Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité[1]

Egle EENSOO[1] Mathieu VALETTE[1]
(1) ERTIM, INALCO, 2 rue de Lille, 75343 PARIS cedex 07
egle.eensoo@inalco.fr, mathieu.valette@inalco.fr

Résumé. Cet article entend dresser, dans un premier temps, un panorama critique des relations entre TAL et linguistique. Puis, il esquisse une discussion sur l'apport possible d'une sémantique de corpus dans un contexte applicatif en s'appuyant sur plusieurs expériences en fouille de textes subjectifs (analyse de sentiments et fouille d'opinions). Ces expériences se démarquent des approches traditionnelles fondées sur la recherche de marqueurs axiologiques explicites par l'utilisation de critères relevant des représentations des acteurs (composante dialogique) et des structures argumentatives et narratives des textes (composante dialectique). Nous souhaitons de cette façon mettre en lumière le bénéfice d'un dialogue méthodologique entre une théorie (la sémantique textuelle), des méthodes de linguistique de corpus orientées vers l'analyse du sens (la textométrie) et les usages actuels du TAL en termes d'algorithmiques (apprentissage automatique) mais aussi de méthodologie d'évaluation des résultats.

Abstract.
A method of corpus semantics applied to opinion mining and sentiment analysis: the impact of dialogical and dialectical features on the expression of subjectivity.
This paper first aims to provide a critical overview of the relationship between NLP and linguistics, and then to sketch out a discussion on the possible contribution of corpus semantics in an application-based context based on several subjective text mining studies (sentiment analysis and opinion mining). These studies break away from traditional approaches founded on the detection of axiological markers. Instead, they use explicit criteria related to the representation of actors (dialogical component) and argumentative or narrative structures (dialectical component). We hope to highlight the benefit of a methodological dialogue between theory (text semantics), meaning-oriented methods of corpus linguistics (i.e. textometrics) and NLP current practices in terms of algorithmic (machine learning) and assessment methodology.

Mots-clés : textométrie, sémantique de corpus, fouille d'opinion, analyse des sentiments
Keywords: textometry, corpus semantics, opinion mining, sentiment analysis

il est déjà publié dans les actes de TALN 2015 :

Eensoo, Egle, Mathieu Valette (2015) « Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité », Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015), Caen (France)

http://www.atala.org/taln_archives/TALN/TALN-2015/

http://www.atala.org/taln_archives/TALN/TALN-2015/taln-2015-long-010.pdf

[1] Cet article a été publié en premier dans :
Eensoo, Egle, Mathieu Valette (2015) « Une méthodologie de sémantique de corpus appliquée à des tâches de fouille d'opinion et d'analyse des sentiments : étude sur l'impact de marqueurs dialogiques et dialectiques dans l'expression de la subjectivité », Actes de la 22e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2015), Caen (France).
http://www.atala.org/taln_archives/TALN/TALN-2015/
http://www.atala.org/taln_archives/TALN/TALN-2015/taln-2015-long-010.pdf

1 Introduction

Avec l'essor dans le TAL des méthodes par apprentissage automatique et la relative désaffection pour les méthodes symboliques à base de règles linguistiques formelles dans le monde académique[2], les linguistes sont aujourd'hui contraints de repenser leur rôle dans un contexte où dominent les méthodes mathématiques. Si l'annotation requise pour la constitution des données d'apprentissage nécessite un savoir-faire et une connaissance experte parfois adossée à des présupposés théoriques, les spécialistes de la fouille de textes, par exemple, montrent peu d'intérêt pour les théories linguistiques, vraisemblablement à raison, tant se creuse le fossé entre les préoccupations minutieuses mais *ad hoc* de certains linguistes et celles des talistes, guidées par un principe de réalité : la masse de données textuelles accessibles.

Cet article propose un panorama critique des relations entre TAL et linguistique et esquisse, au moyen d'exemples commentés issus d'applications en fouille de textes (analyse de sentiments et fouille d'opinions), une discussion sur l'apport possible d'une réflexion linguistique dans ce contexte applicatif. Nous souhaitons en particulier mettre en lumière le bénéfice potentiel d'un dialogue méthodologique entre des méthodes de linguistique de corpus orientées vers l'analyse du sens (la textométrie), l'exploitation de concepts de la sémantique textuelle (Rastier, 2001, 2011) et les usages actuels du TAL en termes d'algorithmiques mais aussi de pratiques évaluatives.

L'article est construit en quatre parties. Le paragraphe 2 offre une lecture optimiste des relations qu'entretiennent le TAL et la linguistique et de leur réunion possible autour de l'objet *texte*. Le paragraphe 3 procède à l'examen en miroir des outils et méthodes nécessaires à l'établissement d'une sémantique instrumentée, en mettant notamment en vis-à-vis la textométrie et le TAL. Le paragraphe 4 présente les concepts linguistiques et la méthodologie adoptés par les auteurs pour une tâche de fouille de textes subjectifs. Enfin, le paragraphe 5 présente, à des fins illustratives, trois expérimentations adossées à la méthodologie décrite dans le paragraphe précédent.

