant un objet du Web tels que les traits multimédia, les traits présentationnels et structurels, et les traits morpho-syntaxiques. Notre méthode utilise deux approches : la sélection des traits pertinents qui consiste à éliminer les traits *mots* qui sont peu porteurs de sens ou qui ne seraient pas discriminants, et la construction de nouveaux traits qui consiste à rassembler les traits pertinents en macro-traits, les *thèmes*, pour obtenir une granularité plus grosse de la caractérisation. Nous présentons la méthode ainsi qu'une analyse des résultats obtenus. Nous avons identifié deux types de thèmes, les thèmes de domaine d'activité et les thèmes fonctionnels. Les premiers correspondent au domaine d'activité présenté par l'auteur du site, et les deuxièmes correspondent à ses messages de recommandation et de navigation liés au média Internet.

Mots-clés : catégorisation thématique, sélection de traits, réduction des dimensions, corpus Web.

Abstract:

In this paper, we present a feature selection method for semantic features within a corpus made of personal Web pages. Semantic features are those which represent the meaning of the Web page, and are part of the set of salient features of a Web object such as multimedia, appearance, structural, and morphosyntactic features. We use two complementary approaches: the selection of relevant features by removing *word* features with poor meaning support or with poor discriminating power, and the construction of new features by clustering relevant features to produce *topics*. We present the method as well as an analysis of our results. We identified two types of topic, topics from a domain and functional topics. The first ones correspond to the domain presented by the Web site author, and the second ones correspond to her/his information or navigation messages due to the Internet media.

Keywords: thematic analysis, clustering, feature selection, dimension reduction, Web corpus.

1. Introduction

Le travail présenté a été réalisé dans le cadre du projet RNRT SensNet[1] sur la *Catégorisation sémantique des usages et des parcours sur le Web*. L'objectif général du projet est une analyse sémantique du Web à partir de ses usages effectifs, et l'un des points qui en dépend est l'identification et le traitement des traits pertinents pour caractériser les objets du Web. Au début du projet, plusieurs catégories de traits ont été identifiés : les traits sémantiques, les traits multimédia, les traits présentationnels et structurels, et les traits morpho-syntaxiques. A chaque type de trait correspond un type de caractérisation d'une page Web : les traits sémantiques[2] s'attachent à renvoyer le sens de ce qui est présenté par l'auteur de la page, les traits multimédia représentent l'image et le son, les traits présentationnels et structurels représentent la mise en page ainsi que les liens intra et inter-pages, et les traits morpho-syntaxiques contribuent à représenter le style de la page. Dans cet article, nous nous intéressons plus particulièrement au traitement des traits sémantiques.

---

[1] La description du projet est sur le site www.telecom.gouv.fr/rnrt/projets/res_01_39.htm. Les partenaires du projet sont France Télécom R&D, le mesureur d'audience sur le Web Nielsen//NetRatings www.nielsen-netratings.com, le LIMSI-CNRS, et l'université Paris III.

[2] En linguistique structurale, le trait sémantique, plus souvent appelé *sème*, est « l'unité minimale de signification entrant, comme composant, dans le sens d'une unité lexicale » (Grand Larousse Universel). Nous prenons ici la notion de trait sémantique dans un sens plus général d'unité de signification attachée à une entité, la page Web. Cette unité de signification peut être un mot, un thème, un domaine.

Une des particularités des corpus issus de l'Internet est leur très grande taille, et les traits qui peuvent contribuer à les caractériser sont trop nombreux pour permettre d'en saisir de manière globale, synthétique, le contenu. Le corpus que nous avons étudié comporte 674 687 mots différents regroupant 20 472 811 occurrences. Les algorithmes d'analyse de données ne traitent pas de telles dimensions. Une solution souvent utilisée en fouille de textes est de réduire les dimensions de l'espace des traits, en s'efforçant de ne retenir que les traits les plus pertinents au regard de la caractérisation du corpus. Plusieurs méthodes ont été proposées pour résoudre ce problème. Elles se regroupent en deux approches principales : la sélection des traits pertinents qui consiste à éliminer les traits qui sont peu porteurs de sens ou qui ne seraient pas discriminants [YANG97], et la construction de nouveaux traits qui consiste à rassembler les traits pertinents en macro-traits pour obtenir une granularité plus grosse de la caractérisation [BEKK03]. Nous avons utilisé ces deux approches pour développer un ensemble de méthodes qui contribuent à la réduction des dimensions de l'espace des traits sémantiques d'un corpus de pages Web. La réduction des dimensions de l'espace des traits se fait en deux étapes. Dans un premier temps, nous effectuons, au niveau du corpus, une élimination des mots non pertinents à l'aide d'une liste de mots vides et d'un seuil minimum d'occurrence, puis au niveau de la page Web, une élimination des mots non discriminants à l'aide d'un critère fréquentiel tf*idf, classique en recherche d'information [SALT88]. Dans un deuxième temps, nous construisons des nouveaux traits, les *thèmes*, regroupant des ensembles de mots pertinents. La détection des thèmes est effectuée par un algorithme inspiré de la méthode des mots associés, issue du domaine de la scientométrie[3] et décrite dans [CALL86]. Nous indexons ensuite les documents par les thèmes trouvés.

