Sémantique interprétative appliquée à la détection automatique de documents racistes et xénophobes sur internet

Mathieu Valette

SÉMANTIQUE INTERPRÉTATIVE APPLIQUÉE À LA DÉTECTION AUTOMATIQUE DE DOCUMENTS RACISTES ET XÉNOPHOBES SUR INTERNET

Mathieu Valette

Centre de Recherche en Ingénierie Multilingue, INaLCO

Version légèrement étendue d'un article paru dans :

Approches Sémantiques du Document Numérique, Actes du 7e Colloque International sur le Document Electronique, 22-25 juin 2004, Patrice Enjalbert et Mauro Gaio, éd., 2004, pp. 215-230.

Résumé :

La demande pressante des institutions en matière de protection des usagers contre les contenus illicites ou préjudiciables sur Internet (racisme, xénophobie, pédophilie) invite à dépasser les systèmes de filtrage automatique conventionnels basés sur des listes de mots-clés ou des annuaires d'adresses préétablies, peu efficaces et exigeant de fréquentes mises à jour. *PRINCIP*, la plate-forme multilingue de détection de pages Web racistes dont nous présentons quelques aspects, met en jeu une analyse sémantique globale, multicritère, et différentielle des documents. Elle repose à la fois sur les propositions théoriques de la sémantique interprétative et les possibilités offertes par l'implémentation dans un système multi-agents, tout en se démarquant des approches ontologiques classiques.

Abstract:

The authorities' pressing needs regarding Web-users' protection against illegal or abusive content on the Net -racism, xenophobia, paedophilia- have implied setting aside conventional key-word-based filtering systems as well as black lists, given their lack of efficiency and the need for frequent updating. PRINCIP, the multilingual platform for filtering racist pages on the Web is based on a global, multi-criteria differential semantic analysis of Web pages based on the breakthroughs of interpretative semantics as well as the opportunities arising from implementation in a Multi-Agent System, in contrast to conventional ontological approaches.

1. Problématique

Le projet PRINCIP (Plate-forme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet, http://www.princip.net) est un système de détection automatique des pages Web racistes, xénophobes développé conjointement par plusieurs laboratoires de recherche européens[1]. Il repose sur une critique des systèmes de filtrage actuels, et notamment sur ceux qui recourent à de simples listes de mots-clés (CyberSitter, CyberPatrol). Ceux-ci témoignent en effet d'une approche naïve du texte raciste, suggérant qu'il y a des mots racistes et des mots qui ne le sont pas, sans considération pour leur mise en texte. Autrement dit, ces systèmes reposent sur un préjugé ontologique discutable, comme si le racisme était une langue de spécialité avec une terminologie stable et univoque : il y aurait des concepts racistes et des mots leur correspondant.

Pourtant, l'analyse des textes racistes montre une toute autre réalité : d'une part, en tant qu'expression d'une opinion, le racisme n'est pas un discours référentiel, mais relève davantage de la rhétorique[2] ; d'autre part — et en conséquence — sa caractérisation et sa détection impliquent la prise en compte de l'intertextualité inhérente au Web, manifestée, dans le cas présent, par la présence de sites *sur* le racisme, c'est-à-dire antiracistes, qui partagent avec les textes racistes une part non négligeable de leur vocabulaire. En bref, l'idée de *mots*-clés racistes s'avère peu pertinente : les traits sémantiques caractéristiques du texte raciste se situent en-deçà, ou au-delà de ces mots-clés, privilégiés par l'approche ontologique.

Sans prétendre décrire de manière exhaustive l'ensemble des stratégies mise en œuvre pour détecter le racisme dans le cadre de PRINCIP, nous exposerons dans le présent article différents aspects de l'approche sémantique non ontologique que nous privilégions. Nous aborderons le problème de l'intertextualité et du choix théorique qui en découle (§ 2) ; puis nous présenterons les différents critères sémantiques retenus pour la caractérisation des documents racistes sur Internet selon les trois paliers de description du texte proposés par la sémantique interprétative de François Rastier (§ 3) ; enfin, nous décrirons brièvement les options retenues en termes d'implémentation dans un système multi-agents (§4).

2. L'intertextualité de l'Internet

2.1. Racisme et antiracisme, frontières et recouvrements

De précédent travaux d'analyse du discours, notamment ceux de Simone Bonnafous [BONN89, 91] et Pierre-André Taguieff [TAGU88], ont mis en évidence la dialectique qui oppose et lie tout à la fois les auteurs antiracistes aux auteurs racistes. Pour ces raisons, détecter les pages Web racistes n'est envisageable qu'à la condition de prendre en compte l'*intertextualité* avec d'autant plus d'attention qu'elle est massive et généralisée sur Internet, où les contenus ne sont pas qualifiés ni hiérarchisés par les moteurs de recherche.

Globalement, les modalités de l'intertextualité des documents racistes et antiracistes relèvent de la citation (les antiracistes citent les textes racistes) et de l'appropriation (les racistes s'approprient le vocabulaire antiraciste).

---

[1] Financé intégralement par la Commission Européenne, dans le cadre du *Safer Internet Action Plan*, le consortium PRINCIP comprend notamment le Centre de Recherche en Ingénierie Multilingue de l'Institut National des Langues et Civilisations Orientales de Paris, le Laboratoire d'Informatique de l'Université Paris 6-Pierre et Marie Curie, l'Institut für Germanistik de l'université Otto-von-Guericke à Magdebourg, la School of Applied Language and Intercultural Studies de la Dublin City University.

[2] Les fondements culturels de cette rhétorique ont fort bien été étudiés par Denis Blondin [BLON95] : en bref, elle repose sur l'opposition Nous *vs.* les Autres.

La rhétorique antiraciste consiste en effet à déconstruire l'argumentation des textes racistes, de sorte qu'une large place est faite aux citations, celles-ci pouvant aller du simple mot au paragraphe, voire davantage. Par conséquent, les lexies les plus stables et les plus ancrées dans le vocabulaire des auteurs racistes, c'est-à-dire celles qui feraient de bons candidats *a priori* à la constitution d'une liste de mots-clés, sont celles dont les auteurs antiracistes vont faire un usage critique privilégié. En somme, le discours rapporté fausse sensiblement les statistiques sur corpus. Par exemple, les lexies *« Race blanche »*, ou *« bougnoule »*, réputée racistes, sont en fait, dans environ deux tiers des cas, actualisées dans des textes antiracistes. Le phénomène est sensiblement le même pour le vocabulaire xénophobe d'extrême droite (*« immigrationisme »*, *« immigration-invasion »*, *« complot judéo-maçonique »*, etc.).

