LA RECHERCHE D'INFORMATIONS PAR ACCÈS AUX CONTENUS SÉMANTIQUES :
VERS UNE NOUVELLE CLASSE DE SYSTEMES DE RECHERCHES D'INFORMATIONS ET DE MOTEURS DE RECHERCHE (ASPECTS LINGUISTIQUES ET STRATEGIQUES)[1]

JEAN-PIERRE DESCLÉS, BRAHIM DJIOUA

Abstract. Current Information Retrieval Systems (IRS), which operates primarily on terms (by measuring their presence and co-presence or their absence) and without using semantic relations, increases the number of documents thus causing an unquestionable noise that is difficult to control by the users. How to discover *rare or hidden document*, which contains genuine information? How to select relevant information for precise goals? To introduce more semantics in information retrieval paradigm for textual documents we propose a new approach based on a new linguistic and computational technique of Exploration of the Context, developed at the LaLIC laboratory, which is more complex than the identification of regular expressions. It comes to modeling and carrying out a new system generation for Information Retrieval Systems (IRS) which is founded on automatic annotations of texts. We present in this paper, the interaction with an indexing process throw building a prototype which results from a coupling of a two engines: (i) an annotation system EXCOM and (ii) a second indexing system MOCXE. The consequences of these IRS are economically fully serviced. The various fields of survey (strategic, economic, innovation...) have obvious needs for powerful data-processing tools which would enable them to detect rare information rather than to consolidate only known information of all the specialists. The innovation indeed supposes the detection of signals which indicate, for example, the emergence of new concepts, new methods, new molecules, new techniques, the unknown relationships for companies, and the nomination of such new leader, new potential markets, new requests for users, a beginning of epidemic or a crisis.

1. IMPORTANCE ECONOMIQUE ET CULTURELLE DE L'ACCÈS AUX INFORMATIONS

Celui qui a accès, au moment où il en a besoin, aux informations qui lui sont utiles, en particulier aux informations que la plupart ne connaissent pas encore, détient un réel pouvoir, car alors il peut adapter son action en utilisant ces

[1] Il s'agit d'un projet que nous menons en ce moment dans le laboratoire LaLIC de l'Université de Paris-Sorbonne, avec le soutien de partenaires externes. Ont participé à l'élaboration de cet article : F. Le Priol, L. Nait-Baha et les doctorants M. Alrahabi, M. Bertin, A. Blais, J. Garcia-Flores.

RRL, **LII**, *1–2*, p. 5–54, Bucureşti, 2007

nouvelles informations, encore largement inconnues pour ses interlocuteurs et concurrents. Il est certain que posséder toutes les informations qui seraient déjà connues par la très grande majorité ne donne pas toujours de très grands avantages. En revanche, avoir obtenu une information rare, cachée car difficile à trouver, donne à celui qui la détient des atouts nouveaux. Cela se confirme aussi bien dans le secteur de la veille économique, technologique, écologique ou sécuritaire, que dans les secteurs de l'innovation scientifique et technique ou dans les réseaux commerciaux, boursiers et de marketing… (Pedauque 2006). Par ailleurs, comme le signale avec justesse Jean-Noël Jeannene[2], (Jeanneney 2005), il serait dangereux que les fonds très riches des grandes bibliothèques, en particulier celles du continent européen, soient devenus pratiquement inaccessibles car n'étant pas couverts par les modes d'indexation des moteurs de recherche les plus utilisés sur la planète. Dans ce cas, la recherche des informations et le savoir constitué qui en résulte, seraient entièrement dominés par quelques grandes bibliothèques anglo-saxonnes, jetant ainsi un voile d'obscurité sur tous les autres savoirs qui sont archivés ailleurs. Il faut désormais se rendre compte que la recherche des « bonnes » informations s'inscrit dans une nouvelle « guerre mondiale technologique » dont les enjeux sont non seulement économiques mais aussi scientifiques, stratégiques et culturels et, par conséquent, ils conditionnent directement l'avenir des sociétés industrielles et commerciales, le pouvoir d'action des états et même des régions du monde[3] (Cohen 2004) … Il est certain que les états se mobilisent pour répondre à ce nouveau défi : gagner la bataille d'accès aux informations afin de conserver une capacité d'action plus autonome et contrôlée. Il est évident que les États Unis, avec Google et Yahoo, ont pris un avantage certain mais, depuis peu, l'Europe, avec le projet « Quaero »[4], et le Japon (Ministère de l'Economie, du Commerce et de l'Industrie), ont lancé des programmes concurrents destinés à développer des outils nécessaires à l'indexation de contenus multimédias en voulant ainsi développer des moteurs et des services destinés non plus seulement au grand public (avec des approches généralistes) mais aussi aux professionnels de différents secteurs comme des spécialistes de la veille sécuritaire et économique, de la recherche scientifique et technologique, de la communication (journalistes, analystes, enseignants, chercheurs en sciences sociales, étudiants, …; Ihadjadene 2004).

[2] « *Pour retrouver une boutade de Colette sur Jean Sébastien Bach, il suffit de taper « colette bach ». Pas de problème, un « Quid » commode. Mais « si je cherche plus vaguement à comprendre qu'une question lourde, par exemple si la démocratisation favorise ou non l'égalité, j'aurai des centaines de milliers de pages à parcourir ... ».* Cité par Jean Noël Jeanneney, *Quand Google défie l'Europe*, 2005, 64.

[3] Sur ce thème, on peut consulter, entre autres, l'ouvrage de Corine Cohen, *Veille et intelligence stratégiques*, Hermès, Paris, 2004, en particulier le chapitre 1 : « Surveiller l'environnement, une nécessité absolue pour les entreprises ». On peut rappeler : « se faire battre est excusable, se faire surprendre est inexcusable. » (Napoléon Bonaparte), 16.

[4] Projet franco-allemand au départ, il est devenu trop centré, selon nous, sur des approches audio-visuelles, négligeant ainsi les approches centrées sur les textes qui sont pourtant essentiels dans toute recherche d'informations.

Trop d'informations tue l'information. En effet, avoir accès à toutes les informations disponibles sur le Web ne donne pas toujours les avantages que l'on croit car alors on risque d'être noyé par ce trop grand flux d'informations, plus ou moins brutes, plus ou moins fiables[5] et plus ou moins sélectionnées et ordonnées par des algorithmes[6] (Brin *et al.* 1998), dont les coefficients sont éventuellement biaisés pour guider systématiquement l'utilisateur vers des sites privilégiés, en lui faisant éviter également d'autres sites, sous la pression de critères commerciaux et parfois politiques ou idéologiques[7]. Sans des outils sélectifs performants et plus affinés, on risque de devenir assez vite démuni pour trouver et sélectionner les « bonnes » informations fiables, pertinentes et utiles[8] pour se laisser uniquement guider par les réseaux dominants d'influence et par les idées les plus « à la mode », en tuant ainsi toute capacité d'innover, notamment dans les secteurs technologiques et scientifiques, et en laissant ce soin à quelques grands Centres qui, eux, seraient détenteurs des « bonnes informations » et sauraient les utiliser. Or, ceux qui ne détiennent pas à temps ces « informations utiles » deviennent assez vite incapables de fournir les réponses vraiment adaptées aux défis qu'ils doivent affronter, étant engloutis de plus en plus sous des masses d'informations désorganisées, donc difficiles à interpréter, à classer et à archiver. En effet, les informations ne deviennent des connaissances utiles que lorsque ces informations ont été sélectionnées auprès de sources réellement fiables, puis ensuite catégorisées, structurées et reliées les unes aux autres ; par ce biais, elles deviennent interprétables. Il faut donc réussir à construire des technologies plus performantes que les actuelles technologies pour extraire les informations pertinentes et utiles des grands réseaux comme le Web, ou d'autres fonds documentaires archivés, par exemple ceux des bibliothèques, en éliminant le plus possible le bruit qui les brouille, afin de les organiser en véritables réseaux de connaissances diversifiées, effectivement accessibles et donc utilisables au moment où l'on en a besoin.

2. ÉTAT ACTUEL DES MOTEURS DE RECHERCHE

Comme la presse s'en fait souvent l'écho, un constat doit être dressé : Google est un moteur de recherche qui a su complètement supplanter tous les autres moteurs de recherche (Altavista, Excite, …)[9], donnant ainsi une suprématie

[5] Dans les secteurs hautement compétitifs, il est parfois lancé volontairement des informations volontairement erronées pour détourner l'attention des concurrents éventuels.

[6] Par exemple le célèbre algorithme itératif PageRank de Brin et Page (1998), qui exploite des hyperliens entre pages. Cet algorithme s'appuie sur un postulat que la présence d'un hyperlien dénote un appariement sémantique entre deux pages. Or, cette hypothèse n'est pas toujours vérifiée.

[7] C'est le cas de l'utilisation de Google et de Yahoo en Chine.

[8] Sur la notion d'information pertinente et la valeur d'une information, voir Corine Cohen, *Veille et intelligence stratégiques*, Paris, Hermès, 116-120.

[9] Pour des raisons techniques évidentes mais aussi pour des raisons d'innovation en marketing en s'assurant un concours financier de la publicité pour des entreprises commerciales qui paient une redevance uniquement quand leur site est consulté.

complète à ce type de moteur dans le monde entier. Un certain nombre de hauts responsables[10] ont signalé depuis longtemps l'importance culturelle et évidemment économique de ne pas dépendre uniquement d'indexations opérées sur quelques grandes bibliothèques du monde (Stanford, Chicago, Berkeley, Oxford, …) dont seraient exclues les grandes bibliothèques du monde non anglo-saxon, ce qui rendraient ces dernières pratiquement inutiles car étant devenues non accessibles aux moteurs de recherche les plus utilisées. Cela aurait pour effet de mutiler complètement le savoir et d'aller même contre « la mondialisation du savoir » tant de fois proclamée et réclamée. De plus, les informations déposées dans les fonds bibliothécaires devraient être pratiquement accessibles, grâce aux technologies informatiques, au plus grand nombre, y compris dans les pays encore peu développés sur le plan économique. Or, les moteurs actuels filtrent implicitement l'accès à certaines informations et bloquent l'accès à d'autres informations par un choix subtil de mots clés, présents ou absents, et par un jeu de coefficients introduits dans les formules statistiques de l'algorithme de recherche. L'accès aux informations, qu'elles soient économiques, culturelles ou scientifiques, devrait-il dépendre uniquement de groupes commerciaux qui, selon les fluctuations de la Bourse ou les changements politiques, orienteraient les utilisateurs vers des sites privilégiés au détriment de sites devenus muets car pratiquement inaccessibles par l'algorithme de recherche ?

2.1. Principales caractéristiques des travaux actuels

La prise de conscience des européens, avec, entre autres, le programme Quaero, montre bien l'enjeu du problème, même si la réponse n'est peut-être pas réellement adaptée au défi qu'il lui faudrait relever. Les actuels systèmes de recherche d'informations (SRI) (Salton 1971) (Van Rijsbergen 1975) (Grossman *et al.* 1998) (Frakes *et al.* 1992) fonctionnent à l'aide d'un certain nombre d'hypothèses et de modèles implicites. Les modèles sous-jacents restent fortement inspirés par celui des systèmes documentaires où la langue est plutôt considérée comme un système de nomenclatures dans lesquels les termes linguistiques (essentiellement des noms ou des expansions de noms en syntagmes nominaux) représentent des objets, qualifiés parfois de « concepts », d'un monde référentiel constitué essentiellement d'objets et de classes d'objets, les objets appartenant aux classes et les classes étant liées les unes dans les autres par des relations d'inclusion, les objets d'une classe héritant alors des propriétés de la classe supérieure. Les systèmes documentaires se sont développés à partir de larges thésauri et d'index terminologiques portant sur des domaines précis, bien circonscrits et fermés. Si les indexations ont montré leur utilité dans les recherches documentaires, surtout lorsqu'elles sont informatisées, elles font apparaître, en même temps, certaines limites dans la recherche d'informations qui portent sur les contenus et non plus

[10] Par exemple, Jean-Noël Jeannenet, Président de la Bibliothèque Nationale de Franc (BNF).

uniquement sur la seule identification d'ouvrages et de documents sélectionnés par des mots clés. La plupart des moteurs de recherche ont repris implicitement ce modèle documentaire en partant de mots clés et d'index qui, par des voies essentiellement statistiques, vont rechercher tous les documents qui seraient censés répondre le mieux à une requête réduite, la plupart du temps, à une combinaison booléenne de mots clés. La recherche s'effectue alors par un appariement entre les mots clés de la requête et les mots clés qui indexent les documents. Le nombre de documents rapatriés sur des dépôts comme le Web est souvent très important (en moyenne des milliers, voire des centaines de milliers de réponse) et ainsi, dans la pratique, ingérable par l'utilisateur qui se contente souvent de consulter uniquement les dix ou vingt premières réponses. Certes, des algorithmes contribuent à ordonnancer les réponses mais c'est là qu'intervient un jeu de coefficients, plus ou moins transparents et non publics, qui orientent considérablement la recherche à partir des critères statistiques (fréquence des termes, degré d'importance de référence par liens hypertextes à des documents, degré de pertinence pour une communauté constituée selon des critères d'usage…), en prenant en compte parfois d'autres facteurs, pas toujours explicités. Ainsi, l'importance de la publicité joue un jeu d'orientation puisque les entreprises peuvent acheter des mots clefs à Google, ce qui oriente systématiquement chaque utilisateur vers certains sites ou lui fait éviter d'autres sites devenus pratiquement inaccessibles sous des pressions politiques[11]. Pour concurrencer Google sur son propre terrain, d'autres moteurs se mettent en place. Par exemple, en France, le moteur Exalead offre certaines extensions qui proposent des liens de voisinage autour des mots clés de la requête avec des améliorations dans la présentation des résultats ; si un tel moteur tend à présenter certains avantages, il n'y a pas cependant une véritable rupture technologique et théorique dans la recherche d'informations puisque celle-ci s'effectue toujours avec des termes linguistiques, qui identifient des objets ou des classes d'objets, et non pas avec des relations sémantiques plus générales, exprimées par les langues. En particulier, aussi bien dans les systèmes documentaires que dans les extensions proposées actuellement par les moteurs de recherche, il y a une distinction importante entre les mots dits « vides » (de toute signification) et les mots dits « pleins », qui seraient les seuls à porter une information ; les mots vides correspondent aux unités grammaticales (articles, préposition, lemmatisation), les mots pleins aux lemmes des unités lexicales. Les recherches s'effectuent uniquement sur les mots pleins qui seraient les seuls à porter des informations. Or, une telle distinction, qui est mise en œuvre dans les traitements, est hautement contestable du point de vue de l'analyse sémantique des langues. En effet, les mots dits vides sont des unités linguistiques qui structurent, organisent et donnent une plus ou moins grande pertinence sémantique aux informations extraites des textes.

[11] On peut citer, par exemple, la décision de Google de ne pas indexer des documents qui seraient inopportuns pour certains gouvernements.

Beaucoup de recherches dans le monde s'orientent vers le Web Sémantique destiné à offrir de nouveaux services aux utilisateurs (Berners-Lee *et al.* 2001) (Gruber 1993) (Aussenac-Gilles 2005). Un des courants principaux du Web Sémantique fait appel à des ontologies de domaines. Une ontologie[12] est une description formelle d'un domaine sous la forme d'un réseau de concepts (par exemple, le domaine de la restauration, le domaine des voitures, un sous-domaine de la médecine, les domaines de la géologie ou de l'astrophysique…). Or, la plupart de ces ontologies reposent sur des réseaux de concepts identifiés par des termes linguistiques, les éléments structurant de la langue ayant été systématiquement évacués car étant considérés comme des termes « vides de toute signification », ce qui ne correspond pas, comme nous venons de le dire, aux analyses syntaxiques, grammaticales et sémantiques de la linguistique contemporaine. La constitution des ontologies se heurte à plusieurs obstacles. En tout premier lieu, il faut remarquer que construire une ontologie implique un coût économique important puisque il faut recourir à des experts[13] de chaque domaine, cet expert devant être lui-même accompagné souvent par un informaticien – ou un spécialiste des réseaux sémantiques développés en IA – pour élaborer le réseau des concepts intégrés dans l'ontologie du domaine étudié ; il faut également prévoir un spécialiste qui doit être capable d'assurer la maintenance de l'ontologie au fur et à mesure de l'évolution et des développements du domaine. Ensuite, dès que les ontologies deviennent complexes pour tenir compte réellement des domaines modélisés, l'utilisateur, qui est un spécialiste d'un de ces domaines, n'a pas vraiment accès à la complexité du réseau, à moins d'être bien formé ou d'être accompagné, dans chacune de ses recherches et consultations, par un informaticien ; en effet pour lui, qui est un spécialiste du domaine modélisé par l'ontologie, il n'y a plus une « continuité sémantique » entre la sémantique formelle sous jacente à l'ontologie et la sémantique qui est la sienne[14]. En revanche,

[12] « Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts – p.ex. entités, attributs, processus –, leurs définitions et leurs interrelations. On appelle cela une conceptualisation (…) Une ontologie peut prendre différentes formes mais elle incluera nécessairement un vocabulaire de termes et une spécification de leurs spécification. » in T.R. Gruber, « A translation approach to portable ontology specifications », *Knowledge Acquisition*, 5,199-220, 1993, cité (127) par Roger T. Pédauque, *Le document à la lumière du numérique*, C&F éditions, 2006.

[13] La prise d'expertise sera bien meilleure, et la modélisation qui en résulte également, lorsque l'expert est très compétent. Or, un expert compétent est bien peu disponible, car appelé à de multiples tâches, peu enclin à faire part de son domaine d'expertise et de son savoir faire, pas toujours apte à expliquer le réseau de concepts et de procédures qu'il met en œuvrer dans ses décisions et pratiques, pas toujours intéressé à valider des ontologies écrites dans un langage abstrait dont il ne possède pas toujours bien la grammaire et la syntaxe. Etant peu disponible, il est donc « cher » et ce coût augmente le coût général de la construction d'une ontologie. Le recours à des experts de moins grande qualité conduit souvent en contre partie à des constructions qui doivent être sans cesse révisées avant de devenir opérationnelles.

[14] Par exemple, le moteur Swoogle présente ce genre d'inconvénients.

un contact direct avec les textes relatifs au domaine étudié ou avec des informations textuelles portant sur ce domaine permettrait à cet utilisateur de maintenir cette « continuité sémantique » entre les connaissances qu'il recherche et vise à organiser et les modélisations et représentations formelles qui sont nécessaires aux recherches automatiques par les moteurs. Or, les ontologies, du moins telles sont constituées actuellement, ont tendance à rompre toute continuité sémantique. Il est vrai que maintenant, beaucoup de recherches destinées à construire les ontologies de chaque domaine, prennent pour point de départ des annotations manuelles ou semi-automatiques de segments textuels ou d'images, issues de travaux collaboratifs. Des algorithmes, fondés essentiellement sur l'apprentissage, ont alors pour tâche, dans un second temps, à construire automatiquement les ontologies. Cependant, cette approche se heurte à une difficulté majeure : celle de la diversité des annotations manuelles, les critères n'étant pas toujours les mêmes, les annotateurs produisant parfois des erreurs qu'il faut savoir détecter pour les corriger.