Après une brève présentation du corpus étudié, nous introduirons les méthodes que nous avons utilisées pour la sélection des traits, pour la construction des thèmes, et enfin pour la catégorisation thématique des pages Web. Nous présenterons ensuite les résultats obtenus. L'évaluation globale des résultats est une tâche difficile, car elle requiert un corpus de référence indexé manuellement. Nous avons effectué une évaluation locale d'un thème du corpus. C'est une évaluation partielle, mais elle indique des tendances et les difficultés d'une telle évaluation. Après une analyse des travaux connexes, nous conclurons sur les apports des méthodes présentées, les difficultés rencontrées dans leur application, et les perspectives de ce travail.

[3] Méthodes statistiques permettant de mesurer la dynamique de l'activité scientifique.

2. Le corpus de pages Web

L'étape de constitution du corpus du projet SensNet n'étant pas complètement terminée, nous avons développé nos méthodes sur un corpus issu d'un précédent projet sur l'analyse des usages d'Internet, TypWeb, décrit dans [BEAU02]. Ce corpus provient du recueil, par une entreprise spécialisée dans la mesure d'audience, des parcours[4] sur le Web d'un panel de 1 140 internautes français. Il comporte 56 984 pages Web qui font partie de sites hébergés par des fournisseurs d'accès. Il est constitué en grande partie de pages personnelles, mais aussi de pages d'associations, de villes touristiques, ou de très petites entreprises.

Une page source Web contient un grand nombre d'informations textuelles, ainsi que des informations, étiquetées en langage HTML, sur les images, la mise en page et la navigation Internet. Pour pouvoir traiter séparément ces informations, il est nécessaire d'extraire séparément les différents traits de la page, traits textuels d'une part et étiquettes HTML d'autre part. Un tel outil d'extraction a été réalisé dans le cadre du projet TypWeb décrit dans [BEAU02], et le corpus est rendu disponible sous un format normalisé XML.

Le corpus étudié comporte 674 687 mots qui totalisent 20 472 811 occurrences. Le nombre moyen de mots par page Web est de 142, chiffre assez élevé qui est dû aux 10 % de documents qui contiennent un grand nombre de mots. La médiane est à 59 mots, la moitié des pages contient donc moins de 59 mots. En fait, une grande partie des pages contient peu de mots, et il y a une grande disparité dans le nombre de mots par page.

3. Sélection des traits textuels pertinents

Les traits retenus doivent satisfaire deux critères pour représenter la signification de la page Web dans le corpus : d'une part le trait retenu doit être porteur de sens, ce qui nous a conduit à l'élimination de certains types de traits, et d'autre part il doit être discriminant, c'est-à-dire particulariser la page par rapport à l'ensemble des pages Web du corpus.

3.1 Traits porteurs de sens

Tout ce qui constitue un élément textuel, visualisable sur Internet, de la page a été retenu comme trait textuel. Parmi ces traits textuels, nous n'avons gardé que les traits lexicaux. Par exemple, les traits constitués uniquement de tirets qui correspondent à une barre de séparation dans la page Web, ont été supprimés. En revanche, le vocabulaire du Web a été gardé (topsite, http, www, freeware,

[4] Ensemble de pages vues par un internaute au cours d'une session, une session étant un temps de connexion qui ne comporte pas d'interruption de plus de 30 minutes.

shareware, etc.). Les mots grammaticaux (articles, conjonctions, pronoms, etc.) et les adverbes précisent le sens d'un mot ou l'articulent avec d'autres mots. Isolés, ils ne prennent sens que pour définir un style. De la même manière, les nombres ont une signification lorsqu'ils sont reconnus en tant qu'entité numérique (date, grandeur physique, montant financier, etc.), mais nous n'avons pas actuellement fait cette reconnaissance sur le corpus. Nous avons donc choisi d'éliminer les nombres et les mots grammaticaux comme étant sans signification hors contexte.

3.2 Traits discriminants

Un mot dans un document peut être plus ou moins représentatif du sujet traité dans le document. On appelle trait discriminant un trait qui permet de distinguer un groupe de documents dans un ensemble plus grand. Le degré de discrimination d'un mot sera plus ou moins important suivant son taux de présence dans les documents du groupe et d'absence dans le reste des documents ; complètement discriminant, il sera présent uniquement dans les documents du groupe et absent dans les autres.

Notre but est de déterminer un nombre restreint de traits sémantiques pouvant caractériser les pages Web. Les mots très peu fréquents ne caractérisent qu'un très petit nombre de documents. Ils sont importants dans une perspective de recherche d'information où la spécificité d'un document doit pouvoir être retrouvée. Mais ils sont discriminants à un niveau de granularité très fin, et sont alors peu utilisables dans les regroupements des mots en thèmes. Nous avons donc éliminé les mots de fréquence inférieure à un seuil donné sur l'ensemble du corpus. L'ensemble des mots peu fréquents constitue la grande masse des mots éliminés. En effet, environ la moitié des mots du corpus sont des hapax, présents dans une seule page.

