Sémantique des textes et recherche d'information

Mathieu Valette, Monique Slodzian

Mathieu Valette, Monique Slodzian (2008) « Sémantique des textes et Recherche d'information », *Extraction d'information : l'apport de la linguistique*, A. Condamines & Th. Poibeau, éd., *Revue Française de Linguistique Appliquée*, volume XIII-1 – juin 2008), 119-133.

version intermédiaire soumise [7 mars 2008] – merci de se reporter à la version publiée

Titre : Sémantique des textes et Recherche d'Information

Mathieu Valette[2, 1] et Monique Slodzian[1]

[1] ERTIM (INaLCO, Paris)

[2] ATILF (CNRS – Nancy Université)

1. Introduction

On admet depuis Sparck-Jones (1999) que les connaissances linguistiques sont susceptibles d'apporter une contribution significative à la Recherche d'Information (RI), actuellement subdivisée en trois domaines, l'extraction d'information, le résumé automatique et les systèmes de question-réponse. La rapide montée en puissance du *Word Wide Web* a fait évoluer ces questions en moins de dix ans. L'un des effets en est l'émergence d'un nouveau sous-domaine, la classification de textes et, à travers elle, d'un intérêt nouveau pour les notions de texte et de corpus (Rajman, 2000 ; Jackson et Moulinier, 2002). Parallèlement, on a vu s'imposer des Systèmes de Recherche d'Information (SRI) plus performants, intégrant des modèles statistiques puissants qui combinent des méthodologies mixtes spécifiquement développées naguère pour chacun des sous-domaines. Ce mûrissement de la problématique s'est accompagné de campagnes d'évaluation sur le long terme qui ont fourni un certain nombre de données sur l'apport respectif des informations linguistiques de types divers. En particulier, les campagnes TREC (Text Retrieval Conference) et Amaryllis qui se sont succédé depuis 1993 balisent significativement les rapports que la RI entretient avec le Traitement Automatique des Langues (TAL). On retiendra en particulier les expériences de classification réalisées sur TREC-5 (Bellot et Elbèze, 2000) et TREC-6 (Gaussier et al. 1998). Dans une première partie, une rétrospective rapide de l'impact du linguistique - à travers les techniques TAL - dans le domaine de la RI, sera l'occasion de faire état à la fois des résultats acquis et des approches standard de la dimension linguistique dans la problématique RI. On s'intéressera en particulier à la lente émergence de la problématique textuelle qui accompagne l'expansion du Web. Nous comptons montrer en quoi l'attention croissante suscitée par la linguistique textuelle correspond à un véritable tournant dans la problématique de la RI sur le Web ; en quoi l'approche par catégorisation des textes constitue une rupture avec les méthodes précédentes.

La deuxième et la troisième partie approfondiront les conditions d'une linguistique textuelle appliquée à la RI. Nous présenterons des méthodologies expérimentées dans le cadre d'un projet de filtrage des textes racistes sur Internet, puis, nous présenterons certaines des recherches actuellement menées en Analyse des Données Textuelles (ADT) dans ce sillage qui sont susceptibles, à plus ou moins court terme, d'améliorer les méthodes de la RI.

2. Traitement Automatique des Langues et Recherche d'Information

Le passage de l'indexation humaine traditionnelle à partir de mots-clés extraits de listes d'autorité à l'indexation automatique des textes à partir des mots contenus dans ces mêmes textes s'est imposé lentement et la vision que l'on en a eue au départ était *a priori* optimiste : il suffirait d'apparier les mots de la requête avec des mots du document par simple comparaison des chaînes de caractères. Comme on pouvait s'y attendre, la langue a défié la logique de la machine en la confrontant à deux de ses fondements, la polysémie et la synonymie. Le principe de polysémie induit pour la RI le risque du bruit (retour de textes non pertinents) et celui de synonymie, le risque du silence (non-reconnaissance d'un texte pertinent parce que le mot de la requête n'y figure pas). La majeure partie des articles de TAL dédiés à la RI durant ces dix dernières années a massivement porté sur ces deux sujets : comment désambiguïser ? Comment enrichir la requête ?

« La RI est-elle une chance pour le TAL ? », se demandait Christian Jacquemin (Jacquemin, 2000). N'est-ce pas suggérer que le TAL a construit sa prospérité dans le domaine de la RI sur ce point aveugle qu'est l'appariement de mots ? Dans un premier temps, on a cherché à enrichir l'indexation des documents au moyen d'analyseurs morphosyntaxiques large couverture. Ainsi s'est-on attaché à améliorer les performances des mots-clés par le recours à la racinisation ou à la lemmatisation : à chaque mot est associé une racine partagée par tous les mots d'une même famille morphologique. Cette stratégie s'adresse au risque de silence. Afin de limiter le bruit, on a fait appel à la syntaxe pour désambiguïser les mots et dépasser le traitement de sacs de mots. On connaît les nombreux travaux sur les termes complexes, les syntagmes nominaux et verbaux qui ont abouti à la création d'outils d'extraction sophistiqués. Au fil des campagnes d'évaluation successives, il est apparu que l'addition systématique de ressources linguistiques de bas niveau (morphologiques et syntaxiques) n'améliorait pas automatiquement les résultats et, en particulier, que les systèmes à base de morphosyntaxe

s'avéraient décevants (Jacquemin 2000). On doit aux travaux menés sur les techniques et outils d'évaluation d'avoir permis ces clarifications en montrant la complexité des interactions entre faits linguistiques et d'avoir ainsi attiré l'attention sur les limites d'une conception logiciste de la « langue naturelle » (Poibeau 2003).

Restait à explorer le niveau sémantique conformément à la vision du linguistique (analyse par couches morphologiques, syntaxiques, etc.) qui caractérise l'informatique linguistique (Bouillon *et al.* 2000). Afin de dépasser les limites des connaissances linguistiques de bas niveau, on a cru qu'il suffirait de s'attaquer au traitement des rapports sémantiques entre les mots, ceux-ci étant considérés comme des atomes désignant des objets du monde réel (classes de référents). En effet, l'hypothèse que la sémantique lexicale pouvait contribuer à lever l'ambiguïté des mots de la requête et du document en leur associant un ensemble de sens non ambigus semblait aller de soi. On présuppose ainsi que l'interprétation d'un mot se réduit à l'identification du concept auquel il se rapporte. Rien d'étonnant dès lors que l'on ait visé les rapports d'ordre conceptuel (synonymie, hyperonymie, méronymie), exprimés essentiellement par des liens de nom à nom, dans le but d'étendre la requête. Le modèle de ce type de ressources est fourni par WordNet (1998) dont les *synsets*, réseaux conceptuels constitués de mots liés par des relations conceptuelles et/ou proches par rapport à une thématique, incluent partiellement et secondairement les verbes. On en attendait une bonne résolution de l'ambiguïté lexicale et une nette amélioration du degré de similarité entre les mots de la requête et ceux du texte. Pourtant, dès 1993, les conclusions de TREC infirment l'hypothèse (Voorhees 1998) : les relations contenues dans WordNet ne fournissent pas l'information nécessaire à la désambiguïsation. La riposte pragmatique a consisté à combiner tous les types de connaissances linguistiques (morphologie, syntaxe, sémantique). Les conclusions d'études récentes sont à cet égard nuancées : « l'intérêt en RI d'exploiter conjointement des informations morphologiques et syntaxiques d'une part ou des informations syntaxiques et sémantiques (…) d'autre part n'a pas été véritablement démontré dans nos expériences, très peu de cas de complémentarité réelle ayant été observés » (Moreau *et al.*, 2007).