2.2. Stratégies classiques du TAL (Traitement Automatique des Langues)

Beaucoup de systèmes de recherche d'informations visent à améliorer les traitements linguistiques de base en introduisant dans les requêtes plus de morphologie et de syntaxe. Certaines recherches se sont orientées vers une expansion morphologique des mots clés en tenant compte des racines, des flexions, des désinences (Poibeau 2003). L'expérience a montré, tant sur l'anglais que sur le français, que ces expansions morphologiques ne donnaient pas les améliorations attendues et espérées et même, dans certains cas, dégradaient les résultats (Ihadjadene 2004). D'autres travaux dans la recherche d'informations se sont inscrits dans une perspective plus générale où le problème revenait d'abord à analyser la requête d'un utilisateur en lui donnant la possibilité de l'exprimer dans une langue naturelle, éventuellement contrainte, de façon à procéder ensuite dans le traitement: 1°) à une analyse morphologique (avec un recours à un dictionnaire informatisé) ; 2°) à une analyse syntaxique pour désambiguïser certaines expressions ; 3°) à la transformation du résultat obtenu en une requête exprimée dans un langage formel directement compatible avec les indexations de la base des documents interrogés. Cette stratégie utilisée dans le TAL reprend le paradigme informatique de la compilation des langages de programmation de haut niveau, paradigme qui a été très utilisé dans les programmes de traduction automatique[15]. Cette approche, aussi intéressante qu'elle soit, nécessite néanmoins des ressources très lourdes : constitution de dictionnaires informatisés exploitables par les analyseurs morphologiques et syntaxiques. Or, malgré plus de cinquante ans de

[15] On peut signaler à ce propos, le modèle de traduction automatique mettant en œuvre des analyses en différents niveaux depuis la langue source jusqu'à un « langage pivot » (une sorte d'interlingua) puis la synthèse vers la langue cible.

travaux dans ces domaines, souvent financés par des projets européens de grande envergure, il n'existe pas encore d'analyseurs morphologiques entièrement fiables (TreeTagger ou l'analyseur de Brill présentent encore de trop nombreuses erreurs sur le français). Quant aux analyseurs syntaxiques, associés à une multiplicité de modèles syntaxiques (grammaires de dépendance, grammaires fonctionnelles grammaires catégorielles, HPSG, TAG, …), ils sont loin de « couvrir » la totalité d'une langue et fournissent souvent, à côté de bonnes solutions, plusieurs analyses parasites qui doivent alors être éliminées par des traitements sémantiques ultérieurs ou situés en amont, ce qui nécessite alors des dictionnaires où la sémantique doit jouer un rôle de plus en plus structurant, et donc nécessite la constitution de ressources de plus en plus lourdes, nombreuses et difficiles à organiser de façon homogène[16]. Il est alors parfois nécessaire d'effectuer des prétraitements manuels ou des post-traitements humains, ce qui augmente le coût d'un système, par exemple de traduction automatique, et réduit son efficacité.

Il est certain que, actuellement, les recherches d'informations portent également sur des supports audio-visuels. Certains pensent même que le textuel aura peu d'importance dans l'avenir. Aussi des recherches s'orientent-elles uniquement vers des recherches multimédias en négligeant totalement l'apport du textuel. Or, il faut remarquer que bien souvent l'indexation d'images est réalisée à partir de documents textuels[17]. De plus, les supports textuels sont et seront toujours, selon nous, un vecteur de l'information et des connaissances. On ne peut pas non plus négliger tous les documents textuels qui sont archivés dans les bibliothèques qu'il faut pouvoir interroger et exploiter. Aussi, en privilégiant trop la dimension audio-visuelle, ne risque-t-on pas de se couper de tout un savoir accumulé et qui deviendrait ainsi pratiquement inaccessible aux recherches automatisables ? Du reste, la réussite incontestable de Google repose plus sur l'accès aux contenus textuels et beaucoup moins sur des contenus multimédias. Certes, il faut maintenant savoir articuler texte et image, par exemple en sachant associer automatiquement les commentaires aux images qui leur correspondent. Quant au traitement direct des images et des représentations figuratives (diagrammes, plans, graphiques, histogrammes), il pose d'énormes problèmes sémiotiques et techniques qui sont de loin beaucoup moins avancés que les traitements sémiotiques et sémantiques opérant sur des textes avec une pratique pratiquement deux fois millénaire. Quant au traitement de la parole, il a fait des progrès incontestables depuis 20 ans. Il en a résulté de nouvelles approches et techniques (réseaux neuronaux, méthodes markoviennes et probabilistes, N-grammes) qui ont permis d'obtenir de très bons résultats dans de nombreuses applications finalisées. Bien

[16] Là encore, le coût de ces ressources est énorme, en particulier dans une approche multilingue.

[17] Voir, à ce propos, la thèse de Shanaz Benhami, soutenue à Paris-Sorbonne. La thèse montre comment on peut apparier automatiquement, dans un texte, des commentaires textuels aux images, diagrammes, icônes. L'indexation peut donc s'effectuer à partir de ces segments textuels, et non pas directement à partir des images….

qu'il demeure encore de nombreux problèmes techniques (par exemple reconnaissance de locuteurs différents), il serait souhaitable d'établir de sérieuses passerelles entre le traitement de la parole et la sémantique. En particulier, il serait intéressant de pouvoir développer, dans le cadre de la recherche d'informations, des programmes informatiques qui permettraient de poser des questions orales, transformées en requêtes formalisées à partir desquelles la recherche pourrait s'effectuer, aussi bien sur des documents textuels que, éventuellement, sur des requêtes orales. Dans d'autres applications, il serait intéressant également d'oraliser certaines des réponses fournies par des systèmes de recherche qui opéreraient sur des documents textuels. Le passage au « tout audio-visuel » risque d'entraîner des désillusions sévères avec des conséquences économiques non négligeables. Les déboires de la bulle Internet sont là pour nous le rappeler.

2.3. La recherche d'informations par les SRI actuels

Comment s'effectue actuellement la recherche d'informations dans de grandes bases de documents textuels ou dans le Web ? Par des moteurs de recherche qui visent à apparier ou à rapprocher, par différents moyens, d'une part, les questions posées et d'autre part, des documents qui sont censés donner les réponses à ces questions. Ainsi, Google qui est utilisé pour rechercher des informations sur le Web, donne en réponse à une question posée sous la forme de termes linguistiques (éventuellement connectés entre eux par des opérateurs logiques comme ET, OU, SAUF) une liste de sites et de documents, classés dans un certain ordre, avec des extraits textuels qui sont des illustrations de réponses censées correspondre à la question posée. L'ordre de classement est lié à des formules mathématiques qui expriment une fonction ayant pour arguments des fréquences des termes composant la requête ainsi que d'autres facteurs non révélés.

Prenons un exemple. Nous voulons savoir, en explorant par exemple les journaux de l'année écoulée : *Qui Chirac a-t-il rencontré en cette période de décembre 2005 ?* Cette question ne peut pas être posée sous cette forme directe. Elle sera posée sous la forme suivante (un ensemble de mots clés) : « Chirac rencontre décembre 2005 », c'est-à-dire sous la forme d'une conjonction de termes linguistiques. Le Système de Recherche d'Informations (SRI) va faire appel à Google qui renverra une liste impressionnante de documents censés donner des réponses à la question posée, c'est à dire en fait les documents dans lesquels on a des occurrences des mots « Chirac », « rencontre », « décembre » et « 2005 ». L'utilisateur doit alors explorer un à un ces documents pour y chercher les réponses souhaitées. Dans la liste des documents proposés, un certain nombre de réponses ne correspondent pas du tout à la question. Par exemple, certains documents renvoient à

« rencontres à Chirac en Corrèze » ;
« clubs de rencontre à Chirac » ;
« clubs de rencontre pour soutenir Chirac » ;
« rencontre de football en présence du Président Chirac »…

mais aussi à des références où *« Chirac »* et *« rencontre »* ne se trouvent pas dans la même phrase ni forcément proche l'un de l'autre dans le texte, comme dans la réponse suivante :

> « ... Point de presse conjoint de M. Jacques Chirac et de Mme Angela Merkel [Voltaire] ... Cela dit, pour être précis, la rencontre entre la Reine Louise et ... »

où *« Chirac »* appartient à une phrase et *« rencontre »* à une autre sans qu'on puisse y voir le lien sémantique qui relierait ces deux phrases. Il s'agit là d'un exemple de la non pertinence de certains documents proposés. Le nombre de documents proposés en comporte souvent plusieurs centaines de milliers. Il s'agit du bruit que tout SRI de qualité se doit de diminuer dans de fortes proportions. Alors que les systèmes de recherche documentaire, dans un fonds fermé, se soucient essentiellement de réduire le silence (ne pas oublier un document !), les systèmes ouverts sur le Web devraient, eux, se concentrer sur la réduction du bruit car, comme nous l'avons déjà dit, trop d'informations tue l'information utile. En effet, un utilisateur n'a pas toujours le temps et la patience de consulter tous les documents proposés pour en évaluer ensuite la pertinence. Or, dans sa recherche des informations sur des bases de documents textuels, l'utilisateur souhaite donc :

- pouvoir poser sa question de façon naturelle et précise ;
- obtenir en réponse non pas seulement des documents qu'il lui faudra ensuite examiner pour exploiter des informations qu'il contiennent, mais également des extraits de documents où il trouve déjà une réponse qui ainsi devient directement exploitable en lui permettant également d'apprécier la pertinence de la réponse ;
- obtenir uniquement des réponses pertinentes, en nombre raisonnable (en évitant éventuellement les redondances) et, dans certains cas, classées selon certains critères explicites que l'utilisateur connaît et peut éventuellement choisir en fonction de ses besoins.

S'intéressant à la recherche d'informations, un utilisateur souhaite parcourir des documents textuels en privilégiant un « point de vue de fouille ». Dans notre exemple, l'utilisateur privilégie le point de vue de la « rencontre », c'est-à-dire u'il recherche l'existence d'une certaine connexion entre des personnes ; plus spécifiquement, il s'agit d'identifier toutes les connexions de *« Chirac »* avec d'autres personnes. L'utilisateur n'est pas donc intéressé par les seules co-occurrences du terme *« Chirac »* et du terme linguistique *« rencontre »*. En effet, il souhaite obtenir également des réponses comme dans la liste (1) :

« *Chirac a déjeuné avec le premier ministre* » ;

« *Chirac a accueilli le premier ministre de Grande Bretagne à l'Elysée* » ;

« Monsieur *Chirac a eu un entretien avec La Chancelière de l'Allemagne* » ;

« *Le Ministre de l'intérieur s'est rendu à un rendez vous du Président Chirac* » …

Ces segments textuels donnent des informations tout aussi importantes que :

« *Rencontre de Chirac avec le Prix Nobel de la Paix* ».

Or, un moteur de recherche « classique » identifie les seules co-occurrences dans des documents des termes linguistiques *« Chirac »* et *« rencontre »* et non pas des termes apparentés (*a déjeuné avec, a accueilli, a eu un entretien avec, s'est rendu à un rendez-vous*) ; il est donc incapable de repérer directement des documents où ces deux termes ne seraient pas présents et donc de fournir des documents qui contiendraient des informations comme celles de l'exemple. Il doit classer ensuite les réponses obtenues selon divers critères liés, pour la plupart, à des fréquences d'emplois (fréquences d'occurrences dans un document, fréquences d'utilisation …). Ce qu'attend un utilisateur intéressé par le point de vue « rencontre » relativement à *« Chirac »*, c'est justement de retrouver des informations comme celles de la liste (1), en posant une question sous la forme d'une instanciation d'un schéma relationnel comme :

Q : *« Chirac a rencontré Qui ? » OU « Qui a rencontré Chirac ? »*

Pour résumer, nous pouvons dire qu'un moteur classique comme Google fonctionne sur des bases essentiellement statistiques avec la seule identification de termes linguistiques (des mots simples ou composés) et l'utilisation des liens hypertextuels en s'appuyant sur le principe général : l'importance d'un document, et donc le poids qui lui est attribué, est fonction du nombre de ses citations dans d'autres documents. Le modèle sous-jacent aux moteurs de recherche actuels fonctionnant sur le Web (système ouvert) est hérité essentiellement de la recherche documentaire fonctionnant sur des systèmes fermés à partir de principes fondés sur une conception positiviste qui remonte à l'école de Vienne des années trente. De telles recherches font alors appel à :

- une reconnaissance de termes linguistiques (des mots, des syntagmes nominaux) désignant la plupart du temps des entités nominales et combinés par des opérateurs booléens ;
- une certaine similarité entre les termes de la question et ceux des termes descripteurs des documents (par exemple des mots clés) ;
- à des organisations de descripteurs (appelées thesauri) reliant des descripteurs plus génériques à des descripteurs plus spécifiques : lorsque les termes de la question correspondent exactement à une suite de descripteurs de documents, ces documents sont censés répondre à la question posée et ils sont sélectionnés.

Ainsi, lorsqu'un document possède dans sa description les deux descripteurs *« Chirac »* et *« rencontre »*, l'adresse de ce document est retournée par le moteur de recherche en réponse à la requête *« Chirac rencontre »*. Cependant, il n'y a pas toujours coïncidence entre les termes linguistiques de la requête et les descripteurs des documents enregistrés. Prenons un autre exemple avec une recherche d'informations sur la définition d'un raz de marée (*qu'est-ce qu'un raz de marée ? Comment définir et caractériser un raz de marée ?*). La question posée, à l'aide d'un moteur

du type de Google pourra être : *« "Raz de marée" définition »*. Dans ce cas, les documents obtenus sont très nombreux (42 100 documents renvoyés) et pas toujours pertinents. En effet, les deux premiers résultats renvoient à des sites encyclopédiques comme « Wikipedia », le troisième résultat n'a aucune relation avec la notion de définition (il renvoie au site www.ffsa.fr/webffsa.nsf/html/rzdemarree).

3. VERS DE NOUVEAUX MODES D'INTERROGATION

Peut-on envisager une autre stratégie de recherche d'informations en faisant appel à une analyse linguistique et à une meilleure compréhension de l'organisation discursive des textes ? Il s'agit d'innover en imaginant, en modélisant et en réalisant effectivement une nouvelle génération de Systèmes de Recherche d'Informations (SRI) fondés sur des *annotations sémantiques automatiques de textes* qui opéreraient en prenant appui sur une théorie sémantique de la linguistique cognitive contemporaine. Ces nouveaux SRI, qu'il faut construire, doivent être à même de :

- mieux satisfaire les utilisateurs en leur permettant de poser des requêtes sous forme de relations sémantiques déterminées par les points de vue de fouille choisis ;
- effectuer des recherches automatiques à partir des « contenus sémantiques » et non pas à partir des seuls « termes linguistiques » (c'est-à-dire des formes linguistiques plus ou moins fréquentes dans les documents) ;
- donner des réponses par extraction des informations pertinentes en rapport avec le point de vue de fouille adoptée et sous forme de segments textuels.

Reprenons notre exemple qui nous sert de support illustratif à notre démarche. En interrogeant l'ensemble des journaux de la presse, nous aimerions poser la question *« Chirac et Poutine se sont-ils rencontrés ?»* en souhaitant, éventuellement, des précisions complémentaires *« Où ? »* et *« Quand ? »*. En posant la question au système MOCXE[18] que nous développons actuellement au laboratoire LaLIC (Desclés 2006, Djioua *et al.* 2006, Desclés, 2007, Djioua *et al.* 2007):

(Q) : *« Chirac, rencontre, Poutine ? »*

le système répond immédiatement à partir d'un corpus de textes annotés donnés en entrée. Nous présentons la copie de l'écran des réponses (voir la copie d'écran 1) [19]. Commentons les réponses obtenues. Le système a sélectionné quatre segments textuels annotés par le point de vue « rencontre » et contenant les termes

[18] Le système MOCXE est un moteur de recherche qui opère avec des annotations construites automatiquement par un moteur d'annotations EXCOM. Nous reviendrons plus loin sur ces deux machines et sur leur architecture informatique.

[19] Voir l'annexe où se trouvent les copies d'écran.

« Chirac » et *« Poutine »*. Ces segments sont regroupés par le document auquel ils appartiennent. La première réponse relative au premier document est détectée par la présence de « *arrivant côte à côte* » ; la seconde par « *se sont rencontrés* » ; la troisième par « *rendre visite* » et la dernière est relative au deuxième document est identifié par « *visite de travail* ». Il est clair que ces éléments font partie de ressources linguistiques constituées avec l'analyse linguistique de la notion de « rencontre ». Il est possible, bien entendu, d'accéder au document annoté à partir des réponses, par un lien hypertextuel. Ainsi, pour la deuxième réponse renvoyée, nous obtenons la copie d'écran 2 où sont indiqués (coloriés) tous les segments annotés. On peut y reconnaître le document qui a été annoté automatiquement par la « rencontre »[20]. Dans ce document sont surlignés et coloriés les segments identifiés à partir de marqueurs linguistiques de la « rencontre ». Certains segments sont relatifs à la question posée (**Q**) : *« Chirac, rencontre, Poutine ? »* et d'autres segments sont uniquement relatifs à la notion générale de « rencontre ». Supposons maintenant que nous posions la question (**Q**) : *« Qui Chirac a-t-il rencontré ?» OU « Qui a rencontré Chirac ?»*. La copie d'écran 3 présente sept réponses sur les neuf réponses à cette question. Le système MOCXE[21] permet d'affiner la question en demandant uniquement les rencontres qui sont relatives à des événements (catégorie « rencontre événementielle »). En posant une telle question, nous obtenons deux réponses extraites d'un seul document, celles-ci sont présentées dans la copie d'écran 4. On remarquera que de telles réponses ne peuvent pas être obtenues à partir de requêtes acceptées par Google ou par des moteurs de recherche équivalents.

L'exemple précédent montre que toute la recherche est articulée d'une part, autour de la notion de « point de vue de fouille » et d'autre part, autour de la notion de « segment textuel annoté » à l'aide de marqueurs linguistiques liés explicitement à l'expression de la notion du point de vue de fouille adopté. Alors que dans les recherches plus traditionnelles, la sélection est opérée par identification d'occurrences de mots clés présentes à la fois dans les documents et dans la requête, notre approche s'appuie sur des expressions linguistiques relevant de champs sémantiques dûment constitués par des points de vue (par exemple le point de vue de la « rencontre »). Il ne s'agit plus maintenant d'identifier seulement les termes d'une requête avec leurs occurrences dans un document mais de déclencher des recherches à partir d'expressions sémantiquement déterminées par le point de vue de fouille, ce qui amène à identifier des segments textuels entiers qui ainsi peuvent être annotés puis indexés. Notre méthodologie n'impose donc pas qu'un thesaurus des termes ait été préalablement constitué ; elle ne nécessite pas de faire appel à une ontologie d'un domaine particulier ; elle ne suppose pas non plus, avant tout traitement, le repérage d'entités nommées spécifiques à un domaine particulier. Dans l'exemple précédent, les termes « Chirac » et « Poutine » ont été introduits

[20] Par la machine EXCOM.
[21] Nous verrons plus loin une présentation de MOCXE.

comme des termes de la requête et non pas comme des entités nommées devant être préalablement identifiées par un processus qui ferait appel à des ressources spécialisées. Remarquons cependant que notre approche de la recherche d'informations reste compatible avec des ressources supplémentaires, à condition toutefois qu'elles soient disponibles, sous forme d'ontologies particulières et de bases d'entités nommées[22]. Ainsi, supposons que nous disposions d'une table d'équivalence entre « Chirac » et ses différentes expressions dénotatives et titres relevant de sa fonction (« Mr. Chirac », « Monsieur Chirac », « Monsieur le Président de la République », « le locataire de l'Élysée », « le chef des armées », « le chef de l'Etat », « le premier personnage de l'Etat », « le garant de la Constitution », « le premier magistrat de France »…), alors la recherche peut être étendue à partir de ces nouvelles expressions, ce qui augmentera le nombre et la pertinence des réponses.

