Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques

Matthieu Vernier* — **Laura Monceaux***

* *Laboratoire d'Informatique de Nantes Atlantique (LINA - UMR 6241)*
Université de Nantes
2, rue de la Houssinière
44322 Nantes cedex 03
{Matthieu.Vernier, Laura.Monceaux}@univ-nantes.fr

RÉSUMÉ. *De récentes considérations en fouille d'opinions, guidées par des besoins applicatifs en rupture avec les approches traditionnelles du domaine, requièrent d'utiliser des ressources lexico-sémantiques quantitativement et qualitativement riches. Il n'existe actuellement pas de telles ressources pour le français. Dans cette optique, nous présentons une méthode d'apprentissage pour enrichir automatiquement un lexique de termes subjectifs. Elle s'appuie sur un oracle, l'indexation des documents du Web par un moteur de recherche, et sur les résultats donnés en réponse à un grand nombre de requêtes. La modélisation de contraintes linguistiques sur ces requêtes permet d'inférer les caractéristiques de subjectivité d'un grand nombre d'adjectifs, d'expressions nominales et verbales de la langue française. Nous évaluons l'enrichissement du lexique de termes subjectifs en mesurant en contexte la qualité de la détection locale des évaluations dans un corpus de blogs de 5 000 billets et commentaires.*

ABSTRACT. *Recent considerations in opinion mining, oriented by real applicative tasks, require creating lexical and semantic resources quantitatively and qualitatively rich. In this context, we present a method to automatically enhance a lexicon of subjective terms. The method relies on the indexing of Web documents by a search engine and large number of requests automatically sent. The construction of these requests, linguistically motivated, can infer the value of semantics and axiological aspect of a large number of adjectives, nouns and noun phrases, verbs and verbal phrases of French. We then evaluate the lexicon enhancement by testing the detection of local evaluation in a corpus of 5 000 blogs notes and comments.*

MOTS-CLÉS : *fouille d'opinions, subjectivité, ressource lexico-sémantique, apprentissage.*

KEYWORDS: *opinion mining, subjectivity, lexical and semantic resource, machine learning.*

TAL Volume 51 n° 1/2010, pages 125 à 149

1. Introduction

Le Web 2.0 en tant qu'espace d'expression libre a littéralement dopé l'intérêt pour la fouille d'opinions. À travers les blogs et les différents sites de réseaux sociaux (Twitter, Facebook), les internautes partagent leurs centres d'intérêts, argumentent leurs points de vue, affirment leur personnalité, voire cherchent volontairement à médiatiser leurs opinions pour influer sur leurs communautés. La richesse informationnelle contenue sur ces différents supports et le pouvoir potentiel induit par la capacité à mesurer, voire à prédire, l'évolution de l'opinion générale font naître des besoins applicatifs nouveaux. Les industriels, ayant un fort besoin relationnel avec leurs clients (EDF, SNCF), les créateurs de nouvelles technologies sensibles aux modes (téléphone et ordinateur portables), les politiciens, les médias et les sociologues sont les principaux utilisateurs finaux de ce type d'application. Ils sont désireux de connaître précisemment ce qui est exprimé à un instant donné autour d'un sujet. En TAL, les blogs sont de plus en plus souvent utilisés comme supports d'étude pour la fouille d'opinions (Mishne et Glance, 2006 ; Conrad et Schilder, 2007 ; Vernier *et al.*, 2009b). Plus complexes à traiter que les textes de critiques selon Liu (2009), ils imposent de nouveaux défis et nécessitent d'autres approches que celles proposées par les très nombreux travaux pionniers en classification de textes (Turney, 2002 ; Pang *et al.*, 2002 ; Aue et Gamon, 2005). À l'image du récent DÉfi Fouille de Texte (DEFT'09[1]), une problématique plus actuelle consiste à repérer et délimiter, avec une granularité fine, des passages subjectifs marqueurs d'opinions, afin d'en catégoriser, par la suite, différents aspects sémantiques : leur axiologie (positive ou négative), leur champ sémantique (hédonique, esthétique, pragmatique, éthique, croyance, etc.), l'engagement du locuteur, etc. Nous proposons ci-dessous un extrait de blog rencontré. Les passages entre crochets correspondent à l'annotation optimale recherchée :

