PluriTAL

**Apports croisés,
complémentaires et pluriels pour le TAL**

Mathieu VALETTE
mvalette@inalco.fr

Analyse sémantique et thématique : repérage sur le Web de sites racistes

Institut National des Langues et Civilisations Orientales

- **Internet et le nouveau statut du texte**
  - La banalisation du support numérique : une opportunité pour les sciences du texte, tant en termes de prospection scientifique (possibilité de traiter informatiquement des corpus numérisés) qu'en termes de débouchés professionnelles – les deux étant liés car les documents Internet constituent de fait des ressources (réservoir d'attestations, corpus) – étudier les textes à partir d'Internet permet d'étendre le champ professionnel des linguistes.

- **Quels métiers ? Quels nouveaux champs de prospection ?**
  - Emergence de nouveaux champs d'investigation liés à la gestion de l'information : trouver la *bonne* information.
  - Une demande social et économique dans le domaine des technologies de l'information. Un rôle majeur à jouer pour les linguistes.

- **Questionner les textes**
  - Développer et éprouver des méthodologies d'analyse des données textuelles.
  - Allier une pratique scientifique (statistiques textuelles) et des théories linguistiques.

- **Vers une problématique de l'*interprétation automatique***
  - Un exemple de recherche appliquée : le projet PRINCIP…

PluriTAL

Le projet PRINCIP (www.princip.net)

Plate-forme pour la Recherche, l'Identification, la Neutralisation des Contenus Illicites ou Préjudiciables

- **Objectif**

  Détection automatique des contenus racistes, antisémites et xénophobes sur le web, en allemand, anglais et français

- **Cadre**

  *Safer Internet Action Plan (*Commission Européenne)

- **Principaux partenaires**
  - Université Pierre et Marie Curie (Paris 6)
  - Institut National des Langues et Civilisations Orientales
  - Universität Otto-von-Guericke Magdeburg
  - Dublin City University

Problématique : le racisme sur Internet (1/4)

- **Variété des textes racistes**

  Des textes les plus explicitement haineux (archives des forums)...

  >> *SALE ARABE ON VA TE GAZER VIVE LA FRANCE MORT AUX IMMIGRÉS*

  … aux plus euphémiques

  >> *Le nécessaire combat contre l'immigration massive qui frappe notre pays peut revêtir différentes formes complémentaires, comme l'engagement politique ou associatif.*

Problématique : le racisme sur Internet (3/4)

- **Intertextualité du web**
  - Les antiracistes citent les textes racistes

>> *A l'université d'été du MEDEF, le président du conseil de surveillance d'Axa a fait une sorte mercredi sur "**la race blanche** (qui) est en train de se suicider" en raison de sa faible démographie. Il a également étalé son mépris à l 'égard du "**crétinisme rampant de certains de nos concitoyens, qui utilisent à peine 200 mots - et encore, je devrais dire 200 borborygmes**".*

http://citoyenfr.lautre.net/breve.php3?id_breve=18

>> *Le jour où une de ses collègues l'a traitée de "**sale bougnoule**", Mme Gherbi en a avisé sa chef de service.*

http://www.sos-racisme.org/article.php3?id_article=273

Résultats sur corpus contrasté raciste / antiraciste

« race blanche »

64,80 % des occurrences sont antiracistes
35,20 % des occurrences sont racistes

« bougnoule »

71,88 % des occurrences sont antiracistes
28,12 % des occurrences sont racistes

Problématique : le racisme sur Internet (4/4)

- **Intertextualité du web**
  - Les racistes s ’approprient le vocabulaire antiraciste à des fins euphémiques

>> *Une rumeur invérifiable cours selon laquelle des bandes de Roubaix prêteraient main-forte aux* ***‘jeunes’*** *des quartiers Nord d'Amiens.*

>> *« je vois de jour en jour cette invasion musulmane qui s'emplifie, les* ***potes*** *qui terorisent les gens honettes et foutent le bordel ds NOTRE BEAU PAYS doivent etre eradiquer sur le champ »*

Résultats sur corpus contrasté raciste / antiraciste

« pote »

31,36 % des occurrences sont antiracistes
61,34 % des occurrences sont racistes

« beur », « beurette »

21,78 % des occurrences sont antiracistes
77,22 % des occurrences sont racistes

Morphologie lexicale

- **Lexies racistes**

  démocrassouille
  licrasseux
  sémitolâtre, islamolâtre
  islamophile, immigrophile
  crouillasse, crouillophile, etc.

- **Lexies antiracistes**

  démocratisation
  accueil des immigrés, etc.