3.1. Systèmes concurrents

On retrouve dans la littérature plusieurs tentatives d'utilisations d'outils TAL dans le processus semi-automatique de peuplement d'ontologies de domaines à partir de textes (Cunningham *et al.* 2002) (Handschuh *et al.* 2004). La plupart des ces outils s'appuient, dans la majorité des cas, sur le repérage d'entités nominales en les reliant par des associations de type « est-un ». Ces nombreuses applications de TAL, reposent sur des méthodes d'apprentissage et des méthodes inspirées du processus de compilation. Mais ces méthodes nécessitent la constitution préalable de corpus étiquetés, ce qui est souvent un obstacle important. Ces applications en vraie grandeur butent souvent sur la question des ressources sémantiques nécessaires à des performances satisfaisantes. Des thésaurus (comme MeSH dans le domaine médical), des réseaux lexicaux (comme WordNet et ses versions européennes), des réseaux sémantiques (comme UMLS) ont souvent été mis à profit par divers systèmes, mais leur intégration ne va pas sans demander des investissements importants. Parmi ces tentatives, nous nous intéressons à deux plateformes déjà utilisées dans des applications industrielles : le couple GATE-KIM avec son ontologie KIMO et le système InFact mis en place sur le site www.globalsecurity.com.

3.1.1. *Plateforme KIM*

La plateforme de KIM (Cunningham *et al.* 2002) (Kiryakov *et al.* 2004) fournit l'infrastructure et les services pour l'annotation, l'indexation, et la recherche d'informations sémantiques de façon automatique. Elle permet de lancer

[22] Ce qui n'est pas toujours le cas, évidemment. Sur le repérage des « entités nommées », souvent préalable à toute recherche d'informations – mais pas dans notre approche –, voir par exemple Thierry Poibeau, *Extraction automatique d'information*, Paris, Hermès, 2003, chapitre 5.

des processus d'extraction basés sur des ontologies de domaines, construites à partir d'annotations effectuées avec le système GATE. Le modèle utilisé dans KIM est basé sur l'ajout massif de métadonnées sémantiques dont a besoin le Web Sémantique. Pour chaque entité, mentionnée dans le texte, KIM fournit les références (URI) (i) à la classe la plus appropriée dans l'ontologie, et (ii) à l'exemple spécifique dans la base de connaissance. En raison de l'annotation sémantique automatique, des métadonnées sont produites et associées à la ressource traitée. Le couple GATE-KIM a pour but de construire automatiquement des instances d'une ontologie par l'extraction d'entités nommées en utilisant le langage à base de motifs réguliers Jape.

Pour illustrer les différences entre les couples GATE-KIM et EXCOM-MOCXE, prenons quelques exemples sur la notion de rencontre :

« Gordon Brown met George Bush during his two day visit. »

« Ten days ago, when Blair was interviewed by the BBC's Jeremy Paxman, the prime minister was asked repeatedly whether he had seen that advice. »

Alors que GATE-KIM va s'intéresser aux entités nommées comme « *Gordon Brown », « George Bush », « Blair » et « Jeremy Paxman »* pour les organiser dans des relations entités-associations liées aux fonctions de président ou de premier ministre ; EXCOM-MOCXE va, à partir de l'étude linguistique du point de vue de la rencontre, repérer automatiquement, dans des segments textuels, les relations de rencontre dans des textes avant de les indexer pour enfin donner la possibilité à un utilisateur d'interroger le moteur de recherche par des catégories sémantiques et non pas par de simples mots clés.

3.1.2. *Système InFact*

Infact (Marchisio *et al.* 2000) est un système fondé sur une approche qui permet de relier un texte avec des informations linguistiques comme les catégories syntaxiques (pat-of-speech), les rôles syntaxiques (sujet, objet, verbes, modifieurs) et les catégories sémantiques (comme personnes, lieux, quantités monétaires). InFact est déjà utilisé dans le moteur de recherche GlobalSecurity.org. L'idée générale de ce modèle est l'indexation d'information syntaxique basée sur des schémas généraux des rôles actanciels régissant les phrases. Ainsi, le module de requête permet de poser des questions suivant une structure d'indexation qui représente la structure <*Subject-Action-Object*>. Une requête comme «George Bush <> * <> *» permet de retrouver des documents comportant des événements reliant Bush comme sujet ou comme objet.

Le modèle utilisé ici repose sur une analyse syntaxique robuste en profondeur. Après l'opération de segmentation de textes en phrases, un processus d'analyse syntaxique profonde est effectué. InFact procède à l'indexation de propositions et repère les catégories syntaxiques et les rôles actanciels de chaque terme. Les relations inter-propositionnelles et des constructions grammaticales sont aussi

utilisées afin de repérer les événements entre les termes. Une des différences fondamentales de ce système par rapport à d'autres comme GATE-KIM est qu'il n'utilise pas de règles ni de motifs préfinis. Il s'appuie sur un analyseur syntaxique de dépendance permettant l'extraction des catégories morpho-syntaxiques (POS Tag) et les rôles grammaticaux de tous les termes dans chacune des propositions des textes analysés. Dans l'opération de découpage de morphèmes grammaticaux pour les verbes, par exemple, InFact normalise en une forme infinitive en gardant les informations temporelles (passé, présent, futur), les valeurs aspectuelles (progressive, perfect) et celles de la modalité (possibility, sybjunctive, irrealist, negated, conditional, causal) afin de les utiliser dans le processus d'indexation. Ensuite vient l'opération de repérage des liens inter-propositionnels à travers (i) marques explicites de conjonctions ou des pronoms qui marquent le lien entre les structures syntaxiques de deux propositions adjacentes dans la même phrase ; et (ii) pointant sur une liste de mots-clés annotés dans la précédente et suivante phrases.

3.2. Architecture informatique pour la recherche d'informations par annotations

La stratégie que nous mettons en œuvre décompose le problème de la recherche d'informations en deux parties : dans un premier temps, nous procédons à l'annotation automatique des documents textuels selon des points de vue de fouille puis ensuite au stockage de ces documents annotés; dans un second temps, nous procédons à l'indexation des segments annotés, de façon à pouvoir ensuite interroger l'index pour fournir, en réponse, des segments annotés qui répondent à une requête et pas simplement des documents susceptibles de contenir les réponses. Le dispositif consiste à coupler les deux machines informatiques (deux moteurs), à savoir la machine d'annotation EXCOM[23] et la machine d'indexation MOCXE[24]. (voir la figure 1) (Djioua *et al.* 2007). La première machine EXCOM prend pour entrée des textes et les annote selon des points de vue de fouille en faisant appel à des ressources informatiques étroitement associées à ces points de vue. Certaines applications directes (comme « le résumé automatique par extraction », « la constitution de fiches selon des points de vue de fouille », « la navigation textuelle ») deviennent alors possibles à partir des textes annotés. Les textes annotés (en particulier ceux venant du Web) peuvent être également des entrées du second système. La deuxième machine MOCXE procède à une indexation des annotations et permet de rechercher les informations à partir de ces annotations et de quelques paramètres donnés en entrée.

[23] Le sigle EXCOM signifie « EXploration COntextuelle Multilingue ».

[24] Les réalisations de la nouvelle machine d'exploration contextuelle EXCOM et celle de la machine à indexer MOCXE sont dûes à Brahim Djioua (2006).

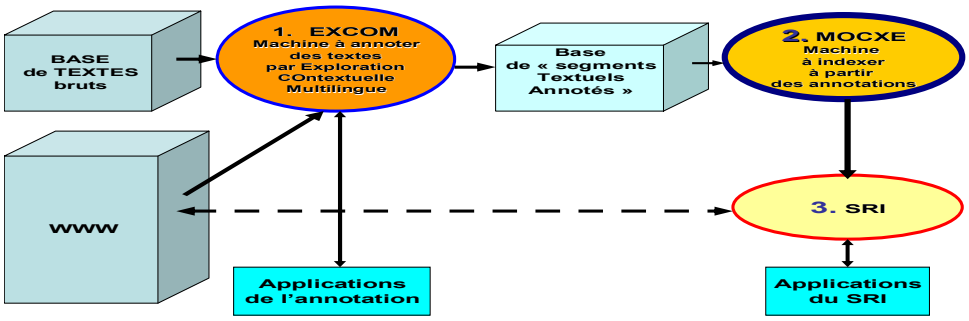


Fig. 1 – Architecture informatique des deux machines EXCOM et MOCXE.

Notre modèle qui prend appui sur des annotations sémantiques automatiques et l'architecture fonctionnelle qui en découle propose un nouveau paradigme d'indexation qui repose sur l'idée centrale : utiliser des « segments clés », identifiés par un processus automatique d'annotations plutôt que des « mots clés ». Cela introduit une certaine innovation par rapport aux techniques classiques d'indexation par les seuls « mots clés ». La structure linguistique minimale d'indexation n'est donc plus, ici, « un terme » – aussi complexe soit-il – mais un segment textuel (une phrase, un paragraphe ou section, un titre…), auquel sont attachées des informations sémantiques discursives et structurées (plusieurs annotations selon différents points de vue, annotation générique d'un point de vue avec, éventuellement, des annotations plus spécifiques …). Pour le point de vue de la « rencontre », nous avons déjà vu un exemple de réponses obtenues à partir d'une indexation par MOCXE (voir les copies d'écran 1, 2 et 4 présentées en annexe).

3.3. Points de vue de fouille sémantique

Comme nous venons de le voir, un utilisateur qui cherche des informations dans un ensemble de documents textuels procède à des fouilles guidées par des points de vue, c'est-à-dire par des lectures focalisées qui privilégient certains segments textuels (par exemple des phrases ou, parfois, des paragraphes entiers) de préférence à d'autres segments. Nous cherchons, dans notre démarche, à reproduire ce que fait naturellement un lecteur humain qui souligne certains segments relatifs à un point de vue particulier qui focalise son attention. Bien entendu, plusieurs points de vue de fouille existent et peuvent co-exister, cela correspond aux diverses focalisations. Tel utilisateur pourra être intéressé par l'identification des « relations de causalité », éventuellement par des relations causales plus spécifiques comme « les causes de la migraine », « les causes d'un raz de marée », « les causes de la grippe aviaire ». Tel autre utilisateur, par exemple un étudiant ou un lycéen de classe terminale, cherchera en fouillant de nombreux textes (encyclopédies, manuels, articles spécialisés) les définitions d'une notion comme par exemple en

sociologie : « classe sociale », en économie : « inflation », en linguistique : « polysémie », en biologie : « génotype »... Un troisième utilisateur cherchera, en consultant la presse des cinq dernières années, à connaître les connexions et les rencontres qui ont pu avoir eu lieu entre deux personnages de l'actualité (par exemple « Poutine » et « Chirac », ou entre « la Secrétaire d'Etat américain » et « le Vice Président des USA »). Un autre utilisateur, par exemple un juge d'instruction, cherchera à établir des connexions, éventuellement par transitivité, entre plusieurs suspects qui déclarent pourtant ne pas se connaître, en établissant, à partir d'informations extraites des textes, que « A connaît B » puisque *A a téléphoné à B* (d'après un certain document de police) et que « A et B connaissent C » puisqu'ils ont *déjeuné ensemble* (d'après un autre document de police)…, ce qui laisse pressentir que « A » pourrait avoir eu quelque connexion avec « C », ces deux individus « A » et « C » étant impliqués, par ailleurs, dans une même affaire de fraude fiscale. Un scientifique dans un laboratoire pharmaceutique cherchera, quant à lui, à extraire des conclusions de divers rapports d'expérience de façon à constituer des fiches consultables ultérieurement, ou encore à déterminer les hypothèses dûment exprimées dans tel ou tel protocole expérimental dont il lit les comptes-rendus. Un journaliste cherchera de son côté à extraire toutes les déclarations d'un personnage politique dont il écrit la bibliographie, ces déclarations étant toutes parues dans la presse de ces dix dernières années.

Ces points de vue de fouille visent une lecture focalisée et l'automatisation d'une annotation éventuelle des segments textuels qui correspondent à une recherche guidée afin d'en extraire les informations recherchées. Chacun des points de vue est explicitement indiqué par des marqueurs linguistiques identifiables dans les textes. Par exemple, nous aurons des marqueurs comme :

- pour les relations de « causalité » : *conduit à* / *favorise* / *entraîne* / *débouche sur* / *a pour conséquence* / *a pour origine* … ;
- pour les relations « définitoires » : *ce qui signifie* / … *est, par définition*, … / *on définit ... par* …, *... veut dire ;*
- pour les relations de « rencontre » : *a rencontré* …/ *a déjeuné avec* / *a été accueilli par* / *a téléphoné à* … ;
- pour les « annonces conclusives » : *nous sommes arrivés à la conclusion que* … / *la conclusion est que*… / *pour terminer.../, pour résumer, ...* ; (Blais *et al*., 2006)
- pour la recherche d'hypothèses : *Mon hypothèse est que* … / *Notre hypothèse est la suivante*… / *Nous avons supposé que ... ;* (Blais *et al*., 2006)

De tels marqueurs linguistiques, comme on peut le voir, restent indépendants des domaines[25]. Ils fonctionnent aussi bien en géologie qu'en mathématiques, dans

[25] Cette caractéristique nous distingue de la plupart des Systèmes de Recherche d'Information actuels qui sont souvent spécifiques à un domaine particulier ou à un genre précis de texte (articles scientifiques, reportages, ouvrage didactique, article de dictionnaire, article d'encyclopédie). En effet,

des textes de journaux, dans des rapports techniques, en sociologie, en économie…(Alrahabi *et al.* 2006) (Bertin *et al.* 2006) (Le Priol *et al.* 2006) La recherche d'informations peut alors s'appuyer sur l'identification de ces marqueurs linguistiques qui mettent en relation des termes linguistiques (qui n'ont pas toujours besoin d'être déterminés et qui peuvent appartenir à différents domaines) de façon à identifier certains segments textuels particulièrement pertinents pour le point de vue de fouille examiné. Ces relations linguistiques expriment des *relations sémantiques* qui structurent le discours selon les intentions du rédacteur et donnent ainsi des instructions de lecture. C'est dans la mesure où telles relations sémantiques laissent des traces discursives dans les documents textuels, qu'il est possible de partir de ces traces pour reconstituer une organisation discursive et en tirer ainsi profit dans le processus de la recherche d'informations extraites des textes. C'est ce principe cognitif et linguistique qui a été mis en œuvre depuis une quinzaine d'années et qui continue à être développé au laboratoire LaLIC pour construire, entre autres, des résumés automatiques de textes.

3.4. Annotations des points de vue

En fouillant un texte, un lecteur humain peut souligner les segments textuels qui correspondent à son point de vue. En cherchant à imiter l'humain, nous pouvons introduire dans un texte des annotations associées aux points de vue de fouille de façon à pouvoir retrouver immédiatement et automatiquement ces informations, tout comme un lecteur humain va pouvoir naviguer entre les segments soulignés par lui (ou par un autre lecteur) et annotés éventuellement dans la marge. L'annotation automatique consiste à introduire dans un texte les annotations qui sont déclenchées par certains marqueurs de point de vue. Le texte s'enrichit ainsi de ces annotations construites à partir d'une analyse sémantique du texte lui-même. Grâce à ces annotations explicites, le texte initial devient beaucoup plus riche car plus exploitable par tout lecteur pressé qui peut ainsi se laisser guider par les annotations reconnues. Une machine peut alors prendre la place de n'importe lequel de ces lecteurs et ainsi extraire les informations qui correspondent aux points de vue de fouille[26]. Nous verrons cependant qu'il faut procéder à certains raffinements techniques[27] pour éviter le bruit avec une annotation qui risquerait d'être trop bavarde ou non pertinente. Il est clair cependant qu'un même segment linguistique peut recevoir plusieurs annotations discursives. Prenons un exemple :

ces systèmes opèrent à partir des « entités nommées » qu'il faut savoir identifier avant toute procédure de recherche. Or, ces entités nommées sont relatives à un domaine. Par exemple, le mot « code » entre dans la composition de plusieurs entités nommées, il renvoie en effet à des objets très différents selon les domaines : code de la route, code informatique, code pénal, code génétique.

[26] Il s'agit, comme le lecteur le reconnaîtra, d'une réalisation partielle du test de Turing fondé sur l'imitation.

[27] Il s'agit de la technique de l'exploration contextuelle qui sera présentée plus loin.

Dans son dernier ouvrage, qui est très important pour comprendre sa démarche, l'auteur a expressément exprimé son hypothèse ainsi : « (...) ».

Un tel segment textuel recevra trois annotations correspondantes à trois points de vue de fouille :

- annotation : « Hypothèse de l'auteur » déclenchée par le marqueur – *son hypothèse ;*
- annotation : « Citation de l'auteur » déclenchée par le marqueur – *a exprimé ... : « ... » -*
- annotation « soulignement du rédacteur » déclenchée par le marqueur – *très important pour comprendre*

La machine EXCOM est destinée à annoter automatiquement des textes en faisant appel à des ressources linguistiques associées à chaque point de vue (voir la figure 2). Un premier programme doit cependant préalablement segmenter les textes en procédant à un découpage qui permet d'identifier les titres, les phrases, les paragraphes[28]…. Le texte étant segmenté, la machine EXCOM annote le texte en utilisant les occurrences des marqueurs linguistiques des points de vue examiné et en tenant compte de l'insertion contextuelle de ces occurrences[29] .

Dans la copie d'écran 5, nous donnons un exemple de sortie avec un texte annoté selon plusieurs points de vue. Les segments annotés, ici des phrases, sont coloriés avec des couleurs différentes pour qu'un utilisateur qui souhaiterait exploiter les différentes annotations puisse les identifier immédiatement et évaluer dans le document la densité d'un type d'annotations. Il s'agit de l'annotation automatique d'un texte scientifique[30]. Le système automatique d'annotation a identifié les phrases qui correspondaient aux points de vue de fouille « annonce thématique » (c'est-à-dire identification des grands thèmes abordés par le texte), « remarques conclusives» (identification des segments textuels qui apportent des remarques conclusives, ces dernières ne se trouvant pas forcément à la fin du texte), « organisation discursive » (segments textuels qui expriment certaines articulations du texte mises en évidence par l'auteur). Ces points de vue peuvent être retenus et utilisés pour construire un résumé automatique du texte obtenu par extraction des seules phrases annotées[31]. On peut chercher également à sélectionner des segments

[28] Ce premier programme SegaText, présenté dans la thèse de Ghassan Mourad (2001) avec une version intégrée à EXCOM par B. Djioua et F. Le Priol, (2006) fonctionne en faisant appel à la technique d'exploration contextuelle.

