# SEMANTIQUE DE CORPUS POUR LES HUMANITES NUMERIQUES
# SEMANTICS OF CORPUS FOR DIGITAL HUMANITIES

Mathieu Valette
Institut National des Langues et Civilisations Orientales
mathieu.valette@inalco.fr
egle.eensoo@inalco.fr
INALCO – 2 rue de Lille, 75007 Paris

**RÉSUMÉ**. Cet article entend faire le point sur la contribution possible d'une sémantique de corpus (Rastier, 2011) aux humanités numériques, en s'appuyant sur différents exemples de traitement de fouilles de textes issus du Web social. Nous entendrons donc les humanités numériques, dans cet article, dans l'acception émergente liées l'analyse des données sociales issues du Web. Nous nous intéresserons en particulier aux productions textuelles d'internautes exprimant des valeurs et des points de vue clivés, ce qui nous permet d'élaborer le concept d'agoniste. **Mots-clés** : Sémantique de corpus ; humanités numériques ; Web social.

**ABSTRACT.** This article aims to review the possible contribution of a corpus of semantics (Rastier, 2011) to the digital humanities, relying on several examples of texts excavated treatment from the social Web. We hear the digital humanities, in this article, in the emerging sense related analyzing social data from the Web. We will focus in particular on textual productions of online expressing the values and points of view cleaved, allowing us to develop the concept of agonist. **Keywords**: corpus of semantics; digital humanities; social web.

## 1. INTRODUCTION

### *1.1. Dématérialisation et immatérialité des documents*

Les humanités numériques constituent aujourd'hui une mutation importante des sciences humaines et sociales. La banalisation du support numérique et les grands chantiers de dématérialisation des textes anciens offrent de nouvelles opportunités, non seulement en termes d'accès aux données, mais aussi en termes d'analyses renouvelées desdites données. L'utilisation d'outils informatiques de traitement permet en effet la création et l'observation de nouveaux objets sémiotiques.

S'inscrivant dans une problématique de *dématérialisation des documents*, les premiers pas des humanités numériques, relevaient d'ambitions à la fois patrimoniales, éditoriales et documentaires. Beaucoup d'initiatives consistaient en collectes, numérisations et collations de documents. Les documents étaient ainsi interrogeables au moyen d'outils d'analyse des données textuels souvent rudimentaires (concordanciers, par exemple). Les projets ont ensuite porté sur la normalisation des bases textuelles avec l'établissement de formats d'échange (Extensible Markup Language, XML) et de normes

d'encodage (Text Encoding Initiative, TEI). Cette normalisation a facilité les travaux d'annotation philologique et d'étiquetage (morphosyntaxique, lexical) et a permis de complexifier les outils d'interrogation.

Aujourd'hui, les gisements inépuisables de nouvelles formes d'expressions liées à Internet font l'objet de travaux approfondis en traitement automatique des langues (TAL), notamment en fouille de texte (*text mining*) et en fouille de données *(data mining)*. Toutefois, les humanités numériques (*digital humanities*) n'en bénéficient pas encore. Or, non seulement ces nouveaux documents numériques augurent de profondes mutations sociales et culturelles, mais ils sont une chance pour les sciences de l'homme et de la société, par exemple pour la sociologie, la psychologie sociale. En somme, ils sont de nature à inscrire les humanités numériques non plus dans une problématique de la dématérialisation des documents mais dans une problématique de l'*immatérialité des documents*. Les nouvelles formes textuelles et les nouveaux objets culturels, qu'ils soient sociaux (web 2.0) ou artistiques, témoignent de transformations sociales auxquelles les humanités numériques ont à se confronter. A cet égard, les méthodologies du TAL doivent être interrogées.

*1.2. Une linguistique, science-pivot des humanités numériques*

Actuellement, la linguistique offre principalement aux humanités numériques un ensemble de propositions philologiques. Elles participent notamment à l'établissement des textes existants dématérialisés. Cette linguistique est *ressourciste*, c'est-à-dire qu'elle est pourvoyeuse de corpus normalisés, de lexiques et d'annotations. La linguistique est donc, de plus en plus, une linguistique de *production* de données et non d'*analyse* des données. La théorie y est secondaire.

Dès lors, l'utilisation, l'adaptation et la création de nouvelles méthodologies d'analyse textuelle constituent un défi important pour les humanités numériques. À la puissance des algorithmes statistiques qui s'imposent aujourd'hui dans l'analyse des données textuelles (ADT) et dans le traitement automatique des langues (TAL), il est possible d'associer un appareil méthodologique proprement linguistique en matière d'analyse sémantique des textes.