Morphèmes racistes et antiracistes

démocr- *licra-*
immigr- *Islam*
crouill-

Morphèmes utilisés dans le lexique raciste

-ouill- *-crass-*
-âtr- *-phil-*

Morphèmes utilisés dans le lexique antiraciste

-ation *asile*
-accueil

Lexie :
« démocrassouille »

crass-

-ouille

démocra-

-ouille	Rappel	Précision
Antiraciste	01.24	14.45
Raciste	06.09	70.68
Neutre	01.28	14.86

-crass-	Rappel	Précision
Antiraciste	00.49	19.62
Raciste	01.95	76.72
Neutre	00.09	03.60

démocra-	Rappel	Précision
Antiraciste	28.92	45.46
Raciste	31.95	50.21
Neutre	02.74	04.31

Lexie :
« Licrasseux »

crass-

-crass-	Rappel	Précision
Antiraciste	00.49	19.62
Raciste	01.95	76.72
Neutre	00.09	03.60

LICRA

licra	Rappel	Précision
Antiraciste	02.24	50.55
Raciste	02.19	49.45
Neutre	00.00	00.00

Analyse thématique
Contraintes textuelles exercées sur le lexique

Environnement raciste du lemme *étranger*				Environnement antiraciste du lemme *étranger*			
Écart	corpus	extrait	mot	Écart	corpus	extrait	mot
69.54	1162	200	étranger	252.11	4039	3388	étrangers
63.86	4039	356	étrangers	125.77	1162	910	étranger
18.98	57	12	**naturalisation**	79.50	1379	651	séjour
13.83	47	8	**naturalisés**	47.94	305	181	**irrégulière**
13.61	1019	42	vote	37.93	231	464	situation
11.78	204	15	délinquants	34.92	892	247	entrée
11.43	1197	40	nationalité	32.69	199	101	**régulière**
11.39	271	17	turcs	31.74	486	161	emplois
11.08	600	26	devenir	31.14	6537	781	droit
10.87	292	17	venus	30.56	184	91	éloignement
10.19	147	11	installés	28.23	1225	250	territoire
9.74	6090	102	pays	28.22	263	103	rétention
9.41	141	10	marocains	27.99	1263	253	titre
9.23	120	9	prosélytisme	27.83	664	172	carte
8.77	581	21	sol	27.43	178	81	résidant
8.43	88	7	**illégaux**	26.97	505	143	**régularisation**
7.02	119	7	**illégal**	26.69	458	134	circulaire

Analyse thématique
Contraintes textuelles exercées sur le lexique

Environnement raciste d'*immigration*				Environnement antiraciste d'*immigration*			
écart	corpus	extrait	mot	écart	corpus	extrait	mot
216.73	1035	995	immigration	215.45	1752	1743	immigration
32.14	23	22	**incontrôlée**	43.10	130	97	**clandestine**
30.97	145	55	issus	31.24	183	86	issus
26.16	58	29	**clandestine**	25.11	3888	436	politique
25.78	305	69	**insécurité**	23.33	283	84	**flux**
23.31	638	95	immigrés	19.61	846	138	**frontières**
21.97	146	40	**massive**	17.87	99	37	zéro
21.27	268	54	**intégration**	16.83	200	52	**migratoires**
18.58	221	43	**invasion**	15.71	116	36	Weil
18.16	166	36	**colonisation**	14.80	48	21	issues
18.10	12	9	ratée	14.78	211	48	**insécurité**
16.77	47	17	**peuplement**	13.34	541	78	**intégration**
16.05	164	32	**chômage**	13.20	593	82	matière
15.67	53	17	issues	13.19	283	52	**fermeture**
15.37	33	13	naissances	12.93	378	61	**chômage**
15.21	98	23	extra	12.85	962	110	immigrés
14.81	395	49	problèmes	12.18	117	29	maîtrise
14.39	803	73	population	12.17	84	24	Amsterdam
14.01	39	13	**regroupement**	11.89	114	28	émigration
13.44	111	22	démographique	11.58	1257	123	**asile**
13.07	588	56	musulmane	11.58	1197	119	question