Parallèlement, les auteurs racistes s'approprient certaines lexies antiracistes notoires. Par exemple *« pote »*, emblème lexical de l'association SOS-Racisme (cf. le slogan *« Touche pas à mon pote »*), s'il n'est plus guère utilisé par celle-ci que dans des lexies composées figées (par exemple, les associations de quartiers *« les maisons des potes »*), est remotivé par les auteurs racistes, qui l'emploient à des fins euphémiques. Il en est de même pour le verlan *« beur »* (ou *« beurette »*), également popularisé par la lutte contre le racisme du début des années 80 (*La marche des beurs* en 1983) : dans notre corpus d'analyse, 77,22% des occurrences relèvent en fait des textes racistes.

Cette intertextualité trouve d'autres formes de manifestations plus problématiques encore, parce qu'elles ressortissent à une rhétorique de la page Web. Ainsi, tel site raciste reproduira *in extenso* un article de la presse non raciste (*L'Express*, *Le Monde*) s'il traite d'un fait de société qui intéresse son propos xénophobe (par exemple, les *tournantes*, viols collectifs commis dans les quartiers défavorisés, thème alors associé à la purification ethnique). Dans ce cas, l'euphémisation est maximale, car le Webmestre n'a pas à ajouter le moindre commentaire : le péritexte (sommaire, liens connexes) suffit aux lecteurs pour mesurer son intention.

La prise en compte de l'intertextualité impose donc de dépasser l'idée qu'il existe des concepts racistes et antiracistes (ou non racistes) actualisés de part et d'autre d'une frontière idéologique. Le matériau lexical raciste s'avère un point d'accès à la problématique, mais ne suffit pas, loin de là, à sa détection. Le racisme est l'expression d'une opinion, non la description d'un univers conceptuel.

2.2. Une approche différentielle des textes

Si, comme nous l'a enseigné Saussure et à sa suite, la sémantique structurale, la valeur linguistique est définie par des oppositions, il apparaît légitime d'adopter une approche différentielle des textes racistes et antiracistes. La sémantique différentielle apporte la solution théorique adéquate à ce cas de figure, en décrivant les éléments signifiants de la langue dans des systèmes d'oppositions et non sur un mode référentiel.

Nous présenterons dans cet article une interprétation et une mise en application de quelques unes des propositions théoriques de François Rastier émises dans le cadre de la sémantique interprétative (cf. [RAST94, 01]). À la différence d'autres travaux récents auxquels on pourra légitimement comparer notre approche ([BEUS98], [TANG97], [THLI98]), nous faisons un usage opportuniste de la théorie, n'en retenant que certains aspects jugés particulièrement adéquats à la problématique de la détection du racisme. L'objectif de PRINCIP en effet est de détecter convenablement les documents racistes sur Internet, non d'évaluer l'applicabilité de la sémantique interprétative.

En l'occurrence, nous présenterons dans cet article une interprétation et une exploitation de l'opposition *fond sémantique* vs. *forme sémantique*. Dans la sémantique interprétative, le fond sémantique est assimilé à une certaine catégorie d'unités textuelles : les *isotopies*,

organisées en faisceaux (une isotopie est l'effet de récurrence d'un même sème) tandis que les formes sémantiques correspondent à une autre catégorie d'unités textuelles que sont les *molécules sémiques* (groupe stable de sèmes non nécessairement lexicalisé).

L'hypothèse principale qui préside à PRINCIP est que les textes racistes et antiracistes partagent un même fond commun mais qu'ils se distinguent par la *saillance* de formes sémantiques soit racistes, soit antiracistes. Ce sont donc les notions générales de fond et de formes sémantiques que nous retiendrons, plutôt que les unités sémantiques auxquelles elles correspondent théoriquement.

Nous aborderons cette question aux trois niveaux d'analyse du texte définis par François Rastier :

1. le niveau microsémantique, où nous étudierons les règles de constitution des lexies racistes ou antiracistes ;

2. le niveau mésosémantique, où seront abordées les unités textuelles non lexicalisées, ou n'ayant pas de lexicalisation privilégiée : isotopies sémantiques, molécules sémiques) ;

3. le niveau macrosémantique, celui des discours et des genres textuels déterminés par un ensemble hétérogène d'indices d'expression.

3. Les critères sémantiques pour la caractérisation

3.1. Niveau macrosémantique : le global et le local

Alors que le filtrage par mots-clés repose sur un seul palier de la description linguistique, la détection multicritère mise en place par PRINCIP s'appuie sur plusieurs paliers de complexité textuelle : lexie, période ou section, et texte, ce dernier jugé primordial dans le cadre de la sémantique interprétative dans la mesure où il détermine le sens des unités de paliers inférieurs (cf. [RAST94, 01]).

La thèse défendue par François Rastier dans sa sémantique interprétative, selon laquelle le global (le texte) détermine le local (le signe) apparaît en effet particulièrement adaptée au filtrage automatique des textes d'opinion, même à un niveau d'analyse relativement rudimentaire. Les données locales, dans les textes racistes, relèvent des lexies susceptibles d'être citées par les antiracistes. Les données quantitatives non spécifiquement lexicales, conditionnées par le genre textuel, seront assimilées à des données globales.

Nous avons distingué deux types de données globales :

1. celles, proprement textuelles, relevants des genres et des discours dans lesquels sont actualisés les textes racistes (ou antiracistes) ;

2. celles, infratextuelles, qui ressortissent à une sémiotique plus générale des documents Web (images, polices de caractères, code couleurs, etc.).

Si mettre au même niveau deux types de données *a priori* fort différents peut surprendre, le rôle interprétatif des données de structuration du document HTML apparaît pourtant, comme nous allons le voir, déterminant, au même titre que les enluminures médiévales. La page Web structurée en HTML, quel qu'en soit le contenu, est soumise à des contraintes intertextuelles fortes qui déterminent la forme du document et les formes du texte. Autrement dit, une page Web, même « vide », présente déjà un fond structurel commun à toutes les pages du sites, que ce soit au niveau des étiquettes HTML elles-mêmes (structuration de la page, métadonnées) ou du matériel lexical affiché à l'écran (par exemple

le *péritexte* : sommaire, rubrique, etc.). L'ensemble constitue ce que nous appellerons la *signature sémiotique* du site.