[29] Les ressources linguistiques sont constituées de classes de marqueurs linguistiques (indicateurs) auxquels sont associés des règles d'exploration contextuelle. Nous y reviendrons plus loin avec la technique linguistique et informatique d'exploration contextuelle.

[30] Il s'agit de la préface de Maurice Gross à l'ouvrage *Notes du cours de syntaxe* de Zellig Harris – Editions du Seuil, Paris, 1976 – annoté automatiquement par EXCOM.

[31] Il est nécessaire de définir les points de vue qui sont nécessaire pour un résumé puis de se donner une stratégie d'ordonnancement des points de vue, en tenant compte notamment des positions des phrases annotées dans le texte (une annonce thématique est en général plus pertinente dans les tout premiers paragraphes; certaines annotations en début de paragraphe annoncent un thème traité

textuels qui contiennent un terme précis, par exemple « *grammaire* », dans le contexte de l'extraction construite pour un résumé. (voir la copie d'écran 6). Le système a annoté également les phrases où certains mots du titre apparaissent (dans notre exemple : *Harris*), toujours dans le contexte du résumé. (voir la copie d'écran 7).

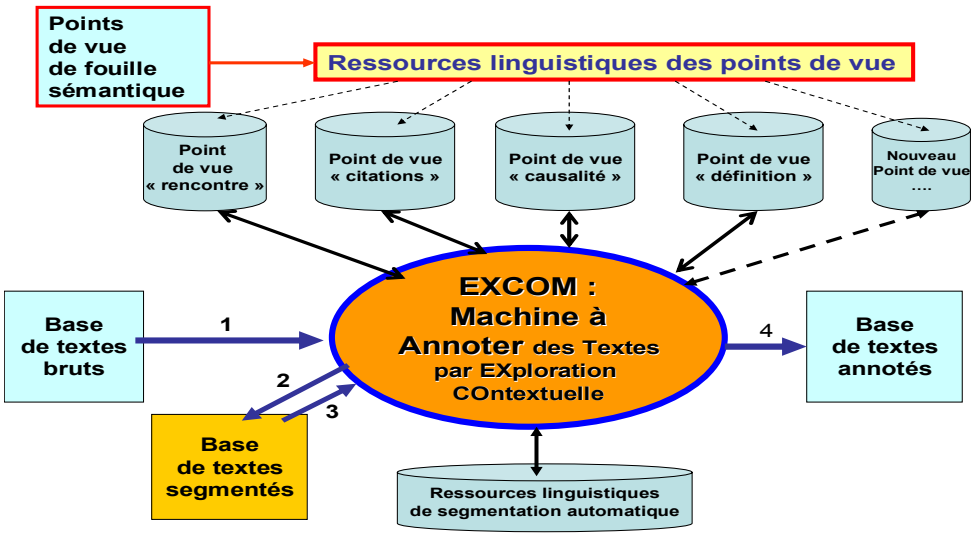


Fig. 2 – Architecture fonctionnelle de la Machine EXCOM à annoter les textes selon des points de vue de fouille

3.5. Cartes sémantiques associées aux points de vue : exemples

Nous avons vu un exemple de recherche d'extraction dans les textes à partir du point de vue ou de la notion de « rencontre » avec le terme « Chirac » choisi par l'utilisateur. Comme nous l'avons déjà dit, le système ne répond pas uniquement sur les co-occurrences des formes linguistiques « *rencontre* » et « *Chirac* » puisqu'à la notion de « rencontre » est associé un ensemble de marqueurs linguistiques (comme : *a déjeuné avec*, *a accueilli*, *a téléphoné à* …) à partir desquels l'annotation des segments textuels peut s'effectuer. On peut cependant affiner la recherche en faisant appel à ce que nous appelons une « carte sémantique » qui, pour une notion donnée, comme la « rencontre » ou «l' annonce

dans le paragraphe…) et du taux de condensation désiré (10% ou 15% du texte initial). Il faut ensuite procéder à un nettoyage automatique du texte obtenu en introduisant certaines anaphores ou, dans d'autres cas, en introduisant les antécédents selon des stratégies heuristiques, de façon à rendre le résumé plus lisible. L'approche du résumé automatique à partir des annotations fournies par EXCOM bénéficie des acquis des précédentes réalisations de résumé automatique SERAPHIN, SAFIR et ContextO.

thématique », va établir des sous catégorisations avec des notions de plus en plus précises, certaines modalités (réalisée/ non réalisée/possible …) et des indications de fiabilité (fort/ faible) dépendant de la force des marqueurs linguistiques. Ainsi, pour la notion « rencontre », on peut avoir les « rencontres physiques », les « rencontres événementielles », les « rencontres par la parole uniquement ».

Nous donnons l'exemple de la carte sémantique associée à la « rencontre » (voir la figure 3) ainsi que la carte sémantique associée au point de vue du « repérage » (voir la figure 4). Dans cet exemple, on considère un schème général « X est repéré par rapport à Y » où X est une entité qui est « repérée » par rapport à un « repère » Y, beaucoup mieux déterminé, par exemple du point de vue référentiel. Plusieurs valeurs spécifient les différentes sortes de repérage[32].

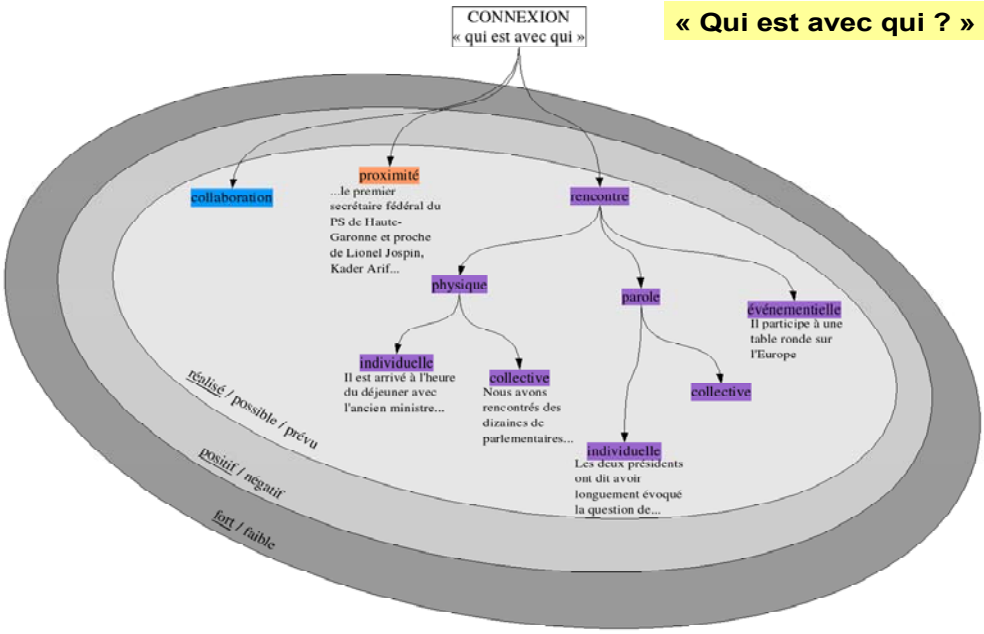


Fig. 3 – Carte sémantique du point de vue de la « rencontre »
(Djioua *et al.* 2006)

[32] La notion de « repérage » est très utilisée dans l'approche du linguiste Antoine Culioli. Nous l'avons adaptée à nos fins en tant que schème général constitutif de nombreuses catégories grammaticales. Le relateur de repérage, qui relie le repéré et le repère, prend trois valeurs principales : identification, différenciation et ruption (sorte de « différenciation forte » conduisant à sortir de la catégories objets identifiables à un prototype ou pouvant s'en différencier par quelque propriété saillante), définies par des propriétés mathématiques différentes (voir Desclés, 1980, 1987). Le schème de repérage sert, entre autres, à analyser les valeurs de la copule « est » et le converse « avoir » (selon E. Benveniste : « avoir » = converse de « est-à »). On retrouve des schèmes analogues dans différents modèles sémantico-cognitifs, par exemple chez R. Langacker avec l'opposition entre « landmark » (repère) et «trajector » (repèré mobile).

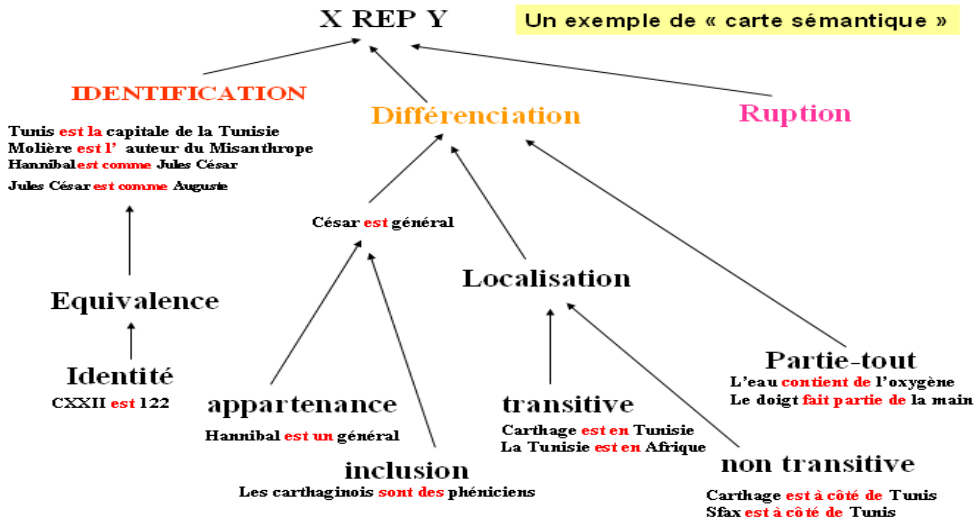


Fig. 4 – Carte sémantique du point de vue du « repérage »
(Desclés 1987, Le Priol *et al.* 2006)

3.6. Technique linguistique et computationnelle d'exploration contextuelle

Etant donnée une notion située dans une carte sémantique, il lui est associé une classe de marqueurs linguistiques qui servent à l'exprimer. Cependant, l'identification des seuls marqueurs dans un texte n'est pas suffisante car alors nous risquerions d'augmenter considérablement le bruit. En effet, la polysémie, où un réseau de plusieurs significations sont associées à une même forme linguistique, est un fait linguistique majeur et nous devons en tenir compte et même savoir l'exploiter dans une approche du traitement automatique des langues. Prenons, par exemple, le marqueur « *conduit à* » qui peut d'un côté, signaler une relation de causalité (*l'augmentation du bruit conduit à une moins bonne information*) et d'un autre côté, renvoyer à une relation non causale (*Marie conduit à l'école sa petite fille*) étant simplement agentive[33]. Nous avons mis au point une technique d'exploration contextuelle (Desclés *et al.* 1991, Desclés 1997, Desclés *et al.* 2005), qui possède une pertinence linguistique [34] et, également, une dimension opérationnelle [35] . Cette dernière consiste à rechercher dans le contexte de

[33] Nous avons dans une construction agentive un seul événement où l'agent est constitutif de la réalisation de cet événement alors que dans une construction causale nous avons toujours deux événements, dont l'un est la cause ou l'explication de l'autre.

[34] (…) au même titre que l'analyse transformationnelle ou l'analyse distributionnelle qu'elle généralise.

[35] (…) en reprenant le formatage des règles déclaratives en « SI conditions ALORS décision » de la programmation déclarative en informatique.

l'occurrence d'un marqueur d'une notion (ce marqueur est appelé « indicateur ») d'autres indices linguistiques (en fait des formes linguistiques non interprétées) qui confirmeront ou infirmeront la prise de décision : l'occurrence de l'indicateur et des indices trouvées dans un espace de recherche – phrase, paragraphe, proposition … – est bien la trace linguistique de la valeur sémantique que l'on peut associer au segment textuel où ces indices apparaissent, ou ce n'est pas le cas. Les indicateurs et indices appartiennent à des classes de marqueurs ; ils sont liés entre eux par des règles, dites d'exploration contextuelle, qui conduisent à prendre des décisions en attribuant, ou en n'attribuant pas selon les indices rencontrées, des annotation sémantiques relevant des points de vue étudiés. L'exploration contextuelle correspond à un processus de raisonnement par abduction[36], au sens du sémioticien C. S. Peirce.

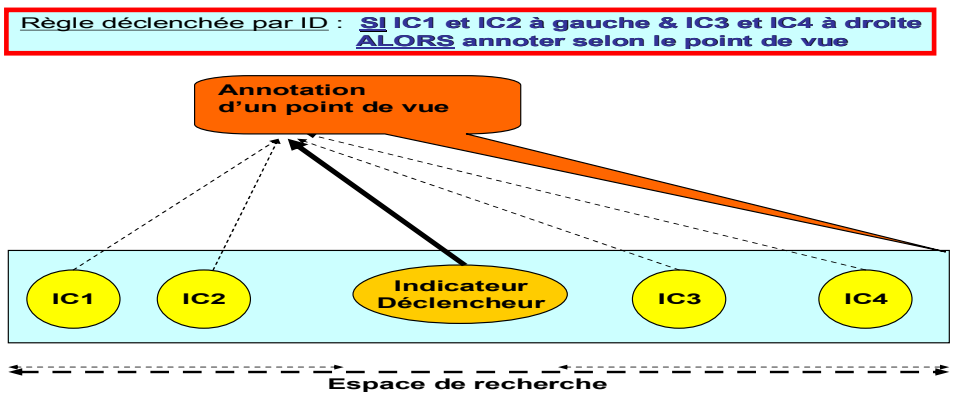


Fig. 5 – Schéma d'une règle d'exploration contextuelle

La technique d'exploration contextuelle est beaucoup plus puissante que la technique des automates et des transducteurs à états finis ou des expressions régulières de Kleene qui permettent de définir des schémas (des « patterns »), décrits par des graphes (Silbersztein, 1993)[37], de co-occurrences d'unités linguistiques autour d'un marqueur linguistique (par exemple un « mot clé »). En effet, plusieurs caractéristiques de l'exploration contextuelle rendent cette

[36] Si H est une hypothèse (annotation d'un segment) et IC1, IC2, …ICn des indices linguistiques de cette hypothèse H, alors si on trouve effectivement les indices IC1, IC2, …et ICn dans le contexte d'un indicateur déclencheur ID de la règle (dans un espace de recherche approprié et spécifié par la règle), alors il est possible de remonter, par abduction, à l'hypothèse H elle-même, qui ainsi devient plausible et permet d'annoter le segment textuel où se trouvent indicateur et indices. Certains indices sont très puissants et rendent très forte la plausibilité de l'hypothèse H d'annotation. La multiplicité des indices renforce la plausibilité de l'hypothèse H. Certains indices négatifs (non présence d'indices) peuvent conduire à accepter la plausibilité de H.

[37] Voir par exemple le système INTEX (M. Silberszstein : *Dictionnaires électroniques et analyse des textes, le système INTEX*, Masson, Paris, 1993).

technique plus flexible et mieux adaptée aux phénomènes linguistiques des textes. Citons en quelques unes : (i) possibilité de tenir compte d'indices « à très longue distance » (par exemple tenir compte au cours de la lecture d'un texte de la présence de certains indices linguistiques du titre) ; (ii) possibilité de définir des « indices négatifs » (absence d'indices dans l'espace de recherche conduisant à une décision) ; (iii) hiérarchisation entre les indicateurs et les indices contextuels, ce qui correspond à des contraintes linguistiques et cognitives pertinentes, alors que, pour un automate fini, toutes les unités sont de même niveau ; (iv) puissance formelle des règles d'exploration contextuelle, lorsque l'on accepte des appels récursifs de règles : on peut reconnaître par des systèmes de règles d'exploration contextuelle des agencements linguistiques appartenant aux langages de type 1 (langages engendrés par des grammaires dont les règles sont sensibles au contexte) dans la hiérarchie de Chomsky[38] (Desclés 2006).

La carte sémantique d'une notion se voit donc associer : (a) la classe des marqueurs linguistiques susceptibles d'identifier l'annotation sémantique des segments textuels où ils ont des occurrences, ces marqueurs étant des indicateurs qui déclenchent le processus d'exploration contextuelle ; (b) ensemble de règles d'exploration contextuelle associé à chaque marqueur déclencheur, ou indicateur ; ces règles permettent de lever l'indétermination polysémique du déclencheur et, éventuellement, de préciser l'annotation qui en résulte.

Donnons un troisième exemple d'un fragment de carte sémantique relative au point de vue de « l'annonce thématique », avec les classes de marqueurs associés.

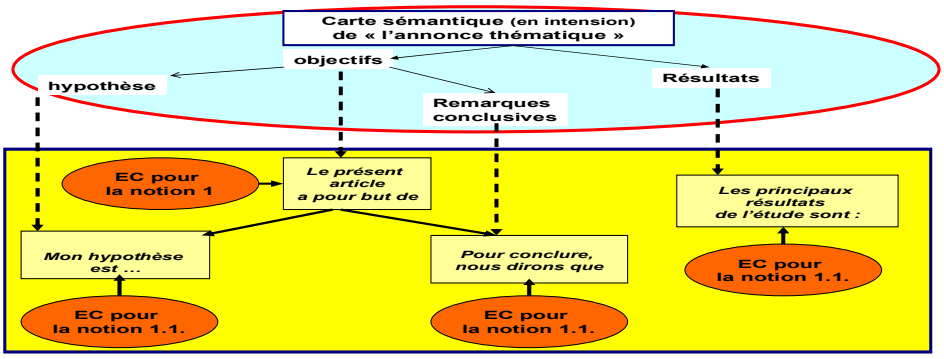


Fig. 6 – Carte sémantique de l'annonce thématique avec les ensembles de règles d'exploration contextuelle (EC) associés à chaque indicateur déclencheur

[38] Ainsi, nous avons démontré que le langage $\{a^n b^n c^n ; n>1\}$ est reconnu par un système très simple de règles d'exploration contextuelle alors qu'un tel langage, comme cela est bien connu dans la théorie des langages formels, ne peut pas être reconnu par un automate fini et qui, de ce fait, n'est pas une expression régulière. Un système de règles d'exploration contextuelle est différent d'une grammaire formelle de type 1 (grammaire contextuelle) car les indices linguistiques qui sont dans la prémisse d'une règle ne sont pas nécessairement contigus et peuvent se trouver dans différentes positions de la chaîne syntagmatique.

4. SYSTÈME DE RECHERCHE D'INFORMATIONS À PARTIR DES ANNOTATIONS DE POINTS DE VUE

Les textes étant annotés automatiquement (par la machine ECOM) ou étant annotés « à la main » par un travail collaboratif, il devient possible de procéder à une indexation des segments annotés et de construire une nouvelle classe de SRI fonctionnant sur ces segments annotés[39].