Analyse thématique
Contraintes textuelles exercées sur le lexique

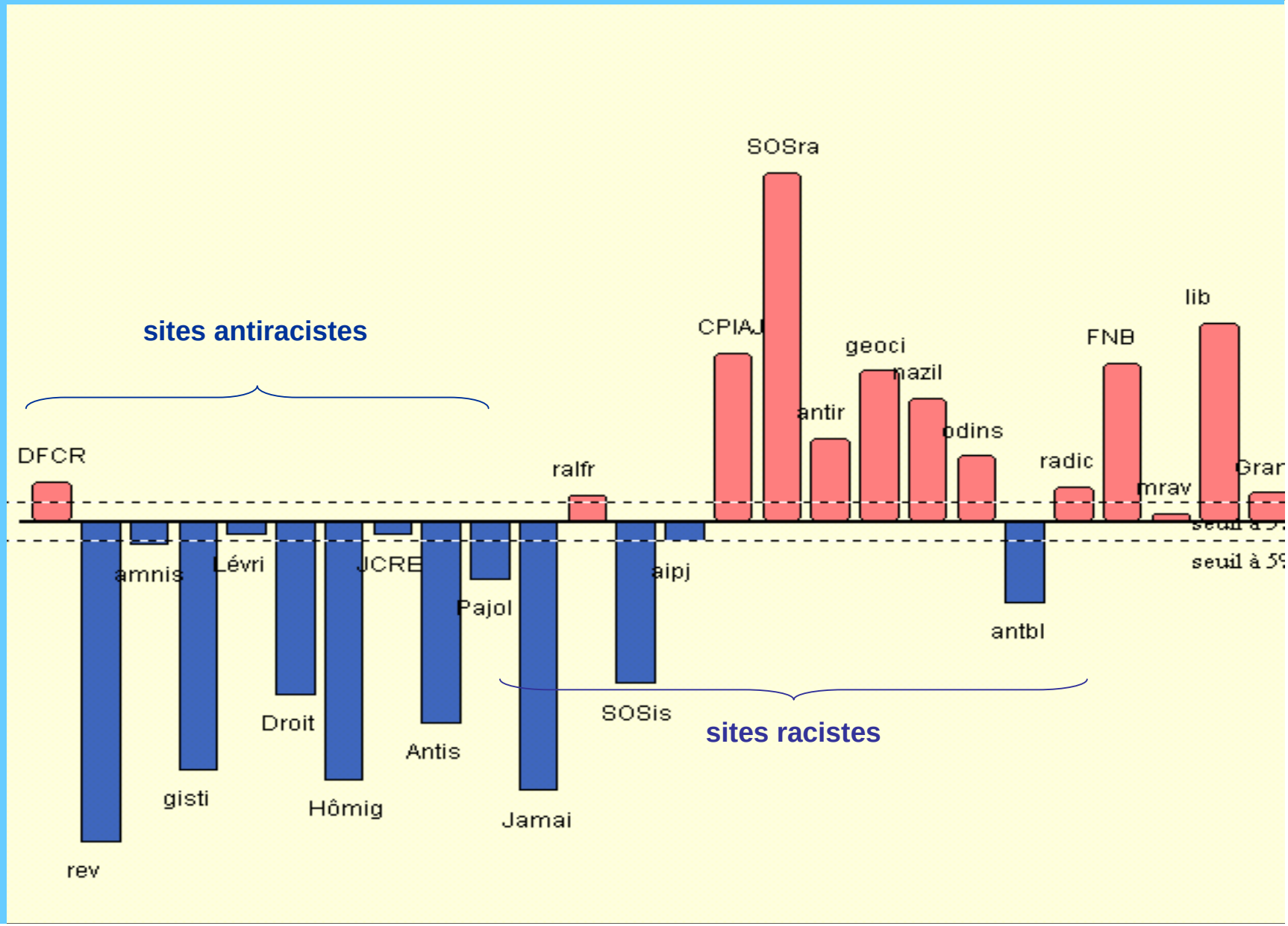
immigration

corrélats racistes	corrélats antiracistes
incontrôlée, croissante, invasion, invasive, colonisation , de peuplement, etc.	flux migratoires fermeture des frontières
problèmes	questions
Le Pen, Chirac, etc. Guillaume Faye	lois Debré, Chevènement, Pasqua, Rapport de Patrick Weil, etc. traité de Maastricht, traité d'Amsterdam, etc.
extra-européens, afro-arabes, afro-maghrébins, musulmans	turque, italiens, portugais, algérienne, etc.
lobby, média, patronat racisme, etc.	Conseiller municipal, lois, débat, mesures, propositions, pression, favoriser, etc.

