

Caractérisation de textes à contenu idéologique : statistique textuelle ou extraction de syntagme ? l'exemple du projet PRINCIP

Mathieu Valette^{1,3}, Natalia Grabar^{1,2}

¹CRIM – INaLCO, 2 rue de Lille, F-75343, Paris cedex 07 – France

²STIM/DIAM, DSI AP-HP, Paris 6, 91 bd de l'Hôpital, F-75634 Paris cedex 13 – France

³UMR 7114 MoDyCo, Paris 10, 200 av. de la République, F-92001 Nanterre Cedex – France

mathieu.valette@inalco.fr
ngr@biomath.jussieu.fr

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. The purpose of PRINCIP, the multilingual platform for filtering racist and revisionist pages on the web is to implement a global, multi-criteria differential semantic analysis of web pages based on textual statistics, phrase extraction, and the theoretical proposals of François Rastier's interpretative semantics. In this paper, we present the results obtained by combining the use of two different NLP tools: Lexter (Didier Bourigault) and Hyperbase (Étienne Brunet).

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. L'objectif de la plate-forme multilingue de détection de pages web racistes et révisionnistes PRINCIP est de mettre en œuvre une analyse sémantique globale, multi-critères, et différentielle des documents reposant à la fois sur les statistiques textuelles, l'extraction de syntagmes, et les propositions théoriques de la sémantique interprétative de François Rastier. Nous présentons ici les résultats obtenus dans cette optique en combinant l'utilisation de deux outils distincts, Lexter (Didier Bourigault) et Hyperbase (Étienne Brunet).

Mots-clés. Lexter, extraction de syntagmes, Hyperbase, statistique textuelle, sémantique interprétative, détection, filtrage, classification automatique, texte idéologique.

1. Problématique

PRINCIP (Plate-forme pour la Recherche, l'Identification et la Neutralisation des Contenus Illégaux et Préjudiciables sur l'Internet) est un projet de détection automatique des pages web racistes et xénophobes développé conjointement par plusieurs laboratoires de recherche européens¹. Il repose sur une critique des systèmes de filtrage fondés sur des listes de mots-

¹ Financé par la Commission Européenne, dans le cadre du *Safer Internet Action Plan*, les partenaires linguistique du projet PRINCIP sont le Centre de Recherche en Ingénierie Multilingue (CRIM) de l'Institut National des Langues et Civilisations Orientales de Paris (INaLCO), l'Institut für Germanistik de l'université

clés qui témoignent d'une approche naïve du texte raciste, en suggérant qu'il y a des mots racistes et des mots qui ne le sont pas, sans considération pour leur mise en texte (ou condition d'énonciation). 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.

Or, en tant qu'expression d'une opinion, le racisme n'est pas un discours référentiel, mais il relève de la rhétorique. Par ailleurs, sa caractérisation et sa détection impliquent la prise en compte de l'*intertextualité* inhérente au web, manifestée, dans le cas présent, par la présence sur le réseau de sites *sur* le racisme, c'est-à-dire antiracistes, qui partagent avec les textes racistes une part non négligeable de leur vocabulaire. Enfin, pour des raisons qui tiennent à sa délictuosité, du moins au regard du droit français, le texte raciste fait la part belle à l'euphémisme.

Ainsi, « *bougnoule* », dans une perspective référentielle, c'est-à-dire détaché de ses conditions d'énonciation, est considéré par les locuteurs français comme un mot raciste ; mais il n'apparaît que de façon très marginale dans le vocabulaire raciste. Plus encore, l'analyse de nos différents corpus montre qu'il est trois fois plus fréquent sur les sites antiracistes que sur les sites racistes. Les auteurs racistes, dans la plupart des cas, préféreront gommer tout particularisme lexical et parler d'un « *jeune des cités* », ou d'un « *jeune des quartiers* », voire, simplement d'un « *jeune* ». Dès lors, les traits sémantiques caractéristiques du texte raciste se situeront en-deçà, ou au-delà du dictionnaire des mots-clés tels que « *bougnoule* ».

Pour caractériser les textes racistes, le recours aux statistiques s'est rapidement imposé comme cadre d'investigation. Mais quelles statistiques ? Une récente étude (Vinot *et al.*, 2003), menée en collaboration avec l'École Nationale Supérieure des Télécommunications (ENST, Paris), a permis d'évaluer les performances des algorithmes de classification automatique utilisés notamment dans les systèmes de filtrage du courrier non sollicité, ou *spam*. Ces algorithmes, qui n'ont évidemment pas de préjugés référentialistes, fonctionnent sur un mode contrastif, à partir de deux sous-corpus catégorisés (raciste et antiraciste) : en bref, il s'agit soit de calculer la distance euclidienne qui sépare la représentation vectorielle d'un document (tf*idf) des autres documents du corpus (algorithme k-PPV) ou d'une classe de documents (algorithme Rocchio), soit de distribuer tout nouveau document de part et d'autre d'un hyperplan séparant les données des deux sous-corpus (algorithme SVM).