Jadis *science-pilote* des sciences humaines, la linguistique peut prétendre aujourd'hui au statut de *science-pivot* des humanités numériques en étant notamment prescriptrice de méthodologies. La linguistique bénéficie en effet d'une expérience du texte, tant d'un point de vue théorique (linguistique textuelle, analyse du discours,

sémantique interprétative, philologie) que pratique, avec la linguistique de corpus. Complémentairement, la linguistique doit prendre part et position face aux nouveaux enjeux théoriques et méthodologiques naissants, et de ne pas laisser à d'autres disciplines (sciences de l'information, sciences de la communication, informatique) le soin de décrire, seules, ces nouveaux objets sémiotiques.

### *1.3. Objectif de l'article*

Cet article entend faire le point sur la contribution possible d'une sémantique de corpus (Rastier, 2011) aux humanités numériques, en s'appuyant sur différents exemples de traitement de fouilles de textes issus du Web social. Nous entendrons donc les humanités numériques, dans cet article, dans l'acception émergente liées l'analyse des données sociales issues du Web. Nous nous intéresserons en particulier aux productions textuelles d'internautes exprimant des *valeurs et des points de vue* clivés, ce qui nous permet d'élaborer le concept d'*agoniste*.

L'article est composé de trois parties. Dans un premier temps, nous dressons panorama critique des relations entre TAL et sémantique. Puis, nous esquissons présentons rapidement les conditions d'une sémantique de corpus outillées pour la fouille de textes. Enfin, nous illustrons notre propos au moyen de quelques exemples issus d'applications.

Nous souhaitons en particulier mettre en lumière le bénéfice potentiel d'un dialogue méthodologique entre une théorie (la sémantique textuelle), des méthodes en textométrie et les usages actuels du TAL en termes d'algorithmiques mais aussi de pratiques évaluatives.

## 2. Mise en œuvre d'une sémantique de corpus

### *2.1. Sémantique et Traitement Automatique des Langues*

Nous proposons d'envisager la sémantique en premier lieu par rapport à son instrumentation en traitement automatique des langues. Les rapports entre le TAL et la linguistique sont inconstants. Longtemps unis par des objets formels similaires sinon communs (la proposition, la phrase) et un positionnement référentialiste persistant, leur relation s'est appauvrie depuis une quinzaine d'années. Les modèles théoriques de la sémantique formelle se sont avérés inadaptés à la prise en compte de l'évolution rapide de la demande applicative à laquelle le TAL a été confronté. Jusqu'au début des années 2000, la plupart des applications concernaient la thématique, le lexique ou la

terminologie. La plupart des tâches nécessitant une automatisation (résolution d'anaphore, désambiguïsation lexicale, l'identification des parties du discours) relevaient d'une sémantique de la *phrase*. Rapidement, les technologies de l'information et la *redocumentarisation* du monde (Pédauque 2007) ont réactualisé le statut scientifique du texte – statut que la linguiste ne lui accorde encore que marginalement et au sein de courants minoritaires (analyse du discours, sémantique textuelle en France). Des tâches telles que la classification de textes et la fouille de textes ont émergé, rendant nécessaire une approche macroscopique et à grande échelle des productions langagières plus en phase avec l'unité *texte* qu'avec l'unité phrase. Les modèles formels de la sémantique de la phrase, avec leurs analyses « profondes » mais très locales apparaissent moins efficaces pour l'analyse de grands corpus, bien qu'elles proposent encore des solutions pertinentes pour l'extraction d'informations ciblées (Zweigenbaum et al. 2008).

L'essor, dans le courant des années 2000, des applications en fouille de données subjectives (fouille d'opinion, analyse des sentiments, détection des émotions, etc.) implique également une évolution des tâches en sémantique. Alors que le TAL sémantique se focalisait sur des concepts ou, d'une manière générale, des unités référentielles (entités nommées, termes, thèmes), il est aujourd'hui confronté à des *valeurs*.