A l'inverse, certains mots sont trop également répartis sur les pages du corpus et auront donc un degré de discrimination trop faible. Nous avons donc donné un poids à chaque mot indexant une page Web suivant son pouvoir discriminant pour cette page. Pour cela nous avons utilisé un critère tf*idf [SALT88]. Nous avons éliminé de chaque page Web les mots dont le poids tf*idf était inférieur à 1, les considérant comme moins représentatifs de la page. Ce seuil a été déterminé empiriquement. L'effet de cette dernière sélection est de diminuer le nombre de mots indexant une page, sans toutefois diminuer le nombre global de mots gardés. En effet, ne garder que les mots les plus représentatifs de la page diminue l'écart existant entre le nombre de mots indexant une page très longue, et le nombre de mots indexant une petite page.

3.3 Conclusion

Les résultats de la sélection des mots pertinents sont présentés dans le tableau 1. On voit sur ce tableau que la majorité des documents restent indexés lorsqu'on supprime les mots non informatifs et les mots de fréquence faible. En effet, 12 % de l'ensemble initial des mots suffit à représenter 98 % du corpus. La majorité des mots éliminés a une faible fréquence. L'élimination par le poids tf*idf intervient au

niveau de la page et non du corpus dans son ensemble, et de ce fait supprime peu de mots globalement. Les mots supprimés le sont au niveau local.

Type de sélection	**Nombre de mots**	**Nombre de pages avec des mots**
Corpus initial	674 687	56 984
Elimination des mots non informatifs	646 154 (96 %)	56 298 (99 %)
Elimination des mots de fréquence inférieure à 5	89 672 (13 %)	56 140 (98,5 %)
Elimination des mots de tf*idf inférieur à 1	82 207 (12 %)	55 913 (98 %)

Tableau 1 : Résultats de la sélection des traits pertinents

4. Catégorisation thématique

La sélection des traits pertinents a déjà considérablement réduit le nombre de traits sémantiques. Nous sommes en effet passé d'une matrice de 674 687 mots x 56 984 pages à une matrice de 82 207 mots x 56 984 pages. La construction de macro-traits *thèmes* à partir de cet ensemble devrait réduire encore ces dimensions et rendre les traits sémantiques plus lisibles.

4.1 Construction des thèmes : l'analyse des mots associés

Pour déterminer automatiquement les thèmes sous-jacents au corpus de pages Web, nous nous sommes inspirés de la méthode d'analyse des mots associés, décrite dans [CALL86], qui s'appuie sur les relations de cooccurrence entre les mots d'un corpus de documents. Le but initial de cette méthode était d'identifier la structure d'un domaine de recherche à partir du contenu des articles publiés dans ce domaine. Elle a été principalement utilisée dans des études en scientométrie ([COUR89], [CALL91], [DING00]).

La méthode consiste à construire des agrégats de mots (ou *thèmes*) à l'aide d'un indice statistique d'équivalence entre deux mots au sein d'un corpus de documents. Chaque mot est considéré comme un caractère statistique d'un document. La fréquence de chaque mot dans le document n'est pas prise en considération, les deux modalités retenues sont la présence et l'absence d'un mot dans l'indexation d'un document. La fréquence de la modalité *présence d'un mot* est égale au nombre de documents contenant ce mot. L'indice d'équivalence est défini de la manière suivante.

$$e_{ij} = (c_{ij} / c\ i) * (c_{ij} / c_j)$$

Avec :

- c_{ij} : le nombre de documents contenant à la fois les mots i et j,
- c_i : le nombre de documents contenant le mot i,
- c_j : le nombre de documents contenant le mot j,
- e_{ij} : l'indice d'équivalence entre les mots i et j.

Le terme (c_{ij}/c_i) représente la fréquence conditionnelle de la présence du mot j étant donné la présence du mot i, dans des documents du corpus. Symétriquement, le terme (c_{ij}/c_j) représente la fréquence conditionnelle de la présence du mot i étant donné la présence du mot j. Cet indice est normalisé : sa valeur est comprise entre 0 (si les deux mots ne sont jamais trouvés ensemble dans les documents) et 1 (si les deux mots sont toujours trouvés ensemble). Lorsque les mots i et j sont indépendants l'un de l'autre, l'indice d'équivalence est égal à la fréquence marginale de la présence des deux mots i et j dans les documents (c_{ij} / N, N étant le nombre de documents du corpus). L'indice d'équivalence est assez proche de l'indice d'information mutuelle décrit dans [MAN99]. Cependant, une faiblesse de ce dernier est d'introduire un biais en favorisant les termes rares. En effet, à probabilité conditionnelle égale, la valeur de l'indice d'information mutuelle est plus grande pour les termes peu fréquents [YANG97].

L'algorithme classiquement utilisé dans l'analyse des mots associés, décrit dans [MONA00], construit séquentiellement des agrégats de mots à partir de la liste de tous les couples de mots du corpus rangés dans l'ordre décroissant de l'indice. Un seuil minimum de cooccurrence limite le nombre de mots pris en compte. Un premier agrégat est construit par l'ajout au couple de mots de plus fort indice de l'ensemble des mots qui leur sont liés. Les agrégats suivants sont construits de la même manière à partir du couple de mots de plus fort indice parmi les mots non encore agrégés. Un nombre maximum de liens limite la taille des agrégats. Une fois tous les agrégats construits, les liens restant dans la liste sont ajoutés pour relier les agrégats entre eux.