On objectera que la communauté TAL a tenu compte des critiques adressées aux ressources de type ontologique et que l'on a vu évoluer les modèles en question, qu'il s'agisse de WordNet ou plus encore de la théorie du Lexique génératif de Pustejovsky. Les uns et les autres ont cherché à exploiter les liens syntagmatiques pour désambiguïser à l'aide du

contexte, en travaillant en particulier la relation nom-verbe (Bouillon et al. 2000). Ces tentatives sont formalisées entre autres par Pustejovsky (1995) dont le lexique génératif vise à définir les mots dans des ensembles structurés de prédicats (structures de *qualia*) et à obtenir automatiquement les relations sémantiques qui lient des paires NV présentes dans un corpus de textes.

Résultant de la prise en compte forcée du syntagmatique (syntagmes nominaux et verbaux), le surgissement de la notion de cooccurrence constitue une première rupture en dépassant la notion de vocabulaire, pivot de la pensée onomasiologique. L'intérêt de l'étude des mots en contexte syntagmatique vient de ce que celui-ci apporte une meilleure différentialité des sens que le mot seul. Mais cela n'introduit pour autant qu'une version faible du « contexte » puisqu'elle ne se définit qu'au plan quantitatif et topologique : la fenêtre est en effet l'empan, l'espace textuel dans lequel on note les cooccurrences. Elle peut correspondre à cinq mots de part et d'autre de l'item concerné, à une phrase, un paragraphe, voire un document. Incapable de caractériser un texte à elle seule, elle informe sur le local (contraintes syntaxiques, par exemple) et, dans le cadre de la RI, elle ne peut tout au plus qu'informer sur la thématique, non sur les autres composantes du texte. À cet égard, elle constitue un filtre sémantique limité. L'étude des cooccurrences continue de relever d'une linguistique de la phrase et non du texte, d'une théorie de la signification et non du sens (Slodzian 2000). On cherche à désambiguïser des mots et non à caractériser des textes globalement pour les différencier d'autres textes qui, tout en contenant les mêmes mots, ne sont pas forcément pertinents pour la requête.

Dans une certaine mesure, la linguistique de corpus entretient un flou comparable sur l'usage de la cooccurrence. Le parallèle entre le rôle de la linguistique sur corpus pour l'étude des langues et celui du télescope pour celui du ciel est éclairant. Les recherches sur les collocations, les fréquences et les phraséologies y sont souvent menées dans le cadre d'études descriptives sur une langue particulière (Sinclair 1991), par exemple dans la perspective de la lexicographie bilingue pour introduire dans les dictionnaires le palier du syntagme.

L'émergence d'un domaine nouveau, la fouille de textes (Text Mining), résulte de la montée en puissance d'applications nouvelles liées au World Wide Web, en particulier la publication en ligne et les bibliothèques numériques. L'enjeu n'est plus simplement de trouver ou de classer des documents, mais de construire des relations entre des documents et des collections de documents (Jackson et Moulinier 2002).

La fouille de textes (telle qu'elle est publicisée dans les technologies de RI) en tant qu'elle suppose la segmentation d'unités textuelles et intertextuelles requiert une théorisation du texte qui donne accès à une sémantique du texte. À l'objectif de désambiguïsation des mots se substitue celui de caractérisation sémantique des textes.

Dans cette perspective, l'étude des occurrences sera considérée comme une technique nécessaire et non suffisante à la constitution d'unités sémantiques et au profilage sémantique des textes. Un mot du texte (lexème) n'est qu'une lexicalisation privilégiée du thème et l'on pourrait fort bien rencontrer des thèmes sans lexicalisation privilégiée, avertit Rastier (2001).

3. Unité et éléments du texte

L'annonce d'une application de la sémantique des textes en corpus à la RI et au profilage des textes pose la question de la définition de ce qu'est un texte et l'information. Ni sac de mots, ni enchaînement de phrases, un texte présente une cohésion propre qui en fait une unité linguistique et s'interprète en fonction d'autres textes qui relèvent de ce qu'on appelle l'intertexte (concrétisé par le corpus).

La linguistique des textes, équipée par la linguistique de corpus et l'Analyse des données textuelles (ADT), a su appliquer à l'unité-texte la proposition différentialiste de Saussure. Les propriétés remarquables d'un texte émergent en effet d'analyses différentielles. La RI (classification de texte) pratique ce type d'analyses et est même, du fait de la forte prégnance des mathématiques appliquées dans le domaine, prescriptrice de méthodes. Celles-ci peuvent être classificatoires comme les méthodes non supervisées dites de *clustering* telles que la Classification Hiérarchique (Hartigan 1975), ou discriminantes comme les méthodes supervisées (par exemple les Machines à Vecteurs de Support, Joachims 1998, ou K Plus Proches Voisins, Yang 1997) et enfin, l'Analyse Sémantique Latente (LSA) dont l'une des originalités est d'être dédiée au traitement automatique du langage (Deerwester 1990, Landauer *et al.* 1998).

Les linguistes savent gré aux mathématiciens de mettre à leur disposition autant d'outils et parfois même, comme pour le cas de la Sémantique Latente, de les configurer pour leur objet. Toutefois, les individus de ces modèles statistiques, lorsqu'ils ne sont pas manipulés par des

linguistes, demeurent en général des mots[1] ou, dans le meilleur des cas, des racines, voire des lemmes, bien qu'ils soient moins économiques à produire. C'est qu'en tant que discipline appliquée, la RI est guidée par des impératifs d'efficacité et de rendement. Or, les propositions du TAL et, par son biais, de la linguistique, souvent coûteuses à mettre en œuvre et pas toujours robustes, n'offrent pas une plus-value suffisante pour intéresser la RI. Les ambitions de l'ADT et, dans une moindre mesure, de la linguistique de corpus sont autres : les linguistes sont davantage intéressés par l'instrumentation de leur discipline, la création de méthode d'analyse et *d'« observables »* (Habert 2005) que par la production de systèmes finalisés.