Certes, les méthodes d'extraction et de classification n'ont guère évolué : naïvement, on s'imagine que les adjectifs sont aux données subjectives ce que les substantifs sont aux concepts (Strapparava et Valitutti 2004) et on applique aux premières les méthodes qui ont fait leur preuve sur les secondes. Dépasser le « lexicalisme » du TAL est un des enjeux de la sémantique. L'inventaire des objets de la sémantique susceptibles d'être appréhendés par le TAL est, en effet, loin d'être clos. Il est possible que les contraintes de genres, de discours, que la structure actancielle des textes, que le schéma de la communication, soient utiles à l'interprétation des émotions, sentiments ou des opinions.

En somme, tout se passe comme si les questions qui se posent au TAL sémantique évoluaient d'une problématique logico-formelle dominée par le primat référentiel et le choix historique de la phrase comme unité d'analyse, vers une problématique rhétorique/herméneutique dont l'objet est la réception et l'interprétation des textes considérés comme des unités de sens complexes portées par un projet de communication. La proposition a été formulée par (Rastier 2001) et oppose, *in fine*,

deux paradigmes, la linguistique des langues et la linguistique des textes. Ce moment d'incertitude paradigmatique est l'occasion d'esquisser des méthodes fondées non pas sur les présupposés théoriques du paradigme logico-grammatical mais sur le paradigme rhétorique/herméneutique, peu exploré encore en TAL. Concrètement, nous proposons de combiner une approche inspirée de la sémantique textuelle (Rastier 2011) et les méthodes de l'analyse statistique des données textuelles.

L'analyse statistique des données textuelles (ou textométrie) est un ensemble particulier de pratiques relevant du champ général de la linguistique de corpus. Elle compte des traitements statistiques (analyse factorielle des correspondances, spécificités fondées sur le modèle hypergéométrique, cooccurrences, etc.), mais aussi des outils de visualisation des corpus (nuages de mots, histogrammes, etc.) et documentaires (concordanciers) destinés à l'aide à l'interprétation des textes.

*2.2. Vers une sémantique de corpus : la textométrie et la sémantique des textes*

Les affinités de la textométrie et de la sémantique des textes ont été identifiées précocement (Rastier, éd. 1995). La plupart ont été explicitées par (Mayaffre 2008) et de façon systématique par (Pincemin 2010) à laquelle nous renvoyons le lecteur.

*Le texte ne fait l'objet d'aucune préconception réductrice* – Les signes qui composent le texte ne sont pas hiérarchisés (les substantifs ne sont pas préférés *a priori* aux mots grammaticaux ou aux signes de ponctuations) et ne sont pas substituables par des constructions artificielles (en particulier si elles sont de haut niveau, tels les concepts, les hyperonymes, les synonymes). Or, l'annotation de corpus au moyen de ressources variées est non seulement très courante en TAL mais ne fait guère l'objet de réflexion critique. Pourtant, même le traitement basique qui consiste à lemmatiser un corpus, parce qu'elle en factorise les formes, fait l'objet de débats circonspects en textométrie (Brunet 2001) comme en sémantique des textes (Bourion 2001).

*Le retour au texte est la condition de l'interprétation* – L'analyse en textométrie comme en sémantique textuelle repose sur une itération entre l'analyse des sorties logicielles et la consultation des textes ; en d'autres termes, la connaissance des textes est une condition nécessaire à leur analyse, elle est notamment génératrice d'hypothèses interprétatives. Comme on l'a vu dans le paragraphe précédent, les données linguistiques qualitatives sont, de plus en plus fréquemment, exclues des articles de TAL. On leur préfère des données quantitatives.

*Le contexte global construit par le corpus de référence joue un rôle déterminant dans l'interprétation des faits sémantiques* – C'est le principe souvent répété de détermination du global sur le local, qui périme nombres des problématiques linguistiques relevant d'unités inférieures, comme la phrase. Du côté de la textométrie, la constitution de corpus de référence, de travail et d'élection, dont il a été question précédemment en est une mise en œuvre.

*«Dans la langue, il n'y a que des différences»* – Héritée de la tradition saussurienne, le différentialisme fonde la sémantique interprétative et est sans doute un aspect remarquable de la textométrie dans le contexte général de la linguistique de corpus. Le succès jamais démenti des mesures de spécificités (tests $\chi^2$ ou d'écart réduit, modèle hypergéométrique) destinées à contraster une partie d'un corpus avec une autre de manière à en faire émerger les singularités, en atteste.