Notre algorithme de construction des agrégats diffère de l'algorithme classique en deux points principaux. Tout d'abord, Les couples de mots sont pris dans l'ordre de cooccurrence décroissante, et non dans l'ordre de l'indice décroissant, pour construire d'abord les agrégats autour des mots les plus fréquents du corpus. Ensuite, l'algorithme ne s'arrête pas au premier niveau d'agrégation (les mots liés au couple initial), mais poursuit récursivement en prenant les mots les plus fortement liés à l'ensemble des mots déjà agrégés, afin de regrouper des agrégats qui resteraient séparés dans la méthode classique. L'agrégat est alors limité par deux seuils, un seuil minimal de l'indice pour le couple initial, et un seuil sur la somme des indices des liens entre le mot à ajouter et les mots déjà agrégés. Nous pouvons ainsi former des agrégats de tailles très différentes reflétant les répartitions inégales des thèmes du corpus.

4.2 Catégorisation des pages Web par les thèmes

Nous avons indexé les documents par les thèmes (agrégats) en attribuant un poids au thème suivant son degré de similarité avec le document. Nous avons testé deux indices simples de similarité :

$$S_1 = n / N \qquad S_2 = (\text{Somme}_{i = 1 \text{ à } n}\, tf*idf_i) / N$$

Avec :

- n : nombre de mots du thème dans le document,
- $tf*idf_i$: poids du mot i du thème dans le document,
- N : nombre de mots du document.

Un document peut ainsi être indexé par plusieurs thèmes qui seront rangés dans l'ordre décroissant de l'indice de similarité. Nous n'avons pas mis de seuil sur le poids d'un thème mais nous n'avons conservé que les deux premiers thèmes indexant chaque document.

4.3 Evaluation

Pour pouvoir évaluer les thèmes trouvés, leur justesse et leur recouvrement par rapport aux centres d'intérêt des internautes, il faudrait que le corpus ait été manuellement indexé suivant des domaines répertoriés. Ne disposant pas actuellement de ces données, nous avons donc effectué une première évaluation manuelle, très ponctuelle, de la fiabilité de l'indexation par les thèmes. Pour cela, nous avons choisi un thème de petite taille, *forsythias_rodhos* qui comporte 26 mots. Nous avons pris arbitrairement comme nom d'un thème le couple de mots à partir duquel le thème a été construit. Cependant, ses mots les plus fréquents donnent une indication plus juste sur son contenu. Les mots les plus fréquents du thème sont : *jardin* (présent dans 221 pages Web), *fleurs* (140), *mailto* (136), *plantes* (130), *hiver* (128), *franck* (102), *printemps* (97) *été* (96). Les autres mots du thème sont : *forsythias* (6), *rodhos* (5), *agrumes* (7), *automne* (72), *azalées* (5), *cerisiers* (9), *colline* (38), *escholzias* (5), *lys* (33), *narcisses* (5), *pommiers* (7), *primevères* (5), *floraison* (18), *plante* (61), *engrais* (22), *graines* (43).

Les mots du thème *forsythias_rodhos* comportent des termes génériques concernant les plantes (*fleurs, plantes, plante, agrumes*), des représentants de la classe des fleurs (*forsythias, rodhos, azalées, escholzias, lys, narcisses, primevères*,), de la classe des arbres fruitiers (*cerisiers, pommiers*), de la classe des lieux extérieurs (*jardin, colline*), la classe des saisons (*printemps, été, automne, hiver*), ainsi que des mots du vocabulaire de la culture des plantes (*floraison, engrais, graines*). Les mots « étrangers » sont *mailto* et *franck*. Ils soulèvent le problème de la prise en compte automatique à un niveau très fin de la structure d'une page Web, dans laquelle sont présents en même temps différents niveaux d'information que l'internaute, pour sa part, sépare aisément. L'interprétation d'un thème en termes

génériques (ou plutôt en sèmes génériques si on se réfère à [RAST01]) est délicate. Les classes sémantiques que nous avons énumérées sont incomplètes, mais elles se rattachent néanmoins aux domaines des *plantes* (prédominant si on considère les fréquences de récurrence dans les documents) et de la *culture des plantes*.

Nous avons indexé les 56 984 pages du corpus avec les thèmes trouvés pour les indices de similarité S_1 et S_2 (voir le paragraphe 4.3), pour un ensemble de mots restreint par page (on ne conserve pour chaque page que les mots dont le tf*idf est au moins égal à 1) ou étendu (on conserve tous les mots de chaque page). Ceci nous a donné trois indexations : la première avec l'indice S_1 et un ensemble de mots restreint par page, la deuxième avec l'indice S_1 et l'ensemble étendu de mots par page, et la troisième avec l'indice S_2 et l'ensemble étendu de mots par page. Ces trois indexations nous ont donné trois ensembles de pages pour le thème *forsythias_rodhos* que nous avons vérifiées manuellement. Les pages indexées par le thème ont été considérées comme pertinentes lorsque les plantes constituaient le thème prédominant de la page. Le nombre de pages susceptibles de se rattacher à ce thème dans le corpus est de 169. C'est un nombre approximatif, calculé à partir des pages jugées pertinentes dans les sites trouvés à partir des trois indexations.