Ainsi, si les mathématiciens de la RI ont facilement intégré le peu dépaysant principe différentiel des linguistes, la notion d'*unité texte*, qui suppose de ne considérer le mot que comme un niveau de granularité textuelle particulier parmi d'autres, demeure peu explorée. Mais les mathématiques n'ont pas de préjugé sur ce qu'est un texte – et les mathématiciens, s'ils en ont parfois, sont en général disposés à adopter ceux des linguistes, à condition qu'ils apportent une valeur ajoutée. Le linguiste se doit donc d'être à l'écoute des demandes de la RI et de prouver que, sur un certain nombre d'applications, l'analyse linguistique peut s'avérer bénéfique. Ce sont donc des préjugés de linguiste que nous exposerons ici.

Sommairement, l'apport de la linguistique à la RI peut s'énoncer en deux temps : (i) analyse et caractérisation des textes au moyen des outils de l'ADT et d'une théorie linguistique idoine (ii) sélection et production de critères de filtrage. Dans le paragraphe suivant, nous relatons une expérience de filtrage du racisme inspirée par cette approche[2].

Comment filtrer des textes racistes sur Internet, lorsque que l'on sait que d'une part, les racistes et les antiracistes partagent un vocabulaire commun dans la mesure où ils se citent et parfois se répondent, et d'autre part, que le racisme est davantage une question de rhétorique que de vocabulaire ? Plus précisément : les cibles des textes racistes évoluent (Juifs, Noirs, Arabes, « *racailles* de banlieue » ou « jeunes des quartiers », etc.) mais les modalités discriminatoires sont homogènes (nous *vs.* les autres, surnombre, combat, passage à l'acte, théorie du complot, etc.). Comment filtrer le racisme, enfin, quand la loi en punit le

[1] Des *« termes »*, selon les praticiens qui font peu de cas de l'opposition terme/lexie propre aux linguistes terminologues.

[2] Projet europén PRINCIP « Plateforme pour la Recherche, l'identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'internet », du programme Safer Internet, 2002-2004.

prosélytisme et le pousse à des stratégies de contournement, autrement dit à être euphémique, policé, en bref, présentable ? C'est à ces questions que les linguistes du projet PRINCIP ont répondu.

3.1. Forces et faiblesses de la classification automatique

Dans Vinot *et al.* (2003), une expérience de classification sur corpus a été menée pour évaluer les performances de trois algorithmes de classification automatique utilisés notamment dans les systèmes de filtrage du courrier non sollicité. Dans cette expérience, ces algorithmes fonctionnaient sur un mode contrastif à partir de deux sous-corpus catégorisés (raciste et antiraciste). En bref, il s'agissait de calculer la distance euclidienne qui sépare la représentation vectorielle d'un document (tf*idf) des autres documents du corpus (dans le cas de l'algorithme k-PPV) ou d'une classe de documents (dans le cas de l'algorithme Rocchio 1971). Ou encore, de distribuer tout nouveau document de part et d'autre d'un hyperplan séparant les données des deux sous-corpus (dans le cas de l'algorithme SVM).

Les algorithmes ont présenté de bons résultats en ce qui concerne la classification des documents issus de sites racistes dédiés, parce que ce sont les « signatures lexicales » de ces sites qui ont été discriminantes (sommaires, slogan, etc.). Mais ils restent peu efficaces lorsqu'il s'agit de détecter un alinéa ou une incise raciste dans un texte qui ne l'est pas dans son entier, ou lorsque le document est isolé (page « perso »). Par ailleurs, il suffit que la connexion lexicale entre un texte antiraciste et le sous-corpus raciste d'apprentissage soit un peu trop élevée pour que ledit texte soit classé comme raciste. Mais à l'inverse, les erreurs de classement des textes racistes ne sont pas dues au vocabulaire proprement dit, mais à des stratégies énonciatives subtiles (rhétorique de l'euphémisme, etc.).

D'une manière générale, il semble que les algorithmes aient mieux classé les documents antiracistes que les documents racistes. C'est l'indice que le discours antiraciste est d'une relative homogénéité, tandis que le discours raciste se manifeste de façon beaucoup plus variée, d'une part parce que ce dernier n'est pas lié à un système de références (lexicales, voire ontologiques) stables et est donc impossible à représenter par une unité lexicale, d'autre part, parce qu'il est actualisé dans différents discours et genres (discours idéologique, politique, genres pamphlétaire, essayiste, journalistique, etc.).

Les textes antiracistes mal classés sont essentiellement des textes littéraires, c'est-à-dire des textes qui ne répondent pas au style argumentatif caractéristique de l'antiracisme, et des textes

où, à des fins rhétoriques, les auteurs recourent abondamment à l'antiphrase et à la citation. Les documents racistes mal classés, plus nombreux, sont le plus souvent des textes politiques et idéologiques où le racisme n'est pas le thème principal et se trouve enchâssé dans une rhétorique de l'euphémisme. Les mesures statistiques sur lesquelles reposent les algorithmes de classification sont, en conséquence, inefficientes.

En résumé, on peut dire que devant deux ensembles de documents très différents quant à leur structure et aux modalités énonciatives, les algorithmes de classification et les linguistes adoptent une stratégie semblable en ce qui concerne le discours antiraciste, et différente avec le discours raciste. Les algorithmes privilégient une approche globale du document et, d'une certaine façon, négligent la dimension « énonciative » du texte raciste ; tandis que les linguistes délaissent le péritexte (sommaires, rubriques, titres) et se focalisent sur les modalités d'énonciation. L'expérience n'a pas été réalisée sur Internet directement, mais elle nous a permis de déterminer l'apport possible de la linguistique dans un projet de filtrage.

3.2. Filtrer selon les paliers de la description sémantique

Il est possible de pallier les limites des algorithmes de classification au moyen de règles linguistiques. Nous avons mis en application deux propositions théoriques de la sémantique textuelle (Rastier 1994, 2001) : (i) une interprétation et une exploitation de l'opposition *fond sémantique* vs. *formes sémantiques* (ii) une extension de cette opposition à plusieurs niveaux de complexité documentaire[3].