### *3.3. Synthèse*

Prenons acte (i) de l'évolution du TAL vers une problématique rhétorique/herméneutique intéressée par l'interprétation des textes et non plus seulement par l'extraction des données discrètes qu'ils recèlent; (ii) de l'inadéquation des modèles linguistiques dominants, préoccupés par des phénomènes relevant de la langue et non du texte ; (iii) de l'affinité invétérée entre la textométrie et la sémantique textuelle. À partir de cet inventaire, nous formulons le projet de jeter un pont entre la sémantique textuelle et le TAL par le truchement de la textométrie, afin de mutualiser les avantages d'une association entre celles-ci et les standards du TAL en fouille de textes. Nous illustrerons notre propos à partir d'une tâche de fouille de données subjectives.

## 3. Méthodologie de sémantique de corpus: identifier les opinions, les sentiments

### *3.1. Concepts théoriques empruntés à la sémantique textuelle*

Il s'agit d'évaluer l'hypothèse selon laquelle les discours porteurs de valeurs et de subjectivité se construisent par des interactions entre différentes *composantes sémantiques* ne relevant pas du strict vocabulaire des valeurs. Nous proposons ci-dessous une synthèse basée sur des études récentes, (Eensoo et Valette, 2012, 2014ab). Nous avons montré dans le cadre de deux tâches d'analyse des sentiments et de fouille d'opinion, par méthodes d'apprentissage, que les descripteurs classifiants les plus efficaces pouvaient être organisés selon les différentes composantes sémantiques telles qu'elles sont théorisées par (Rastier 2001). Il s'agit de :

- *la composante dialectique* (représentation du temps et du déroulement aspectuel, des rôles et des interactions entre acteurs),
- *la composante dialogique* (représentation des acteurs, modalités notamment énonciatives),
- *la composante thématique* (contenus et univers sémantiques exprimés dans les textes).

L'expression des états privés apparaîtrait comme un phénomène hétérarchique mettant en jeu plusieurs niveaux de la textualité, au delà d'un vocabulaire subjectif.

Par ailleurs, nous avons mis en évidence l'élaboration, par le biais de faisceau de critères sémantiques et textuels, d'acteurs stéréotypiques que nous avons nommés des *agonistes,* adaptant à notre objet le concept que (Rastier 2001) emprunte lui-même à la critique littéraire. Nous définirons l'agoniste comme *une classe d'acteurs stéréotypés correspondant à une position ou à la défense d'une valeur (ou d'un ensemble de valeurs)*. L'agoniste est une *construction textuelle* reposant sur une combinaison d'éléments relevant des composantes sémantiques. Avec l'agoniste, nous nous démarquons des travaux en analyse du discours qui stipulent l'existence d'éléments dévolus au discours évaluatifs tels que les unités lexicales ou des segments prédicatifs.

*3.2. Élaboration textométrique des critères de catégorisation*

Notre objectif est de trouver des critères de classification linguistiquement explicables et suffisamment robustes pour servir comme descripteurs aux méthodes d'apprentissage supervisé. Nous conformant à la caractérisation du textomètre effectué précédemment, nous faisons l'hypothèse que les critères de classification *interprétables* sont plus performants que les descripteurs trouvés par des méthodes d'apprentissage, souvent non signifiants d'un point de vue textuel et incidents au corpus d'apprentissage (ex: présence de fautes d'orthographe non pertinentes par rapport aux catégories de classification). Ainsi, lors de l'étape de sélection de critères, le textomètre écarte les critères liés à l'échantillon du corpus et choisit les critères textuels cohérents avec les composantes sémantiques (thématique, dialogique, etc.) actualisées dans le corpus.

Pour nos expérimentations, nous avons utilisé différents types de critères : (i) unités isolées : un choix de formes, lemmes ou catégories morphosyntaxiques ; (ii) collocations de taille variée (de 2 à 4 unités); (iii) cooccurrences phrastiques multiniveaux (combinant les éléments de différents niveaux de description linguistique: formes, lemmes ou catégories morphosyntaxiques). Tous les critères sont sélectionnés

selon quatre principes: leur caractère spécifique à un sous-corpus, leur répartition uniforme dans le sous-corpus, leur fréquence et leur pertinence linguistique.