Ainsi, sur un corpus comprenant la totalité des pages d'un site raciste donné[3], nous avons mesuré que sur le texte seul, 24,75% des occurrences de formes appartenaient à ces informations péritextuelles communes à toutes les pages du site. Sur la source HTML, étiquettes et péritexte confondus, ce pourcentage atteint 47,45% ; – c'est dire le poids de ces données souvent oubliées en linguistique textuelle. Lié à la question des genres du Web, leur statut herméneutique reste à approfondir.

3.1.1.Données globales et genres textuels

La catégorisation manuelle des corpus d'apprentissage a permis de dresser l'inventaire des genres et des discours dans lesquels s'inscrivent la plupart des textes racistes. On a relevé principalement des discours littéraires (textes de chansons, récits, témoignage), politiques (tract, discours, programme) et journalistiques ou idéologiques (article, pamphlet, opinion, faits-divers).

Mais l'un des genres privilégiés des auteurs racistes est le pamphlet ou le libelle. Cela se manifeste par des informations textuelles caractérisant la diatribe et la polémique : points d'exclamation, adverbes de négation ou d'évaluation dénotant un style outré ou hyperbolique (*« jamais »*, *« rien »*, etc.), pronom et désinence de la deuxième personne du pluriel, morphèmes dépréciatifs (*« -âtre- »*) ou vulgaires (*« foutr- »*), etc.

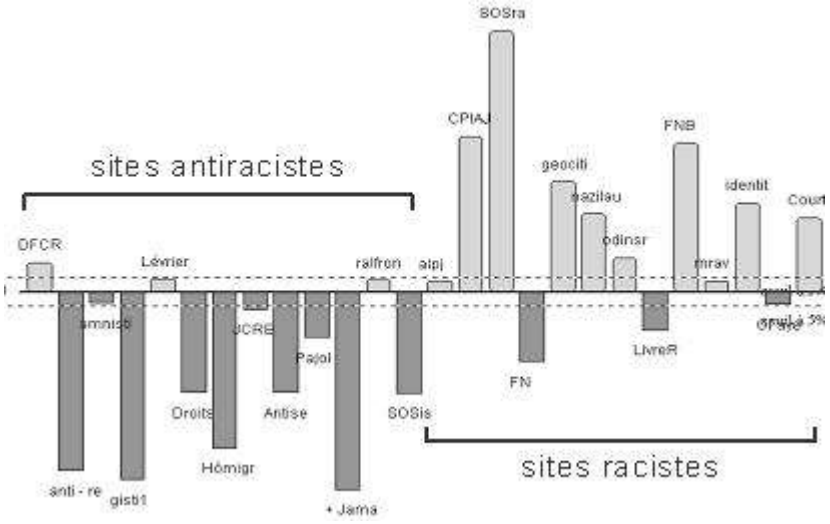

Figure 1. Fréquences relatives d'un ensemble de formes caractérisant la diatribe dans un corpus constitué de 13 sites antiracistes (à gauche) et 13 sites racistes (à droite).

Comme les textes antiracistes sont rarement pamphlétaires, ces critères d'expression sont sensiblement plus fréquents dans les documents racistes qu'antiracistes, comme l'illustre la figure 1 où sont présentées les fréquences relatives cumulées d'une sélection de formes évaluatives fréquentes dans les pamphlets (*« jamais »*, *« rien »*, *« peu »*, *« tout »*, *« trop »*, et le point d'exclamation) dans un ensemble de 13 sites antiracistes et 13 sites racistes. Les colonnes sombres marquent un déficit, les colonnes claires un excédent par rapport à une fréquence théorique indiquée par la ligne médiane. De cet écart quantitatif, la plate-forme PRINCIP infère une différence sémantique suffisante pour catégoriser adéquatement les documents.

[3] SOS-racaille.org qui a été interdit en 2003, mais dont il existe quelques avatars aujourd'hui.

Ainsi, lorsque PRINCIP aura à traiter un document comprenant, par exemple, une occurrence de la lexie raciste *« immigration-invasion »*[4], elle évaluera l'opinion de l'auteur à partir des critères d'expression dits de *bas niveau* (i.e. hors lexies racistes) présentes dans le texte et calculera le « taux » de racisme et le « taux » d'antiracisme du document. En d'autres termes, les données globales (comme le genre, qui conditionne les critères d'expression de bas niveau) ont une incidence sur les données locales (lexies) dans la mesure où elles leur donnent un *sens* raciste ou antiraciste.

Parce qu'ils sont *a priori* sans lien sémantique avec les critères lexicaux de haut niveau, et que, par conséquent, ils demeurent en-deçà de la conscience des auteurs et des lecteurs des textes, ces critères de bas niveau relèvent d'un *implicite inconscient* (déterminé par le choix du genre par l'auteur)[5]. Ce défaut de conscience s'avère crucial dans la mesure où il assure la pérennité du système : si les lexies se périment, les genres, eux, s'avèrent beaucoup plus stables dans le temps.

Les textes antiracistes présentent eux aussi des critères d'expression de bas niveau. Légalistes, ils comportent les indices du genre ; par exemple, certaines entités nommées et dates anciennes y sont fréquentes et témoignent, par-delà la problématique des genres, d'une mémoire des événements (textes de loi, action historique contre le racisme, etc.) complètement absente des textes racistes.

3.1.2.Données globales infratextuelles

Les critères globaux retenus par PRINCIP ne sont pas seulement linguistiques, et relèvent aussi de la structuration du document numérique. Le code HTML fournit de précieux critères d'expression.

Il apparaît que globalement, les auteurs antiracistes ont davantage recours aux étiquettes de mise en forme que les antiracistes, qui se distinguent quant à eux par un usage plus poussé des possibilités multi-modales du HTML. En d'autres termes, les antiracistes mettent en ligne des textes quand les racistes produisent des documents Internet. Quelques exemples en donneront une bonne illustration.

L'organisation et la hiérarchisation des parties d'un texte en titres et sous-titres (indiquées par les balises <H1>, <H2>, <H3>, etc.) est banale dans les textes antiracistes mais rarissime dans les textes racistes : 45,14% des textes antiracistes y recourent contre seulement 1,67% des textes racistes. Les taux de précision sont respectivement de 96,42% et 3,58%[6]. Les listes structurées (balises <UL>, <OL>, <LI>) sont également une spécificité antiraciste (42,89% des documents antiracistes en contiennent contre 14,15% des documents racistes, et le taux de précision est de 75,19% au bénéfice des antiracistes).