2 Le statut contemporain du texte dans le TAL

Longtemps unis par des objets formels similaires sinon communs (la proposition, la phrase) et un même positionnement référentialiste, la linguistique et le TAL ont vu leurs rapports se distendre depuis une quinzaine d'années. Les modèles théoriques de la linguistique formelle se sont en effet avérés peu adaptés 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. Les tâches correspondantes nécessitant une automatisation (résolution d'anaphore, désambiguïsation lexicale, 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 actualisé le statut d'objet scientifique du *texte* – statut que la linguistique ne lui accorde encore que marginalement et au sein de certains courants seulement (analyse du discours, linguistique textuelle). 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, notamment en termes de rappel, bien qu'elles proposent encore des solutions pertinentes pour l'extraction d'information *précise*, liée aux applications telles que l'interface homme/machine (système de question-réponse, ou réponse à des questions formulées en langue dite naturelle) (Zweigenbaum *et al.*, 2008). Par ailleurs, les méthodes symboliques sont plébiscitées dans l'industrie où beaucoup d'applications nécessitent un haut taux de précision sans que le rappel soit déterminant. Enfin, la tendance actuelle est à l'hybridation dans le monde académique comme dans l'industrie. Le couplage de données produites à partir de méthodes à base de règles et de technique apprentistes permet d'améliorer les performances de systèmes de manière significative (Villena-Román *et al.,* 2011).

L'essor, dans le courant des années 2000, des applications en fouille de textes subjectifs (fouille d'opinion, analyse des sentiments, détection des émotions, etc.) implique également une évolution des tâches : alors que le TAL privilégiait les unités *référentielles* et souvent lexicales (entités nommées, concepts, termes, thèmes), il est aujourd'hui confronté à des *valeurs*. Certes, les méthodes d'extraction et de classification n'ont guère évolué : dans beaucoup d'applications, les adjectifs sont aux textes subjectifs ce que les substantifs sont aux concepts (Strapparava & Valitutti, 2004) et on a tendance à appliquer aux premières les méthodes qui ont fait leur preuve sur les secondes. Dépasser le « lexicalisme » du TAL est un des enjeux de la linguistique car l'inventaire des objets de la linguistique susceptibles d'être appréhendés par le TAL est, en effet, loin d'être clos. Il est par exemple probable 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[3].

[2] (Tanguy, 2012) relate plusieurs études (Church, 2011, Hall *et al.*, 2008) où a été observé que la proportion d'articles de l'*Association for Computational Linguistics* intégrant une section statistique a progressé de 30 à 90 % du début des années 90 à la fin des années 2000.
[3] On pourra utilement consulter (Micheli *et al.*, 2013) à ce propos.

En somme, tout se passe comme si les questions qui se posent au TAL évoluaient d'une problématique logico-formelle dominée par le primat référentiel et le choix historique de la phrase (et son avatar : l'énoncé) comme unité d'analyse, vers une problématique herméneutique et interprétative dont l'objet est la réception et l'interprétation des textes considérés comme des unités de sens complexes déterminées par un projet de communication. La proposition a notamment été formulée par (Rastier, 2001) et oppose, *in fine*, deux paradigmes, la linguistique des langues et la linguistique des textes. Ce moment de flottement paradigmatique est l'occasion d'esquisser des méthodes fondées non pas sur les présupposés théoriques du paradigme logico-grammatical mais sur un paradigme herméneutique et interprétatif peu exploré encore en TAL.

Dans le paragraphe suivant, nous procédons donc à l'examen contrastif des présuppposés épistémologiques et méthodologiques du TAL (et plus particulièrement de la fouille de textes) d'une part, et d'une sémantique de corpus dédiées à l'interprétation des textes d'autre part.

3 La textométrie, ou l'interprétation assistée par ordinateur[4]

L'analyse statistique des données textuelles (ADT), ou *textométrie*, est un ensemble particulier de pratiques relevant du champ général de la linguistique de corpus. Elle comprend des traitements statistiques (analyse factorielle des correspondances, spécificités fondées sur le modèle hypergéométrique, etc.) et des outils de visualisation des corpus (nuages de mots, histogrammes, etc.) et documentaires (concordanciers) destinés à l'aide à l'interprétation des textes[5].

3.1 La textométrie et le TAL

La linguistique de corpus et la textométrie ne relèvent pas du TAL. En dépit de quelques traits communs (les corpus numériques, les algorithmes mathématiques informatisés) et d'affinités intercommunautaires ponctuelles, elles se distinguent à tous les égards. On dresse ci-après l'inventaire de ces différences.

Du point de vue épistémologique – Le TAL, fondamentalement, vise l'automatisation des processus, l'élimination de la part de l'humain dans les traitements, tandis que la textométrie repose sur une itération entre l'analyse des sorties logicielles et la consultation des textes ou de fragments ; en cela, il s'agit davantage d'une linguistique assistée par ordinateur. Par ailleurs, le TAL est utilitariste et a pour finalité des applications informatiques, ce qui implique une recherche de performance et d'optimisation ; la textométrie a des objectifs épistémiques : accroître les connaissances et participer à l'interprétation d'un corpus. Enfin, à la différence du TAL où la mise en place d'un protocole d'évaluation est indispensable, l'évaluation et la reproductibilité ne sont pas problématisées par la textométrie. Les études textométriques sont validées par homologation, c'est-à-dire par l'assentiment d'une communauté qui, dans le meilleur des cas, est distante (par exemple, communauté de la critique littéraire pour l'analyse textométrique de textes littéraires), mais parfois n'est peut-être qu'un avatar du « jugement d'acceptabilité » contre lequel s'est pourtant dressée la linguistique de corpus.