L'analyse du corpus et l'extraction des critères a été effectué avec deux logiciels textométriques – Lexico 3 (Salem *et al.* 2003) et TXM (Heiden *et al.* 2010) – qui implémentent les algorithmes de spécificités (Lafon, 1980) et de cooccurrences (Lafon, 1981). Nous avons choisi les deux premiers types de critères selon la procédure suivante:

1. calcul des spécificités des items isolés (formes, lemmes et catégories morphosyntaxiques) et de leurs n-grammes (fonction «Segments Répétés» de Lexico 3) pour chaque sous-corpus;
2. analyse des contextes d'apparition des items spécifiques (au moyen de concordances textuelles) afin de s'assurer de leur pertinence textuelle et de l'unicité de leur fonction (les critères ayant une seule fonction et signification ont été privilégiés);
3. vérification de la répartition uniforme des items dans le sous-corpus (fonctionnalité «Carte de Sections» de Lexico 3);
4. sélection des items spécifiques pertinents dans le sous-corpus.

La sélection des cooccurrences s'est fait comme suit:

1. calcul des cooccurrences des items spécifiques fréquents et uniformément repartis sur la totalité du corpus;
2. analyse des contextes d'apparition de ces cooccurrences;
3. sélection des cooccurrences spécifiques à un sous-corpus.

Dans les deux cas, les critères de classification pour chaque texte sont des fréquences ou des valeurs booléennes (présence/absence) des items sélectionnés.

*3.3. Classification par apprentissage supervisé*

La deuxième étape consiste à utiliser des algorithmes d'apprentissage supervisé pour classer les textes. Nous en avons expérimenté plusieurs, chacun d'une famille différente : les arbres de décision (J48), *Naive Bayes* et les Machines à Vecteurs de Support (SMO). L'objectif est d'observer les différences et similitudes au niveau des performances en changeant la nature et la quantité des critères. Dans le présent article, nous ne mentionnerons que les résultats des deux algorithmes les plus efficaces pour les tâches choisies.

## 4. Deux études de sémantique de corpus[1]

### *4.1. Agonistes dysphoriques et euphoriques dans un corpus d'ego-documents*

Cette étude[2] s'insère dans le contexte de la veille sanitaire et intéresse la psychologie sociale. Il s'agit d'identifier les marqueurs linguistiques de la tristesse (*dysphorie*) et de la joie (*euphorie*). Nous disposons d'un corpus de 300 ego-documents de langue française (témoignages, récits d'histoires vécues) postés par les internautes sur différents forums de discussion à dominante médico-sanitaire (aufeminin.com, doctissimo.fr, etc.) et catégorisé en deux classes[3]: les textes dysphoriques et les textes euphoriques. Nous avons ainsi identifié et inventorié 70 critères sémantiques à partir de l'analyse textométrique puis nous les avons caractérisés en fonction des composantes sémantiques. Il en résulte la construction de deux agonistes.

L'*agoniste dysphorique* est construit sur la noyau sémique /inaccompli/ + /dysphorique/. D'un point de vue dialogique, l'acteur-énonciateur apparaît égocentré (surreprésentation de la 1e personne du singulier) et enclos sur son univers intime, il exprime un univers impressif et non factuel («*Je ne sais pas*[4] comment cela va évoluer»). Du point de vu dialectique, on constate une excentration de l'action (+/passivité/) : («*On me dit* que les causes de cette maladie ne sont pas encore précises»).

L'*agoniste euphorique* est élaboré sur un noyau sémique inverse : /accompli/ + /euphorique/. Du point de vue de la composante dialogique, c'est un acteur-énonciateur altruiste qui s'adresse à un tiers (surreprésentation de la 2e personne du singulier) («Alors *tu* vois il faut avoir espoir»). L'agoniste euphorique construit des univers alternatifs en faisant part de son expérience à des fins d'édification («Je tenais à faire part de *mon expérience*») et en intertextualisant son témoignage (« Je te file une adresse : *http://www*. …». Le caractère le plus remarquable des textes euphorique réside au niveau de la composante dialectique. À la différence de l'agoniste dysphorique, l'agoniste euphorique élabore un texte séquencé, descriptif ou argumentatif («*Par contre* j'étais soignée à l'homéopathie»).

---

[1] Pour un exposé complet des résultats, nous invitons le lecteur à se reporter aux publications correspondantes : (Eensoo et Valette, 2014a) pour la première étude, (Eensoo et Valette, 2014b) pour la seconde.

[2] Lire Eensoo et Valette 2014a pour un développement.

[3] La catégorisation, autrement dit l'annotation manuelle, a été réalisée par un prestataire de la société SAMESTORY qui nous a confié ce corpus à des fins de recherche.

[4] Désormais, tous les éléments en italique sont des exemples de critères de catégorisation.