L'hypothèse principale présidant à l'opposition fond/formes est que des textes qui traitent d'un même domaine, ou d'un objet voisin, partagent un même fond commun mais qu'ils se singularisent par la *saillance* de formes sémantiques distinctes. Dans le cadre du projet de filtrage PRINCIP, ces formes sont soit racistes, soit antiracistes. Ce sont donc les notions générales de fond et de formes sémantiques qui ont été retenues, plutôt que les unités sémantiques auxquelles elles correspondent théoriquement. Techniquement, la collecte des textes racistes ou antiracistes, aussi bien pour le corpus d'apprentissage que pour la plateforme de filtrage, a été réalisée à partir d'une collection de mots-clés identifiés comme

[3] Dans la sémantique interprétative, le fond sémantique est assimilé à une certaine catégorie d'unités textuelles : les *isotopies*, organisées en faisceaux (une isotopie est l'effet de récurrence d'un même sème) tandis que les formes sémantiques correspondent à une autre catégorie d'unités textuelles que sont les *molécules sémiques* (groupe stable de sèmes non nécessairement lexicalisé). Nous faisons un usage étendu de cette proposition. On trouvera des développements dans Valette (2004).

étant susceptibles d'apparaître dans des textes racistes comme dans des textes antiracistes. Ces mots-clés sont en général des indicateurs de thématiques (*« immigration »*, *« races »*) ou des entités nommées (idéologues du racisme, associations, etc.). D'un point de vue théorique, ils correspondent au fond sémantique. On distingue donc deux étapes : (i) recrutement des textes racistes et antiracistes, (ii) discrimination des textes. Ces deux étapes nécessitent la production de critères de collection et de critères différentiels (cf. Valette & Grabar 2004).

Rastier observe trois niveaux d'analyse sémantique : (i) le niveau *microsémantique*, où nous étudierons les règles de constitution des lexies racistes ou antiracistes ; (ii) le niveau *mésosémantique*, où seront abordées les unités textuelles non lexicalisées, ou n'ayant pas de lexicalisation privilégiée : isotopies sémantiques, molécules sémiques ; (iii) le niveau *macrosémantique*, celui des discours et des genres textuels déterminés par un ensemble hétérogène d'indices d'expression.

3.2.1. Niveau macrosémantique : le global et le local

Le principe herméneutique selon lequel le global (le texte) détermine le local (le signe) apparaît particulièrement adapté au filtrage automatique des textes d'opinion. Les données locales, dans les textes racistes, relèvent des lexies susceptibles d'être citées par les antiracistes. Les variables quantitatives non spécifiquement lexicales, conditionnées par le genre textuel, seront assimilées à des critères globaux. Nous avons distingué deux types de critères globaux : des critères documentaires et des critères textuels.

a) Critères documentaires. Ces critères « infratextuels » relèvent d'une *sémiotique des documents*, en l'occurrence, des documents numériques du Web. L'expérience rapportée ci-dessus (Vinot *et al.* 2003) montre l'importance des données de structuration. La page Web structurée en HTML, quel qu'en soit le contenu, est soumise à des contraintes intertextuelles fortes qui déterminent la forme du document et les formes du texte. Une page Web, même « vide », présente déjà une assise structurelle commune à toutes les pages du site, que ce soit au niveau des étiquettes HTML elles-mêmes (structuration de la page, métadonnées) ou du matériel lexical affiché à l'écran. Il apparaît que globalement, les auteurs antiracistes ont davantage recours aux étiquettes de mise en forme (alinéas, énumérations, citations, etc.) que les racistes, qui se distinguent quant à eux par un usage plus poussé des possibilités multi-modales du HTML. En forçant le trait, on pourrait dire que les antiracistes mettent en ligne

des *textes* quand les racistes produisent des *documents* Internet[4]. Si les antiracistes sont gens de l'écrit, comme semble l'attester leur sens de la composition, les auteurs de textes racistes se sont appropriés le potentiel multimodal du Web avec davantage d'adresse. Nicinski (2004) a montré notamment que les codes couleurs ne sont pas les mêmes dans les documents antiracistes et dans les documents racistes. Chez ces derniers, ils reposent notamment sur un contraste clair-obscur (l'ennemi est représenté dans des nuances sombres tandis que sa « victime » est lumineuse) qui fait écho à la rhétorique que l'on rencontre dans les textes (nous *vs.* les autres). Cette analyse manuelle a été corroborée par l'analyse statistique et convertie en critères de filtrage[5].

b) Critères textuels. Les critères textuels relèvent des genres et des discours dans lesquels sont actualisés les textes racistes (et antiracistes). Une analyse manuelle des corpus d'apprentissage a permis de dresser l'inventaire des genres et des discours dans lesquels s'inscrivent la plupart des textes racistes. On a relevé principalement des discours littéraires (textes de chansons, récits, témoignages), politiques (tracts, discours, programmes) et journalistiques ou idéologiques (articles, pamphlets, textes d'opinion, faits-divers). Mais l'un des genres privilégiés par les auteurs de textes racistes est le pamphlet ou le libelle qui se manifestent par des informations textuelles caractérisant la diatribe et la polémique : points d'exclamation, adverbes de négation ou d'évaluation dénotant un style hyperbolique (*« jamais »*, *« rien »*, etc.), pronom et désinence de la deuxième personne du pluriel, morphèmes dépréciatifs (*« -âtre- »*) ou vulgaires (*« foutr- »*), etc. Comme les textes antiracistes sont rarement pamphlétaires, ces critères d'expression sont sensiblement plus fréquents dans les documents racistes qu'antiracistes. De la même façon, une étiquette morphosyntaxique ou une partie du discours présentent un potentiel discriminant propre, indépendamment du lemme auquel elles correspondent. Ainsi, on a pu constater que les

[4] Sur l'opposition texte *vs.* document, on pourra lire Pédauque (2006).

[5] Par exemple, l'opposition entre le rouge et noir apparaît très caractérisante pour une approche différentielle racisme/antiracisme. Le rouge domine dans les sites racistes : les 2/3 des occurrences des balises correspondant au rouge (92 % pour le rouge sang) relèvent du sous-corpus raciste. De même, 80,28% des images JPEG de notre corpus proviennent des sites racistes. Elles sont par ailleurs présentes dans 44,5 % des pages racistes et dans seulement 10,97 % des pages antiracistes. Enfin, si l'hypertextualité (liens internes au site) semblent bien maîtrisée par les deux parties, les racistes, là encore prennent un léger avantage en ce qui concerne la connectivité (liens externes) : 71,32% des documents racistes contiennent au moins un lien tandis qu'une moitié seulement des documents antiracistes s'ouvre vers la Toile.

substantifs étaient significativement plus fréquents dans les textes antiracistes que dans les textes racistes ; la tendance s'inverse en ce qui concerne les verbes[6].

c) Articulation des critères macrosémantiques et de l'opposition fond sémantique / formes sémantiques. Faisons l'hypothèse que le système traite un document comprenant *localement* une occurrence de la lexie raciste *« immigration-invasion »*[7], il évalue l'opinion de l'auteur à partir des critères documentaires ou textuels présents dans le texte à un *niveau global*. Le système calculera le « taux » de racisme et le « taux » d'antiracisme du document. En d'autres termes, les données globales (comme le genre, qui conditionne les critères d'expression de bas niveau (c'est-à-dire ce qui n'est pas lexical : pontuations, étiquettes morphosyntaxiques, etc.) ont une incidence sur les données locales (*i.e.* lexicales). Ces critères de bas niveau s'avèrent cruciaux dans la mesure où ils assurent la pérennité du système : si les lexies se périment, les genres et les critères qui y sont associés, eux, s'avèrent beaucoup plus stables dans le temps.