Ces algorithmes ont présenté de bons résultats en ce qui concerne la classification des documents domiciliés sur des 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 »).

2. Objectifs

La méthodologie mise en place dans le cadre du projet PRINCIP tient compte à la fois de l'inadéquation des approches référentialistes et des limites des approches statistiques, ou, plus positivement, il s'agit d'en combiner les avantages de façon à en neutraliser les défauts.

2.1. *Lexies racistes et mise en texte*

Otto-von-Guericke à Magdebourg, la School of Applied Language and Intercultural Studies (SALIS) de la Dublin City University (DCU).

Si l'on peut difficilement parler d'Internet comme d'un corpus (Rastier, sous presse), on peut néanmoins parler de son *intertextualité* massive. Dans le cadre de notre problématique, celle-ci prend naissance dans la dialectique qui oppose les auteurs antiracistes aux auteurs racistes. La rhétorique antiraciste consiste en effet à déconstruire l'argumentation des textes racistes. Une large place est allouée aux citations (que ce soit des lexies, des phrases, des paragraphes). 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 des auteurs antiracistes vont faire un usage critique privilégié parce qu'elles sont facilement identifiables (voir l'exemple ci-dessus, « *bougnoule* », ou le vocabulaire de l'extrême droite). Parallèlement, les auteurs racistes s'approprient certaines lexies antiracistes notoires. Par exemple « *pote* », emblème lexicale de l'association SOS-Racisme, s'il n'est plus guère utilisé par celle-ci que dans des lexies composées figées (ex. les associations de quartiers « *les maisons des potes* »), est remotivé par les auteurs racistes et utilisé à des fins euphémiques.

2.2. Critères de collection, critères différentiels

Finalement, le matériel conceptuel raciste constitue un point d'accès à la problématique du racisme. Il permet de collecter *tous* les documents relevant de cette problématique, qu'ils soient racistes ou antiracistes. On appellera ce matériel conceptuel les *critères de collection*. Ce sont les conditions de leurs actualisations qui seront discriminantes et permettront de distinguer les documents racistes des documents sur le racisme. Les critères discriminants qui permettent de contraster les deux sous-corpus sont appelés les *critères différentiels*.

En termes d'analyse de corpus, un nombre non négligeable de critères de collection ont été obtenus à l'aide du logiciel Lexter (Bourigault, 1994) – un analyseur syntaxique superficiel réalisé pour le compte d'EDF, originellement dédié à l'extraction de syntagmes (nominaux et adjectivaux) à partir de corpus spécialisés, dans une perspective d'acquisition terminologique. Les critères différentiels, purement statistiques, ont été obtenus à l'aide du logiciel Hyperbase conçu par Étienne Brunetⁱⁱ – un outil de lexicométrie et d'analyse documentaire. Prévu initialement pour traiter des textes littéraires (le dictionnaire de référence utilisé est le TLF issu de la base de données textuelles Frantext de l'INaLF, mais certaines fonctions s'affranchissent de cet héritage), Hyperbase est un outil d'analyse statistique accompagné d'un concordancier.

Ces deux logiciels ont été détournés de leur fonction première : d'une part, le racisme, comme nous l'avons vu, n'est pas une langue de spécialité et l'idée d'une terminologie raciste apparaît complètement infondée ; d'autre part, les textes racistes ne sont pas des textes littéraires mais relèvent, dans l'immense majorité des cas, du discours politique ou journalistique.

3. La constitution des corpus

L'outil final PRINCIP de détection de contenus illicites est destiné à traiter directement des documents collectés sur le net. Les corpus de travail ont donc également été construits à partir de données existant sur le Web. Nous avons utilisé les moteurs de recherche généraux que nous interrogeons avec des mots clés « sensibles » (« *bande allogène* », « *gangs ethniques* »),

ⁱⁱ UMR6039 Bases, Corpus et Langage, Université de Nice (<http://ancilla.unice.fr>).

« *racaille black* », « *Sieg Heil* », etc.). La constitution du corpus a été faite en deux étapes : (i) collecte massive de documents ; (ii) catégorisation manuelle.