*Évaluation des critères* – Au total, 70 critères ont été construits: 30 critères relevant de la composante dialectique; 16 relevant de la composante dialogique; 17 critères relevant de la composante thématique (non décrits ici) et 6 critères thymiques (idem). L'évaluation de la capacité classificatrice des critères qualifiés dans le paragraphe précédent, a été réalisée au moyen d'une classification de textes effectuée en utilisant un algorithme d'apprentissage automatique de la famille des *Machines à vecteurs de support* – SMO (Platt, 1998).

| **Types de critères** | **Exactitude** |
|---|---|
| Mots simples (10700 critères) | 68 % |
| Cr. dialogiques (16 critères) | 64 % |
| Cr. dialectiques (30 critères) | 73 % |
| Cr. dialectiques + dialogiques (45 critères) | 77 % |
| Tous les critères (70 critères) | 84 % |

Tableau 1 : résultat de la classification, agonistes dys- et euphoriques

Le tableau 1 donne à voir quelques résultats de la classification. Notre ligne de comparaison (*baseline*) est la classification sur formes simples, qui permet d'obtenir un taux d'exactitude de 68 %. En bref, on notera que c'est le cumul des 45 critères dialectiques et dialogiques qui nous permet de nous élever significativement au dessus de notre ligne de comparaison (77 %). Ce résultat est particulièrement intéressant car ce sont ces composantes qui se démarquent le plus nettement des pratiques en fouille de textes, lesquelles, en général, privilégient des descripteurs thématiques ou thymiques. Enfin, la totalité de nos 70 critères issus d'une analyse textométrique permettent d'atteindre une classification réussie à hauteur de 84 %, soit 16 points de plus que la ligne de comparaison, ce qui est un résultat très encourageant.

*4.2. Agonistes pro-Roms et anti-Roms dans un corpus de commentaires d'articles*

Cette seconde étude[5] relève du champ de l'étude des positions idéologiques dans le discours médiatiques. Il s'agit d'identifier les marqueurs linguistiques du racisme et de la xénophobie. Notre corpus est constitué de 644 commentaires d'articles de presse écrits par les lecteurs-internautes relatif à la situation des Roms en France (faits divers, initiative politique, etc.). Ils proviennent de quatre quotidiens français: *Le Monde*,

[5] On lira Eensoo et Valette (2014b), pour un développement.

*Libération*, *Le Figaro* et *Le Parisien*. Ces commentaires ont été classés en deux supercatégories, *hostile* et *non hostile*, elles-mêmes divisées en cinq catégories plus fines: *raciste*, *xénophobe* et *défavorable distancié* d'une part, et *compassionnel* et *favorable distancié* d'autre part. Nous avons inventorié 143 critères sémantiques (90 critères thématiques, 42 critères dialectiques et 11 critères dialogiques) à partir de l'analyse textométrique effectuée sur ce corpus.

L'*agoniste compassionnel* se caractérise par une élaboration égocentrée, relevant de la zone anthropique identitaire[6]. D'un point de vue dialogique, il exprime à la première personne du singulier (*Je*, *J'*) son opinion personnelle, en s'adressant à une communauté d'interlocuteurs (*vous*). Les critères thématiques s'organisent en trois thèmes: un thème proximal exprimant l'empathie (évocation des mendiants: *femme*, *enfants, misère*) et deux thèmes distaux. Le premier est relatif aux opposants, au sens actanciel, il est incarné par les *mafias*, les *réseaux*, *organis*ant la *migration* des Roms. Le second thème distal concerne les adjuvants (charité *chrétien*ne). Le texte apparaît, en revanche, peu élaboré d'un point de vue dialectique. Les marqueurs de structuration sont particulièrement rares. Cette pauvreté dialectique concorde logiquement avec l'importance des critères proximaux. On a vu précédemment que les critères dialectiques argumentatifs relevaient de constructions sémantiques distales.

L'*agoniste favorable distancié* n'investit pas la zone anthropique identitaire. Il s'efface au profit de son interlocuteur et surtout, de grandes thématiques générales. Les critères dialogiques sont rares: anaphore et adresse interlocutoire (*tu*). Les critères thématiques relèvent de deux thèmes sémantiques principaux: les *valeur*s humanistes de citoyenneté (*insertion*, *éducation*, *formation*) et de *respect*. Sont dénoncés les *propos racistes*, le *racisme* en général, la *haine*. Les noms propres sont particulièrement nombreux dans cette catégorie et sont peut-être l'indice d'un fort ancrage politique et sociétal. Les critères argumentatifs (*Mais*, *comme*, *comment*, *dont*), caractéristiques de la composante dialectique sont ici statistiquement significatifs. C'est l'indice d'un ancrage dans la zone anthropique distale (construction intellectuelle, abstraction, mise à distance).