Du point de vue des applications – Comme on l'a vu, les applications ne sont que marginalement un enjeu en textométrie, même si certains travaux sont susceptibles d'applications (constitution de ressources par exemple) alors que la demande socio-économique détermine dans une large mesure les tâches auxquelles le TAL s'attelle. Cette demande implique le renouvellement des problématiques de recherche : il y a 15 ans, l'extraction d'informations lexicales ou syntagmatiques destinées à alimenter des bases de connaissances (mémoires de traduction, terminologies de métier, système de question-réponse, etc.) structurait le champ. Puis, avec l'essor des réseaux sociaux sur le web, des applications en fouille d'opinion, analyse des sentiments, analyse du buzz, etc. se sont développées. La traduction automatique, historiquement liée au TAL symbolique, connaît également un regain d'intérêt motivé par l'efficacité des méthodes statistiques.

Du point de vue des documents – Les pratiques de la textométrie et celles du TAL opposent les notions de *sources* et de *ressources* : les documents analysés en textométrie sont variées et souvent caractérisées avec précision. À la fin des années 1990, les œuvres littéraires dominaient (romans, poésie, théâtre) mais on étudiait aussi des enquêtes ouvertes, des textes politiques, syndicaux, etc. Au milieu des années 2000, les nouveaux genres de l'Internet font leur apparition (mails, puis forums de discussion, tweets). On retrouve en partie ces types documentaires en TAL (très rarement les textes littéraires), mais les textes à vocation technique ou encyclopédique (telles que Wikipédia) apparaissent privilégiés. Surtout, davantage que des

[4] Les observations faites dans ce paragraphe s'appuient en partie sur une l'analyse contrastive des actes de deux conférences communautaires francophones emblématiques : les Conférences en Traitement Automatique de la Langue Naturelle (TALN) et les Journées internationales d'Analyses statistiques des Données Textuelles (JADT). L'étude, menée sur un échantillon de 8 volumes d'actes de TALN et 8 volumes d'actes de JADT (de 1999 à 2014) donnera lieu à une publication ultérieure.

[5] Les actes des Journées Internationales d'Analyses statistiques des données textuelles (JADT) donneront au lecteur un aperçu des pratiques textométriques : http://lexicometrica.univ-paris3.fr/jadt/.

sources précises (i.e. des œuvres ou des éditeurs électroniques, des sites web, etc.), ce sont des ressources générales qui sont désignées : Internet, Web, Google, Google Books, Facebook. Les corpus, en TAL sont avant tout des réservoirs d'objets linguistiques infratextuels (termes, structures prédicatives, etc.). L'établissement philologique du corpus en TAL est souvent réduit à quelques valeurs quantitatives (nombre d'occurrences de mots, nombre de textes) quand les textomètres présentent leur corpus de manière plus qualitative (description des auteurs, des genres textuels, etc.).

Du point de vue des méthodes – c'est probablement au niveau des méthodes d'analyse que la différence entre la textométrie et le TAL est la plus visible. A la différence des talistes, les textomètres ne sont pas des informaticiens mais, en règle générale, des utilisateurs finaux de logiciels dotés d'interface graphique (Hyperbase, Lexico 3, TXM, Iramuteq, TextObserver, etc.) lesquels s'adossent de plus en plus souvent aux outils que les talistes développent ou utilisent pour leurs propres tâches : bibliothèques de traitements linguistiques (par exemple NLTK, Stanford NLP), langages de programmation (par exemple, Perl ou Python pour la manipulation de textes ; R ou Matlab pour le calcul), etc. En bref, les textomètres sont dépendants d'outils qu'ils conçoivent et parfois qu'ils implémentent. On a là une différence de culture remarquable : l'essentiel des efforts en matière de création d'outils en textométrie se porte actuellement sur l'ergonomie logicielle et la visualisation des données. Les méthodes mathématiques employées, qui satisfont le plus grand nombre, évoluent peu depuis 30 ans mais les heuristiques et les savoir-faire analytiques sont déterminants. Souvent les talistes s'étonnent du peu de variété des méthodes statistiques des textomètres, et leur opposent d'impressionnantes bibliothèques de traitements. C'est qu'ils ne prennent pas la mesure des tâches herméneutiques qui font la spécificité de l'ADT. L'interprétation des résultats d'analyse, en TAL, est non cruciale et souvent occultée au profit de deux types de commentaires : commentaires sur les performances de la méthode utilisée d'une part, commentaires sur les résultats d'évaluation qui suivent des méthodes normalisées. L'évaluation des performances du système repose en effet sur les mesures de congruence entre le résultat de la classification et le corpus de test annoté (taux d'exactitude, précision, rappel, f-score, etc.). Or, comme l'observe (Yvon 2006, 41), d'autres évaluations sont possibles (analyse sémantique des valeurs discriminantes sélectionnées par l'algorithme, adéquation avec une théorie linguistique, plausibilité cognitive, etc.) mais les alternatives sont rares et peu valorisées en termes académiques. Mieux encore, les données langagières proprement dites sont jugées encombrantes et, pour des raisons éditoriales sans doute, mais peut-être par manque d'outils intellectuels pour les appréhender, on ne les montre guère (Hall *et al.*, 2008).

Du point de vue de ce qu'est un corpus – L'inclination apprentiste qu'a suivi le TAL ces dernières années a profondément accentué les différences liées à l'utilisation et la fonction du corpus. Les méthodes d'apprentissage automatique dit « supervisé », lesquelles sont encore privilégiées en TAL, consistent à créer un modèle reproduisant la configuration optimale des données du corpus, quelles qu'elles soient. Si, dans une tâche de classification de textes par exemple, un corpus est composé de deux classes, l'entraînement du modèle consistera à sélectionner les critères (par exemple les mots-formes) qui caractérisent de façon appropriée les textes d'une classe par rapport à l'autre, quand bien même ces critères ne seraient nullement interprétables d'un point de vue linguistique.