3.2.2. Niveau mésosémantique : les unités textuelles

Il s'agit de privilégier des unités textuelles qui ne sont pas les lexies et qui du point de vue du traitement, correspondent à des *cooccurrences* de morphèmes ou de mots[8], et dans une perspective strictement sémantique, de traits récurrents et de groupes de traits. À la différence des mots isolés, les unités textuelles peuvent donc être discontinues. Leur actualisation ne dépend pas de la présence de la totalité des items qui la composent, elle est graduelle. Parmi ces unités, les *thèmes*, qui sont des formes sémantiques, ont été particulièrement étudiés, dans une perspective que nous allons présenter maintenant. Le thème sémantique consiste en un « groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier » (Rastier 1994, 223). En l'absence d'un dictionnaire sémique (cf. néanmoins *infra*, le paragraphe 4.2), nous avons déterminé plusieurs façons d'exploiter la notion de thème telle que la conçoit la sémantique interprétative. La plus productive consiste à isoler les cooccurrents d'une lexie relevant du fond sémantique dans des contextes racistes puis antiracistes, de manière à en identifier les spécificités. Ces cooccurrents sont rebaptisés des

[6] On pourrait sans doute trouver plusieurs explications à ce phénomène et à ses conséquences, mais nous verserions vers une problématique « analyse de discours » qui serait ici hors de propos.
[7] Sur la collecte des lexies racistes, cf. Valette & Grabar (2004, 1109-1110).
[8] Dans l'acception qui est la nôtre, deux items sont en cooccurrence lorsqu'ils sont actualisés dans une même fenêtre prédéterminée (paragraphe, alinéa, etc.). Nous opposons ainsi la cooccurrence à la *collocation*, où les items sont immédiatement voisins.

corrélats lorsqu'ils sont jugés qualifiants sémantiquement et qu'ils sont *saillants*[9]. Ils relèvent alors des formes sémantiques (racistes ou antiracistes). Pour isoler les corrélats (formes saillantes) d'une lexie appartenant au fond sémantique, nous avons utilisé les sorties du logiciel d'ADT Hyperbase[10]. À l'aide d'un test d'écart réduit, on isole les spécificités positives des contextes du mot-pôle (issu du fond sémantique) et on en conserve les plus discriminantes. Définir la tâche Ainsi, un mot tel que *« étranger »* est considéré comme un fond sémantique d'une part, parce que, sur notre corpus de test, les mesures de rappel antiraciste et raciste sont respectivement de 55,96% et 44,04%, et d'autre part, parce que ce mot est actualisé dans 59,24% des textes dans leur ensemble. Autrement dit, le mot *« étranger »* est très fréquent dans les textes racistes et dans les textes antiracistes (dans plus d'un texte sur deux) mais il n'est pas discriminant dans la mesure où il apparaît à peu près autant dans les deux sous-corpus. L'analyse montre en revanche que, parmi les cooccurrents les plus saillants de la lexie *« étranger »* dans le sous-corpus raciste, on trouve *« illégalité »*, *« naturalisation »*, *« délinquants »* tandis que dans le sous-corpus antiraciste, on peut identifier *« régularisation »*, *« emplois »*, *« droit »*. Une fois la règle linguistique tenant compte de cette observation construite, le gain en termes de performance est convaincant : la rappel raciste atteint 90,41%, *i.e.* 9 textes racistes sur 10 sont correctement identifiés grâce à ce seul critère ; et la rappel antiraciste atteint 64,71%, *i.e.* un peu plus de 6 textes sur 10, ce qui n'est pas exceptionnel mais il faut garder à l'esprit que le filtrage s'effectue à partir de plusieurs dizaines de critères.

3.2.3. Niveau microsémantique : la composition des lexies

L'opposition fond/formes sémantiques appliquée au niveau au niveau morphématique s'est imposée au vu de la très grande créativité lexicale des auteurs de textes racistes. Nous avons constaté que beaucoup de lexies racistes étaient composées d'un ou de quelques morphèmes du fond sémantique (*« -judéo- »* , *« -démocr- »* , etc.) et d'un ou de quelques morphèmes des formes sémantiques racistes (*« -phil- »*, *« -âtr- »*, *« -crass- »*, etc.), par exemple *« judéophilie »*, *« crouillophile »*, etc.. Nous avons ainsi constitué plusieurs dictionnaires morphémiques assortis d'informations statistiques (précision, rappel) et nous avons étudié les principales règles de constitution des lexies. Comme pour les règles du palier mésosémantique,

[9] Pour une discussion, on lira Rastier (2001, 211-213).
[10] Hyperbase a été conçu et est maintenu par Étienne Brunet, université de Nice, http://ancilla.unice.fr/.

la précision d'un morphème du fond sémantique doit tendre vers 50% raciste, 50% antiraciste. Quant à son rappel, sans être déterminant, il doit être le plus élevé possible. Pour les formes sémantiques, on a choisi des morphèmes ayant un taux de précision élevé, puisque c'est la précision qui permet de différencier les deux formes sémantiques. Le système conçu détectait ainsi des patrons morphématiques comme, par exemple, *« licrasse »* qui est composé d'un morphème du fond sémantique *« LICRA »*[11] (49,45% d'occurrences antiracistes, 50,55% d'occurrences racistes, 0% d'occurrence neutre) et d'un morphème *« crass »* (76,72% d'occurrences racistes contre seulement 19,62% d'occurrences antiracistes et 3,60% d'occurrences neutre). L'objectif de cette approche est à la fois d'identifier des critères fortement caractérisants et d'anticiper sur la néologie raciste. Ainsi, si les lexies *« licrasse »* ou *« licrasseux »* ont été repérées lors de tests sur Internet en avril 2004, ce n'était pas le cas de la lexie **« licrassouille »* qui pourtant pourrait être actualisée un jour. Le taux de rappel de cet outil de filtrage microsémantique est faible, mais sa précision est très élevée et avoisine souvent les 100%.