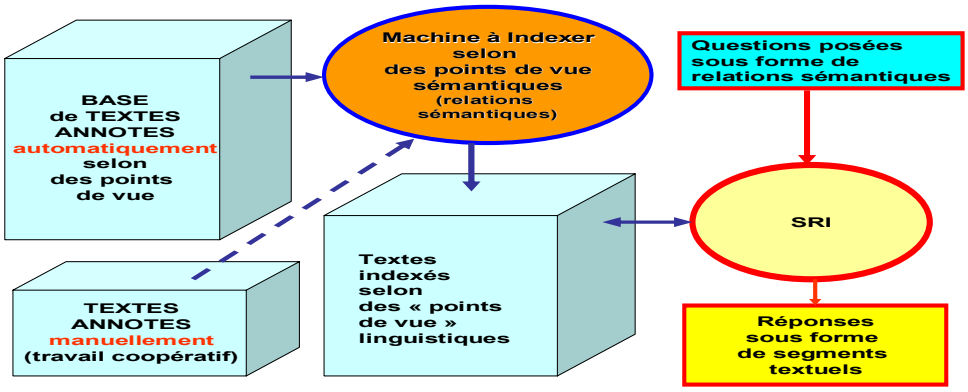


Fig. 7 – Recherche d'informations à partir d'annotations

4.1. Indexation multilingue de segments annotés

La méthodologie d'annotation automatique selon des points de vue est évidemment transposable à d'autres langues (Alrahabi *et al.* 2006). En effet, ayant déterminé les principaux concepts constitutifs du point de vue, ces derniers s'inscrivent dans un réseau de concepts, que nous appelons « carte sémantique », chacun des concepts de la carte étant relié à d'autres concepts tout particulièrement selon des relations de spécification/généralisation (Desclés 1987). A chaque concept correspond une classe ou un ensemble de classes d'expressions linguistiques qui en sont les marqueurs linguistiques. Ayant construit, pour une langue, la carte sémantique relative à un point de vue de fouille (par exemple pour la notion de « rencontre » ou celle de « repérage » ou encore celles de «l'annonce thématique » ou des « remarques conclusives »), nous pouvons l'utiliser pour une autre langue en recherchant les expressions linguistiques correspondantes aux

[39] A notre connaissance, il n'existe pratiquement pas de SRI qui opèrent avec des segments textuels annotés selon des points de vue sémantiques. Notre approche est donc, sur ce seul critère, innovante. D'autres recherches abordent ce genre de problèmes, par exemple chez James Pustejovsky (avec l'identification des modalités) mais la technique employée ne fait pas appel à une exploration systématique du contexte des marqueurs linguistiques identifiés, elle met plutôt en œuvre des schémas exprimables par des expressions régulières.

mêmes concepts de la carte, ce qui amène parfois à reconsidérer légèrement la carte sémantique de départ pour l'adapter et l'ajuster à la langue d'arrivée. Le schéma de l'interrogation multilingue est maintenant le suivant : une question relative à un point de vue de fouille est posée par exemple en français. Cette question est ensuite traduite par une question formulée, par exemple, en coréen. A condition bien entendu d'avoir annoté les textes en coréen selon ce même point de vue avec des ressources linguistiques appropriées, le même moteur d'indexation va rechercher les segments annotés (en coréen) qui sont des réponses plausibles à la question posée en français. Une traduction (automatique ou humaine) de ces réponses est alors retournée à l'utilisateur initial. L'économie de la démarche est évidente[40]. En effet, plutôt que de traduire tous les documents coréens en français (cette opération est coûteuse !) et de procéder ainsi à une recherche d'informations sur les documents traduits, la recherche s'effectue directement sur les documents coréens. Les erreurs de traduction (en particulier par la traduction automatique) ne se cumulent pas puisque sont traduits, selon notre approche, seulement les segments textuels pertinents pour la question posée, aussi une « erreur » de traduction ne se propage-t-elle pas aux autres documents[41].

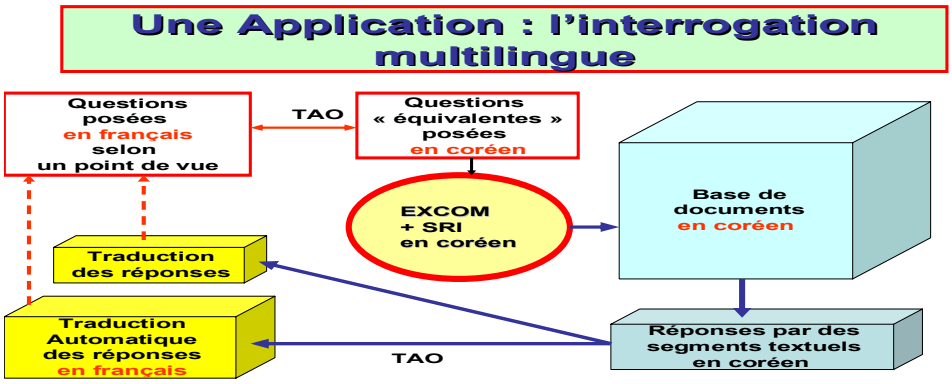


Fig. 8 – Interrogation multilingue en français de textes en coréen

Pour de telles interrogations multilingues, la constitution des ressources selon différents points de vue sémantiques doit être entreprise pour chaque langue. Il s'agit là d'un travail linguistique, certes minutieux mais qui n'implique pas des moyens trop importants puisqu'il exige ni analyseur morphologique ni analyseur

[40] On pourra comparer notre approche aux approches plus traditionnelles des « SRI multilingues » (rédigé par Christian Fluhr), chapitre 6 de *Méthodes avancées pour les systèmes de recherche d'informations*, Hermès, Paris, 2004.

[41] On sait que c'est la multiplicité des erreurs locales de traduction dans un texte qui rend la traduction globale plus ou moins exploitable en traduction automatique. Une erreur ou une approximation de traduction d'une réponse peut rendre opaque cette dernière mais la globalité des réponses traduites n'est pas nécessairement en jeu.

syntaxique préalables. En effet, le « transfert de langue à langue » reste assez économique puisque pour chaque point de vue la « carte sémantique » est relativement indépendante des marqueurs linguistiques qui l'expriment dans chaque langue. Aussi, le travail d'élaboration d'une carte sémantique étant effectué à partir de corpus dans une langue, est-il rapidement transposable dans une autre langue. A titre d'exemple, le transfert vers le coréen et vers l'arabe de quelques points de vue associés à « l'activité résumante »[42] de quelques textes, nous a demandé quelques semaines de travail avec des collègues coréens et arabes. Nous donnons ici un échantillon des annotations obtenues avec le même moteur EXCOM (voir les copies d'écran 7 et 8). Le dispositif informatique (mêmes moteurs informatiques EXCOM et MOCXE, mêmes algorithmes) et la méthodologie générale restent les mêmes quelque soient les langues, seules changent les ressources.

On peut rechercher des informations dans un texte au sujet d'un terme introduit par l'utilisateur dans le contexte d'une annotation. A titre d'exemple, nous pouvons chercher à relever les occurrences du terme *« Harris»* dans le contexte du point de vue de « l'annotation thématique » dans le texte de la préface des *Notes du Cours de syntaxe* de Z. Harris. Le résultat de la recherche apparaît sous la forme présentée dans la copie d'écran 9. Le processus d'indexation est suffisamment général pour s'adapter immédiatement aux annotations effectuées sur d'autres langues. Il est donc possible de rechercher, à partir d'une interrogation relative à une même notion, par exemple celle du point de vue « annonce thématique », des segments textuels annotés dans plusieurs documents de différentes langues (coréen, arabe et français) comme nous le voyons dans les copies d'écran 9 et 10. Il est évidemment possible d'affiner la recherche avec des questions qui croisent à la fois un même point de vue et des termes spécifiques donnés par un utilisateur dans sa propre langue. Le système fonctionnera à condition que, toutefois on puisse disposer d'un traducteur (automatique) des termes ajoutés à la notion choisie. Par exemple, ici sans faire de traduction, on peut introduire une recherche sur le terme « Harris » dans le contexte de l' « annonce thématique ». La traduction n'a pas été nécessaire ici pour une obtenir une réponse puisque le terme latin « Harris » a été conservé dans le texte coréen comme le présente la copie d'écran 9.

5. VERS UNE GÉNÉRICITE D'APPLICATIONS DE L'ANNOTATION AUTOMATIQUE

Ne pas traiter une application relative à un domaine précis mais avoir plutôt une approche générique pour une famille d'applications avec une même méthodologie, une même architecture informatique (même moteurs d'annotation et d'indexation), les mêmes ressources associées aux points de vue, telle est la démarche que nous préconisons et avons entreprise. En effet, le couplage entre

[42] La carte sémantique de l'activité résumante a été construite à partir de l'analyse de textes en français.

annotation automatique et indexation nous a permis de réaliser, depuis plus de dix ans, des applications précises qui exploitent des annotations sémantiques et discursives[43]. Nous avons donc acquis une expérience considérable (en linguistique textuelle et en informatique), ce qui nous a permis de justifier la pertinence de la méthode d'Exploration Contextuelle. En tenant compte des retours d'expérience, ce modèle nous conduit à la construction de la nouvelle plateforme EXCOM[44] qui prend la suite des plateformes SERAPHIN et SAFIR pour le résumé automatique puis ContextO pour la fouille de textes[45] (Desclés *et al.* 2005). La possibilité d'indexation que nous venons de développer avec la machine MOCXE vient maintenant compléter un dispositif général qui nous oriente vers une nouvelle approche beaucoup plus sémantique des SRI. Parmi les applications que nous avons déjà entreprises, mais que nous devons maintenant intégrer dans la nouvelle architecture d'EXCOM, citons :

- Fouille sémantique de bases de textes selon des points de vue comme : la recherche des définitions d'un terme technique ; la recherche des citations de telle personnalité (politique, scientifique, du monde culturel…) ; la recherche des causes d'un phénomène physique (telle catastrophe par exemple), d'une maladie, d'un risque d'épidémie (vache folle, grippe aviaire) …(Le Priol *et al.* 2006) (Mourad 2001)
- Synthèses automatiques et résumés automatiques de documents textuels avec filtrage distribué dans des fiches formatées, stockables et exploitables ultérieurement (Blais *et al.* 2006);
- Structuration des connaissances et constructions d'ontologies à partir de textes contenant des informations portant sur des domaines spécifiques ;
- Articulations entre textes et images pour annoter les images par des annotations textuelles, puis les indexer ;
- Bibliométrie avancée réalisée par des critères de catégorisation (positive, négative ; adhésion ; refus ; choix des hypothèses ; reprise des résultats…) des citations d'auteurs, conduisant ainsi à une « bibliosémantique » qualitative, beaucoup plus informative et nuancée [46] que les résultats quantitatifs des actuels travaux de bibliométrie… ; (Bertin *et al.* 2006)
- Ordonnancement temporel des événements dans un texte ;

[43] Ces applications ont déjà été présentées dans des thèses universitaires et dans des colloques et revues internationales. Elles ont permis également des collaborations étroites avec des grandes entreprises (EDF-DR, France Telecom) et des PME (EDIAT, Pacte Novation, Lingway, Bureau Van Dijk, …).

[44] Rappelons que le sigle EXCOM signifie « Exploration Contextuelle Multillingue ».

[45] Voir Desclés et Minel, 2000.

[46] La classification des universités dans le mode de Changaï est un exemple de classification, fondée essentiellement sur le quantitatif, qui reste peu nuancée et fortement contestable, en particulier pour les sciences humaines.

- Spécifications informatiques à partir de textes …

Il est certain que les textes annotés selon certains points de vue ajoutent une valeur ajoutée aux textes bruts. Par ailleurs, les demandes de l'environnement professionnel et social réclament de tels services fondés faisant appel aux annotations sémantiques, là où les SRI actuels ne permettent pas de fournir des résultats satisfaisants. Citons quelques secteurs où notre approche apporterait des solutions :

- « Rencontre » : renseignement ; police ; sécurité ; veille économique et technologique ;
- « Citations » : journalistes, avocats, chercheurs pour les fouilles dans des documents de sciences humaines et sociales (sociologie, économie, histoire, philosophie), étudiants, lycéens, enseignants ;
- « Définitions » : chercheurs, documentalistes, journalistes, juristes, étudiants, lycéens…
- « Causalité » : veilleurs, chercheurs (biologie, médecine, pharmacologie, écologie, sciences humaines…), étudiants…
- « Annonces thématiques » : veille stratégique et économique, analyse des brevets ; résumés et synthèses automatiques (chercheurs, étudiants) ;
- « Bibliosémantique » : veille stratégique, recherche, outil d'évaluation de la recherche …
- « Articulations textes-images » : résumé et synthèse automatiques (médecine, économie, musée ; indexation d'images dans l'audiovisuel …)
- Spécifications informatiques à partir de textes descriptifs : informaticiens, biologistes…

On peut imaginer un service qui aurait pour fonction (commerciale) d'annoter des textes et des documents textuels selon certains points de vue (par exemple : identification des définitions, recherche des citations, classification sémantique des références, annonces thématiques…). Ces textes annotés ont alors obtenu une plus grande valeur car ils peuvent être des entrées de machines à indexer et donc être interrogés selon les points de vue de l'annotation. Une bibliothèque qui annoterait son fonds selon certains points de vue de recherche deviendrait plus attractive pour ses utilisateurs qu'une bibliothèque qui ne posséderait que des documents non enrichis par des annotations. Il y a là un domaine de réflexion important avec des visées politique, scientifique, culturelle et commerciale non négligeables.

5.1. Enjeux économiques

Les SRI actuels, qui fonctionnement sans faire appel à des relations plus sémantiques, ont tendance à accumuler des informations toujours plus nombreuses et pas assez catégorisées, provoquant ainsi un bruit certain qui lasse finalement les utilisateurs. Ces derniers acceptent trop facilement que certains critères

d'ordonnancement des informations opèrent, à leur place une sélection dans la présentation des documents. Or, nous l'avons déjà dit, cette présentation, par des jeux de coefficients statistiques et de réglages, plus ou moins cachés, de certains paramètres, tend à privilégier certains documents et à reléguer des informations à des rangs très éloignés, ce qui ne leur donne guère de chance d'être trouvées et exploitées. Comme nous l'avons déjà remarqué, la « bonne information », ou l'information pertinente, n'est pas celle que tout le monde connaît au même instant mais c'est une information souvent plus cachée, peu connue et qui, une fois reliée à d'autres informations, deviendra une nouvelle connaissance qui donnera alors un nouveau pouvoir à celui qui l'aura acquise. La recherche d'informations est maintenant confrontée aux problèmes suivants :

– Comment découvrir l'information rare ou cachée, celle qui contient une réelle information pertinente que la plupart ignorent ?

– Comment, dans le bruit ambiant, sélectionner les informations pertinentes pour les objectifs que l'on poursuit ?

– Comment augmenter les possibilités opérationnelles donnant accès aux contenus des informations, à la source de constructions de connaissances ?

Actuellement, de nombreuses approches techniques dans les laboratoires de recherches et les Centres Recherche et Développement des grandes entreprises s'orientent vers des approches qui se voudraient capables de fournir des réponses opérationnelles à ce genre de questions : il s'agit de construire des moteurs de recherche qui « intégreraient du contenu sémantique » et qui n'opéreraient plus seulement avec les seules unités linguistiques plus ou moins fréquentes. Le recours aux ontologies générales et aux ontologies des domaines accompagnées de procédures inférentielles qui impliquent des probabilités, est une voie qui est activement explorée. Les recherche actuelles du TAL ont développé des modèles qui ont été conçus sur le paradigme de la compilation informatique des langages de programmation de haut niveau : on entreprend, dans un premier temps, une analyse lexicale puis, ensuite, une analyse syntaxique pour construire enfin des représentations sémantiques exprimées dans un langage abstrait, souvent de nature logique ou apparentée qui serait une sorte d'interlingua…

Or, ces recherches n'ont pas toujours donné les résultats finalisés que l'on était en droit d'espérer au bout de soixante ans d'efforts. On a donc essayé, en particulier en traduction automatique, d'éviter la construction des représentations sémantiques en procédant à des « transferts directs » entre niveaux morphologiques puis entre niveaux syntaxiques. On a également modifié les modèles linguistiques initiaux pour les rendre beaucoup plus compatibles avec les exigences et techniques informatiques actuelles comme, par exemple, la reconnaissance de schémas syntaxiques réguliers, la mise en place de procédures de d'appariement avec un schéma (« pattern matching »), de ramener les descriptions aux formats déterminés par les « attributs-valeurs » d'une catégorie – morphologique, syntaxique, sémantique.

C'est le cas du modèle HPSG, qui a su adapter le modèle initial des Grammaires Syntagmatiques de Noam Chomsky, aux contraintes informatiques des langages de programmation actuels les plus répandus. La plupart de ces modèles, tout comme le modèle plus ambitieux « Sens-texte » de Igor Mel'chuck (avec sept niveaux depuis la structure superficielle des phrases jusqu'aux représentations sémantiques sous forme de graphes acycliques et d'importants dictionnaires sémantiques) ou le modèle de la « Grammaire Applicative Universelle » de Sebastian K. Shaumyan, ou encore le modèle de la « Grammaire Applicative et Cognitive » développée dans le laboratoire LaLIC, sont très intéressants sur le plan de la recherche fondamentale car ils font découvrir des propriétés des langues, mais ils nécessitent, pour être mis en œuvre, la constitution de très nombreuses ressources, entre autres la constitution de dictionnaires informatiques, sous la forme de bases structurées de données lexicales, de la langue générale et des différentes langues de spécialité, puis la construction d'analyseurs morphologiques et syntaxiques.

Ces derniers analyseurs doivent cependant être « robustes » pour pouvoir s'adapter non seulement aux textes relativement « normés » (par exemple : textes narratifs plus ou moins littéraires, articles de journaux, articles scientifiques et techniques publiés…) mais aussi aux documents textuels beaucoup moins normés (dépêches d'agences, compte rendus médicaux, notes de service, compte rendus d'expériences, rapports techniques d'ingénieurs, spécifications en langue naturelle de systèmes techniques et informatiques, manuels de montage et de maintenance…) où des images, des diagrammes, des figures viennent s'insérer à l'intérieur de segments textuels. Les besoins de la société de la communication, notamment dans la recherche d'informations dans de grandes bases textuelles multilingues, fermées ou ouvertes (Web), sont devenus de plus en plus importants pour les développements de l'économie et les modèles plus « classiques » de l'informatique linguistique n'y répondent pas toujours avec efficacité.