Dans la mesure où, comme nous l'avons vu, les textes antiracistes donnent une large place aux citations, les balises qui leur sont spécialement dédiées (<CITE>, <BLOCKQUOTE>) y apparaissent sensiblement plus fréquemment (dans 18,45% des cas, et seulement dans 4,84% des textes racistes), et ce avec une précision de 79,22%.

Les polices de caractère peuvent également être très discriminantes. Dans notre corpus contrasté, la police Verdana apparaît très spécifique aux pages racistes (le taux de précision raciste est de 92 % et le taux de rappel antiraciste de seulement 3 %). Les balises emphatiques (<EM>, <STRONG>), qui tendent à se substituer aux classiques italiques et gras, sont privilégiées par les antiracistes (dans 56,85% des cas, avec une précision de

[4] Sur la collecte des lexies racistes, cf. [VALE04], pp. 1109-1110.

[5] Nous savons gré à François Rastier d'avoir porté notre attention sur ce point.

[6] Le *rappel* est le rapport du nombre de documents pertinents sélectionnés au nombre total de documents du sous-corpus considéré. La *précision*, dans l'acception qui est la nôtre, est le rapport du nombre de documents sélectionnés pertinents au nombre total de documents sélectionnés dans l'ensemble du corpus. Le cumul des précisions équivaut donc à 100%.

73,87%), et ignorée des racistes dont la préférence semble aller vers le souligné (<U>) : 38,36% des documents racistes en contiennent contre 12,96% des documents antiracistes ; pour une précision raciste de 74,73%.

Si les antiracistes sont des gens de l'écrit, comme semble l'attester leur sens de la composition, les racistes se sont appropriés le potentiel multimodal du Web avec davantage d'adresse. Dans ces travaux de description sémantique des images racistes sur Internet réalisés pour PRINCIP, Monica Nicinski [NIC04] a montré que les codes couleur ne sont pas les mêmes chez les antiracistes et chez les racistes. Chez ces derniers, ils reposent en partie sur un contraste clair-obscur qui fait écho à la rhétorique que l'on rencontre dans les textes (nous *vs.* les autres)

En effet, l'opposition entre le rouge et noir, sans évidemment être une exclusivité du racisme, apparaît très caractérisante pour une approche différentielle racisme/antiracisme. Le rouge, notamment domine dans les sites racistes : nous avons en effet mesuré que dans un corpus constitué de pages HTML racistes et antiracistes de 20 millions de caractères, les deux tiers des occurrences des principales étiquettes correspondant à cette région de la palette chromatique (rouge primaire : #FF0000; rouge profond : #990000, rouge sang : #CC0000) se trouvent dans les pages racistes, avec un pic à 92 % pour le rouge sang.

De même, 80,28% des images JPEG de notre corpus proviennent des sites racistes. Elles sont par ailleurs présentes dans 44,5 % des pages racistes et dans seulement 10,97 % des pages antiracistes.

Si racistes et antiracistes font un usage semblable (statistiquement parlant) des image au format GIF, lorsqu'une image, quel que soit son format, est placée en arrière-plan (balise <BODY BACKGROUND=>), il s'agit dans 77,7 % des cas d'une page raciste. La présence de bannières est également un critère discriminant. Sur notre corpus de test, la totalité des étiquettes <BANNERS> se trouve dans les pages racistes.

Enfin, si l'hypertextualité (liens internes au site) semblent bien maîtrisée par les deux parties, les racistes, là encore prennent un léger avantage en ce qui concerne la connectivité (liens externe) : 71,32% des documents racistes contiennent au moins un lien, et parmi ces documents, la moyenne dépasse les trois liens par page (3,1 liens/page) ; tandis qu'une moitié seulement des documents antiracistes (50,62%) s'ouvre vers la Toile, en proposant, en moyenne, à peine plus d'un lien (1,07 liens/page). Enfin, les Webmestres racistes offrent plus volontiers la possibilité d'un contact par courrier électronique que leur détracteurs : 76,73% des occurrences de l'étiquette correspondante sont le fait des textes racistes.

3.2. Niveau mésosémantique : les unités textuelles

Si PRINCIP relativise l'importance des concepts et des mots-clés qui y sont associés dans son approche du texte raciste, cela ne signifie pas pour autant que les *unités* textuelles y sont négligées, bien au contraire ; il s'agit de privilégier d'autres unités textuelles, qui du point de vue du traitement, correspondent à des *cooccurrences* de morphèmes ou de mots[7], et dans une perspective strictement sémantique, de traits récurrents et de groupes de traits. À la différence des mots isolés, les unités textuelles peuvent donc être discontinues. Leur actualisation ne dépend pas de la présence de la totalité des items qui la composent, elle est graduelle, de sorte qu'il est possible de faire évoluer le seuil de présence des items à partir duquel une unité est considérée comme actualisée. Il peut être relativement bas si le document contient déjà beaucoup d'indices.

[7] Dans l'acception qui est la nôtre, deux items sont en cooccurrence lorsqu'ils sont actualisés dans une même fenêtre prédéterminée (paragraphe, alinéa, etc.). Nous opposons ainsi la cooccurrence à la *collocation*, où les items sont immédiatement voisins.

En bref, d'un point de vue quantitatif, une unité textuelle n'est pas soit présente, soit absente, elle est plus ou moins présente. Parmi ces unités, les *thèmes* (ou *molécules sémiques*) ont été particulièrement étudiés, dans une perspective que nous allons présenter maintenant. D'autres pistes de recherches ont été également explorées, notamment, les *isotopies sémantiques*. Ci-après, on abordera brièvement la question des isotopies pour ensuite exposer nos travaux sur les molécules sémiques.

3.2.1. Les isotopies sémantiques

Une isotopie sémantique est un « effet de la récurrence d'un même sème »[8] dans un texte, selon un empan pouvant aller du syntagme au texte tout entier. On s'est attaché à quelques isotopies très discriminantes et lexicalement stables, telle que, par exemple, l'isotopie *'animalité'* assimilant l'Autre à une sous-espèce humaine ou une espèce animale. Elle se trouve signifiée par *« femelle »*, *« mâle »*, *« bipède »*, *« macaque »*, *« chien »*, *« rat »*, *« cafard »*, *« cloporte »*, *« ramper »*, *« peste »*, *« choléra »*, *« vermine »*, *« proliférer »*, *« grouiller »*, *« puer »*, etc.[9] Voir, par exemple, l'extrait du site anti-Palestinien CPIAJ, ci-dessous :

> Sans doute attirés par l'*odeur infecte* des amoncellements de déchets locaux, une bande de *rats* d'Iran serait venue faire bombance au Liban [...]. Les *animaux* ont été signalés au nord de la frontière. En attendant, ils manifesteraient clairement l'intention de *nuire*. L'embêtant, c'est que maintenant il va falloir les détruire vite et bien jusqu'au dernier. Bref, depuis (plus de vingt ans) que le Shah n'est plus là, les *rats* dansent » (site CPIAJ).