Les sous-thèmes trouvés dans les pages jugées pertinentes sont les cartes postales sur les plantes, la vente des plantes, leur culture, la botanique et la santé par les plantes. Les sites de cartes postales ou de photos ou dessins représentant des plantes ont été considérés comme pertinents lorsque le sujet central était les plantes. En revanche, les paroles de chansons et les galeries de peintures n'ont pas été considérées comme pertinentes, même si certains tableaux ou certaines paroles évoquaient des plantes. Les locations d'appartement ne sont pas pertinentes même si le jardin est mentionné. Les photos de mode (automne, hiver) ne sont pas pertinentes.

Type d'indexation	**Rappel**	**Précision**
S_1 et ensemble restreint de mots par page	45,86 %	65,59 %
S_1 et ensemble étendu de mots par page	35,34 %	68,11 %
S_2 et ensemble étendu de mots par page	71,43 %	56,21 %

Tableau 2 : Evaluation du thème forsythias_rodhos

Les résultats de l'évaluation sont montrés dans le tableau 2. Les chiffres relativement faibles sont en partie dus aux particularités d'une page Web. Nous avons vu que beaucoup de pages contiennent peu de mots, et c'est le cas en particulier des pages d'accueil et de sommaire. Ces dernières sont souvent mal indexées par les thèmes. Elles le sont cependant mieux par la deuxième indexation (qui garde plus de mots) que par les deux autres. Les pages de contenu, qui comportent donc plus de mots du domaine présenté par l'auteur de la page, sont en revanche mieux indexées par la première indexation (qui garde moins de mots).

Cette différence entre les pages structurantes (accueil, sommaire) qui contiennent peu de mots, et les pages à contenu, qui en contiennent souvent beaucoup plus, est une source de difficulté dans la catégorisation. La troisième indexation, qui fait intervenir la pondération des mots de la page, est moins précise mais retrouve un plus grand nombre de documents corrects.

Cette évaluation nous a donné quelques indications intéressantes, en particulier sur l'influence du type de page, page sommaire ou page à contenu, sur l'indexation, mais il faudra maintenant comparer avec d'autres méthodes et des données de référence pour obtenir une évaluation globale.

5. Résultats

Nous avons obtenu sur le corpus étudié 399 thèmes de tailles très différentes. Le nombre moyen de pages par thème est de 139, mais la médiane est à 24 pages. Cela signifie que pour un corpus d'environ 60 000 pages Web, la moitié des thèmes comporte moins de 24 pages. Nous avons donc quelques thèmes de taille importante qui correspondent à des grandes tendances du corpus, et un grand nombre de petits thèmes qui montrent la grande variété des centres d'intérêts des internautes.

Les résultats présentés ont été obtenus avec un seuil de 0,5 pour l'indice d'équivalence des deux mots qui initialisent un agrégat, et un seuil de 0,004 x N pour l'ajout d'un mot à un agrégat, N étant le nombre de mots de l'agrégat au moment de l'ajout. Nous avons donc mis un seuil élevé pour l'initialisation de l'agrégat, et un seuil faible pour l'ajout. L'indexation des thèmes a été faite avec l'indice de similarité S_2 (voir paragraphe 4.3).

5.1 Les thèmes de domaine

Deux thèmes sont prédominants dans le corpus étudié : nous les avons intitulés *photos de femme* et *généalogie* à la fois d'après les mots qui les décrivent et les contenus observés des documents qu'ils indexent[5].

Le thème *casta_laetitia* (chaque thème prend arbitrairement pour nom le couple de mots initial de l'agrégat, mots qui ne sont pas forcément les plus fréquents du thème) est le thème le plus important sur les photos de femme. Il comporte 438 mots, regroupe 5 640 pages Web, et l'ensemble des mots qui le compose indique qu'il est centré sur les photos d'actrices et le sexe. Les mots les plus fréquents de ce thème sont : *photos* (2 266 pages), *gallery* (637), *pictures* (464), *membre* (392), *francophone* (370), *filles* (348), *stars* (344), *sexy* (291). Ce thème comporte un grand nombre de noms d'actrices.

[5] Nous n'avons pas observé tous les documents mais un certain nombre pris au hasard.

⇨ **Exemple du contenu normalisé en XML d'une page Web indexée par ce thème :**

<PAGE>membres.tripod.fr/…/page1.htm</PAGE>
<DUMPTEXT>
Sarah Michelle Gellar
Biographie
Filmographie
Photos
Autres **Célébrités**
Home
</DUMPTEXT>

Les mots mis en gras *Sarah, Michelle, Photos, Célébrités*, sont des mots du thème *casta_laetitia*.

Le domaine de la généalogie est principalement représenté par le thème *féminin_masculin*, qui comporte 626 mots et regroupe 3 354 pages. Les mots les plus fréquents du thème sont : *france* (1 682 pages), *marie* (1034), *famille* (941), *principale* (808) *sexe* (801), *naissance* (795), *décès* (530), *père* (508). Ce thème comporte également un grand nombre de prénoms et de noms de famille. Beaucoup de documents sont des fiches généalogiques.