Pour essayer de prendre en compte les contenus sémantiques, on a cherché à y adjoindre des descriptions sémantiques sous forme de vastes réseaux sémantiques (type Wordnet), de graphes conceptuels (par exemple les graphes de Sowa) et de systèmes de représentations des connaissances mis en place par l'IA… Il faut cependant reconnaître que la constitution de toutes ces ressources est très lourde à mettre en place et à maintenir, donc très coûteuse. De telles ressources sont évidemment nécessaires pour obtenir des traductions automatiques de qualité et l'on peut comprendre pourquoi la plupart des projets européens ont vivement soutenu, pour résoudre ces problèmes, la constitution de ressources multilingues diversifiées. La constitution de grandes ressources linguistiques et d'analyseurs assez sophistiqués a conduit à la construction de logiciels ou de pro-logiciels fort coûteux. Il apparaît maintenant que ces derniers ont pu être vendus à quelques très grosses sociétés (comme les constructeurs d'avions par exemple…) qui ont accepté de les payer plusieurs centaines de milliers de dollars. Ainsi, certains logiciels,

avec leurs ressources volumineuses, n'ont été vendus qu'à un très petit nombre d'exemplaires, voire à l'unité. Aussi, le marché du TAL et de l'ingénierie linguistique est-il resté très fragile, malgré les analyses des besoins potentiels de la société de la communication, surtout depuis l'arrivée d'Internet. De nombreuses sociétés, ont réussi à survivre uniquement parce qu'elles avaient reçu la « manne » de projets – par exemple de projets européens – en s'associant avec des laboratoires universitaires ou des organismes de recherche financés par les états, ou encore parce qu'elles s'étaient vues rachetées par des groupes multinationaux qui imposaient alors une direction de recherche et un type de produit à diffuser prioritairement, sans tenir compte des innovations possibles ou des recherches qu'il conviendrait de poursuivre..

Pour introduire plus de sémantique dans la recherche d'information sur les textes, quatre voies peuvent être envisagées :

- soit extraire des informations sémantiques par des modèles d'apprentissage opérant sur des corpus d'entrainement à partir de critères statistiques;
- soit se placer dans le paradigme du TAL classique inspiré par le paradigme de la compilation informatique avec une hiérarchie de niveaux pour atteindre des représentations sémantiques, traitées alors comme des représentations de connaissances ;
- soit construire des ontologies pour chaque domaine et s'en servir pour apparier les questions et les documents recherchés ;
- soit avoir beaucoup plus recours aux informations contextuelles en privilégiant des points de vue de fouille sémantique.

Les trois premières voies de développement supposent, comme nous l'avons dit, la constitution de ressources très lourdes, en particulier pour les ontologies qui doivent être spécifiques à chaque domaine. La dernière voie suppose en revanche la constitution de ressources plus légères car elles tendent à être indépendantes des domaines. Cette nouvelle voie devrait donner des résultats intéressants à court terme avec un coût économique relativement faible. Elle suppose cependant de bonnes analyses linguistiques pour constituer les ressources nécessaires. Évidemment, il ne faut pas négliger les approches qui exploiteraient plus directement les informations multimédia (son, image, vidéo), mais il nous semble difficile d'ignorer complètement l'apport complémentaire du textuel, par exemple sous forme de commentaires ou de descriptions textuelles des images, icônes, sons. L'abandon du textuel au profit du seul multimédia, comme certaines approches semblent prendre cette voie, serait très certainement une erreur stratégique.

L'approche par annotation sémantique au moyen de la technique d'Exploration Contextuelle puis par indexation fondée sur ces annotations permet de construire des SRI d'un nouveau genre, afin de fournir un accès réel aux contenus sémantiques et à leur exploitation. Nous avons déjà évoqué quelques types d'applications. Les différents domaines de la veille (veille stratégique, veille

économique, veille pour l'innovation, veille sécuritaire…) ont des besoins évidents d'outils informatiques performants aptes à détecter les informations rares et cachées. L'innovation suppose la détection, avant les concurrents réels ou potentiels, de signaux qui seraient des indices de l'émergence de nouveaux concepts, de nouvelles méthodes, de nouvelles molécules, de nouvelles techniques, de la fusion encore dissimulée d'entreprises, de la nomination de tel nouveau dirigeant, de nouveaux marchés en expansion ou potentiels, de nouvelles demandes, d'un début d'épidémie ou l'annonce d'une catastrophe, ou encore d'un début de campagne contre un produit.

L'annotation automatique sémantique des documents textuels ajoute, comme nous l'avons déjà remarqué, une plus grande valeur[47] à ces documents puisqu'elle leur ouvre une voie pour une indexation par annotations et une recherche plus ciblée. En revanche, l'annotation manuelle, qui est actuellement effectuée dans de nombreuses approches, par exemple du Web sémantique, n'est pas du tout réaliste, même si on espère que des procédures par apprentissage pourront donner une voie d'accès à l'automatisation ultérieure des annotations. En revanche, l'annotation automatique selon des points de vue de fouille est, dès maintenant, une voie opérationnelle dont il a été a prouvé la faisabilité et une interaction étroite avec une indexation de relations plus sémantiques.

5.2. Deux stratégies de recherche et de développement.

L'approche que nous préconisons, notamment par l'annotation sémantique automatique des documents conduit à la construction de logiciels qui devraient, du fait de la légèreté des ressources et de la modularité des points de vue de fouille, avoir un prix de vente relativement modique, les rendant ainsi accessibles à un très grand nombre d'utilisateurs. Ces logiciels, s'ils sont vendus à un prix relativement faible (au plus cent euros), seront diffusés et vendus en grand nombre auprès de PME (cabinets d'avocats, journalistes, documentalistes, services de veille), des bibliothèques, des centres documentaires, ainsi qu'auprès de particuliers (étudiants, lycéens, enseignants).

Actuellement, dans la recherche des informations[48], et même dans le domaine du TAL, les approches statistiques (méthodes vectorielles, indices de similarité, modèles markoviens, N-grammes, algorithmes génétiques, réseaux de neurones formels), avec apprentissage sur des corpus dominent considérablement. Si ces

[47] Elle réalise en partie le souhait: « Il s'agit ensuite de développer, de l'intérieur, les capacités de repérage des contenus, en enrichissant la structure intime des documents de tous les repères possibles, pour permettre le maximum d'utilisations diverses par chacun, annotations spécifiques à l'utilisateur, gloses de toute sorte, citations, renvois … » cité par Jean-Noël Jeannenet *Quand Google défie l'Europe*, 2005, 64.

[48] Voir Madjid Ihadjaden, *Méthodes avancées pour les systèmes de recherche d'informations*, Paris, Hermès, 2004.

approches présentent des intérêts évidents avec des succès non négligeables dans des domaines très particuliers, elles ne permettent pas de répondre à des besoins réels de la part d'utilisateurs potentiels. De plus, les résultats obtenus par des algorithmes qui opèrent avec des fréquences, avec des poids attribués à certaines unités de traitement, avec des indices numériques assez sophistiqués ne permettent pas aux utilisateurs de savoir réellement « comment ça se passe » et donc d'avoir une garantie sur la fiabilité des réponses obtenues, ce qui pose de réels problèmes épistémologiques, voire éthiques[49]. Certes, les méthodes statistiques ne réclament pas de compétences en linguistique et c'est pourquoi le TAL et l'ingénierie linguistique sont en train de devenir un des domaines privilégiés d'ingénieurs informaticiens qui n'ont reçu aucune formation sérieuse en linguistique et qui, par conséquent, ignorent les autres solution[50] et modèles plus linguistiques et mieux fondés sur une meilleure connaissance de la nature du langage, sur des structures discursives des textes, sur des approches résolument sémantiques et sur une prise en compte systématique et raisonnée des catégorisations (grammaticales, lexicales et discursives) des langues, autant de paramètres qui peuvent être avantageusement exploités dans de nouveaux systèmes d'informations avec une analyse plus poussée et renouvelée des nouveaux besoins.

L'autre approche dominante est celle des ontologies des domaines et du Web-sémantique qui lui est très liée. Là encore, les réflexions sémantiques, y compris les solutions opérationnelles, qui viendraient du mode des sciences humaines (linguistique en particulier) sont écartées beaucoup trop vite par les ingénieurs et chercheurs en informatique qui pensent avant tout « système informatique » avec des formatages trop rigides qui risquent de bloquer toute innovation. Or, le coût économique, comme nous l'avons déjà dit, de construction et de maintenance des ontologies des domaines est prohibitif, si bien que certains groupes importants dans le monde sont en train de renoncer à cette approche qui nécessiterait trop d'années-hommes avant d'être devenue vraiment opérationnelle sur les domaines complexes (domaines du matériel ferroviaire, domaine de l'avionique…). Cette approche par ontologies ne permet pas toujours de maintenir la « continuité sémantique » entre les ontologies formelles (sous forme de graphes complexes) et les habitudes plus textuelles des utilisateurs. On assiste, depuis quelques années, à des approches qui tentent de construire des ontologies pour chaque domaine[51] à partir des textes mais, là encore, à de très rares exceptions, ce

[49] Toute activité scientifique se doit de justifier la méthode qui a été employée pour obtenir un résultat ou pour prendre une décision.

[50] Dans les comités d'évaluation ou dans les grands centres de Recherche et Développement, certains décideurs, essentiellement des ingénieurs informaticiens, faute d'une bonne information sur l'évolution de certains travaux en linguistique, peuvent faire obstacle à des projets d'un nouveau paradigme, préférant des systèmes où seule l'informatique intervient plutôt que des systèmes qui nécessiteraient une meilleure analyse des objectifs avec une approche plus interdisciplinaire.

[51] Notre approche (Desclés, Flairs'07) des ontologies nous amène à considérer trois niveaux d'ontologies : (i) les ontologies des domaines avec des concepts-de-premier-niveau ; (ii) des

sont essentiellement des informaticiens qui construisent et réalisent les ontologies à partir de méthodes statistiques ou, lorsque la linguistique est convoquée, en identifiant essentiellement les termes nominaux et pratiquement pas les relations sémantiques entre termes, ce qui revient à réduire finalement une langue à être essentiellement un système de nomenclatures d'objets.

Deux stratégies dans la recherche des informations et la construction des connaissances pour leur exploitation, apparaissent de plus en plus nettement. Pour les évaluer et décider laquelle doit être préférée, il faut tenir compte des coûts économiques, des demandes d'utilisateurs potentiels, des possibilités techniques apportées par non seulement l'informatique mais aussi par une nouvelle approche, plus sémantique de la linguistique et du multilinguisme. Comparons les caractéristiques principales de ces deux stratégies.

Stratégie A	Stratégie B
Constitutions de ressources lourdes « universelles » et non sélectives (Dictionnaires informatisés ; Analyseurs morphologiques, syntaxiques Représentations sémantiques par graphes conceptuels …).	Constitution de ressources légères selon des points de vue sémantiques (ressources modulaires selon les points de vue de fouille constituées par la méthode de l'exploration contextuelle).
Ontologies par domaines.	Compatibilité possible avec des ontologies des domaines.
Utilisables dans des domaines particuliers, là où les ontologies ont déjà été construites.	Utilisables dans tous les domaines.
Applications au Web sémantique par le biais des ontologies.	Applications au Web sémantique par le biais des catégorisations sémantiques et discursives.
Prix de revient : élevé.	Prix de revient : peu élevé.

ontologies sous forme de cartes sémantiques accompagnées de classes de marqueurs linguistiques ; (iii) des ontologies avec des concepts-de-troisième-niveau (« Upper ontologies ») qui décrivent la sémantique profonde des concepts-de-second-niveau des cartes sémantiques construites. Les concepts-de-second-niveau des cartes sémantiques (ii) sont les éléments constitutifs (des notions sémantiques) qui permettent de construire effectivement les ontologies du premier niveau (i) en tenant compte des expressions linguistiques qui ont des occurrences dans les textes relatifs à divers domaines et également de «peupler» ces ontologies par des instances (ou désignations d'objets des domaines traités).

Coût élevé => prix de vente élevé	Coût relativement faible => prix de vente peu élevé
Diffusion faible seulement auprès de très grosses entreprises avec des produits chers.	Diffusion élevée auprès des PME et particuliers avec des produits peu chers ; Démocratisation des outils de recherche d'information. Multiplication des ventes.
Utilisation difficile et contraignante (problème de la continuité sémantique). Annotation souvent manuelle.	Utilisation intuitive (lien maintenu avec la dimension textuelle). Annotation automatique.

La stratégie A est implicitement adoptée par la majorité des secteurs de Recherche et Développement qui cherchent à construire des moteurs de recherche d'informations fondés sur des méthodes plutôt statistiques en souhaitant néanmoins donner un accès au « contenu » par des recherches plus ciblées, spécifiques à des domaines particuliers avec une diminution sensible du bruit. La plupart des équipes qui travaillent selon cette stratégie sont composées essentiellement d'informaticiens avec, éventuellement, quelques linguistes chargés alors d'annoter manuellement des corpus afin que ces annotations puissent être utilisées dans des programmes d'apprentissage.

Quant à la stratégie B, que nous préconisons, elle nécessite des efforts dans l'analyse linguistique des corpus avec des linguistes de très bon niveau et spécialement formés à la constitution des ressources selon les points de vue sémantiques et discursifs, points de vue qu'il faut apprendre à développer en fonction des besoins et des objectifs de recherche. Elle doit donc s'appuyer sur des équipes interdisciplinaires où d'un côté, les informaticiens respectent la problématique linguistique de départ et d'un autre côté, les linguistes s'adaptent aux contraintes opérationnelles de programmes finalisés. La satisfaction de nouveaux besoins est réellement envisageable par la stratégie B et beaucoup moins, ou du moins pas aux mêmes échéances, par la stratégie A.

De grands groupes de Recherche et Développement se tourneraient tout naturellement vers la stratégie B dès lors qu'une voie opérationnelle qu'il leur serait proposée en leur permettant de répondre effectivement aux nouveaux besoins qui apparaissent[52] et qui se développeront très vite dans les années qui viennent au

[52] Par exemple : identifier et extraire dans des textes, journaux, par exemple, toutes les citations d'une personnalité ; identifier dans une série de textes d'un fonds documentaire, toutes les définition d'une notion qu'on souhaite analyser…

sein de « la société de l'information et de la communication ». Aussi, si la stratégie B n'est pas actuellement soutenue en Europe, nous pouvons prédire que, dans deux ou trois ans au plus, ce type d'approche, par annotations sémantiques automatiques compatibles avec des indexations ultérieures, arrivera dans les pays européens avec des nouveaux logiciels et plateformes informatiques [53] mais ces logiciels et plateformes fonctionneront exclusivement sur des textes formulés en anglais. La Communauté Européenne devra alors, une fois de plus, s'adapter et acheter les licences de ces nouveaux logiciels de recherche d'informations, avec leurs ressources spécifiques à l'anglais, pour tenter de ne pas trop perdre de terrain dans la recherche, l'acquisition, la structuration et la diffusion des connaissances. Les documents dans les différentes langues d'Europe des grandes bibliothèques[54] seront certes toujours accessibles aux moteurs de recherche (type Google ou versions ultérieures de Google ciblées vers de communautés spécifiques donc formatées dans un savoir *a priori*) mais sous une forme dégradée car n'étant pas annotés initialement par des points de vue de fouille sémantique et discursive, alors que d'autres documents textuels des bibliothèques dans les grandes universités américaines (Yale, Standford, Harvard, Chicago ..) et bientôt les bibliothèques de Chine et de l'Inde seront, eux, annotés, et donc accessibles aux les plus recherches avancées et compatibles avec de très nombreux services. Il y a là non seulement des enjeux économiques mais également des enjeux culturels qui peuvent compromettre ultérieurement et durablement les développements scientifiques et technologiques de l'Europe et son rayonnement, en freinant la diffusion de ses propres recherches, et en rendant plus difficile l'accès aux documents textuels de ses propres centres de diffusion (bibliothèques universitaires, Centres de documentation, laboratoires …).

BIBLIOGRAPHIE

Alrahabi, M., A. H., Ibrahim, J.-.P. Desclés, 2006, "Semantic Annotation of Reported Information in Arabic", *FLAIRS-19,* Florida*,* May 11–13.

[53] Ces logiciels seront comparables à ceux dont nous avons déjà largement prouvé la faisabilité opérationnelle. Il nous est nécessaire de passer à de très grands volumes afin d'avoir des tests d'évaluation sur ces grands volumes de textes. Ce genre d'évaluation nécessite des moyens supplémentaires qu'un laboratoire universitaire n'a pas ; Il est donc nécessaire de développer des partenariats avec des entreprises qui accepteraient de réorienter légèrement leurs activités en acceptant des propositions plus linguistiques. Avec les tels partenariats, il devient possible de développer de « business plans » alliant marketing, financements, diffusion et recherche.

[54] Il est très intéressant et bien entendu nécessaire que les grandes Bibliothèques Européennes s'orientent vers une numérisation de leur fonds mais il faut prévoir, en même temps, des services supplémentaires qui exploitent les documents numérisés. Il y a là pour l'Europe un domaine où elle pourrait prendre une avance certaine et ouvrir ainsi un accès vraiment multilingue aux recherches des informations. Il est dommage que le projet Quaero ne soit pas allé dans cette direction, préférant faire l'impasse dur les documents textuels numérisés et d'un développement de nouveaux services, au profit des documents audio-visuels.

Aussenac-Gilles, N., S. Dagobert, 2005, "Text analysis forontology and terminology engineering", *Applied Ontology, An Interdisciplinary Journal of Ontological Analysis and Conceptual Modelling*, vol. 1, 1, 35–46.

Berners-Lee, T., J. Hendler, O. Lassila, 2001, *The Semantic Web: A new form of Web content that is meaningful to computers will unleach a revolution of new possibilities,* Scientific Americain.

Bertin, M., J.-P., Desclés, B. Djioua, K. Yordan, 2006, "Automatic Annotation in Text for Biblimetrics Use", *FLAIRS-19*, Florida, May 11–13.

Blais A., J.-P. Desclés, B. Djioua, 2006, "Le résumé automatique dans la plate-forme EXCOM", Paris, *Digital Humanities*, 5–9 juillet.

Brin S., L. Page, 1998, *The anatomy of a large-scale hypertextual Web search engine*, Stanford University.

Cohen, C., 2004, *Veille et intelligence stratégiques*, Paris, Hermès.

Cunningham, H., D. Maynard, K. Bontcheva, V. Tablan, 2002, "GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications", *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics* (ACL'02), Philadelphia.

Desclés, J.-P, 1987, "Réseaux sémantiques : la nature logique et linguistique des relateurs", *Langages*, 87, 55–78.

Desclés, J.-P., 1990, *Langues naturelles, langages applicatifs et cognition*, Paris, Hermès.

Desclés, J.-P., C. Jouis, Oh H-G. D., Reppert, 1991, "Exploration Contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte", in: D. Herin-Aime, R. Dieng, J.-P. Regourd, J. P. Angoujard (eds), *Knowledge modeling and expertise transfer*, Amsterdam, 371–400.

Desclés, J.-P., 1997, *Systèmes d'Exploration Contextuelle. Co-texte et calcul du sens*, (ed. Claude Guimier), Presses Universitaires de Caen, 215–232.

Desclés, J.-P, J.-L. Minel, 2005, "Interpréter par exploration contextuelle", in: F. Corblin, C. Gardent (eds), *Interpréter en contexte*, Paris, Hermès, 305–328.

Desclés, J.-P., 2006. "Contextual Exploration processing for Discourse automatic annotations of texts", *FLAIRS-19,* Florida, May 11–13.