Il est à noter que l'isotopie conjointe *'maladie'* (*« infecter »*, *« gangrener »*, etc.) donne également de bons résultats, mais on la rencontre également dans certains textes antifascistes.

3.2.2. Les thèmes sémantiques

Alors que l'isotopie se caractérise par la récurrence d'un sème dans un empan donné, le thème sémantique (ou *molécule sémique*) consiste en un « groupement stable de sèmes, non nécessairement lexicalisé, ou dont la lexicalisation peut varier » ([RAST94], p. 223). Comme nous n'avons pas envisagé de constituer des dictionnaires sémiques, nous avons déterminé plusieurs façons d'exploiter la notion de thèmes telle que la conçoit la sémantique interprétative. La plus productive consiste à isoler les cooccurrents d'une lexie relevant du fond sémantique dans des contextes racistes puis antiracistes, de manière à en identifier les spécificités.

Ces cooccurrents sont rebaptisés des *corrélats* lorsqu'ils sont jugés qualifiants sémantiquement et qu'ils sont *saillants* (pour une discussion, lire [RAST01], pp.211-213). Ils relèvent alors des formes sémantiques (racistes ou antiracistes).La seconde façon consiste à distinguer les différents éléments lexicaux d'un micro-récit fréquent dans les textes racistes ou antiracistes, comme par exemple les viols collectifs chez les racistes ou l'organisation de manifestations (ressortissant à la pratique de la vie associative) chez les antiracistes.

Pour isoler les corrélats (formes saillantes) d'une lexie appartenant au fond sémantique, nous avons utilisé les sorties du logiciel de lexicométrie Hyperbase (Étienne Brunet, université de Nice, http://ancilla.unice.fr/). La chaîne de traitement est la suivante :

1. relevé de tous les contextes de la lexie étudiée (ou mot-pôle, *mp*) dans le sous-corpus d'apprentissage raciste (*scr*), puis dans le sous-corpus antiraciste (*sca*),

[8] Cf. [RAST94], p. 223.

2. raboutage de l'ensemble de ces contextes de manière à constituer deux nouveaux textes, un raciste (*ntr*) et un antiraciste (*nta*),

3. mesure à l'aide d'un test d'écart-réduit (loi normale) des spécificités de *ntr* et *nta* par rapport à *scr* et *sca*,

4. sélection des corrélats dans la liste des spécificités (cooccurrents) obtenue.

A. Exemple n°1 : Formes sémantiques associées au vocable « immigration ».

Le tableau de la figure 2 représente une sélection réalisée sur deux sous-corpus, l'un composé de textes racistes (à gauche), l'autre de textes antiracistes (à droite), à partir du mot-pôle *« immigration »*.

Mot-pôle : « immigration » *(fond sémantique)*			
Environnement raciste du mot-pôle (forme sémantique raciste) :		*Environnement antiraciste du mot-pôle (forme sémantique antiraciste) :*	
écart-réduit	*forme*	*écart-réduit*	*forme*
32.14	incontrôlée	43.10	clandestine
26.16	clandestine	25.11	politique
25.78	insécurité	23.33	flux
21.97	massive	19.61	frontières
21.27	intégration	17.87	zéro
18.58	invasion	16.83	migratoires
18.16	colonisation	15.71	Weil
18.10	ratée	14.78	insécurité
16.77	peuplement	13.34	intégration
16.05	chômage	13.19	fermeture
15.21	extra	12.93	chômage
14.81	problèmes	12.18	maîtrise
14.39	population	12.17	Amsterdam
14.01	regroupement	11.89	émigration
13.44	démographique	11.58	asile
13.07	musulman	11.58	question

Figure 2 : thème sémantique d'« immigration » (extrait)

Il s'ensuit la neutralisation des cooccurrents peu ou non pertinents du mot-pôle (par exemple, dans le cas présent : *« clandestine »*, *« insécurité »*) et sa qualification *sémantique* par delà sa *signification* propre (le concept d'immigration).

Ainsi, les corrélats racistes d'*« immigration »* participent de lexies composées telles que *« immigration incontrôlée », « immigration croissante », « immigration-invasion », « immigration-colonisation », « immigration de peuplement »*, etc., tandis que les corrélats antiracistes suggèrent une problématique des *« flux migratoires »* et de la *« fermeture des frontières »*.

Les noms mentionnés par les racistes sont des personnalités politiques (Le Pen, Chirac) ou, plus incidemment, des idéologues de la lutte contre l'immigration, tandis que les « entités nommées » antiracistes relèvent de documents législatifs (rapport Weil, lois Debré, traité d'Amsterdam, etc.).

Enfin, les populations sont qualifiées de façon continentale, géographique, ethnique ou confessionnelle dans les textes racistes (« *extra-européens », « afro-arabes », « afro-*

[9] Nous avons rapporté ici des lemmes, mais on rencontre souvent plusieurs formes pour un lexème. Ainsi, *« vermine »* donne *« verminerie »*, *« vermineux »*, etc. Cette remarque a valeur générale.

maghrébins », « musulmans ») et par leurs origines nationales par les antiracistes (*« turque »*, *« italiens »*, *« portugais »*, *« algérienne »*, etc.).

B. Exemple n°2 : Formes sémantiques associées au lemme « étranger ».

L'exemple suivant repose sur le même calcul de spécificités que celui étudié ci-dessous. Il met en évidence les principaux corrélats de la lexie *« étranger »* dans nos différents sous-corpus (cf. figure 3).

Il apparaît un intéressant parallèle entre *irrégularité* et *régularisation* d'une part, et *illégalité* et *naturalisation* d'autre part. Pragmatiques, les antiracistes traitent de la mise en conformité avec la loi des immigrés en situation irrégulière, tandis que les racistes se focalisent sur le caractère hors-la-loi de certains immigrés (clandestins), avec un amalgame fort banal (les étrangers *« délinquants »*), et, plutôt que d'envisager la possibilité de leur régularisation, les racistes évoquent directement la possibilité d'une naturalisation des étrangers, ce qui ménage un accès aux thématiques récurrentes du métissage (racial, ethnique ou culturel) et de l'altération de l'identité nationale.