⇨ **Exemple du contenu normalisé en XML d'une page Web indexée par ce thème :**

<PAGE>….free.fr/.../dat1.htm</PAGE>
<DUMPTEXT>
Informations généalogiques

Retour à la page principale

SIMEON, Emile **Sexe: Masculin**
Naissance: 06 février 1876 à Hombourg-Budange, 57920, Moselle, **France**
Décès 19 février 1956 à Valmestroff, 57970, Moselle, **France**
Parents:

Père: SIMEON, Stéphan
Mère: DAME, Catherine
Famille:
…
</DUMPTEXT>

Les mots mis en gras *Sexe, Masculin, Naissance, France, Décès, Parents, Père, Mère, Famille*, sont tous des mots qui appartiennent au thème *féminin_masculin*.

Le reste des thèmes représente une grande variété de domaines. Parmi les plus fréquents, certains sont orientés vers la diffusion des connaissances tels que *etats_unis* (économie et géographie, 1 187 pages), ou *grammar_odds* (exercices et plaisanteries en anglais, 233 pages). D'autres rassemblent des pages sur les outils de programmation (*basic_visual*, 1 422 pages), les jeux vidéos (*raider_tomb*, 789 pages), les recettes de cuisines (*poivrez_salez*, 626 pages), les plantes (*forsythias_rodhos*, 169 pages), l'enseignement (*cognitivo_émotionnelles*, 490 pages). Le tourisme est également présent, ainsi que les séries télévisées, les chanteurs, etc.

5.2 Les thèmes fonctionnels

Nous avons appelé thème fonctionnel tout thème centré sur une fonction plutôt que sur un domaine d'activité. Parmi les thèmes de taille importante, trois rassemblent un vocabulaire qui se rapporte à une fonction dans la page Web plutôt qu'au sujet que l'auteur de la page, ou du site, cherche à diffuser. Ce sont les thèmes *click_moved, droits_réservés, et onmouseout_onmouseover*.

Le thème *click_moved* comporte 102 mots et rassemble 10 853 pages. Les mots les plus fréquents de ce thème sont *page* (6 541 pages), *click* (3 191), *moved* (2 710), *http* (2 627), *www* (2 008), *fr* (1 654), *accueil* (1 377), *text* (1220). L'ensemble des mots qui le composent semble indiquer que ce thème rassemble des pages de redirection (« page has moved, click here »), des pages d'accueil et de liens (comme l'indiquent les mots *http*, *www*, *fr* et *accueil*), des indications de navigation dans le site (comme l'indiquent le mot *précédente*, présent dans 346 pages, et le mot *suivante*, présent dans 504 pages).

⇨ **Exemple du contenu normalisé en XML d'une page Web indexée par ce thème :**

<PAGE> ….free.fr/Agenda/archives.htm</PAGE>
<DUMPTEXT>
FRAME: left
FRAME: right
Cette **page utilise** des **cadres**, mais votre **navigateur** ne les **prend** pas en **charge**.
</DUMPTEXT>

Les mots mis en gras *page, utilise, cadres, navigateur, prend, charge*, sont tous des mots qui appartiennent au thème *click_moved*.

Le deuxième thème fonctionnel est *droits_réservés* qui comporte 204 mots et rassemble 4 928 pages Web. Les mots les plus fréquents de ce thème sont *france* (1 682 pages), *sites* (1 157), *net* (797), *visiteurs* (648), *copyright* (578), *droits* (529), *arial* (520), *gratuit* (516). Ce thème semble rassembler essentiellement des pages

Web qui comportent une indication sur les droits d'auteur sur ce qui est présenté sur la page. Les termes les plus fréquents de ce thème tels que *droits, réservés, gratuit, copyrigh*t, mais aussi des termes moins fréquents comme *reproduction* (380), *interdite* (309), *reserved* (254), *rights* (254), *accès* (375), *auteurs* (194), *autorisation* (149), *licence* (129), *audios* (56), *confidentialité* (56), *consultables* (53), *exclusifs* (51), signalent une préoccupation des auteurs de pages personnelles sur les droits, soit vis-à-vis de leur propre production, soit vis-à-vis des documents qu'ils mettent à disposition.

Le troisième thème fonctionnel important est *onmouseout_onmouseover*. Ce thème comporte 128 mots et rassemble 4 666 pages Web. Les mots les plus fréquents du thème sont *http* (2 627 pages), *www* (2 008), *index* (1 415), *src* (1 216), *images* (1 124), *href* (1 016), *html* (880), *image* (835). Ce thème est composé en majorité d'étiquettes HTML (*src, href, alt, align, bgcolor*), et de mots se rapportant à des produits ou à des techniques spécifiques d'Internet (*banners, macromedia, shockwaveflash, scrollbars, javascript, topsites*). Une partie de ces mots provient d'erreurs de syntaxe des auteurs de sites personnels : lorsque la syntaxe HTML n'est pas respectée dans la page Web, l'outil qui extrait les traits textuels et HTML et qui normalise le corpus peut prendre une étiquette HTML pour un mot. Nous avons alors des étiquettes HTML qui viennent se mélanger aux mots. Pour le reste, les mots du thème concernent les outils macromedia, les scripts java et les hit-parades des sites. Il semble que, dans les pages Web, ces éléments interviennent plutôt en bannières, dans des frames, en messages d'avertissement, qu'en tant que sujet principal de la page.

5.3 Discussion

Le regroupement en thèmes du vocabulaire du corpus de pages Web étudié fait ressortir une difficulté majeure dans l'analyse des centres d'intérêts des internautes, l'existence dans la page Web de deux niveaux d'information très différents, le niveau du domaine et le niveau fonctionnel.