La collecte de documents fut réalisée de deux manières : interrogation manuelle et automatique de pages et de sites et leur rapatriement, à l'aide des outils (Grabar & Berland, 2001) et Unix Wgetⁱⁱⁱ.

Les données présentées ci-après ont été obtenues essentiellement à partir d'un corpus de sites contrasté raciste/antiraciste. Une typologie fine n'étant pas l'objet de cet article, on se contentera d'en décrire rapidement les grandes tendances :

Sites antiracistes

- explicitement dédiés à la lutte contre le racisme (*Dire et Faire contre le racisme, Jeune Contre le racisme en Europe, Antisémitisme Info, SOS-Racisme*, etc.) ;
- dédiés à des causes qui confinent à l'antiracisme : contre l'extrême-droite (*Plus jamais ça, Ras l'front*) ou pour la défense des migrants (*Groupe d'information et de soutien des Immigrés, Droit Humain, Hommes et Migrations*, etc.).

Sites racistes

- agressifs et haineux, explicitement « racialisé » (*AIPJ, CPIAJ, SOS-Racaille, NaziLauck, OdinsRage, Racist National Library*, etc.).
- plus policés, ciblant parfois selon des critères socioculturels plutôt que raciaux ou ethniques (*Unité radicale, les Identitaires, le Libre journal de la France courtois*, etc.).

4. Typologie des critères et technique d'obtention

Lexter (Bourigault, 1994) est un outil du Traitement Automatique des Langues Naturelles (TALN), dédié à l'analyse syntaxique superficielle des documents provenant d'un domaine de spécialité. Son but premier consiste donc en constitution de ressources terminologiques de ce domaine. Lexter traite des documents déjà étiquetés syntaxiquement dans lesquels il cherche les séquences syntaxiques (syntagmes nominaux et adjectivaux) aptes à représenter les *candidats termes* du domaine.

Lexter fonctionne en deux étapes :

- Recherche des frontières syntaxiques entre les candidats termes (verbes conjugués, pronoms, conjonctions, etc.). Les mots qui appartiennent à ces catégories ne peuvent pas faire partie des candidats termes et sont donc ignorés dans la suite des traitements.
- À l'intérieur des segments des phrases délimités, recherche des patrons prédéterminés qui peuvent correspondre à des *candidats termes*. Les suites de type *Nom Adj, Nom Prep Nom, Nom Prep VerbeInfinitif* constituent les patrons potentiels des termes.

Hyperbase est un logiciel d'analyse lexicométrique proposant toute une palette d'outils documentaires (deux concordanciers, dont l'un fournit les alignements et l'autre les contextes d'attestation) et statistiques (combinant des fonctions d'analyse structurale, i.e. des informations sur la distribution du vocabulaire dans un corpus : spécificités, distances, étendue, richesse, accroissement, hautes fréquences, évolution etc. ; et des fonctions

ⁱⁱⁱ Sur la constitution du corpus PRINCIP, Cf. <http://www-poleia.lip6.fr/~princip/PRINCIP-2119-D1.1.pdf>.

statistiques sur listes de formes constituées manuellement (analyses arborées, analyses factorielles).

4.1. Des candidats termes au « candidats lexies »

Les candidats termes sont des lemmes dont on a perdu le lien au contexte. Autrement dit, c'est la substance lexicale du texte raciste. On a compté pas moins de 62 000 items pour notre seul corpus raciste. Ces données brutes de Lexter sont des lexèmes (ou des lexies simples : verbes, noms, adjectifs, adverbes) et des lexies composées (syntagmes nominaux, syntagmes adjectivaux). Toutes les sorties sont lemmatisées. Comme pour les termes de métiers, la tâche consiste alors à expertiser ces lexies de façon à ne conserver que les bons candidats.

De nombreux « candidats lexies » ont été validés à partir des sorties de Lexter. Prenons l'exemple des syntagmes nominaux, c'est-à-dire des sorties comprenant 2 lemmes, généralement un nom et un adjectif. À partir des sorties de Lexter, on peut distinguer grossièrement trois ensembles de candidats lexies (composées) :

4.1.1. Les hautes fréquences, qui relèvent de la *signature lexicale* des sites, et qui le plus souvent proviennent du péritexte de la page web^{iv}. Par exemple « *vrai visage* », « *république islamique* », « *égorgement de femme et d'enfant* », etc. Un coup d'œil dans un concordancier suffit à reconnaître ces lexies récurrentes, aux contextes chaque fois identiques. Si elles constituent un matériau valable pour notre tâche de détection, du point de vue de la caractérisation du texte raciste, elles présentent l'inconvénient de faire écran aux données moins spécifiques à un ou quelques sites particuliers. Ces limites sont à rapprocher de celles observées à propos des algorithmes de classification (cf. § 1)