3.3. Rendre compte de l'hétérogénéité des variables

La plate-forme de détection a été implémentée au moyen d'un système multi-agents développé par le Laboratoire d'informatique de Paris 6 (Aknine *et al.* 2005). Les règles linguistiques de filtrage étaient d'environ 300 par langue. Chacune de ces règles, qu'elle soit micro, méso ou macrosémantique, qu'elle relève de structures du document, de la morphologie, de la morphosyntaxe ou du lexique, avait la capacité d'exprimer une opinion sur les documents qu'on lui présentait. C'est la somme de plusieurs opinions, parfois contradictoires, qui permet de donner un avis général sur chaque document. Après 6 mois de fonctionnement, le taux de précision du système dans son ensemble était de 97% et son rappel de 74% (valeurs pour les documents en français). Ainsi, un quart des documents racistes présentés n'étaient pas identifiés comme tels, ce qui donne à penser que le premier lot de règles linguistiques implémentées (de l'ordre de 300) ne couvrait pas toute la variété expressive du racisme. En revanche, lorsque le système se prononçait, il ne se trompait que rarement, ce qui, dans une perspective applicative et compte tenu de la dimension expérimentale du projet, peut être considéré comme satisfaisant.

[11] Acronyme de la Ligue Internationale Contre le Racisme et l'Antisémitisme.

4. Nouvelles perspectives

Le projet PRINCIP a prouvé que la sémantique de corpus était à même de proposer des solutions de filtrage. Parmi les perspectives ouvertes dans ce champ de recherche, l'exploration de la notion de critères semble la plus prometteuse, notamment avec la prise en compte des propositions récentes de l'ADT.

4.1. Dans la marmite de l'Analyse des Données Textuelles

Tous les éléments d'un texte (définis mathématiquement comme des *variables*), quelle que soit leur nature, sont potentiellement discriminants – et donc signifiants par rapport à une tâche donné, selon le principe saussurien différentialiste. Un signe de ponctuation, une sous-chaîne de caractères, une forme ou un paragraphe, mais aussi l'annotation résultant d'une analyse liminaire (étiquette morphosyntaxique) ou d'une interprétation, ont une valeur discriminante et sont susceptibles de donner lieu à des critères de filtrage, au même titre que les mots.

Ce déplacement vers d'autres éléments discriminants que le mot, hétérogènes ou non, n'est pas à proprement parler nouveau. Il est pratiqué en linguistique de corpus et en ADT depuis plusieurs décennies déjà, avec par exemple les travaux initiés par Biber (1988) sur le profilage générique de textes. Biber relève seize catégories de traits discriminants morphosyntaxiques pour identifier des genres textuels. Plus tard, Kessler *et al.* (1997), qui poursuivent un objectif similaire, mêlent des éléments lexicaux (latinismes, abréviations), des signes de ponctuations et des éléments dérivés (mesures et pondérations). Les travaux de Malrieu & Rastier (2001) et de Beauvisage (2001) procèdent d'une même volonté de caractériser les genres textuels. Ils utilisent pour cela un jeu d'étiquettes morphosyntaxiques très complet[12].

Plusieurs travaux récents font cependant état d'expérimentations ou de prototypages relevant de cette approche générale, mais en la radicalisant. Plutôt que de demeurer dans les frontières disciplinaires héritées du positivisme logique (syntaxe, sémantique, pragmatique), ils proposent d'étudier des corrélations entre des variables relevant de différents niveaux de descriptions. Ces *corrélations* constituent de nouveaux *« observables »*. Ces observables sont aussi bien produits avec des éléments linguistiques simples (typiquement : des mots), des jeux d'étiquettes métalinguistiques, qu'avec des *mesures* effectuées sur des jeux d'étiquettes. On lira notamment les travaux de Bourion (2001), Poudat (2006), Loiseau (2006) et les

[12] Issu de l'analyseur Cordial de la société Synapse.

propositions théoriques qui y sont associées, consignées par Rastier (2005). Tandis que les combinaisons de critères hétérogènes produites dans le cadre du projet de filtrage PRINCIP par le système multi-agents n'étaient pas construites à des fins d'observation, les combinaisons de l'ADT de nouvelle génération sont au contraire un enjeu de recherche théorique. Comme le résume bien Loiseau (2008), il s'agit de considérer de nouvelles formes de contextualité : « la contextualité comme interaction des niveaux de description, voire plus généralement la contextualité comme coexistence d'interprétations différentes réifiées dans les annotations d'un corpus multi-annoté ».

Comme d'autres, nous considérons donc tous les éléments discriminants d'un texte à part égale, quelle que soit leur granularité, quelle que soit leur nature sémiotique et métalinguistique. Ces éléments peuvent être simples et obvies (signe de ponctuation, mot, etc.), construits (à l'aide de procédures et d'outils TAL comme, par exemple, les étiquettes morphosyntaxiques), composés (cooccurrence multi-niveaux combinant différents éléments discriminants d'origines variées), ou encore complexes (associant aux éléments des mesures et des pondérations).

La sélection des éléments discriminants par le linguiste s'effectue au prisme des typologies proposées dans le cadre de la sémantique textuelle (par exemple fonds et formes sémantiques, paliers textuels - lesquels ont fait la preuve de leur caractère opératoire). L'enjeu d'une telle description réside dans la dimension itérative de la production de variables. Il s'agit dans un premier temps d'inventer des *« méta-variables »* construites à partir d'autres variables (n-grammes, annotations, quantification, pondérations, etc.), et dans un second temps, de les convertir en critères de filtrage. Dès lors, un *critère* de filtrage est le résultat de traitements linguistiques réputés caractérisants ou discriminants. Le transfert de ces propositions émanant de la rencontre de la sémantique des textes et de l'ADT vers la RI est en cours, par le biais d'un projet de recherche dont l'objectif est de stabiliser une méthodologie et de produire des outils pour le filtrage de masses documentaires[13].

4.2. Le sème, l'ultime variable ?

Pour parachever ce bref panorama de l'apport des théories et des pratiques textuelles au filtrage documentaire, nous évoquerons ici la réalisation en cours d'un dictionnaire de sèmes

[13] Projet ANR-07-MDCO-002 C-MANTIC (« Méthodologie et outils pour l'application de la sémantique de corpus au filtrage de masses documentaires »), 2008-2011.

(Valette, Estacio-Moreno *et al.* 2006, Valette 2008) dont l'utilisation pour l'annotation de corpus intéresse directement la recherche thématique et la détection de l'innovation sémantique. Il s'agit de convertir un dictionnaire *informatisé* en dictionnaire *électronique* (c'est-à-dire en un dictionnaire pour le traitement automatique)[14].