Parfois, comme nous venons de le voir, les thèmes de domaine et les thèmes fonctionnels sont assez bien séparés, mais certains thèmes comportent des mots qui appartiennent à la fois à un domaine d'activité et à une fonction de recommandation ou de navigation. C'est le cas par exemple du thème *genweb_racinescomtoises*. Ce thème comporte 41 mots dont les uns appartiennent au domaine de la généalogie, et les autres sont liés au fonctionnement du Web (mise à jour, lien avec un fichier, etc.). Les mots qui relèvent de la généalogie, tels que *genweb* (27), *racinescomtoises* (15), *francegenweb*[6] (50), *racines* (93), sont assez peu fréquents. Les mots les plus fréquents, *site* (4 310), *http* (2 627), *ici* (1 482), *jour* (1 440), *être* (1 077), *adresse* (1 064), *mise* (930), *nouvelle* (896), sont de type fonctionnel. Le thème *genweb_racinescomtoises* regroupe 2 084 pages qui semblent en effet appartenir en petite partie seulement au domaine de la généalogie.

[6] francegenweb est un portail de la généalogie en France.

Il est difficile de donner une signification à ce type de thème qui réunit deux niveaux d'information, l'information sur le domaine d'activité et l'information sur les éléments fonctionnels attachés au fonctionnement d'Internet. Donner une signification à un thème construit automatiquement est en fait un problème général mais qui est encore plus manifeste lors du traitement d'une page Web dont le contenu est hétérogène par nature. Il faut pouvoir extraire ce qui a trait au sujet que l'auteur de la page veut mettre à disposition pour pouvoir catégoriser les pages en domaines d'activité. En cela, l'analyse des thèmes fonctionnels peut nous aider en pointant sur les éléments de la page Web qui devraient subir un traitement préalable avant la catégorisation. Une solution envisageable est d'étiqueter les parties des pages qui se réfèrent à une fonction, en utilisant par exemple des patrons construits à partir des mots du thème permettant de les repérer.

6. Travaux connexes

La sélection des traits est un préalable classique à des tâches de classification supervisée, mais elle est moins utilisée pour des tâches de classification non supervisée comme c'est le cas ici. Les critères du CHI2 ou du gain d'information, souvent utilisés, demandent une connaissance préalable des classes dont nous ne disposons pas. Néanmoins, d'après [YANG97], l'élimination des mots peu fréquents est presque aussi efficace et présente l'avantage de ne pas demander de connaissance préalable sur les classes. Par ailleurs, les évaluations de [LIU03] montrent que les méthodes de sélection non supervisée (fréquence, tf*idf) ont des performances comparables aux méthodes de sélection supervisée (CHI2, gain d'information) lorsqu'on supprime jusqu'à 90 % des termes. Au-delà, la sélection supervisée donne de meilleurs résultats. La pondération tf*idf a surtout été testée et utilisée en recherche d'information [SALT88], c'est un critère assez fiable de l'importance d'un mot d'un document par rapport au sujet qu'il traite.

La méthode de catégorisation thématique proposée s'appuie sur la formation d'agrégats de mots, comparable en cela à [BEKK03]. Les méthodes de classification non hiérarchisée, comme la méthode des moyennes mobiles, opèrent par le calcul d'une distance entre documents, en fixant préalablement le nombre de classes. C'est la méthode utilisée par [PAPY03] pour classer des documents Web par type (page informative, page carrefour, etc.) à partir des étiquettes HTML. La difficulté est ensuite de déterminer le bon nombre de classes, celui qui produira les classes de cohésion optimale et qui permettra ainsi d'attribuer plus facilement une étiquette catégorielle à chaque classe. Notre méthode se différencie de ces dernières en cela qu'elle produit une classification non hiérarchisée sans fixer a priori le nombre de classes. En revanche, elle se rapproche de la méthode proposée par [HAN97] basée sur les règles d'association. L'algorithme présenté par les auteurs part de l'hypergraphe construit à partir des plus grands ensembles de mots de support minimal fixé, et fait ensuite une partition de cet hypergraphe. L'algorithme que nous utilisons part aussi de grands ensembles de mots mais ils sont reliés par un indice

statistique représentant la force du lien entre deux mots et non directement par leur cooccurrence, et la partition du réseau en agrégats est faite au moyen de seuils sur la force des liens entre les mots.

D'autres méthodes de classification des pages Web ont été expérimentées. Par exemple, les auteurs de [SERR02] font d'abord un apprentissage automatique d'un ensemble donné de thèmes à partir de textes encyclopédiques, en les modélisant par des chaînes de Markov cachées, puis les utilisent pour indexer un corpus de pages Web pertinentes pour ces thèmes. Mais nous ne connaissions pas a priori les thèmes du corpus. Notre but n'était donc pas de retrouver des catégories d'annuaire, comme le font ces auteurs, mais bien de découvrir les centres d'intérêt des internautes.