Le corpus en textométrie est conçu comme un mode de contextualisation à échelle multiple des phénomènes observables, de la cooccurrence, « forme minimale du contexte » (Mayaffre, 2008) au corpus intégral qui objective l'intertexte (Rastier, 1998) et qui, à mesure qu'il s'élargit, tend vers le contexte extralinguistique qu'il simule. Ainsi, les sous-corpus construits ont toujours une fonction différentielle. On distinguera principalement le *corpus de référence* « constituant le contexte global de l'analyse, ayant le statut de référentiel représentatif, et par rapport auquel se calcule la valeur de paramètres (pondérations...) et se construit l'interprétation des résultats » et le *corpus de travail*, « ensemble des textes pour lesquels on veut obtenir une caractérisation » (Rastier & Pincemin, 1999, 84). Cette approche du corpus, est indubitablement plus sophistiquée en termes d'analyse et d'interprétation des données, mais elle écarte toute instance de validation. En bref, les concepts de corpus en TAL et en textométrie sont fondamentalement distincts.

La textométrie, comme la linguistique de corpus, demeure un ensemble de techniques et d'heuristiques qui nécessite un guidage théorique pour assurer sa pleine mesure. L'analyse de discours (Charaudeau, 1992) en est un exemple. Nous porterons notre attention, pour notre part, sur la sémantique textuelle de (Rastier, 2001, 2011) dont les rapports avec le TAL sont déjà anciens et ont donné lieu, en particulier à la fin des années 90 et au début des années 2000, à plusieurs instanciations[6].

3.2 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. En voici les principaux éléments susceptibles d'alimenter notre discussion.

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 ponctuation) et ne sont pas substituables par des

[6] Par exemple (Beust, 1998), (Thlivitis, 1998), (Perlerin, 2004), (Rossignol, 2005).

constructions artefactuelles, 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, 2000) 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.

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 relativise, sans les exclure, les unités linguistiques inférieures comme la phrase. Du côté de la textométrie, la constitution d'un corpus de référence et d'un corpus de travail en est une mise en œuvre.

« Dans la langue, il n'y a que des différence » – Héritée de la tradition saussurienne (Saussure, 2002), le différentialisme fonde la sémantique textuelle 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 χ^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

Nous prenons acte (i) de l'hypothétique évolution du TAL vers une problématique 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) des hiatus épistémologiques et de la complémentarité méthodologiques observés entre le TAL et la textométrie ; (iv) des affinités entre celle-ci et la sémantique textuelle. Nous formulons le projet général 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, c'est-à-dire l'évaluation à partir de méthode par apprentissage supervisé. Nous illustrerons notre propos à partir d'une tâche de fouille de textes subjectifs.

4 Sémantique de corpus pour la fouille de textes subjectifs

4.1 Principales méthodes du champ applicatif

Nous distinguerons quatre types d'approche en fouille de textes subjectifs : (i) les approches apprentistes, qui ne sont pas spécifiques à la fouille d'opinion ou l'analyse des sentiments mais sont utilisées dans différentes tâches (recherche d'information, traduction automatique, étiquetage morphosyntaxique, classification thématique, etc.). Appliquées à la fouille d'opinion, elles ont tendance à privilégier l'accumulation massive de descripteurs et ne nécessitent pas une connaissance linguistique approfondie des textes (par exemple, Pang *et al.*, 2002 ; Lin & Hauptmann, 2006) ; (ii) les approches cognitivistes, qui font appel à des ressources lexicales supposant l'existence de catégories cognitives préétablies et indépendantes des langues, par exemple des ressources dérivées de Wordnet (Ghorbel & Jacot, 2011 ; Lavalley *et al.*, 2010 ; Kim *et al.* 2010 ; Liu *et al.*, 2003) ou des ressources basées sur la théorie *Appraisal* (Whitelaw *et al.*, 2005, Bloom & Argamon, 2010) ; (iii) les approches opportunistes, qui exploitent des phénomènes linguistiques de surface détectables automatiquement comme des patrons morphosyntaxiques (Turney, 2002 ; Yi *et al.*, 2003), des parties du discours (Hatzivassiloglou & Wiebe, 2000), etc. ; (iv) les approches linguistiques théoriques qui revendiquent un cadre linguistique à des fins heuristiques. (Vernier *et al.*, 2009a, 2009b), par exemple, s'inspirent des catégories évaluatives de (Charaudeau, 1992). Pour un état de l'art plus détaillé, on pourra lire (Eensoo & Valette, 2014b).

C'est dans un cadre méthodologique relevant de cette quatrième approche que nous situons la méta-étude présentée ici.

4.2 Concepts et méthodologie de sémantiques de corpus

Concepts – Nous formulons en effet l'hypothèse que les discours axiologiques se construisent par des interactions entre différentes composantes sémantiques qui ne relèvent pas du strict vocabulaire des valeurs. Nous proposons ci-dessous une synthèse basée sur trois expériences montrant, dans différentes tâches d'analyse des sentiments et de fouille d'opinion, par méthodes d'apprentissage, que les descripteurs classifiants les plus efficaces peuvent ne relever que de deux classes de

valeurs sémantiques appelées *composantes* sémantiques par (Rastier 2001)[7] : la composante dialogique et la composante dialectique.