Desclés, J.-P., 2007, "Ontologies, Semantic Maps and Cognitive Schemes", *FLAIRS-20*, Florida, May 7–9.

Djioua, B., J. Garcia Flores, A. Blais, J.-P. Desclés, G. Guibert, A. Jackiewicz, F. Le Priol, L. Nait-Baha, B. Sauzay, 2006, "EXCOM: an automatic annotation engine for semantic information", *FLAIRS-19*, Florida, May 11–13.

Djioua, B., J.-P. Desclés, 2007, "Indexing Documents by Discourse and Semantic Contents from Automatic Annotations of Texts ", *FLAIRS-20*, Florida, May 7–9.

Frakes, W.B., R. A. Baeza-Yates, 1992, *Information Retrieval: Data Structures & Algorithms* Prentice-Hall.

Grossman, D., O. Frieder, 1998, *Information Retrieval: Algorithms and Heuristics*, Kluwer Academic Publishers.

Gruber, T. R., 1993, "A translation approach to portable ontology specifications", in *Knowledge Acquisition*, 5, 1993, 199–220.

Handschuh, S., S. Staab, 2004, *Annotation for the Semantic Web*, 96, Frontiers in AI and Applications.

Kiryakov, A. *et al*, 2004, "Semantic Annotation, Indexing, and Retrieval", *Elsevier's Journal of Web Semantics*, 1, ISWC2003 special issue (2).

Ihadjadene, M. (ed.), 2004, Méthodes avancées pour les systèmes de recherché d'informations, Paris, Hermès.

Jeanneney, J.-N, 2005, *Quand Google défie l'Europe : Plaidoyer pour un sursaut*, Paris, Mille et une nuits.

Kiryakov, A. *et al.*, 2004, "Semantic Annotation, Indexing, and Retrieval", *Elsevier's Journal of Web Semantics*, 1, ISWC2003 special issue (2).

Le Priol, F., A. Blais, J.-P. Desclés, B. Djioua, J. Garcia Flores, G. Guibert, A. Jackiewicz, L. Nait-Baha, B. Sauzay, 2006, "Automatic annotation of localization and identification relations in platform EXCOM", *FLAIRS-19*, Floride, May 11–13.

Marchisio, G. *et al.*, 2004, "A Case Study Search in Natural Language Based Web Search", in *ACM SIGKDD,* Seattle.

Minel, J.-L., J.-P. Desclés, 2000, « Résumé automatique et filtrage des textes », in: J-M. Pierrel (ed.), *Ingénierie des langues*, Paris, Hermès, 253–270.

Mourad, G., 2001, Analyse informatique de signes typographiques pour la segmentation de textes et l'extraction automatique des citations. Réalisation des applications informatiques : SegATex et CitaRE, Thèse de doctorat, Université Paris-Sorbonne.

Pédauque, R. T., 2006, *Le document à la lumière du numérique*, C&F éditions.

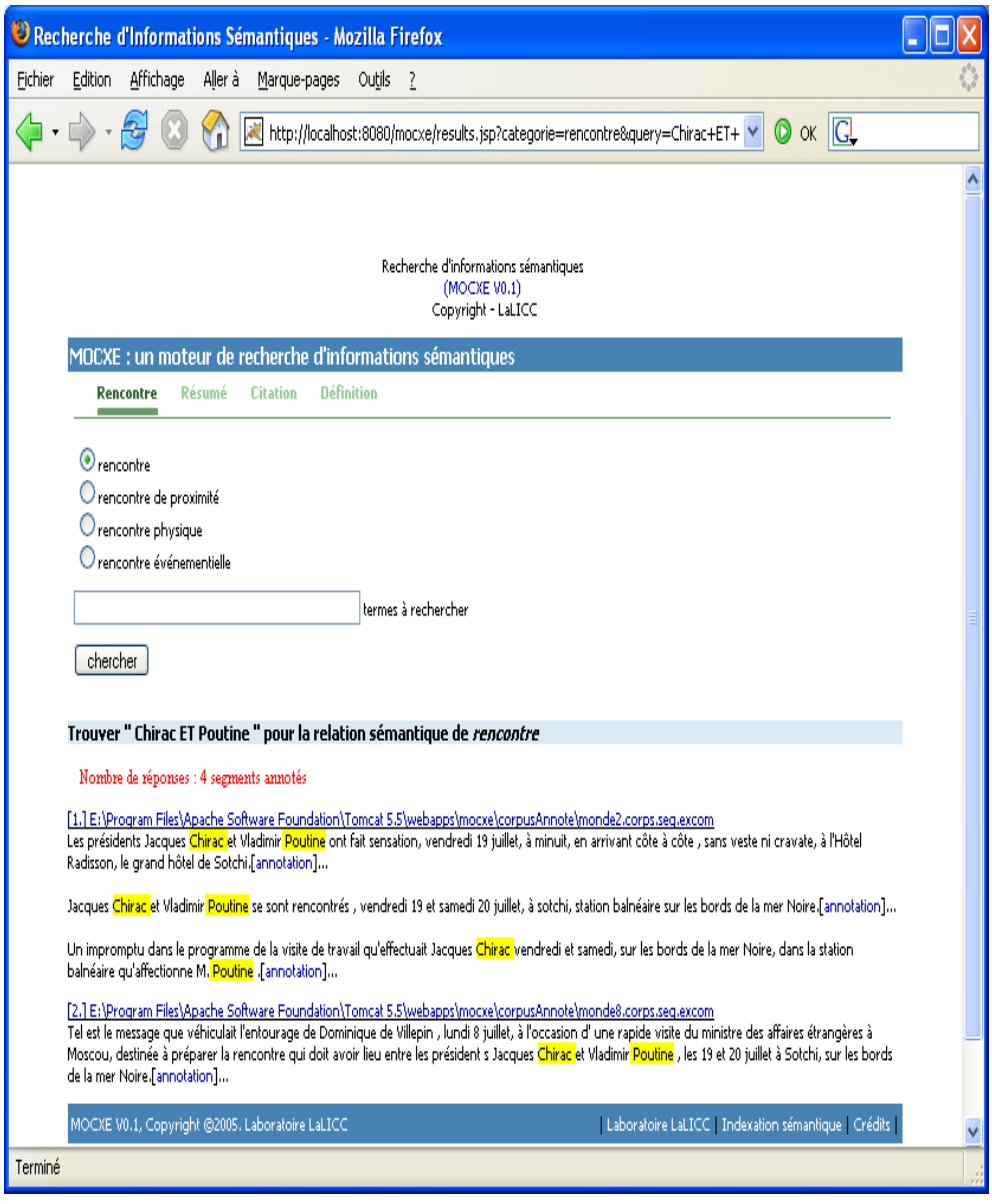
Poibeau, T., 2003, *Extraction automatique d'information, du texte brut au web sémantique*, Paris, Hermès.

Silberszstein, M., 1993, *Dictionnaires électroniques et analyse des textes, le système INTEX*, Paris, Masson.

Van Rijsbergen, C. J., 1975, *Information Retrieval*. London, Butterworth.

Salton G., 1971, "A Comparison between manual and automatic indexing methods", *Journal of the American Documentation,* 20, 1, 61–71.

Annexes



Copie d'écran 1 : Réponses par MOCXE à la recherche : « rencontre, Chirac, Poutine »

Titre: M. Chirac abandonne toute critique de la guerre en Tchétchénie
Auteur: *MARIE-PIERRE SUBTIL*
Edition:*Le Monde 22 juillet 2002, page 2*

Lors d' une rencontre avec Vladimir Poutine sur les bords de la mer Noire, le président français s'est rallié aux positions russes, justifiant les opérations militaires par la nécessité de combattre le terrorisme. Russie et Union européenne " doivent marcher main dans la main ", a-t-il ajouté

SUBTIL MARIE PIERRE

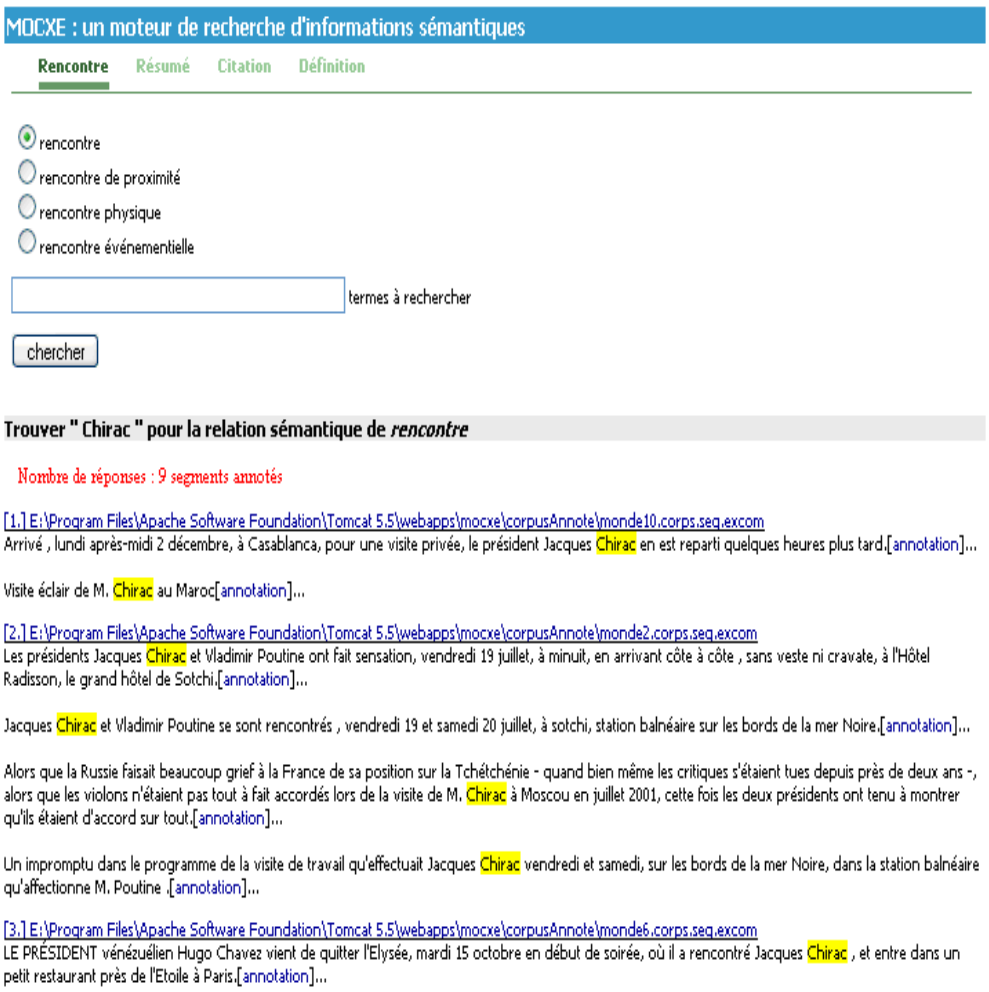
SOTCHI de notre envoyée spéciale

Jacques Chirac et Vladimir Poutine se sont rencontrés , vendredi 19 et samedi 20 juillet, à sotchi, station balnéaire sur les bords de la mer Noire. Ce rapide sommet a été l'occasion de mettre en scène le réchauffement des relations entre les deux pays. M. Chirac , qui avait, en 1999 et 2000, vivement critiqué la guerre menée par Moscou en Tchétchénie, s'est cette fois, dans des termes inhabituels, rallié aux positions russes. Le président français estime qu'un processus politique " est engagé " et a justifié les opérations militaires dans le cadre d'une lutte globale contre le terrorisme. Sur le terrain, les organisations de défense des droits de l'homme dénoncent la poursuite des exactions de l'armée russe. M. Chirac a souligné son souhait de voir " l'Union européenne et la Russie marcher main dans la main. "

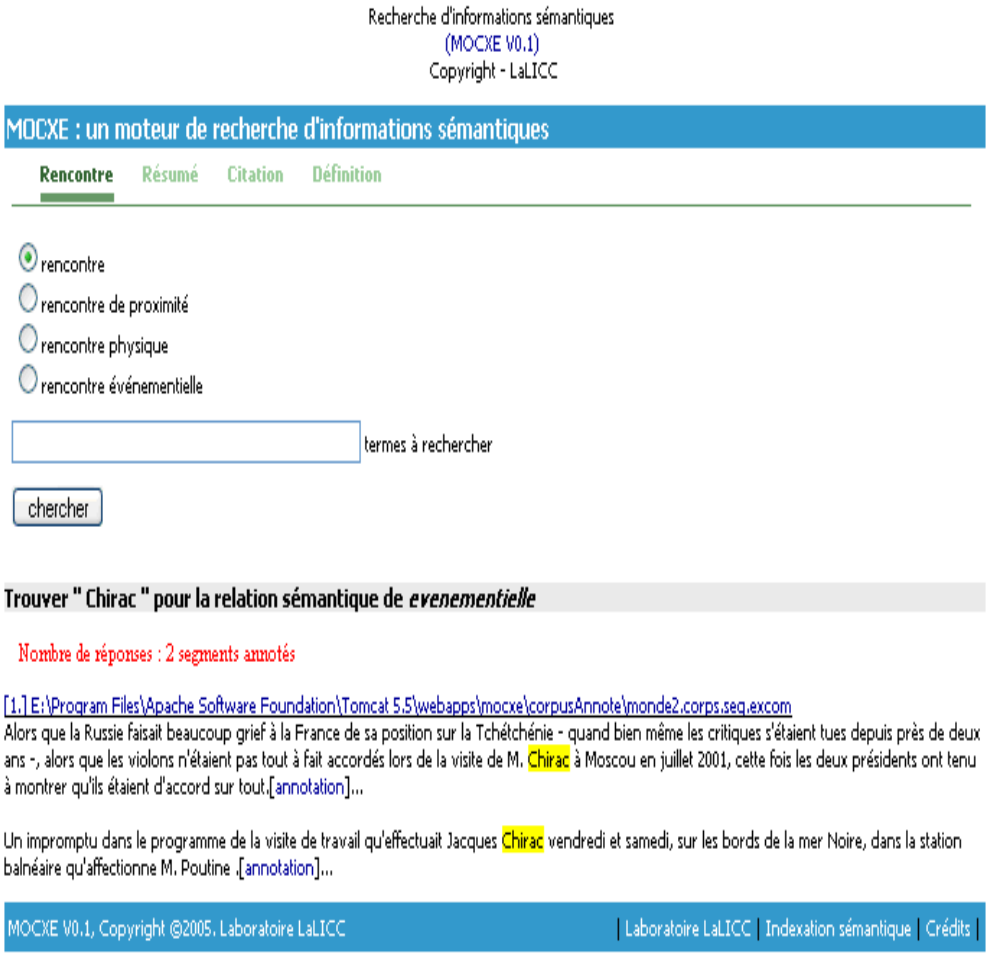
Les présidents Jacques Chirac et Vladimir Poutine ont fait sensation, vendredi 19 juillet, à minuit, en arrivant côte à côte , sans veste ni cravate, à l'Hôtel Radisson, le grand hôtel de Sotchi. Installés au bar devant des eaux gazeuses, ils se sont entretenus pendant trois quarts d'heure et seuls leurs interprètes savent de quoi il fut question. Un impromptu dans le programme de la visite de travail qu'effectuait Jacques Chirac vendredi et samedi, sur les bords de la mer Noire, dans la station balnéaire qu'affectionne M. Poutine .

S'il fallait une image pour montrer que les relations franco-russes vont mieux que jamais, ou du moins mieux que ces dernières années, elle était dans cet aparté imprévu. Alors que la Russie faisait beaucoup grief à la France de sa position sur la Tchétchénie - quand bien même les critiques s'étaient tues depuis près de deux ans -, alors que les violons n'étaient pas tout à fait accordés lors de la visite de M. Chirac à Moscou en juillet 2001, cette fois les deux présidents ont tenu à montrer qu'ils étaient d'accord sur tout. Le président français s'est rallié, dans des termes inusités jusqu'alors, à la position de son hôte sur la Tchétchénie. " La France condamne sans réserve tout acte terroriste quel qu'il soit et considère qu'aucune cause ne peut justifier des actions terroristes ", a affirmé M. Chirac , visant implicitement les séparatistes tchétchènes. La conférence de presse conjointe , organisée dans le parc de la résidence présidentielle, une propriété construite par Nikita Khrouchtchev au milieu des années 1950, ne donna lieu à aucun accroc. Après avoir repris les termes utilisés par le Kremlin, qui justifie la guerre qu'il mène en Tchétchénie en la qualifiant d'" opération antiterroriste ", le président français a répété le credo des Occidentaux : " La seule réponse convenable est de nature politique." Et d'ajouter, ce qui ne pouvait qu'enchanter le président russe : " Je

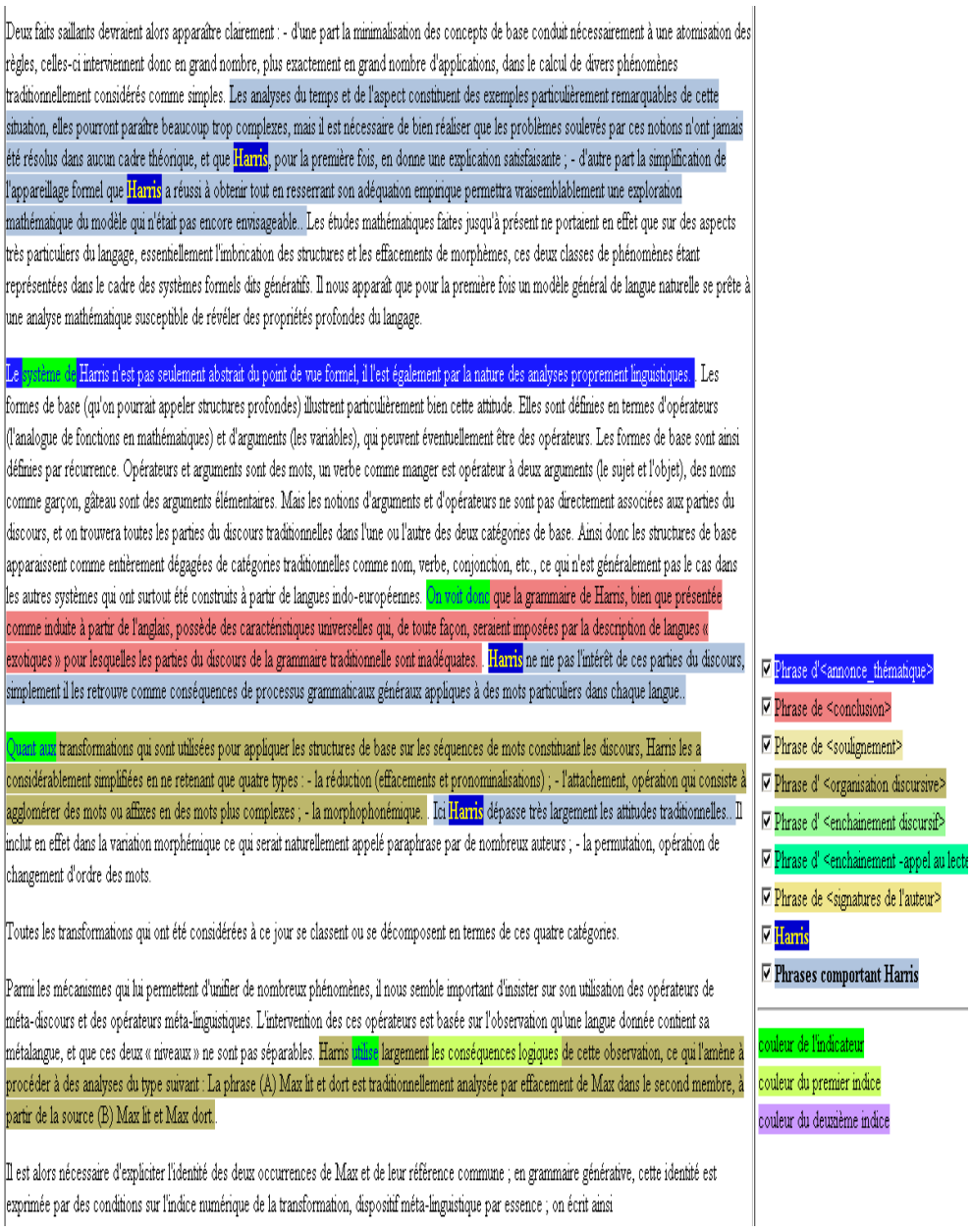
Copie d'écran 2 : document annoté par la machine EXCOM