7. Conclusion

Nous avons présenté une méthode de réduction des dimensions de l'espace des traits sémantiques qui permet une vision plus synthétique du corpus. La méthode permet de passer des quelques 600 000 mots du corpus initial à environ 400 thèmes. La réduction des traits *mots* aux traits *thèmes* conduit à une certaine perte d'information, soit par défaut (certaines pages ne sont plus indexées), soit par imprécision (certaines pages sont mal indexées). Néanmoins cette représentation des pages par des thèmes permet de réduire les dimensions de l'espace d'analyse et de cerner les centres d'intérêt principaux des internautes du panel étudié.

L'évaluation locale et partielle que nous avons effectuée a fait apparaître certaines difficultés liées aux particularités du Web. Tout d'abord, les mots d'une page Web reflètent les différents aspects de son contenu : d'une part le message sur le domaine d'activité que l'auteur de la page présente, et d'autre part les messages de recommandation et de navigation liés au média Internet. Si certains thèmes sont typiques de l'un des deux aspects, ces aspects se mêlent dans d'autres. Il serait donc intéressant de pouvoir séparer l'aspect *domaine d'activité* de l'aspect fonctionnel dans la page Web. Ensuite, les pages elles-mêmes peuvent être séparées en pages fonctionnelles (pages de redirection par exemple), pages structurantes (sommaires) et pages à contenu, comme le montre [BEAU02]. Les thèmes trouvés et la catégorisation sémantique qui s'ensuit reflètent ces disparités. Les pages structurantes sont souvent mal indexées par les thèmes car elles contiennent peu de mots et des mots génériques. Mais c'est le cas aussi des pages qui contiennent des images et peu de mots. Plus généralement, le très petit nombre de mots sur certaines pages rend difficile leur catégorisation dans un thème.

Après ce premier essai, nous projetons de faire une évaluation globale des résultats, ainsi qu'une comparaison avec d'autres méthodes de classification. Pour cela, nous utiliserons des corpus indexés, suivant un ensemble de domaines, par le mesureur d'audience NetRatings. Par ailleurs, l'équipe du projet SensNet travaille sur une catégorisation en types de pages (sommaire, contenu, etc.) qui pourra être utilisée en amont de la catégorisation thématique.

8. Références bibliographiques

[BEAU02] Beaudouin, V., Fleury, S., Habert, B., Pasquier, M. & Licoppe, C. 2002. « Décrire la Toile pour mieux comprendre les parcours. Sites personnels et sites marchands », Réseaux, 2002, 20, 116, 19-51, Parcours sur Internet. Valérie Beaudouin, Christian Licoppe (ed.), FT R&D, Hermès Science Publications, 2002.

[BEKK03] Bekkerman, R., El-Yaniv, R., Tishby, N., & Winter, Y., "Distributional Word Clusters vs Words for Text Categorization". Proceedings of *Special Issue on Variable and Feature Selection of JMLR'2003*, 2003.

[CALL86] Callon, M., Law, J., & Rip, A. *Mapping of the dynamics of science and technology*. Mac Millan , London, 1986.

[CALL91] Callon, M., Courtial, J-P. & Laville, F. "Co-word analysis as a tool for describing the network of interactions between basic and technological research: The case of polymer chemistry". *Scientometrics*, vol. 22, 1, 153-203, 1991.

[COUR89] Courtial, J-P., & Law, J. "A co-word study of artificial intelligence". *Social Studies in Science 19*, 301-311, London: Sage, 1989.

[DING00] Ding, Y., Chowdury, G.G., Foo, S. "Bibliography of information retrieval research by using co-word analysis". *Information Processing & Management*, vol. 37, 817-842, 2000.

[HAN97] E.H. Han, G. Karypis, V. Kumar, and B. Mobasher. "Clustering based on association rule hypergraphs". In *Workshop on Research Issues on Data Mining and Knowledge Discovery*, pp. 9-13, Tucson, Arizona, 1997.

[LIU03] Tao Liu, Shengping Liu, Zheng Chen, Wei-Ying Ma. "An evaluation of Feature Selection for Text Clustering". Proceedings of *ICML'03, International Conference on Machine Learning*, Washington D.C., 2003.

[MAN99] Manning, C., Schütze, H. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge MA., 1999.

[MONA00] Monarch, I. "Information Science and Information Systems: Converging or Diverging?" Actes de *ACSI 2000, Association canadienne des sciences de l'information : Les Dimensions d'une Science de l'Information Globale*, 2000.

[PAPY03] Papy F., Bouhaï N. « Navigation et recherche par catégorisation floue des pages HTML ». Actes des *JFT'2003, Journées Francophones de la Toile*, Tours, France, 2003.

[RAST01] Rastier, F., *Arts et Sciences du texte*. Presses Universitaires de France, Paris, 2001.

[SALT88] Salton, G., Buckley, C. "Term-weighting approaches in automatic text retrieval". *Information Processing and Management*, vol. 24, 513-523, 1988.

[SERR02] Serradura, L., Slimane, M., Vincent, N., Proust, C. « Classification semi-automatique de documents Web à l'aide des Chaînes de Markov Cachées ». Actes de *Inforsid 2002*, Nantes, France, pp. 215-228, 2002.

[YANG97] Yang, Y., & Pedersen, J.O. "A comparative study on Feature Selection in Text Categorization". Proceedings of *ICML'97, International Conference on Machine Learning*, pp 412-420, Nashville, US., 1997.

INDEXATION DES AUTEURS