L'*agoniste raciste*, comme l'agoniste compassionnel, investit la zone anthropique identitaire. On l'observe par l'usage important qu'il fait des pronoms personnels en particulier de première personne. La composante dialectique repose sur

[6] Nous empruntons à Rastier (2001) le concept de zone anthropique.

une rhétorique de l'emphase (*dire que*). Du point de vue thématique, le thème dominant est celui de la spoliation générale: pour cet agoniste, les Roms *viennent* en France *profiter* de l'*argent* des *Français*, Un thème symétrique au thème compassionnel proximal évoqué précédemment est particulièrement intéressant parce qu'il actualise les mêmes traits: l'agoniste raciste se scandalise que l'État français laisse des *Français* vivre *dans la rue*.

L'*agoniste xénophobe* a ceci de particulier qu'il n'est pas caractérisable en termes de composante dialectique ni de composante dialogique. En contrepartie, les thèmes statistiquement caractéristiques qu'il actualise sont lexicalement riches et sémantiquement très homogènes. Il s'agit des thèmes du renvoi dans le *pays d'origine* (*solution*, *renvoyer*, *expulser*, *retour*, *dans leur pays*), de la politique européenne (*libre*, *circuler*, *Europe*, *frontière*) et de l'installation en France (*s'installer*, *insérer, ressources*).

L'*agoniste défavorable distancié* partage plusieurs traits communs avec l'agoniste raciste, notamment en termes de composition thématique, mais son expression diffère. Ce qui le distingue des autres agonistes hostiles, c'est sa rhétorique de l'indignation. Elle s'exprime, d'un point de vue dialectique, par des marqueurs narratifs (*depuis des années*, *puis*), de locution disjonctive (*alors que*), mais aussi par des ellipses (points de suspension) et des marqueurs d'emphase (point d'exclamation). Du point de vue dialogique, l'agoniste défavorable distancié adopte avec une grande régularité phraséologique la posture modale de l'indigné (*je ne comprends pas*). Comme nous l'avons dit, les thématiques qu'il aborde recoupent en partie celles de l'agoniste raciste, mais il a recourt à un vocabulaire différent. Le thème principal demeure donc celui de la spoliation des Français par les Roms, avec la complicité de l'État (*logement*, *charge*, *payer*, *impôt*).

| **Types de critères** | **Exactitude** |
|---|---|
| Mots simples (6075 critères) | 40 % |
| Cr. dialogiques (11 critères) | 38 % |
| Cr. dialectiques (42 critères) | 43 % |
| Cr. thématiques (90 critères) | 47 % |
| Tous les critères (143 critères) | 51 % |

Tableau 2 : résultat de la classification, agonistes pro- et anti-Roms

*Évaluation des critères*[7] – Les 143 critères ont été évalués au moyen de l'algorithme Naïve Bayes Multinomial (A. Mccallum& K. Nigam, 1998). Notre ligne de comparaison demeure la classification sur formes simples, qui permet d'obtenir un taux d'exactitude de 40 %. Ce taux apparemment bas s'explique par la présence de 5 classes: le taux d'exactitude correspondant au hasard est de 20 %. Le gain concernant l'ensemble des critères est de 11 points. Les 11 critères dialogiques seuls sont un peu en deçà de la ligne de comparaison, mais font mieux que les adjectifs; les 42 critères dialectiques sont plus performant que la ligne de comparaison et même que les critères lemmatiques. Enfin, les critères thématiques, les plus nombreux, permettent d'obtenir un score honorable de 47 %, qui n'égale pas toutefois le résultat obtenu avec la classification binaire. Sans doute les frontières entre les différentes catégories sont-elles trop précises pour un certain nombre de critères les subsumant.