Si cette procédure de détection n'est pas, à l'heure où nous écrivons ces lignes, complètement stabilisée, on observe pour le moment un gain en termes de précision de l'ordre de 30% (en moyenne) par rapport aux valeurs du mot-pôle. Ainsi, sur l'exemple *« étranger »*, considéré comme du fond sémantique parce que les mesures de précisions antiracistes et racistes sont respectivement de 55,96% et 44,04%, pour un rappel moyen de 59,24%, on a mesuré, en prenant en compte la classification effectuée par les corrélats, une précision antiraciste égale à 90,41% et une précision raciste égale à 64,71%.

Mot-pôle « étranger » *(fond sémantique)*			
Environnement raciste du mot-pôle (forme sémantique raciste) :		*Environnement antiraciste du mot-pôle (forme sémantique antiraciste) :*	
écart-réduit	*forme*	*écart-réduit*	*forme*
18.98	naturalisation	79.50	séjour
13.83	naturalisés	47.94	irrégulière
13.61	vote	37.93	situation
11.78	délinquants	34.92	entrée
11.43	nationalité	32.69	régulière
11.39	turcs	31.74	emplois
11.08	devenir	31.14	droit
10.87	venus	30.56	éloignement
10.19	installés	28.23	territoire
9.74	pays	28.22	rétention
9.41	marocains	27.99	titre
9.23	prosélytisme	27.83	carte
8.77	sol	27.43	résidant
8.43	illégaux	26.97	régularisation
7.02	illégal	26.69	circulaire

Figure 3 : thème sémantique d'« étranger » (extrait)

C. Exemple n°3 : Les micro-récits

Les micro-récits sont également des avatars des molécules sémiques. À la différence des analyses présentées ci-dessus, ils ne reposent pas systématiquement (mais parfois incidemment) sur l'opposition entre fond et forme sémantique. Par exemple, les « tournantes » font partie des thèmes récurrents dans certains textes du genre « fait divers ». Par analogie avec les viols collectifs commis lors de la guerre de Yougoslavie, ils sont assimilés à une purification ethnique. Les sèmes sont de plusieurs ordres :

- Lieux : avec le sème générique /extérieur/ (l'environnement : *« banlieue »*, *« quartier »*, *« cité »*, etc.) ou /intérieur/ (le lieu du crime : *« cave »*, *« sous-sol »*, *« parking »*, *« chambre »*, *« rame »* etc.)

- Actant : /masculin/ (le bourreau : *« garçon »*, *« copain »*, *« pote »*, *« lascar »*, *« compère »*, *« mâle »*, etc.) ou /féminin/ (la victime : *« jeune fille »*, *« blanche »*, *« française »*, *« gauloise »*, etc.

- Action : /viol/ *(« tourner », « tournante »*, *« violer »*, *« pénétration »*, etc.)

Par exemple :

> Le petit rat qui deale sa came et tourne des gauloises en compagnie d'autres colons exotiques (nous soulignons).

> Au printemps 2001, une pré-*adolescente* alors agée de *douze ans* devient la cible des *agressions sexuelles* d'une *bande* de « jeunes ». Elle a, comme la plupart du temps, le profil type de la *victime* de *tournante* : c'est une petite *blanche*, issue d'un milieu modeste, particulièrement vulnérable, contrainte d'évoluer au sein d'une population majoritairement issue de l'immigration afro-maghrébine devenue plus typique de Roubaix que les bons vieux Ch'tis de ch'nord.» (nous soulignons)

Liée à la pratique de la démocratie et à la structuration associative de l'antiracisme, l'organisation de manifestations relève également de ce qu'on peut appeler des micro-récits : les antiracistes convoquent des assemblées générales, organisent des débats, coordonnent des actions militantes, etc. Ce thème comprend par exemple les corrélats suivants : *« organisation »*, *« manifester »*, *« mobilisation »*, *« rendez-vous »*, *« 14 heures »*, *« place de la République »*, *« banderole »*, etc. Par exemple :

> Mercredi 14 juin 2000 *manifestation* à Lille. En *soutien* aux grévistes de la faim qui sont dans un état critique, grande *manifestation* à Lille à *18h place de la République* (site Pajol, nous soulignons)

3.3. Niveau microsémantique : la composition des lexies

L'opposition fond/formes sémantiques appliquée au niveau lexical nous a été inspirée par la très grande créativité lexicale des auteurs racistes. Nous avons constaté que nombres des lexies racistes étaient composées d'un ou quelques morphèmes du fond sémantique et d'un ou quelques morphèmes de la forme sémantique racistes.

Une illustration exemplaire peut-être donnée par le couple d'antonymes *« judéophobie »* et *« judéophilie »* : tous deux partagent un morphème du fond sémantique (*« judéo- »*) et s'opposent par leurs suffixes : *« -phobie »* relève de la forme antiraciste tandis que *« -philie »* est une forme raciste. En collaboration avec Anne-Laure Jousse (INaLCO), nous avons ainsi constitué plusieurs dictionnaires morphémiques et étudié les principales règles de constitution des lexies.

La constitution du dictionnaire morphémique est déterminée par le taux de rappel et de précision de différents morphèmes (*« euro- »*, *« franc- »*, « *démocr- »*, etc. ou des entités nommées telles que *« LICRA »*, etc.). La précision d'un morphème du fond sémantique doit tendre vers 50% raciste, 50% antiraciste et 0% neutre, quand son rappel, sans être déterminant, doit être le plus élevé possible.

Le lexème *« démocr »*, par exemple, a été retenu pour le fond sémantique parce que ses valeurs sont, de ce point de vue, excellentes (cf. figure 4). Bien que moins exemplaire, l'entité nommée *« LICRA »* présente cependant des valeurs intéressantes (cf. figure 5) : son

taux de rappel est relativement bas, mais il s'agit d'un fond « parfait » dans la mesure où, de par sa signification très restreinte (i.e. *Ligue Internationale contre le racisme et l'Antisémitisme*), elle n'est en pratique pas actualisée dans des textes dit neutres (i.e. ne relevant ni du racisme ni de l'antiracisme).

démocr-	*rappel*	*précision*
Corpus raciste	32,92	49,96
Corpus antiraciste	29,67	45,03
Corpus neutre	3,29	5,01

Figure 4. Mesures (rappel et précision) du morphème démocr- *(fond sémantique).*

LICRA	*rappel*	*Précision*
Corpus raciste	2,19	49,45
Corpus antiraciste	2,24	50,55
Corpus neutre	0	0

Figure 5. Mesures (rappel et précision) du morphème LICRA *(fond sémantique).*

	rappel	*précision*
-mafi(a)- (*mafia*, *mafieux*, etc.)	5,61	61,46
-ouill- (magouille, fripouille, etc.)	6,09	70,68
-man- (*israëlomane*, etc.)	23,65	68,37
-crass- *(crasseux)*	1,96	76,72

Figure 6. Mesures de quelques morphèmes de la forme sémantique raciste[10]*.*

	rappel	*précision*
-phob- (islamophobie, etc.)	22,44	72,48
-circul- (circuler, circulation, etc.)	22,19	67,77
-universit- (universitaire, etc.)	12,46	58,17
résid- (résider, résidants, etc.)	19,45	57,31

Figure 7. Mesures de quelques morphèmes de la forme sémantique antiraciste.