Le sème, unité minimale de signification au statut purement métalinguistique, suscite des sentiments contrastés, entre fascination et rejet (rejet corrélé à celui du structuralisme en général). Élément formateur du signifié, il est un acteur clé de l'analyse sémantique des textes dans la mesure où il est susceptible de participer à toutes les unités textuelles et, évidemment, aux fonds et aux formes sémantiques. Très présent d'un point de vue théorique, son existence en termes applicatifs, c'est-à-dire dans les traitements automatiques est plus marginal, notamment parce qu'il n'existe pas de ressource sémique pour l'annotation[15]. Ainsi, dans l'analyse thématique, les thèmes sont des regroupements des mots du texte alors que ce sont théoriquement certains de leurs sèmes qui, à proprement parler, constituent le thème. Le linguiste annote alors les thèmes construits en fonction d'une analyse sémique manuelle des mots qui les composent. Or, disposer d'une ressource sémique permettrait d'étendre considérablement les capacités de détection (qu'il s'agisse de fonds ou de formes sémantiques). Concrètement, si les mots *« émeutier »* et *« banlieue »* sont en cooccurrence dans une fenêtre donnée, il s'agira d'une forme sémantique telle que l'ADT est actuellement en mesure de la détecter à partir d'un test d'écart réduit ou telle qu'elle a été détectée au palier mésosémantique dans le projet PRINCIP. Ce thème est notamment composé des sèmes /ville/ et /émeute/. Si, ailleurs dans un même corpus, *« échauffourée »* et *« quartier »* sont en cooccurrence, l'ADT n'a actuellement pas la possibilité d'associer cette cooccurrence à la première (*« émeutier »* + *« banlieue »*). Pourtant, les sèmes /ville/ et /émeute/ y sont à nouveau actualisés. Pour la RI, l'enjeu est donc d'être capable de détecter des formes sémantiques statistiquement significatives au moyen des sèmes qui les constituent indépendamment des mots du texte.

Au niveau du fond sémantique, les travaux exploratoires exposés par Grzesitchak *et al.* (2007) montrent qu'il est possible d'observer des récurrences de sèmes (c'est-à-dire des isotopies potentielles) là où les mots ne suffisent pas. Par exemple, dans un article du *Monde*

[14] En l'occurrence, il s'agit du *Trésor de la Langue Française*, cf. Dendien & Pierrel (2003).
[15] Lire toutefois Bommier-Pincemin (1999) et Rossignol (2005).

diplomatique qui traite de l'administration de la ville de Toulon par le Front National, le sème /harceler/ apparaît dans 90% des paragraphes alors que ni le mot *« harceler »*, ni ses dérivés morphologiques ne sont actualisés dans le texte. Encore expérimentale, cette recherche sur les régions peu explorées de l'infralexicalité devrait, à terme, trouver des débouchés applicatifs dans la RI. Une typologie des sèmes et des isotopies devrait améliorer son rendement et permettre notamment davantage de prédictibilité des sèmes pertinents. Ceux-ci sont, pour le moment, en partie sélectionnés manuellement.

5. Conclusion

Peut-on considérer que l'évolution récente des techniques de fouille de textes converge avec la perspective ouverte par la sémantique textuelle pour les applications RI ?

En intégrant la fouille de textes, les moteurs de recherche tels que *Google*, *Exalead* ou *Yahoo !* procèdent aux opérations de base des outils dits de *text-mining*, à savoir l'analyse linguistique pour la tokenisation et la lemmatisation des mots. Ils optent ensuite pour des approches mixtes statistiques et sémantiques en présumant que seule la deuxième approche permettra la prise en compte de certaines spécificités des corpus de textes traités (discours médical, commercial, scientifique ou autre, par exemple). Si l'hypothèse est correcte, la solution envisagée n'est pas à la hauteur des ambitions. La sémantique qu'ils préconisent est conçue comme un référentiel externe offrant des fonctionnalités comme la reconnaissance d'entité, par exemple. En somme, il y a d'un côté des connaissances issues de l'analyse linguistique et traitées statistiquement (mots clés, thesaurus, etc.), et de, l'autre, les annotations issues du référentiel que le moteur va ajouter au texte. Il s'agit donc d'une sémantique totalement extrinsèque au texte. La technologie d'analyse sémantique se résume à la production et à la maintenance de référentiels linguistiques et ontologiques, donc à un problème de gestion des connaissances. On s'en tient au principe de similarité selon lequel des mots qui apparaissent dans des contextes similaires ont des sens voisins et, par conséquent, que la tâche consistant à rechercher le document le plus proche est analogue à celle consistant à rechercher le mot sémantiquement le plus proche.

Nous exprimons toute notre reconnaissance à Evelyne Bourion, Carine Duteil-Mougel et Pierre Zweigenbaum pour leurs lectures de cet article, leurs observations et leurs suggestions.

6. Bibliographie

Aknine, S., Slodzian, A., Quenum, J-G. (2005) :« Web Personalisation for User Protection: A Multi-agent Method », B. Mobasher and S.S. Anand (eds.), *Intelligent Techniques in Web Personalization*, Springer-Verlag, 306–323.

Beauvisage, Th. (2001) : « Exploiter des données morphosyntaxiques pour l'étude statistique des genres : application au roman policier », *TAL*, vol. 42, n° 2, 579-608.

Bellot, P., Elbèze, M. (2000) « Classification locale non supervisée pour la recherche documentaire », C. Jacquemin (ed), *Traitement automatique des langues pour la recherche d'information*, TAL, vol 41, 335-367.

Biber D. (1988) : *Variation across speech and writing*, Cambridge, Cambridge University Press.

Bommier-Pincemin, B. (1999) : *Diffusion ciblée automatique d'informations : conception et mise en oeuvre d'une linguistique textuelle pour la caractérisation des destinataires et des documents*, Thèse de Doctorat, Université Paris IV Sorbonne.

Bouillon P. et al. (2000) : « Apprentissage de ressources lexicales pour l'extension de requêtes », C.Jacquemin ed., *Traitement automatique des langues pour la recherche d'information,* TAL, vol 41, 335-367.

Bourion, E. (2001) : *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Université de Nancy II, disponible sur http://www.texto-revue.net.

Dendien, J., Pierrel, J.-M. (2003) : « Le trésor de la langue française informatisé. Un exemple d'informatisation d'un dictionnaire de langue de référence », *TAL*, vol. 44 n°2, 11-37.