4.1.2. Les fréquences moyennes, sans doute les plus intéressantes, car c'est parmi elles que ce cachent les véritables lexies racistes, susceptibles d'être présentes sporadiquement mais de façon quantitativement remarquable. Par exemple, « *crime raciste* », « *islamiquement correct* », « *peuple français* », etc. Ces candidats lexies sont ensuite étudiés en contexte au moyen d'Hyperbase de façon à évaluer leur qualité en tant que lexies racistes d'une part, et en tant que lexies racistes susceptibles d'être actualisées dans un texte antiraciste d'autre part.

4.1.3. Enfin, les fréquences basses, plus difficiles à traiter, participent davantage à la dynamique investigatrice qu'à l'extraction de candidats lexies. Par exemple, elles permettent, par regroupement et recoupement d'informations, de mesurer le point d'un adjectif particulier, et la variabilité de ses instanciations, ou d'observer une construction néologique récurrente, etc. (par exemple, les différentes qualifications de « *terroriste* » : « *terroriste beur* », « *terroriste musulman* », « *terroriste tchétchène* »).

4.2. Actualisation des candidats lexies

^{iv} 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 commun à toutes les pages du sites, que ce soit au niveau des étiquettes HTML elles-mêmes (structuration de la page, métadonnées) que des formes (lexicales) affichées à l'écran (par exemple le péritexte). L'ensemble constitue la *signature sémiotique* du site. Ainsi, Sur un corpus comprenant la totalité des pages d'un site raciste donné (www.sos-racaille.org, aujourd'hui fermé), nous avons mesuré que sur le texte seul, en moyenne 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%.

Les lexies racistes, comme nous l'avons dit précédemment, sont tout aussi bien actualisées dans des textes racistes qu'antiracistes. Une fois que des candidats lexies ont été retenus, il s'agit donc d'étudier leur actualisation, c'est-à-dire leur comportement en contexte et leur poids statistique. Pour ce faire, nous avons recouru à Hyperbase en particulier pour les lexies simples. Les lexies composées sont moins aisées à traiter au moyen de ce logiciel car d'une part, les fonctions statistiques leur sont inaccessibles et d'autre part, il n'est pas possible de lemmatiser un syntagme entier. Pour quelques calculs statistiques sur lexies composées, nous avons donc utilisé des programmes en Java composés à cet effet.

Pour illustrer le rôle des statistiques textuelles dans PRINCIP, reprenons un exemple de lexie raciste mentionnée ci-dessus : « *crime raciste* ». D'aucun n'auront pas manqué de trouver paradoxale cette lexie qui semble *a priori* (i.e. d'un point de vue référentiel) antiraciste. Certes, le lemme « *crime raciste* » est un peu plus fréquent dans les textes racistes que dans les textes antiracistes (55% des occurrences apparaissent dans les textes racistes, contre 45% dans les textes antiracistes), mais cette proportion, calculée sur un nombre d'occurrences restreint, est assez peu pertinente. En revanche, sitôt que l'on contextualise un peu ce résultat, par exemple – c'est un minimum souvent discuté^v – en distinguant la forme plurielle de la forme singulière, l'écart statistique se creuse : si la répartition de la première évolue peu, celle de la forme singulière stricte s'avère sensiblement plus raciste qu'antiraciste (66,17% contre 33,83%).