Pour les formes sémantiques, on a choisi des morphèmes ayant un taux de précision élevé, puisque c'est la précision qui permet de différencier les deux formes sémantiques. Le taux de rappel, à cet égard, est moins déterminant. Les exemples proposés ci-dessous constituent de ce point de vue de bons spécimens. La figure 6 présente les taux de rappel et de précision de morphèmes racistes, la figure 7, ceux de quelques morphèmes antiracistes.

L'objectif de cette approche morphémique est d'anticiper sur la néologie des racistes. Ainsi, pour s'en tenir aux exemples présentés ici, si les lexies *« licrasse » « licrasseux »* (Fond *« LICRA »* + forme raciste *« crass »*) ont été repérées lors de tests sur Internet en avril 2004, ce n'est pas le cas de **« licrassouille »* qui pourtant, pourrait fort bien être actualisée un jour, d'autant plus que les lexies *« démocrasseux »* et *« démocrassouille »*, par exemple, qui reposent sur le même principe de composition, sont attestées dans nos corpus avec une précision raciste de 100%.

[10] N.B. Pour faciliter la lecture, nous ne donnons pas les mesures de rappel et de précision de ces morphèmes pour les autres sous-corpus. Ils sont évidemment très inférieurs. Par exemple, le taux de rappel antiraciste de *« -ouille- »* est de 1,24 % et sa précision de 14,45 % ; le taux de rappel neutre est de 1,28 % et sa précision de 14,86 %.

4. Analyse multicritère et système multi-agents

La plate-forme de détection est implémentée au moyen d'un système multi-agents développé au sein de l'équipe OASIS (Objets et Agents pour Systèmes d'information et de Simulation) du Laboratoire d'informatique de Paris 6 (cf. par exemple [SLOD03]).

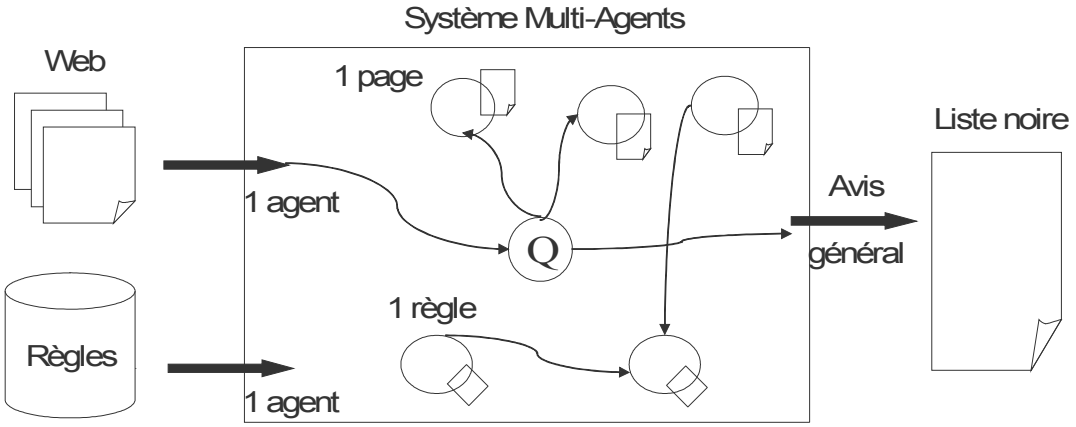

Figure 8. Plate-forme multi-agents pour l'implémentation de PRINCIP.

Les règles linguistiques permettant de détecter le racisme seront environ 300 par langue. Chacune de ces règles a la capacité d'exprimer une opinion sur les documents qu'on lui présente. C'est la somme de plusieurs opinions, parfois contradictoires qui permet de donner un avis général sur chaque document. En entrée du système multi-agents, chaque page Web est associée à un agent et se voit allouée un temps de traitement. Par ailleurs, chaque règle linguistique est associée à un agent.