Deerwester, S., S. Dumais, G. W. Furnas, T. K. Landauer, R. Harshman (1990). « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science* 41 (6): 391-407.

Gaussier, E., Grefenstette, G., Hull, D., Roux, C. (2000), « Recherche d'information en français et traitement automatique des langues », C. Jacquemin ed. *Traitement automatique des langues pour la recherche d'information*, ATALA/Hermès.

Grzesitchak, M., Jacquey, E., Valette, M. (2007) : « Systèmes complexes et analyse textuelle : Traits sémantiques et recherche d'isotopies », *ARCo'07*, *Acta-Cognitica*, 227-235.

Habert, B. (2005) : « Portrait de linguiste(s) à l'instrument ». *Revue Texto ! Textes et cultures*, vol. X, n°4, disponible sur http://www.revue-texto.net, rubrique « Corpus et méthodes ».

Hartigan, J.A. (1975) : *Clustering Algorithms*, New York: John Wiley & Sons, Inc.

Jackson, P., Moulinier, I., (2002) : *Natural Language Processing for Online Applications, Text Retrieval, Extraction and Categorization,* NLP, vol.5, chap.4, John Benjamins.

Jacquemin, C. (2000), Présentation, C.Jacquemin ed. *Traitement automatique des langues pour la recherche d'information,* ATALA/Hermès.

Joachim, M.S. T. (1998) : « Text categorization with support vector machines : Learning with many relevant features ». In *ECML-98*, *Tenth European Conference on Machine Learning*, 137-142.

Kessler B., Nunberg G., Schütze H. (1997) « Automation detection of genre », *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics and the 8th Meeting of the European Chapter of the Association of Computational Linguistics*, San Francisco: Morgan Kaufman,, 32-8.

Landauer, Th. K., Foltz, P. W., Laham, D. (1998) : « Introduction to Latent Semantic Analysis », *Discourse Processes*, n° 25, 259-284.

Loiseau, S. (2006) : *Sémantique du discours philosophique : du corpus aux normes. Autour de G. Deleuze et des années 60*, Thèse de doctorat, Université Paris X-Nanterre.

Loiseau, S. (2008, à paraître) : « Comparaison de trois normes discursives à travers un corpus multi-annoté », Valette, M., éd., *Textes, documents numériques, corpus. Pour une linguistique des textes instrumentée, Syntaxe & Sémantique*, n°9.

Malrieu, D. Rastier, F. (2001) : « Genres et variations morphosyntaxiques », *TAL*, vol. 42, n°2, 548-577.

Moreau F., Claveau V., Sébillot P. (à paraître) : « Intégrer plus de connaissances linguistiques en recherche d'information peut-il augmenter les performances des systèmes ? », *Actes de CORIA'07*, St Etienne.

Nicinski, M., (2004) : « Typologie et description sémantique des images utilisées dans les sites Internet racistes », *Caractérisation des contenus de l'Internet : au-delà du lexique, l'approche sémantique*, journée ATALA, 31 janvier 2004, Paris.

Pédauque, R. T. (2006) : Le document à la lumière du numérique, Caen, C&F éditions.

Poibeau, T. (2003) : *Extraction automatique d'information, du texte brut au web sémantique,* p.184, Hermès-Lavoisier

Poudat, C. (2006) : *Etude contrastive de l'article scientifique de revue linguistique dans une perspective d'analyse des genres*, Thèse de doctorat, Université d'Orléans, disponible sur http://www.texto-revue.net.

Pustejovsky, J. (1995) : *The Generative Lexicon*, Cambridge, The MIT Press.

Rajman, M et al. (2000) : "Le modèle DSIR : une approche à base de sémantique distributionnelle pour la recherche documentaire", *Traitement automatique des langues pour la recherche d'information,* Hermès, Paris

Rastier F. (2001) : *Arts et sciences du texte*, Paris, PUF.

Rastier, F. (2005) : « Enjeux épistémologiques de la linguistique de corpus », in G. Williams (éd.). *La Linguistique de corpus*, Rennes : Presses Universitaires de Rennes, 31-46, disponible sur http://www.texto-revue.net.

Rastier, F., Cavazza, M., Abeillé, A. (1994) : *Sémantique pour l'analyse: de la linguistique à l'informatique*, Paris, Masson.

Rocchio, J. J. (1971) : « The SMART Retrieval System : Experiments in Automatic Document Processing », *Relevance Feedback in Information Retrieval*, Gerard Salton (editor), Prentice-Hall Inc. : New Jersey, 313–323.

Rossignol, M. (2005) : *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*. Thèse de doctorat, Université de Rennes 1, disponible sur http://www.texto-revue.net.

Sinclair, J. (1991) : *Corpus, concordance, collocation.* Oxford University Press.

Slodzian, M. (2000) : « L'émergence d'une terminologie textuelle et le retour du sens ». *Le sens en terminologie,* p. 61-85, Henri Béjoint et Philippe Thoiron eds. Presses universitaires de Lyon.

Sparck-Jones, K .S. (1999) : « The role of NLP in Text Retrieval », in *Natural Language Information Retrieval,* Strzalkowski (eds), Boston, MA, Kluwer, 1-24.

Valette, M. (2004) : « Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur Internet », *Approches Sémantiques du Document Numérique, CIDE 07*, 215-230, disponible sur http://www.texto-revue.net.

Valette, M. (2008, sous presse) :« À quoi servent les lexiques sémantiques ? Discussion et proposition », *Cahiers du CENTAL*.

Valette, M., Grabar, N. (2004) : « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP », *Le poids des mots, Actes des JADT 04*, 1106-1116.

Valette, M., Estacio-Moreno, A., Petitjean, J., Jacquey, E. (2006) « Éléments pour la génération de classes sémantiques à partir de définitions lexicographiques. Pour une approche sémique du sens », *Verbum ex machina, Actes de TALN 06*, P., *Cahiers du CENTAL,* 2.1, Vol. 1, 357-366, disponible sur http://www.texto-revue.net.

Vinot, R., Grabar, N., Valette, M. (2003) : « Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet », *Actes de TALN 03*, 257-284.

Voorhees, E.M. (1998): « Using WordNet for Text Retrieval », *WordNet, An Electronic* Lexical Database, C.Fellbaum ed. The MIT Press.

Yang, Y. (1997) : « An evaluation of statistical approach to text categorization », Technical Report, *CMU-CS-97-127*, Carnegie Mellon University.

Mathieu Valette
ATILF (CNRS, Nancy) - 44, av. de la Libération F-54000 Nancy
Courriel : mvalette@atilf.fr

Monique Slodzian
ERTIM (INALCO, Paris) - 2, rue de Lille F-75007 Paris
Courriel : mslodz@inalco.fr