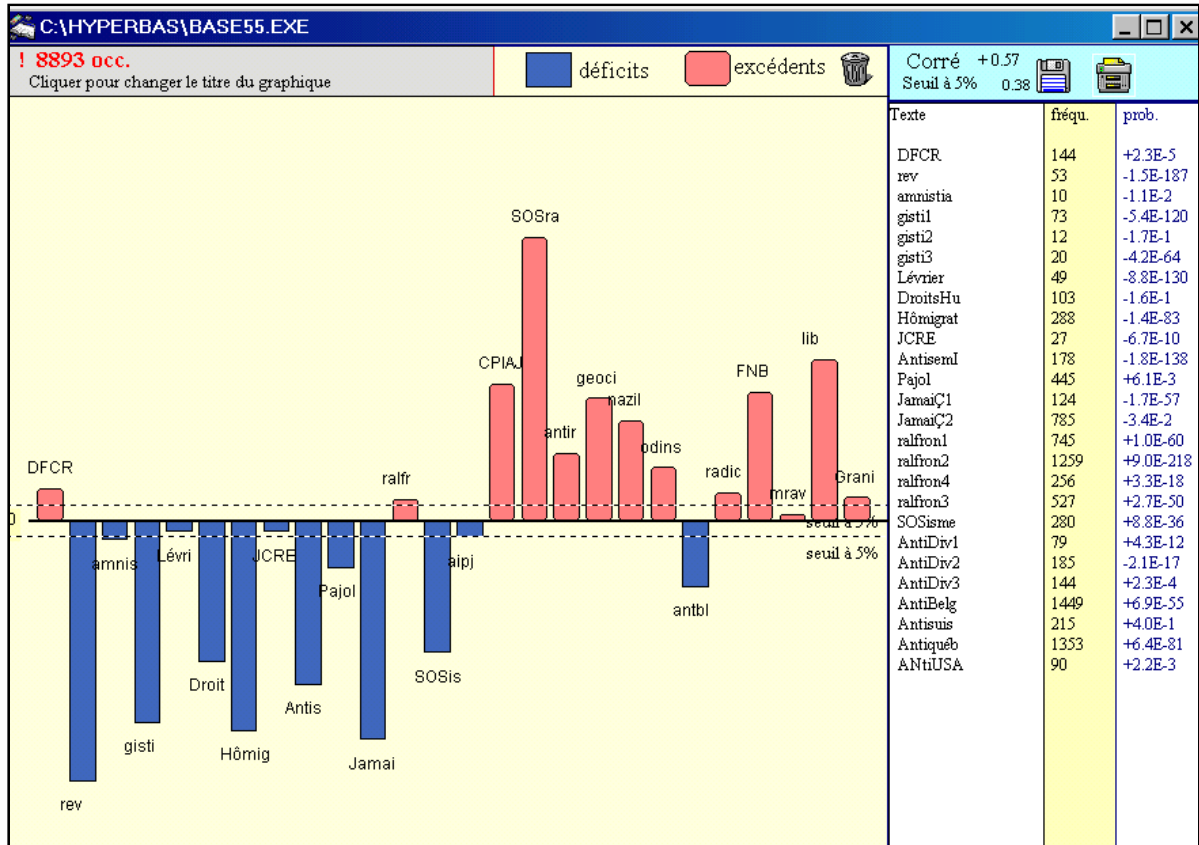
Il ne s'agit pas là d'un cas isolé mais d'une tendance lourde : le singulier est privilégié par les racistes quand, pour un même lemme, les antiracistes adopteront plus volontiers le pluriel. Ainsi, « *la femme* » vs. « *les femmes* », « *le peuple* » vs. « *les peuples* », etc. Un des exemples les plus marquants est « *droite nationale* », version auto-référentielle de ce qu'il est d'usage d'appeler l'extrême droite : 80% des occurrences de la forme singulière sont racistes quand 89% des occurrences de la forme plurielle – sensiblement plus rare en valeur absolue – sont antiracistes.

Pour opposer l'actualisation raciste et l'actualisation antiraciste, nous avons visé tout les paliers de la complexité textuelle, en nous focalisant au maximum sur les items qui ne relèvent pas ontologiquement des « domaines » du racisme. La liste des critères différentiels ci-dessous est évidemment loin d'être exhaustive, elle est indicative :

4.2.1. Les symboles

Les points d'exclamation et les points de suspension sont plus fréquents chez les racistes ; à l'inverse, les antiracistes font un usage légèrement plus important des points-virgules et des parenthèses. L'histogramme ci-dessous (capture d'écran d'Hyperbase), illustre ce contraste avec les points d'exclamation :

^v Cf. récemment Brunet 2000.



Ce graphique illustre l'écart observé entre la fréquence du point d'exclamation et la fréquence théorique attendue^{vi} (le calcul pondère cet écart selon la formule de l'écart réduit). Sur le côté gauche de l'histogramme, on a sélectionné 13 textes antiracistes, et sur le côté droit, 13 sites racistes.

4.2.2. Les morphèmes

Parmi les morphèmes, on distinguera :

(i) les morphèmes rares, mais très discriminants, comme, par exemple, certains suffixes dépréciatifs ou utilisés de façon dépréciative (-ouille, dans « démocrassouille », -âtre, dans « sémitolâtre », -mane, ou -manie, dans « israëlomane »). Nous avons alors mesuré leur pondération en termes de précision et de rappel, par exemple :

	Rappel	précision
-ethn- (<i>ethnie, ethnique, etc.</i>)	25,60	59,23
-mafi(a)- (<i>mafia, mafieux, etc.</i>)	5,61	61,46
-ouill- (<i>magouille, fripouille, etc.</i>)	6,09	70,68
-man- (<i>israëlomane, etc.</i>)	23,65	68,37
-phil- (<i>crouillophile, philomarxiste, etc.</i>)	20,00	57,76

Rappel et précision racistes de quelques morphèmes à partir du corpus de test.