- La *composante dialogique* concerne la représentation des acteurs, le positionnement énonciatif et la distribution des rôles actanciels. Elle actualise essentiellement les pronoms personnels, les pronoms possessifs et certaines entités nommées.

- La *composante dialectique* est une catégorie sémantique dédiée à la représentation du temps et du déroulement aspectuel, des structures argumentatives et de certaines modalités. Le vocabulaire la caractérisant est plus varié. Il peut s'agir de marqueurs de structuration (adverbes tels que *enfin*, *donc*, *cependant*), des verbes modaux (*falloir*, *devoir*, etc.), et des indicateurs rhétoriques (emphases, points d'interrogation, mots interrogatifs, etc.).

Cette grille interprétative a permis à (Eensoo & Valette, 2012, 2014a, 2014b) de mettre en évidence que l'expression subjective pouvait être caractérisée avec un nombre restreint de marqueurs relevant des différentes composantes sémantiques sans nécessairement recourir à un vocabulaire subjectif. Ils élaborent le concept d'*agoniste* comme « une classe d'acteurs stéréotypés correspondant à une position ou à la défense d'une valeur (ou d'un ensemble de valeurs) » (Eensoo & Valette, 2014b, 116). L'agoniste est une construction textuelle (et non psychologique ou cognitive) reposant sur une combinaison d'éléments linguistiques relevant des composantes sémantiques.

Méthodologie générale – La méthodologie de sémantique de corpus adoptée repose sur deux étapes : l'analyse textométrique de corpus préalablement annotés d'une part, la validation par apprentissage supervisé des critères textométriques obtenus et qualifiés sémantiquement au moyen de la grille interprétative adoptée, d'autre part. L'analyse textométrique effectuée en amont de toute classification automatique permet d'identifier des critères de classification linguistiquement explicables et suffisamment robustes pour servir comme descripteurs aux méthodes d'apprentissage supervisé. L'hypothèse est que les critères de classification *interprétables* sont plus robustes que les descripteurs trouvés par des méthodes d'apprentissage, souvent non signifiants d'un point de vue textuel et incidents au corpus d'apprentissage. 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 (ici, la composante dialogique et la composante dialectique) actualisées dans le corpus.

À des fins expérimentales, nous avons écarté de cette étude les critères textuels relevant de la composante *thématique* (*i.e.* les thèmes et les domaines actualisés) de façon à évaluer l'autonomie des positions agonistiques par rapports aux thématiques des trois expériences exposées ci-après : médicaux et sanitaire en 5.1, idéologiques en 5.2 et politiques et législatifs en 5.3.

4.3 Élaboration textométrique des critères sémantiques de catégorisation

Pour les expérimentations présentées dans le paragraphe 5, ont été utilisés plusieurs types de critères : (i) unités isolées : un choix de formes, lemmes ou catégories morphosyntaxiques, (ii) collocations (n-grammes) de taille variée (de 2 à 4 unités) et (iii) cooccurrences phrastiques multiniveaux (combinant les éléments de différents niveaux de description linguistique : formes, lemmes ou catégories morphosyntaxiques). Tous les critères ont été 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'identification des critères ont été effectuées 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). Les deux premiers types de critères sont choisis 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 » du Lexico 3) ;

La sélection des cooccurrences est faite comme suit :

[7] (Rastier, 2001) inventorie quatre composantes sémantiques : dialogique, dialectique, thématique, tactique. La composante thématique est abordée par (Eensoo & Valette, 2012, 2014a, 2014b) mais nous n'en ferons pas état ici.

1. calcul des cooccurrences (fonction « Cooccurrences » de TXM) 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 absolues car d'une part, il a été démontré que les fréquences relatives sont moins performantes que les valeurs booléennes (Pang & Vaithyanathan, 2002), d'autre part, nous avons constaté que les fréquences absolues sont plus performantes que les fréquences relatives.

4.4 Classification par apprentissage supervisé à partir des critères sémantiques élaborés

La deuxième étape consiste à utiliser des algorithmes d'apprentissage supervisé pour classer les textes. En utilisant la plateforme WEKA (Hall *et al.* 2009)[8], plusieurs algorithmes, de familles différentes, ont été testés : les arbres de décision (J48), *Naive Bayes*, *Naive bayes multinomial* 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, ne sont mentionnerons que les résultats des algorithmes les plus efficaces pour les tâches choisies. A l'exception de la troisième expérience (5.3) pour laquelle nous disposons d'un corpus de test (Grouin *et al.*, 2007), les évaluations sont opérées suivant la méthode de la validation croisée sur 10 sections.

5 Trois expériences de sémantique de corpus

5.1 Agonistes dysphoriques et euphoriques dans les forums de discussions médicales et sanitaires[9]

(Eensoo & Valette, 2012, 2014a) disposent d'un corpus de 300 ego-documents (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 : les textes dysphoriques et les textes euphoriques. La référence de la catégorisation est établie par l'agrégateur des documents Samestory[10]. Nous ne disposons pas de guide d'annotation mais en analysant un échantillon du corpus nous avons pu déduire la stratégie d'annotation. Un témoignage « dysphorique » est (i) une histoire qui fini mal, (ii) un témoignage exprimant des doutes, des interrogations, ou sollicitant de l'aide. Un témoignage « euphorique » est (i) une histoire triste qui finit bien, (ii) un témoignage modulant la gravité d'une situation en en soulignant les points positifs, (iii) un conseil. Pour les besoins de l'application d'analyse de sentiments, ils i identifient et inventorient 70 critères sémantiques à partir de l'analyse textométrique puis les caractérisent en fonction des composantes sémantiques.