Un agent de requête (Q, sur la figure 8) introduit l'ensemble des documents à analyser dans le système, ce qui a pour effet de générer autant d'agents-document, lesquels décident des agents-règles à appliquer à leur document, suivant un ensemble de critères complexes (vitesse d'exécution de la règle, nature des informations linguistiques qu'elle contient, fiabilité et adéquation de son jugement au document en présence, etc.). Chaque agent-règle exprime alors un vote sur le document. Lorsque le temps de traitement alloué est atteint, un « dépouillement » des votes est effectué. Si le résultat est jugé satisfaisant, l'agent de requête exprime un avis « raciste », « antiraciste », « ni l'un ni l'autre » ou n'en exprime aucun si le décompte des votes ne le permet pas. Dans ce dernier cas, un complément de temps de traitement peut être dispensé si des indices laissent néanmoins supposer que le document pourrait être raciste. Une catégorie particulière est réservée pour les documents qui apparaissent à la fois racistes et antiracistes.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<rule language="FR" ID="FR_dis:outgroup_myth_destruction_113">
        <active value="true"/>
        <procedure value="java:RulePatternMatch"/>
        <technique value="PatternMatcher"/>
        <clue>désagr[éè]g</clue>
        <clue>désorganis</clue>
        <clue>alt[èé]r</clue>
        <clue>d[ée]labr</clue>
        <description/>
        <function value="dis:outgroup_myth_destruction"/>
        <input value="text"/>
        <level value="substring:root"/>
        <output value="racism_evidence" weight="3" precision="85.88"/>
        <recall value="4.15"/>
        <speed value="232"/>
        <threshold value="0"/>
</rule>
```

(1) Isotopie sémantique composée d'un ensemble de lexèmes

(2) Information sémantique : rhétorique de la destruction

(3) Information sur les performances de la règle : poids, précision, rappel, vitesse d'exécution, seuil

Figure 9. Un exemple de règle sémantique : fragment de l'isotopie /destruction/

La figure 8 présente l'exemple minimal et simplifié d'une règle, en l'occurrence, une règle isotopique. Nous avons mis en évidence trois de ces caractéristiques principales. En premier lieu, un ensemble d'éléments (ici, des lexèmes) qui constituent l'isotopie ou une partie de l'isotopie sémantique retenue (1) : ces différents lexèmes partagent un trait sémique */ altérité/*. L'isotopie relève d'une catégorie préalablement identifiée comme appartenant au mythe de la destruction. Cette qualification est précisée dans un champ prévu à cet effet (2). Elle est utilisée par l'agent de requête lors du choix des règles à appliquer sur un document. Enfin, la règle a été « mesurée » selon un ensemble de critères (3) : précision, rappel, poids, vitesse d'exécution, etc. Tous ces critères, et d'autres non détaillés ici, sont utilisés dans les stratégies coopératives du système multi-agents.

La plate-forme PRINCIP sera opérationnelle en juillet 2004. Ses performances sont actuellement évaluées par la Ligue Belge des Droits de l'Homme.

5. Conclusion

Dominée par l'approche ontologique (le Web dit « sémantique ») et essentiellement cantonnée à la constitution de terminologie et à la veille, la problématique de la détection et de la catégorisation automatique est appelée à connaître quelques inflexions théoriques. Les textes susceptibles d'être traités automatiquement ne sont plus seulement ceux, univoques, des sciences et techniques. Internet, notamment, et la masse considérable de documents qui y circulent, et celle incommensurable de ceux qui y circuleront demain, créent de nouvelles demandes en termes de catégorisation, de classification et de filtrage : ce ne sont plus seulement des outils de *recherche* dont l'utilisateur a besoin, mais des outils d'*interprétation*. Ce sont donc de nouvelles méthodologies d'analyse des textes que les théoriciens doivent proposer aux ingénieurs du traitement automatique du langage.

La plate-forme PRINCIP, théoriquement fondée sur la sémantique interprétative, s'inscrit dans cette évolution. En se posant la question primordiale des genres textuels et de l'intertextualité sur le net, en travaillant avec des outils théoriques proprement sémantiques (et non seulement ontologiques), c'est-à-dire sur des morphèmes, des unités sémantiques non lexicalisées, voire sur des étiquettes HTML en tant qu'elles participent à la structuration du texte, plutôt que sur des concepts, des mots isolés et des phrases, l'équipe du projet PRINCIP entend participer à ce débat.

N.B. Les travaux de synthèse présentés ici ont pour cadre un projet collectif. Je suis redevable aux membres du consortium PRINCIP de l'ensemble de ces réflexions, et notamment à l'équipe du Centre de Recherche en Ingénierie Multilingue de l'INaLCO. J'ai en outre plaisir à remercier Évelyne Bourion, Aurélien Slodzian et François Rastier pour les remarques, suggestions et critiques qu'ils ont formulées afin d'améliorer cet article. Je tiens également à signifier ma reconnaissance à Emmanuel Cohen, Alexander Estacio Moreno et Anne-Laure Jousse pour leur participation à ces recherches.

1. Post-Scriptum

À l'occasion de la mise en ligne de cet article sur *Texto!*, six mois après la fin du projet PRINCIP, il nous semble intéressant de donner quelques informations relatives aux performances générales de la plateforme de détection. Elles tiennent en peu de chiffres : le taux de précision du système est de 97% et son rappel est de 74% (valeurs pour les documents en français). Ainsi, en l'état actuel, plus d'un tiers des documents racistes présentés ne sont pas identifiés comme tels, ce qui donne à penser que le premier lot de règles linguistiques implémentées (de l'ordre de 300) ne couvre pas toute la variété expressive du racisme. En revanche, lorsque le système se prononce, il ne se trompe que rarement, ce qui, dans une perspective applicative et compte tenu de la dimension expérimentale du projet, peut être considéré comme satisfaisant.

6. Bibliographie

[BEUS] Beust, P., *Contribution a un modèle interactionniste du sens. Amorce d'une compétence interprétative pour les machines*, Thèse de doctorat, Caen, 1998.

[BLON95] Blondin, D., *Les deux espèces humaines. Autopsie du racisme ordinaire*, Paris, L'Harmattan, 1995.

[BONN89] Bonnafous et P.A.Taguieff (éd.), *Racisme et antiracisme : frontières et recouvrements*, *Mots* 18, 1989.

[BONN91] Bonnafous S., *L'immigration prise aux mots*, Kimé, 1991.

[NINC04] Nicinski, M., « Typologie et description sémantique des images utilisées dans les sites Internet racistes », *Caractérisation des contenus de l'Internet : au-delà du lexique, l'approche sémantique*, journée ATALA organisée par F. Rastier, N. Grabar, T. Beauvisage, 31 janvier 2004, Paris.

[RAST94] Rastier, F., Cavazza, M., Abeillé, A., *Sémantique pour l'analyse: de la linguistique à l'informatique*, Paris, Masson, 1994.

[RAST01] Rastier, F., *Arts et sciences du texte*, Paris, PUF, 2001.

[SLOD03] Slodzian, A., Aknine, S., « Intelligent Agents for Tracking Racist Documents on the Internet », *Workshop on Intelligent Techniques for Web Personalization (ITWP '03)*, Acapulco (Mexique) 2003.

[TAGU88] Taguieff, P.-A., *La force du préjugé. Essai sur le racisme et ses doubles*, Paris, La découverte, 1988.

[TANG97] Tanguy, L., *Traitement automatique de la langue naturelle et interprétation : contribution à l'élaboration d'un modèle informatique de la sémantique interprétative*, Thèse de Doctorat, Rennes 1, 1997.

[THLI98] Thlivitis, T., *Sémantique Interprétative Intertextuelle : assistance informatique anthropocentrée à la compréhension des textes*, Thèse de doctorat, Rennes 1, 1998.

[VALE04] Valette, M., Grabar, N., « Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP », *Le poids des mots, Actes des 7èmes Journées internationales d'Analyse statistique des Données Textuelles (JADT), 10-12 mars 2004, Louvain-la-Neuve (Belgique),* G. Purnelle, C. Fairon, A. Dister, éds., UCL-Presses Universitaires de Louvain, 2004, p. 1106-1116.