^{vi} Fréquence théorique d'un mot dans un texte = fréquence du mot dans le corpus pondérée par la probabilité p ou part du texte dans le corpus.

(ii) les morphèmes fréquents, tant dans le sous-corpus antiraciste que dans le sous-corpus raciste, mais avec un écart significatif. C'est notamment le cas de certains grammèmes. Par exemple, les prépositions *à* et *de* sont plus courantes dans les textes antiracistes, tandis les adjectifs possessifs relèvent du discours raciste. Ces résultats sont corroborés par un calcul effectué sur une base Hyperbase étiquetée à l'aide de Cordial.

4.2.3. Les parties du discours

Le résultat emblématique de l'étiquetage de notre corpus, sur lequel il y aurait lieu de gloser longuement, concerne la distribution des noms et des verbes : les antiracistes privilégient globalement les premiers, les racistes les seconds. Ce résultat est confirmé par d'autres informations telles que la longueur moyenne des mots. Les mots de l'antiracisme sont plus long en raison de la composition des substantifs (le suffixes *-ation*, plus fréquent, en est un bon indice). On pourrait avancer que le texte antiraciste est réflexif et objectif tandis que le texte raciste relève du passage à l'acte, du procès. Fidèle compagnon du verbe, l'adverbe est également très présent dans les textes racistes.

Si l'on s'intéresse aux étiquettes sémantiques proposées par Cordial, on remarque que les antiracistes manipulent davantage de noms abstraits que les racistes, et que ces derniers, à l'inverse, privilégient les noms liés au corps humain et au règne animal. Cette observation est loin d'être anodine : comparer l'autre à un animal est un trait sémantique fondamental du discours raciste, où autrui est considéré comme un sous-homme, faisant partie d'une sous-race, ou d'une sous-espèce humaine.

Cette observation nous amène à nous intéresser aux lexies du racisme, c'est-à-dire, aux lexies qui constituent le vocabulaire raciste, mais qui ne sont pas, en elles-mêmes, racistes. Celles-ci peuvent être :

- des lexies simples : noms (« *chaos* », « *milliard* », « *sang* », « *destin* », « *larmes* », « *bombes* », etc.) ; verbes (« *falloir* », « *mourir* », « *dévaster* », « *massacrer* », etc.) ; adverbes (« *très* », « *jamais* », « *partout* », « *prétendument* », « *simplement* », « *vraiment* », « *naturellement* ») ; adjectifs (« *grand* », « *immense* ») ; mais aussi des dates (les dates récentes sont actualisées dans les textes racistes, les plus anciennes dans les textes antiracistes), etc.
- des lexies composées : « *coup de couteau* », « *aux mains de* », « *au cœur de* », « *en prise avec* », « *se mettre à genoux* », etc. et des lexies complexes : slogan (« *français d'abord* », « *ni raciste, ni xénophobe* »), entités nommées (« *Front uni anti-système* »), etc.

4.2.4. Les thèmes sémantiques

La fonction « thème » d'Hyperbase (mesure de l'attraction qu'un ou plusieurs mots pôles exercent sur leurs cooccurrents dans un corpus donné, à partir d'un test d'écart réduit) nous a permis de discerner des unités sémantiques complexes (isotopies sémantiques et molécules sémiqes) dont le raffinement (éventuelle lemmatisation, extraction des radicaux) a été réalisé manuellement ensuite.

Citons par exemple l'isotopie 'animalité' dont il a été question ci-dessus. Nous avons isolé un certain nombre d'items (ici lemmatisés) qui relèvent de cette isotopie : « *femelle* », « *mâle* », « *bipède* », « *macaque* », « *bâtard* », « *chien* », « *rat* », « *cafard* », « *cloporte* », « *ramper* », « *peste* », « *choléra* », « *vermine* », « *proliférer* », « *grouiller* », « *puer* », etc. Par exemple :

Sans doute attirés par l'**odeur infecte** des amoncellements de déchets locaux, une bande de **rats** d'Iran serait venue faire bombance au Liban (ah, l'état 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).