Il en résulte la construction de deux agonistes. D'un point de vue dialogique, l'agoniste dysphorique 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*[11] comment cela va évoluer »). Du point de vu dialectique, on constate une excentration de l'action : (« *On me dit* que les causes de cette maladie ne sont pas encore précises »). L'*agoniste euphorique* est élaboré sur un noyau sémique inverse. 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 euphoriques réside au niveau de la composante dialectique. À la différence de l'agoniste dysphorique, l'agoniste euphorique élabore un texte séquencé, descriptif ou argumentatif (« J'avais déjà quelques éruptions qui ont débuté *après avoir pris* la décision de déménager », « *Par contre* j'étais soignée à l'homéopathie »).

Parmi les critères construits, 30 relèvent de la composante dialectique et 16 de la composante dialogique. L'évaluation de la capacité classificatrice des critères qualifiés 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).

[8] http://www.cs.waikato.ac.nz/ml/weka/

[9] Pour un exposé complet des résultats, nous invitons le lecteur à se reporter à l'étude originale correspondante.

[10] http://www.same-story.com/

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

Types de critères	Exactitude *validation croisée*
Mots simples (10 700 critères) *ligne de comparaison*	68 %
Critères dialogiques (16 critères)	64 %
Critères dialectiques (30 critères)	73 %
Critères dialectiques et dialogiques (46 critères)	77 %

TABLE 1 : Résultat de la classification : agonistes dysphoriques et euphoriques

Le tableau 1 donne à voir quelques résultats de la classification. La ligne de comparaison (*baseline*) est la classification sur formes simples (sans changement de casse ni de lemmatisation), qui permet d'obtenir un taux d'exactitude, c'est-à-dire le pourcentage de textes bien classés, de 68 %. Le cumul des critères dialectiques et dialogiques permet de s'élever de 9 points au dessus de la ligne de comparaison (77 %). Ce résultat est 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.

5.2 Agonistes hostiles et non hostiles aux Roms dans un corpus de commentaires d'articles

(Eensoo & Valette, 2014b) étudient un corpus constitué de 644 commentaires d'articles de presse de 2013 ayant pour objet la communauté Rom en France. Les commentaires sont écrits par les lecteurs-internautes. Ils proviennent de quatre quotidiens : *Le Monde*, *Libération*, *Le Figaro* et *Le Parisien*. Ces commentaires ont été classés en deux supercatégories composées de 445 commentaires hostiles pour la première supercatégories et 199 commentaires non hostiles pour la seconde. Les supercatégories ont elles-mêmes été divisées en cinq catégories plus fines que nous n'aborderons pas dans cet article[12]. La catégorisation manuelle des documents a été effectuée par les auteurs de l'étude selon une lecture et interprétation macroscopique des textes excluant l'identification des éléments lexicaux discrets qui pourraient se confondre avec les critères de classification automatique. Par la suite, les auteurs inventorient 42 critères dialectiques et 11 critères dialogiques à partir de l'analyse textométrique effectuée sur ce corpus.

Types de critères	Exactitude *validation croisée*
Mots simples (6075 critères) *ligne de comparaison*	70 %
Critères dialogiques (11 critères)	69 %
Critères dialectiques (42 critères)	71 %
Critères dialectiques et dialogiques (53 critères)	72 %

TABLE 2 : Résultat de la classification : agonistes hostiles et non hostiles aux Roms

Les critères ont été évalués au moyen de l'algorithme Naïve Bayes Multinomial (Mccallum & Nigam 1998). Comme dans l'expérience précédente, la ligne de comparaison demeure la classification sur formes simples, qui permet d'obtenir un taux d'exactitude de 70 %. L'élément marquant ici est que le résultat de la classification, avec seulement 11 critères dialogiques, égale pratiquement la ligne de comparaison, quand les critères dialectiques la dépassent, tout comme la combinaison des deux catégories. On a la démonstration que des marqueurs énonciatifs en très petit nombre – essentiellement quelques pronoms *je*, *nous*, *vous*), des adjectifs possessifs, et le tag *NAM* (pour noms propres, obtenue au moyen d'un étiquetage Treetagger) – peuvent suffire à obtenir une classification, certes perfectible, mais comparable à celle effectuer sur les formes simples. C'est l'indice selon nous que la seule posture énonciative observable dans la sélection des marques de personnes, suffit, dans certaines tâches de classification, à identifier la position idéologique des énonciateurs.

5.3 Agonistes favorable (pour) et défavorable (contre) dans des débats parlementaires

La troisième expérience, réalisée pour les besoins de cet article, s'appuie sur le corpus de débats parlementaires mis à la disposition de la campagne DEFT 2007 (Grouin *et al.* 2007). Ce corpus regroupe 28 832 interventions de députés a` l'Assemblée Nationale extraites des débats. Le corpus d'apprentissage totalise 17 299 interventions. Il est divisé en deux

[12] Voir note 9.

classes : 6 899 interventions favorables a` la loi en cours d'examen ; 10 400 interventions défavorables à la loi en cours d'examen. Le corpus de test, quant à lui, est composé de 11 533 interventions au total, 4 961 intervention favorables et 6 572 défavorables. La référence est établie en considérant le vote effectif (favorable ou défavorable à la loi en examen) des intervenants. L'application de la méthodologie exposée ici a permis d'identifier 26 critères dialogiques et 64 critères dialectiques.