## CONCLUSION

Dans cet article, nous avons tenté de coupler la sémantique textuelle, la textométrie et des méthodes d'apprentissage automatique issues du TAL pour mettre en place une méthodologie générale applicable aux humanités numériques. Il s'est agi en premier lieu de valider certaines des propositions de la sémantique textuelle par le biais du TAL. En cela, notre méthodologie permet d'identifier des segments textuels (et des structures de traits) pertinents et non triviaux pour une tâche de fouille de données subjectives et de les analyser suivant une grille de lecture linguistique. L'analyse résultante permet de comprendre les interactions entre les différentes composantes sémantiques dans la production et l'interprétation de textes d'opinion ou exprimant un sentiment. Par ailleurs, les résultats obtenus montrent que ces segments textuels singuliers donnent de meilleurs résultats que les techniques standard du TAL ; c'est donc une piste possible pour l'amélioration des méthodes de fouille.

*J'ai grand plaisir à remercier Egle Eensoo avec laquelle les recherches exposées ici ont été menées.*

## BIBLIOGRAPHIE

Eensoo Egle et Valette Mathieu, 2014a, « Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments », dans D. Ablali, S. Badir, D. Ducard (éds),

[7] Nous avons ici simplifié la présentation des résultats de l'étude. Pour une analyse plus fouillée, le lecteur voudra bien se reporter à (Eensoo et Valette, 2014b).

*Documents, textes, œuvres. Perspectives sémiotiques.* Presses Universitaires de Rouen, Collection Rivages linguistiques, p. 75-90.

Eensoo Egle et Valette Mathieu, 2014b, « Approche textuelle pour le traitement automatique du discours évaluatif », dans A. Jackiewicz, (éd), *Études sur l'évaluation axiologique*, *Langue française*.

Heiden Serge, Magué Jean-Pierre et Pincemin Bénédicte, 2010, « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement » dans I. C. Sergio Bolasco, editor, *JADT 2010*, vol.(2), p. 1021-1032.

Lafon Pierre, 1980, « Sur la variabilité de la fréquence des formes dans un corpus », *Mots, 1*, 127-165.

Lafon, Pierre, 1981, « Analyse lexicométrique et recherche des cooccurrences », *Mots, 3*, p. 95-148.

McCallum Andrew et Nigam Kamal, 1998, "A Comparison of Event Models for Naive Bayes, Text Classification", dans *AAAI-98 Workshop on 'Learning for Text Categorization'*, p. 41-48

Mayaffre Damon, 2008, « De l'occurrence à l'isotopie. Les cooccurrences en lexicométrie », dans M. Valette (éd), *Textes, documents numériques, corpus. Pour une science des textes instrumentée*, *Syntaxe et sémantique*, 9, p. 53-72.

Pédauque Roger T. (collectif), 2007, *La redocumentarisation du Monde*, Paris, Éditions Cépadues, 213 p.

Pincemin Bénédicte, 2010, "Semántica interpretativa y textometría", dans C. Duteil-Mougel et V. Cárdenas (éds), *Semántica e interpretación, Tópicos del Seminario*, 23, Enero-junio 2010, p. 15-55.

Platt John, 1998, "Machines using Sequential Minimal Optimization", dans B. Schoelkopf, C. Burges et A. Smola (éds), *Advances in Kernel Methods - Support Vector Learning*, Cambridge, MIT Press.

Rastier François (éd.), 1995, *L'analyse thématique des données textuelles : l'exemple des sentiments*, Paris, Didier, collection Études de sémantique lexicale, 270 p.

Rastier François, 2001, *Arts et sciences des textes*, Paris, PUF, 303 p.

Rastier François, 2011, *La mesure et le grain. Sémantique de corpus*, Paris, Honoré Champion, 272 p.

Rastier François et Pincemin Bénédicte, 1999, « Des genres à l'intertexte », I. Kanellos (éd.), *Cahiers de Praxématique*, 33, *Sémantique de l'intertexte*, p. 83-111.

Salem André, Lamalle Cédric, Martinez William., Fleury Serge, Fracchiolla Béatrive, Kuncova André et Maisondieu Aude, 2003, *Lexico3 – Outils de statistique textuelle, Manuel d'utilisation*, Université de la Sorbonne nouvelle – Paris 3. URL: [http://www.tal.univ-paris3.fr/lexico/]

Zweigenbaum Pierre, Bellot Patrice, Grau Brigitte, Ligozat Anne-Laure, Robba Isabelle, Rosset Sophie, Tannier Xavier et Vilnat Anne, 2008, « Apports de la linguistique dans les systèmes de recherche d'informations précises », *Revue française de linguistique appliquée* 1/ 2008 (Vol. XIII), p. 41-62.