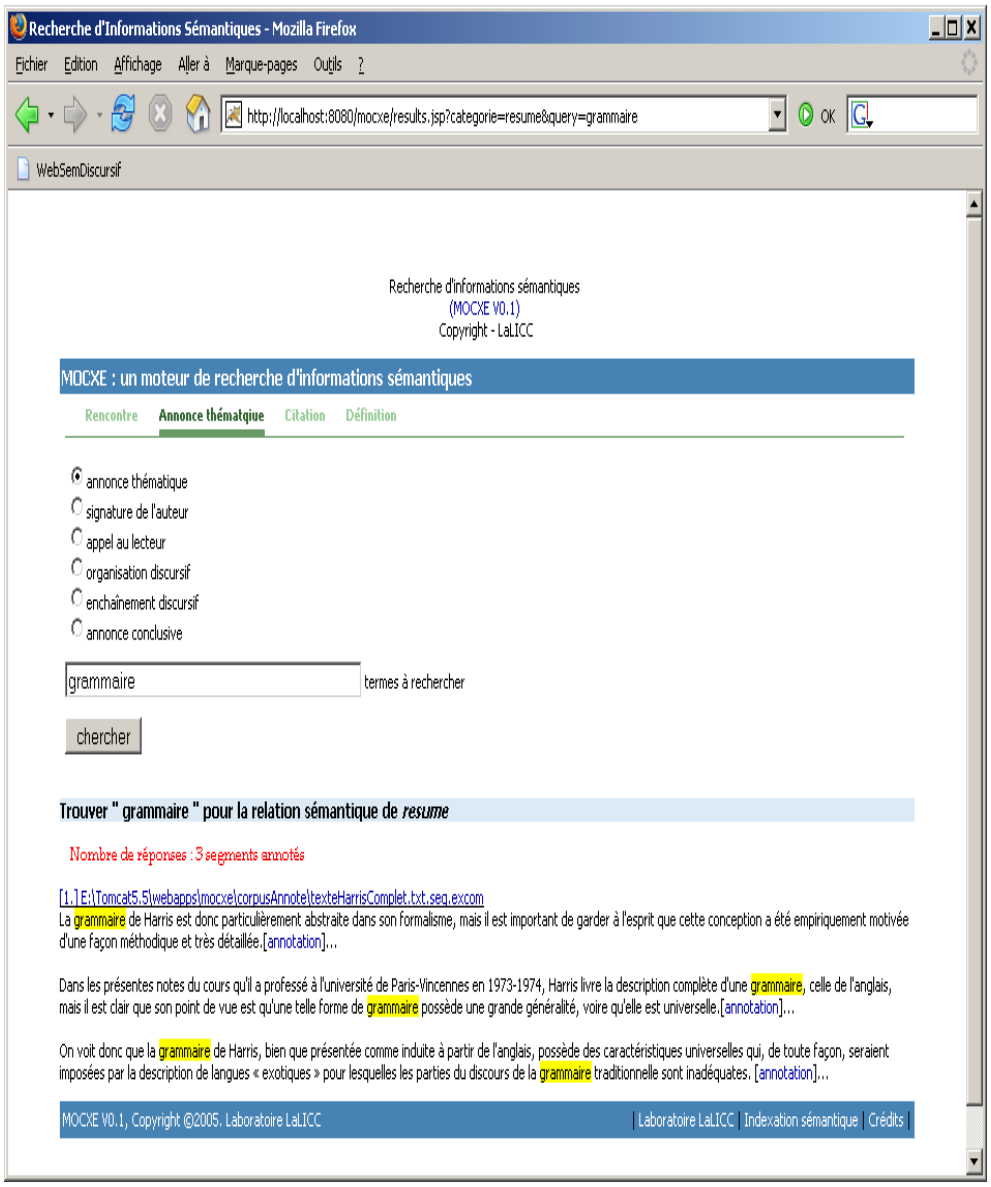
Copie d'écran 3 : recherche avec MOCXE : "Qui a rencontré Chirac ou Chirac a rencontré qui ? "



Copie d'écran 4 : recherche avec MOCXE : "rencontre événementielle, Chirac"



Copie d'écran 5 : Texte annoté (Préface des *Notes du Cours de syntaxe* de Zellig Harris) selon plusieurs points de vue : annonce thématique, remarques conclusives, soulignement de l'auteur, organisation discursive



Copie d'écran 6 : Recherche du terme « grammaire » dans un résumé du texte de la préface de Harris

Nombre de paragraphes : *46*
Nombre de phrases : *126*
Nombre de phrases annotées : *11*

?통사론 강의

(Dans la presentation de Maurice Gross du livre de Zellig Harris

Notes du Cours de syntaxe, publie aux Editions du Seuil)

소개의 글

해리스 (Z.H. Harris)는 본 강의에서 처음으로 그가 생각하는 변형 통사론 개념의 개괄적인 이미지를 구체적으로 제시한다. 지금까지, 그는 수학화1가 가능한 현상들만을 예로 끌어내거나 혹은 영어2에 국한된 언어현상들의 분석을 소개하는 등 개별적인 관점하에서만 그의 이론을 소개 했었다. 1973년부터 1974년에 걸쳐 파리-벵센느 대학 (Universite de Paris-Vincennes)에서 교수한 본 강의록에서, 해리스는, 비록 영어 문법이었지만, 그 문법을 완벽하게 기술한다. 여기서 한 가지 확실한 것은 이러한 문법의 형태가 일반 성을 지니고 있으며 그러므로 더 나아가서 보편적이라고까지 할 수 있다는 것이 해리스가 자신의 문법을 바라보는 관점이라는 것이다.

우리는 해리스가 직접 작성한 강의록을 번역하는데 만족하지 않고, 그가 제시한 분석의 많은 부분을 프랑스어에 적용하는데 주력했다.

이에 따라, 본고의 몇 몇 부분은 해리스 문법 모델의 프랑스어 적용이라는 측면에서 다루어졌는데, 이 과정에서 자연스레 바필 수밖에 없었던 몇몇 문법 이론은 그러므로 저자의 책임이 아니다. 이러한 시도는 독자들이 이 문법 이론을 더 잘 이해하게 하기 위해 저자와의 동의 하에 착안 된 것이었다.

우리는 이와 같은 작업에 도움을 준 해리스에게 감사하며, 아울러, 이 강의 중에 행해진 세미나에서 발표를 통해 우리에게 몇몇 부분의 이해에 도움을 주신 야닉 기어브란트 (Yannick Gheerbrandt)와 아므르 헬미 이브라힘 (Amr Helmy Ibrahim)에게도 감사의 말을 전한다.

생성문법에 좀 더 익숙한 독자들이라면 해리스의 초기 연구로부터 영향을 받아 후에 촘스키(Chomsky)3에서 발전된 형식화 기제를 곧바로 발견할 수는 없을 것이다. 즉, 촘스키는 모든 자연언어의 문법이 논리 수학적인 체계와 유사한 형식 체계로 구성될 수 있다는 가정하에서 그의 모든 이론을 세웠다. 다시 말해서, 문법 규칙은 다시 쓰기 규칙이 되어야만 했다. 반대로, 해리스는 동일한 관점에서 몇몇 시도를 거친 후, 함수적 형식화에 치우친 그러한 가설을 포기했다. 따라서, 해리스의 문법은 복잡한 함수 체계라고 할 수 있는데, 사실상, 그의 문법은 일반적으로 서로 대치되지 않고 또 언제나 결합적이지는 않은 여러 구조에서 나온 다양한 결합체들로부터 생긴 형태들의 조합인 것이다.

따라서, 형식주의 안에서 해리스의 문법은 각별히 추상적이지만, 그는 방법론에 치우치면서 그리고 매우 세세한 방법으로 자신의 개념을 구축했다. 해리스의 강의는 자신의 연구의 통계적인 연보를 나열한 것이 아니었다. 그의 문법 체계에 대한 소개는 처음에는 완전히 공리적이었다. 그 뒤에 바로 그는 자신의 형식화 작업의 대부분을 영어의 언어 현상에 적용했다. 그가 설계한 문법은 전정으로 모든 것을 완전하게 포괄하는 문법으로서, 그 안에서 다루어질 수 있는 통사-의미 체계의 다양성을 세세하게 그리고 있다.

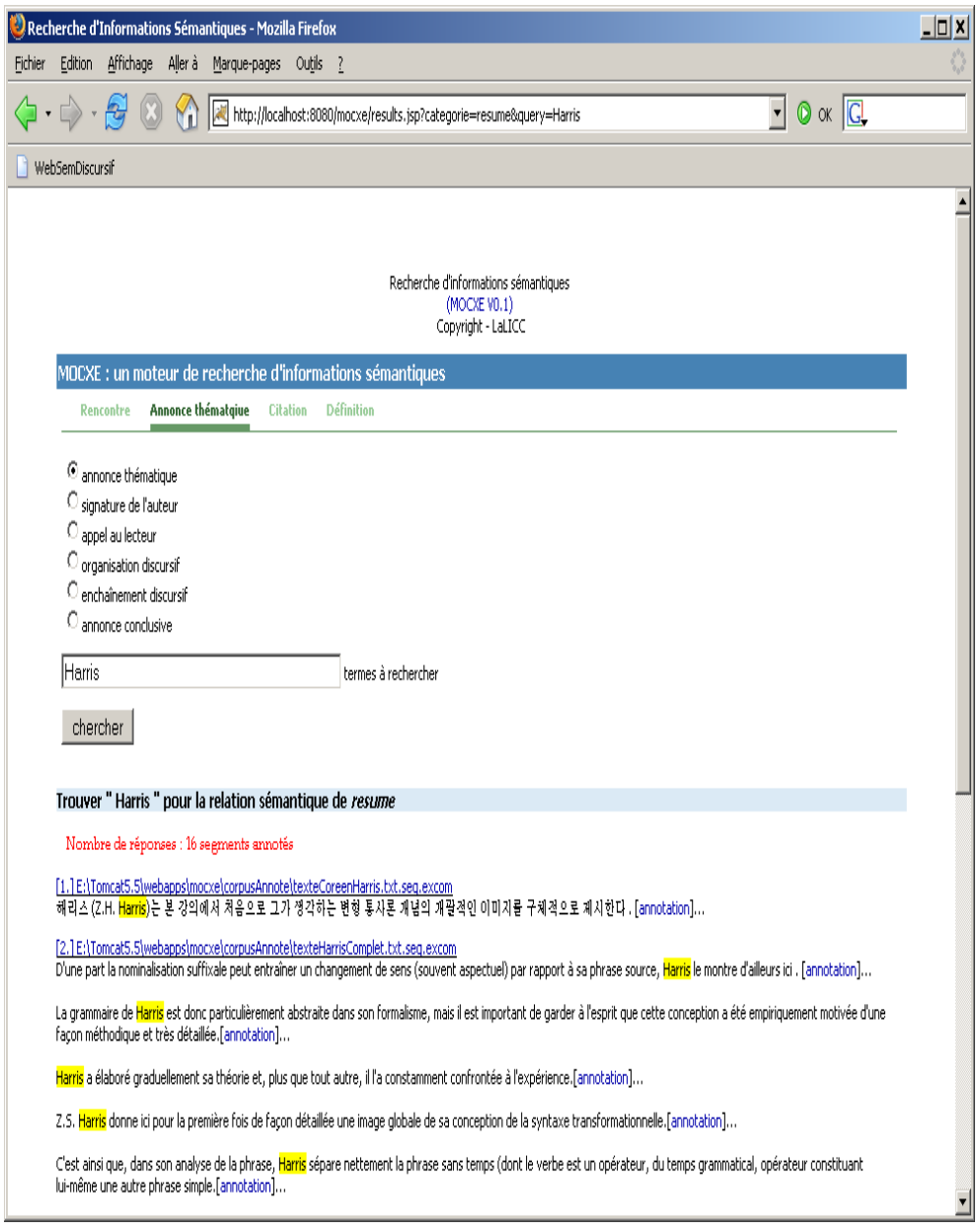
해리스의 분석은 전체적으로 형식화 되어있다. 그러나 그가 선택한 해결방법이 독자들이 다른 곳에서 만날 수 있는 혹은 쉽게 떠올릴 수 있는 다른 방법에 비해 더 바람직할 수 있다는 언어학적인 논거를 항상 제공하지는 않는다. 해리스는 이러한 논거들을 현재 집필 중인 다른 문법서에서 좀더 자세한 예들과 함께 설명하기로 미루어 두었다. 그가 내린 분석들의 언어학적인 몇몇 정의들이 이미 다른 저서에서 소개 된 한편, 그 외의 것들은 설명없이 제시만 되었으므로 어째서 그러한 분석이 도출되었는지에 대해서는 독자들에게 또다른 노력이 요구된다. 그러나, 아무래도 언어 현상들을 얼마나 잘 다룰 수 있는지 가늠해 보기 위해서는 본고의 공리적인 발전 과정 (결과는 늘 성공적이므로)을 따라가 보는 것이 바람직 하리라고 본다.

한편, 해리스는 독자들의 이해를 돕기 위해 어떤 개념에 대한 정의를 내리기전에 그 이전에 나타나는 초반의 정의들로 돌아가는 작업을 보여준다. 다음의 두 가지 두드러진 명백한 사실이 보인다?:

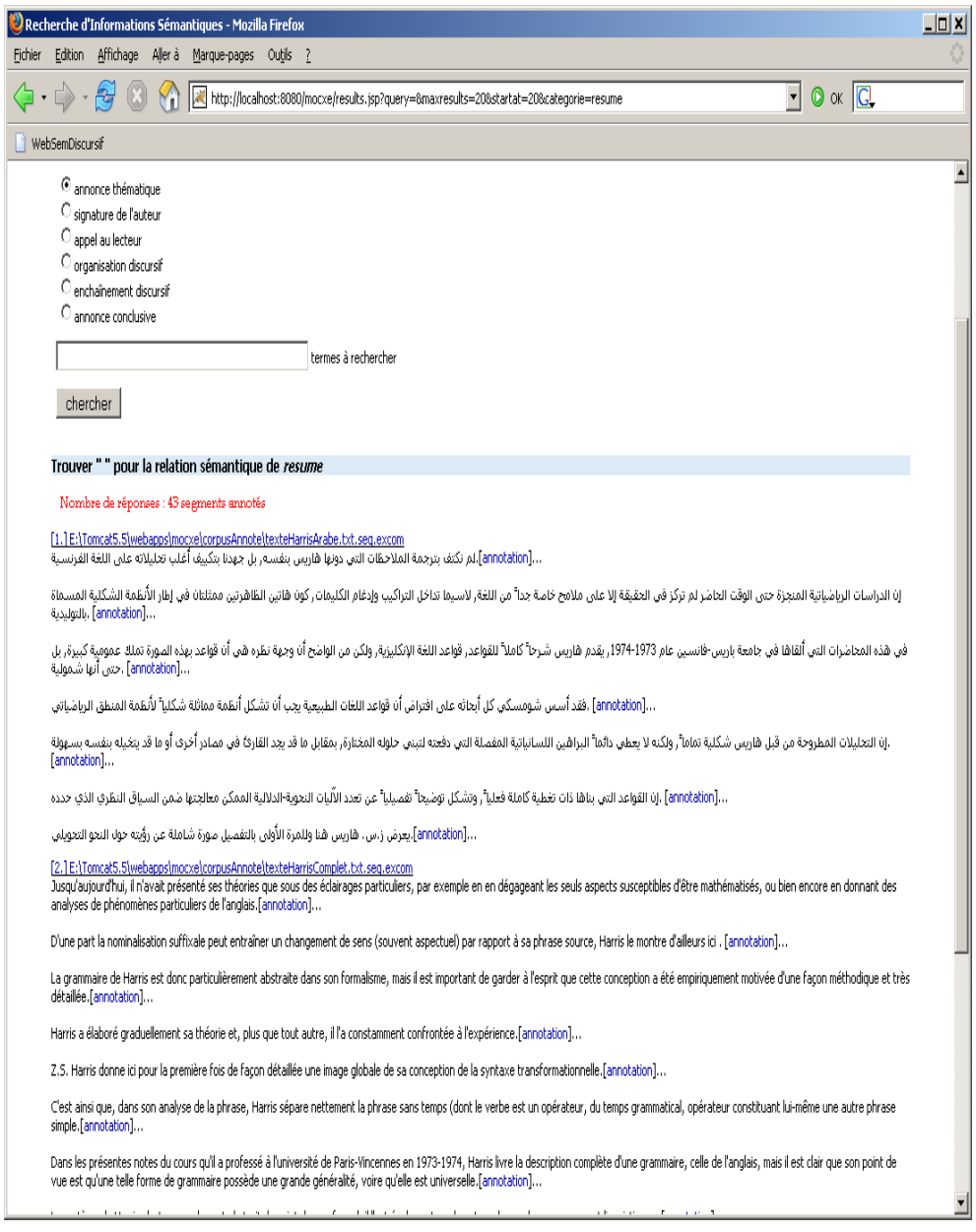
- 먼저, 기본 개념의 극소화는 규칙들의 분해화로 필연적으로 귀결되어 그 숫자가 많아지게 되는데, 이를 더 정확히 말하자면, 이는 그 이전에는 단순하게 다뤄졌던 다양한 현상을 설명하는데 많은 규칙의 적용을 한다는 것이다.

Copie d'écran 7 : Annotations (de la préface du livre de Z. Harris traduite en coréen) à partir de la même carte sémantique élaborée pour le français

Copie d'écran 8 : Annotations (de la préface du livre de Z. Harris traduite en arabe) à partir de la même carte sémantique élaborée pour le français



Copie d'écran 9 : Recherche multilingue autour de l'occurrence de « Harris » dans le contexte des annotations du point de vue de « l'annonce thématique ».



Copie d'écran 10 : Recherche multilingue sur la catégorie « annonce thématique »