Les critères dialogiques favorables à la loi en examen sont le pronom personnel et possessifs de 1e personne (*je*, *mon*, *ma*, *mes*), la mention de partis politiques et des verbes porteurs de la fonction expressive au sens jakobsonien (« C'est pourquoi *nous saluons* le travail accompli par la commission », « ce dont *je me réjouis* », etc.). Les critères dialectiques sont notamment des verbes modaux (*il faut*, *il doit*) et quelques éléments de structuration argumentative, par exemple, énumératifs (« *Enfin*, ce projet répond aux attentes de nos concitoyens », « Il serait *également* souhaitable que, etc. »). Les critères dialogiques défavorables sont les pronoms personnels et possessifs de 2e personne du pluriel (*vous*, *votre*, *vos*) ou encore l'impersonnel *on* qui dénotent des prises de paroles interlocutoires plus marquées que pour les parlementaires favorables. Parmi les marqueurs dialectiques défavorables, on relève une forte saillance des marques de négation (*non*, *ne*, *pas*, *jamais*, *rien*). Des stratégies rhétoriques plus agressives sont également observables via des adverbes d'interrogation (*comment*, *quand*), le point d'interrogation et divers marqueurs argumentatifs d'opposition (*Or*, *Mais*, *pourtant*).

Types de critères	Exactitude *sur corpus de test DEFT 2007*	Exactitude *validation croisée*
Mots simples (5 832 critères) *ligne de comparaison*	70 %	76 %
Critères dialogiques (26 critères)	61 %	65 %
Critères dialectiques (64 critères)	65 %	68 %
Critères dialectiques et dialogiques (90 critères)	66 %	70 %

TABLE 3 : Résultat de la classification : débats parlementaires « pour » et « contre » la loi en examen

La classification effectuée comme précédemment avec SMO donne un taux d'exactitude de l'ensemble de nos critères textométriques de 70 % en validation croisée sur 10 sections et 66% sur le corpus de test fourni par DEFT 2007 ; les résultats sont en deçà des lignes de comparaison mais présentent la particularité de n'avoir été obtenus qu'à partir des formes et des n-grammes de formes, sans lemmatisation, sans étiquetage morphosyntaxique, sans normalisation de la casse, ni recherche de patrons de cooccurrences. On notera que l'écart entre le taux d'exactitude obtenu avec nos critères et celui de la ligne de comparaison est moins important sur le corpus de test que par validation croisée, ce qui témoigne d'une certaine robustesse.

6 Conclusion

En couplant la sémantique textuelle, la textométrie et des méthodes d'apprentissage automatique, nous avons tenté de valider la pertinence applicative du concept de *composante sémantique* dans le cadre de différentes tâches de classification de textes subjectifs. La méthodologie présentée permet d'identifier un très petit nombre de critères textuels de classification qui sont pertinents et surtout non triviaux pour de telles tâches, et de les interpréter suivant une grille de lecture linguistiquement contrôlée. Sans viser le dépassement des différentes méthodes évoquées dans l'état de l'art (paragraphe 4.1.), la méta-étude effectuée apporte la démonstration que les critères relevant des seules composantes dialectique (construction narrative et argumentative) et dialogique (positionnements énonciatifs, acteurs), permettent d'obtenir des résultats de classifications approchant (expérience 5.3), voisinant (expérience 5.2) ou surpassant (expérience 5.1) une ligne de comparaison simulant les techniques apprentistes standard. Mieux encore, l'étude souligne que ces critères textuels classifiants, identifiés selon une méthode d'extraction ascendante (la textométrie), ne ressortissent nullement aux catégories traditionnellement proposées, souvent au moyen de méthodes descendantes par application de modèles cognitifs ou issus de la psychologie, notamment.

Bibliographie

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

Bloom, K. & Argamon, S. (2010). "Unsupervised extraction of appraisal expressions", *Proceedings of the 23rd Canadian conference on Advances in Artificial Intelligence (AI'10)*, Atefeh Farzindar and Vlado Kešelj (Eds.). Springer-Verlag, Berlin, Heidelberg, p. 290-294.

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

Bourion, E. (2001). *L'aide à l'interprétation des textes électroniques*, Thèse de doctorat, Université Nancy 2.

Brunet E. (2000). « Qui lemmatise dilemme attise », *Lexicometrica*, 2, 1-19.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression.* Hachette Education.

Church, K. (2011). "A pendulum swung too far", *Linguistic Issues in Language Technology*, 6.

Eensoo E. & Valette M. (2012). « Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments », dans G. Antoniadis, H. Blanchon, G. Sérasset, Eds, *Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, vol. 2, *TALN, 4-8 juin 2012*, Grenoble, p. 367-374.

Eensoo E. & Valette M. (2014a). « Sémantique textuelle et TAL : un exemple d'application à l'analyse des Sentiments », dans D. Ablali, S. Badir, D. Ducard, Eds., *Documents, textes, œuvres*, Presses Universitaires de Rouen, Collection Rivages linguistiques.

Eensoo E. & Valette M. (2014b). « Approche textuelle pour le traitement automatique du discours évaluatif », dans A. Jackiewicz, (éd.), *Études sur l'évaluation axiologique*, *Langue française*, décembre 2014, 184, p. 107-122.