Pour la constitution du thème de la « tournante » (viol collectif), qui offre un excellent rappel dans les textes du genre fait divers, on peut étudier à titre d'exemple le thème suivant :

Écart corpus extrait	mot	Écart corpus extrait	mot
77.30 235 27	viol	12.80 1495 12	jeunes
62.25 18 6	eiders	12.55 107 3	collectifs
60.95 207 20	collectif	12.26 750 8	femmes
47.39 55 8	tournante	10.30 273 4	agression
32.62 29 4	livrent	9.45 926 7	groupe
28.47 38 4	requins	9.23 192 3	libres
26.88 24 3	répété	9.13 737 6	blanc
26.74 43 4	acquérir	8.34 613 5	femme
25.23 296 10	blanches	8.20 239 3	instant
25.09 76 5	violée	8.17 894 6	souvent
24.87 28 3	sabrina	8.15 639 5	doute
23.83 54 4	vicieux	7.53 733 5	racistes
23.83 54 4	possession	6.89 563 4	noirs
23.83 54 4	enjeu	6.70 592 4	ethnique
22.59 60 4	onze	6.38 2239 8	ans
22.40 61 4	cibles	6.05 702 4	mai
21.07 419 10	blanche	5.93 1088 5	paris
17.35 57 3	blacks	5.80 444 3	prison
17.06 104 4	présenté	5.78 758 4	victimes
16.92 164 5	racial	5.07 935 4	aucun
16.24 343 7	acte	4.68 634 3	avril
15.51 194 5	viols	4.61 1083 4	raciste
15.39 127 4	affirmer	4.57 658 3	lorsque
14.12 150 4	rap	4.24 3890 8	était
13.76 348 6	 cité	4.03 1926 5	fois

*Thème sur la cooccurrence des formes « tournante » et « viol »
sur un corpus composé exclusivement de textes racistes (seuil minimal de 3)*

Si l'on met de côté les éléments trop conjoncturels (par exemple *requins* et *vicieux* se rapportent à un fait divers particulier), on voit se dessiner un thème qu'il nous faut ensuite compléter en croisant les corpus, en analysant les contextes et en recomposant les lexies déstructurées par l'analyse des formes isolées. En résumé, on retiendra 3 catégories de traits sémantiques :

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

– **Actant** : /masculin/ (le bourreau : « *garçon* », « *copain* », « *compère* », etc.) ou /féminin/ (la victime : « *jeune fille* », « *jeune femme* », etc.)

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

À ce thème général, qui correspond à un fait divers banal, si l'on peut dire, et que l'on rencontre dans la presse non raciste, d'autres éléments lexicaux s'ajoutent et font basculer le texte dans la catégorie des racistes : c'est le cas des qualificatifs des bourreaux : « *pote* » historiquement emprunté au discours antiraciste à des fins euphémiques, mais surtout « *mâle* » qui comprend le trait /non humain/ présent dans l'isotopie 'animalité'. Quant à la victime, elle est « *blanche* », « *française* », ou encore « *gauloise* ». Enfin, les informations quantitatives ne doivent pas être négligées, car elles produisent un *effet de réel* : nombre des violeurs, et âge des actants (« *ans* »).

Par exemple :

Une **adolescente de 14 ans** rentre de l'école, cartable au bras. Elle est capturée par 3 **”jeunes”**, de 14, 15 et 17 ans. Ils l'entraînent dans une **cave**, la **violent** tour à tour. Par la porte restée entrouverte, d'autres **adolescents** et enfants regardent, tout simplement.

En résumé, l'actualisation des candidats lexies racistes extraits du corpus par Lexter est évaluée en fonction de critères lexicaux, grammaticaux ou autres (un travail sur les étiquettes HTML a également été mené en parallèle), pourvu qu'ils soient ontologiquement distincts des candidats lexies.

5. Discussion

Les objectifs différents des deux outils (Lexter pour l'acquisition terminologique d'un domaine de spécialité, Hyperbase pour l'analyse lexicométrique des documents littéraires) et, par conséquent, les fonctionnements et les fonctionnalités également différents, conditionnent les résultats qui peuvent être acquis avec chacun de ces outils.

Tout d'abord, les outils repèrent les critères de niveaux différents. Lexter fonctionne au niveau des lexies simples et des lexèmes (« *islam* », « *national* ») et des lexies composées (« *invasion islamique* », « *identité nationale* »). Hyperbase présente la possibilité d'interroger les documents sur les lexies simples et composées, avec la réserve émise précédemment (cf. 4.2.) ou encore les sous-chaînes de caractères (qui peuvent être rapprochées des morphèmes). Lexter ignore les symboles et la ponctuation (parenthèses, points, virgules, points d'exclamation). Hyperbase, pour des raisons de structuration des données, ne traite pas quelques rares symboles (\$, &).