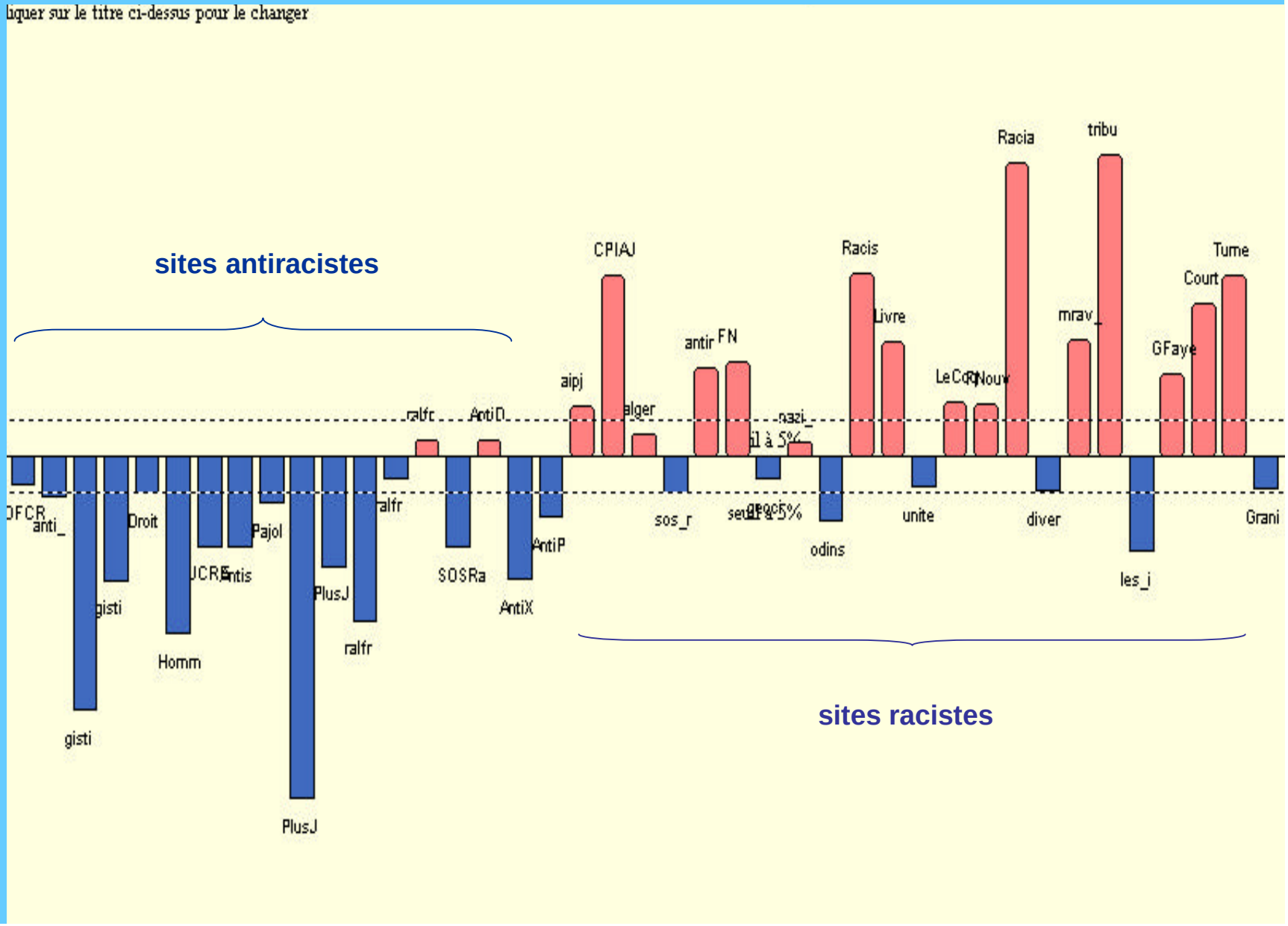
Divers critères d'expression liés à la structure du document

- Etiquettes HTML
  - .jpg : 88,2% racistes sur corpus contrasté raciste/antiraciste
  - body background : 77,7% racistes
  - #990000 #CC0000 #FF0000 (rouge) : 66,2% racistes
- Polices de caractères
  - Verdana, Lucida BlackLetter,
- Signes de ponctuation
  - Point d'exclamation : raciste (voir ci-après)
- Etiquettes morpho-syntaxiques
  - Substantif (antiraciste) vs. Verbe (raciste)
  - Conjonctions de subordination : raciste (voir ci-après)

Fréquence relative des points d'exclamation dans un corpus contrasté raciste/antiraciste



Fréquence relative des conjonctions de subordination dans un corpus contrasté raciste/antiraciste



Critères d’expression liés aux genres des textes

- Exemple : le fait divers
  - Marqueur de narration : première personne du singulier

>> *Il s'approche de **moi** et **me** sort : ” Comment tu parles toi, **je** vais te foutre des claques ! ” Chose qu'il a fait avec beaucoup de succès car **mes** lunettes volèrent **je** ne sais où . Fâché , **je** le prends par le col , et là **j'**entends ” click ”. Oui , c'était le bruit d'un couteau . **Je me** recule pour qu'il y ait une distance de sécurité. Il **m'**insulte.*

- Exemple : le pamphlet
  - **Marque d’évaluation :** trop, assez, tellement, beaucoup, etc.
  - **Négation :** jamais, non, rien, etc.

>> Le peuple Français doit se réveiller avant qu’il ne soit **trop** tard !!

>> L’invasion a déjà **beaucoup trop** avancée !! Il faut dire STOP et passer à l'action maintenant.