Ghorbel H. & Jacot D. (2011). "Further Experiments in Sentiment Analysis of French Movie Reviews", E. Mugellini, P. Szczepaniak, M. Pettenati, M. Sokhn, Eds., *Advances in Intelligent Web Mastering – 3,* Berlin / Heidelberg : Springer, *86*, p. 19-28

Grouin C., Berthelin JB, El Ayari S, Heitz T, Hurault-Plantet M, Jardino M, Khalis Z & Lastes M. (2007). Présentation de DEFT'07. *Actes de l'atelier de clôture du 3ème Défi Fouille de Textes*, 2007. Grenoble, France. p. 1–8.

Hatzivassiloglou V. & Wiebe J. (2000). "Effects of adjective orientation and gradability on sentence subjectivity", *Proceedings of the International Conference on Computational Linguistics (COLING).*

Hall M., Eibe F. Holmes G., Pfahringer B., Reutemann P. & Witten I. H. (2009). « The WEKA Data Mining Software: An Update », *SIGKDD Explorations*, Volume 11, Issue 1.

Hall D., Jurafsky D. & Manning C. D. (2008). "Studying the history of ideas using topic models", *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 363–371.

Heiden S., Magué J.-P. & Pincemin B. (2010). « TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement », I. C. Sergio Bolasco (éd.)*, JADT 2010*, vol. 2, p. 1021-1032.

Kim, S.M., Valitutti, A. & Calvo, R.A. (2010). "Evaluation of unsupervised emotion models to textual affect recognition", *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text (CAAGET '10),* Association for Computational Linguistics, Stroudsburg, PA, USA, p. 62-70.

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

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

Lavalley, R.; Clavel, C. & Bellot, P. (2010). « Extraction probabiliste de chaînes de mots relatives à une opinion », *Traitement Automatique des Langues*, 51, p. 101-130.

Lin W.-H. & Hauptmann, A. (2006). "Are these documents written from different perspectives? a test of different perspectives based on statistical distribution divergence", *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44).* Association for Computational Linguistics, Stroudsburg, PA, USA, p. 1057-1064.

Liu, H., Lieberman, H. & Selker, T. (2003). "A model of textual affect sensing using real-world knowledge", *Proceedings of the 8th international conference on Intelligent user interfaces (IUI '03)*, ACM, New York, NY, USA, p. 125-132.

McCallum A. & Nigam K. (1998). "A Comparison of Event Models for Naive Bayes, Text Classification", *AAAI-98 Workshop on 'Learning for Text Categorization'*, p. 41-48

Mayaffre D. (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.

Micheli R., Hekmat I. & Rabatel A., Eds. (2013). *Les émotions argumentées dans les médias*, *Le discours et la langue*, 4/1, EME Éditions, 222 p.

Pang P., Lee L. & Vaithyanathan, S. (2002). "Thumbs up? Sentiment Classification using Machine Learning Techniques", *Proceedings of EMNLP,* p. 79-86.

Pedauque R. T., Coll. (2007). *La redocumentarisation du Monde*, Paris, Éditions Cépadues, 213 p.

Perlerin, V. (2004). *Sémantique légère pour le document. Assistance personnalisée pour l'accès au document et l'exploration de son contenu*, Thèse de doctorat, Caen.

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

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

Rastier F. (1998). « Le problème épistémologique du contexte et le statut de l'interprétation dans les sciences du langage », *Langages*, 129, p. 97-111.

Rastier F. (2001). *Arts et sciences des textes*, Paris, PUF, 303 p.

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

Rastier F. & Pincemin B. (1999). « Des genres a` l'intertexte », I. Kanellos (éd.), *Cahiers de Praxématique*, 33, *Sémantique de l'intertexte*, p. 83-111.

Rossignol M. (2005). *Acquisition sur corpus d'informations lexicales fondées sur la sémantique différentielle*, Thèse de doctorat, Université de Rennes 1.

Salem A., Lamalle C., Martinez W., Fleury S., Fracchiolla B., Kuncova A. & Maisondieu A. (2003). *Lexico3 – Outils de statistique textuelle, Manuel d'utilisation*, Université de la Sorbonne nouvelle – Paris 3.

Saussure, F. de (2002). *Écrits de linguistique générale*, Paris, Gallimard.

Tanguy L. (2012). *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*, Mémoire d'habilitation à diriger des recherches, Université Toulouse-Le Mirail, Toulouse.

Tanguy L. & Fabre C. (2014). « Évolutions de la linguistique outillée : méfaits et bienfaits du TAL », *L'information grammaticale*, 142, p. 15-23.

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

Vernier M., Monceaux L., Daille B. & Dubreil E. (2009a). « Catégorisation des évaluations dans un corpus de blogs multi-domaine », *Revue des nouvelles technologies de l'information (RNTI)*, p. 45-70.

Vernier M., Monceaux L. & Daille B. (2009b). « DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique », *Actes de l'atelier de clôture de la 5ème édition du Défi Fouille de Textes.*

Villena-Román, J., Collada-Pérez, S., Lana-Serrano, S., González-Cristóbal, J. C. (2011). "Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization", *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference*, p. 323-328.

Whitelaw C., Garg N. & Argamon S. (2005). "Using appraisal groups for sentiment analysis", ACM (éd.), *Proceedings of the 14th ACM International Conference on Information and Knowledge Management*, p. 625-631.

Yvon F. (2006). *Des apprentis pour le traitement automatique des langues*, Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris.

Zweigenbaum P., Bellot P., Grau B., Ligozat A.-L., Robba I., Rosset S., Tannier X ? et Vilnat A. (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.