Lexter effectue une analyse systématique des documents et la recherche de séquences correspondant aux patrons terminologiques. Avec Hyperbase, si l'utilisateur peut utilement recourir aux fonctions d'analyse de la structure du corpus, la recherche de critères est davantage amorcée d'une manière intuitive et ciblée par l'utilisateur. Le comportement de chaque critère est donc analysé suite à une requête le concernant. De là provient l'opposition entre le fonctionnement onomasiologique de Lexter (guidé par les données des documents) et sémasiologique d'Hyperbase (guidé par les idées et les intuitions de l'utilisateur étant donné ses connaissances et attentes vis-à-vis du corpus des documents).

Une dernière différence apparaît du fait que Lexter effectue une méta-analyse, ancrée sur les catégories syntaxiques des mots des documents. Il ne prend pas en compte les formes lexicales, alors que l'analyse effectuée avec l'Hyperbase est ancrée sur des chaînes lexicales données.

6. Synthèse

Nous avons présenté dans ce papier les travaux autour de l'établissement d'une base de critères pour la détection de contenus préjudiciables et racistes sur l'Internet. Les critères purement lexicaux n'étant pas suffisants, nous les complétons avec des critères d'autres niveaux :

- les caractères (symboles, ponctuation).
- les sous-chaînes de caractères (pouvant être rapprochées des morphèmes)
- les chaînes de caractères (mots ou lexies simples),
- la collocation de chaînes de caractères (lexies composées ou complexes).

La découverte et le recrutement de ces critères sont effectués avec deux outils. Lexter permet d'acquérir systématiquement tous les critères qui se trouvent aux niveaux des chaînes de caractères et de collocations (lexies). Hyperbase permet d'interroger le corpus sur les chaînes de caractères, les collocations, les sous-chaînes de caractères, etc. (lexies, morphèmes, graphèmes) prédéterminés par l'utilisateur, ainsi que sur les étiquettes morpho-syntaxiques, dans une version aménagée du logiciel.

D'autres types d'indices (liens hypertexte, code HTML, etc.) ont fait l'objet d'études séparées, avec des outils développés à cet effet. La qualité (précision) et la couverture (rappel) des indices est en cours d'évaluation sur un corpus de test. Les indices obtenus seront ensuite intégrés dans un système multi-agents développé par le Laboratoire d'Informatique de Paris 6.

Remerciements

Nous remercions Didier Bourigault d'avoir mis à notre disposition Lexter, Etienne Brunet d'avoir procédé à l'étiquetage de notre corpus pour Hyperbase, Monique Slodzian et François Rastier pour les discussions tout au cours de ce travail.

Bibliographie

- Bourigault D. (1994), *Lexter, un logiciel d'extraction de terminologie. Application à l'acquisition des connaissances à partir de textes*. Thèse en informatique linguistique, École des hautes Études en Sciences Sociales, Paris.
- Brunet, É. (2000), « Qui lemmatise dilemme attise », *Scolia, 11^{ème} rencontres linguistiques en pays rhénan*, 13, p. 7-32.
- Grabar, N. & Berland, S. (2001). Construire un corpus web pour l'acquisition terminologique. In *Terminologie et intelligence artificielle*, p. 44-54, Nancy.
- Malrieu, D., Rastier F. (2001) « Genres et variations morphosyntaxiques » *Linguistique de corpus, Traitement Automatique des Langues*, B. Daille et L. Romary, eds. vol. 42, n°2, 2001, pp. 548-577.
- Muller, Ch. (1964) Essai de statistique lexicale. "L'illusion comique", de Pierre Corneille, Paris, Klincksieck.
- Rastier, F. Cavazza, M, Abeillé, A. (1994) *Sémantique pour l'analyse : de la linguistique à l'informatique*, Paris, Masson.

Rastier, F. (sous presse) « Enjeux épistémologiques de la linguistique de corpus », Actes des deuxièmes Journées de linguistique de corpus de Lorient, Williams, G., éd., Presses Universitaires de Rennes.

Valette, M. 2003 « « Détection et interprétation automatique de contenus illicites et préjudiciables sur internet : le projet PRINCIP », *Texte ! Sémantique des textes* (<http://www.revue-texto.net>), François Rastier (dir), rubrique *Dits et Inédits*.

Vinot, R., Grabar N., Valette M. (2003) « Application d'algorithmes de classification automatique pour la détection des contenus racistes sur l'Internet », actes du colloque *TALN 2003*, 11-14 juin 2003, Batz sur Mer, pp. 257-284.