> *Ce livre,* [***j'avais hâte***] *de lire car outre le fait de dévoiler la réalité sur Fadela Amara, il m'a permis de découvrir une partie du combat de* [***ma chouchou !!***] [...] *l'*[***une des personnalités préférées des Français***]*, [...] et qui* [***n'a pas froid aux yeux***]. [...] *Le parcours de Fadela Amara* [***tient*** *à la fois* ***du conte de fées***] *et de la construction médiatique et politique.* [...] [***Quelle déception !!***] [...] [***Je n'avais pas vraiment une bonne image***] *des hommes politiques en général* [...] [***Une belle cause***]*,* [***un beau combat***] [***détournés pour les ambitions***] *de certains. Que de* [***mensonges***]*, que de* [***manipulation***]*... Comment, au nom du pouvoir, peut-on ainsi* [***briser l'élan***] *de personnes qui* [***se battent avec leurs tripes***] *pour* [***soutenir des gens qui souffrent***] [...]. *C'est* [***à vomir !***]
>
> Texte extrait de la plateforme Over-blog - 09/09/2009

Cet exemple illustre les problématiques posées lorsque l'on se confronte à la

1. http://deft09.limsi.fr/

détection et la catégorisation des opinions (ou évaluations[2]) à un niveau de granularité plus fin comme :

– la création et l'amélioration des ressources afin d'appréhender la richesse lexicale du langage évaluatif. De très nombreuses expressions comme *ne pas avoir froid aux yeux, tenir du conte de fée, se battre avec ses tripes* ne sont pas prises en compte par les ressources existantes ;

– la modélisation du langage évaluatif pour mieux détecter et tenir compte des tournures négatives, des marqueurs d'engagement du locuteur, de son degré de certitude (***je n'avais pas vraiment*** *une bonne image*) par rapport à l'évaluation exprimée. Typiquement, les outils standard détectent des marqueurs simples (*préféré*, *déception*, *bonne*, *beau*, *mensonge*, *manipulation*, *souffrir*) en leur associant un unique trait sémantique (positif ou négatif) et ne modélisent pas, ou peu, la croyance, l'intensité et l'engagement ;

– la détection des sujets évalués pour traiter les textes qui prennent position sur plusieurs concepts simultanément (Josef Ruppenhofer et Wiebe, 2008 ; Stoyanov et Cardie, 2008). C'est le cas des textes de blogs, ce qui réfute l'idée de catégoriser un texte de blog dans son ensemble. Dans l'exemple donné, et afin de conserver la qualité de l'analyse, il faut pouvoir filtrer les évaluations qui portent sur trois cibles (*Fadela Amara*, *son livre*, *les hommes politiques*). Cette tâche a récemment fait l'objet de campagnes d'évaluations internationales[3] mais reste très peu explorée dans le traitement du français.

Quelques outils permettant de suivre en temps réel l'évolution de l'opinion des internautes *via* leurs activités sur les réseaux sociaux (Tweetfeel[4], Moodviews[5], etc.) sont très perfectibles d'un point de vue qualitatif sur ces aspects. Les solutions utilisées sont généralement très naïves : la modélisation du langage évaluatif est réduite à un axe positif/négatif, les ressources lexicales sont limitées à une liste de termes simples (essentiellement des adjectifs), les tournures négatives ne sont pas prises en compte, les sujets évalués sont les mots qui *cooccurrent* dans la phrase, etc.

Dans ce contexte, une première version de l'outil Apopsis (Vernier *et al.*, 2009b) a été proposée pour permettre de détecter au sein d'un texte les différents passages exprimant une opinion et de catégoriser leur polarité axiologique, leur rôle discursif (appréciation, jugement, accord, désaccord, etc.) et la stratégie énonciative du locuteur : Assume-t-il sa subjectivité ou cherche-t-il à la dissimuler ? Quel est son degré de croyance par rapport à l'évaluation qu'il exprime ? La première version de l'outil s'appuie, entre autres, sur des ressources lexicales construites manuellement. Apopsis a été évalué qualitativement et quantitativement sur un corpus de 100 blogs

2. D'un point de vue terminologique, nous préférons choisir le terme d'évaluation et de langage évaluatif en adoptant le point de vue des théories linguistiques anglo-saxonnes (Martin et White, 2005) et françaises (Charaudeau, 1992 ; Galatanu, 2000).
3. TREC : http://trec.nist.gov/
4. Langue traitée : *anglais* - http://www.tweetfeel.com/
5. Langue traitée : *anglais* - http://moodviews.com

et a participé à la campagne d'évaluation DEFT'09 (Vernier *et al.*, 2009a) où il a obtenu les meilleurs résultats. Il a par ailleurs été utilisé dans plusieurs scénarios de tests applicatifs à échelle réelle pour mesurer, par exemple, l'évolution de l'opinion des internautes à propos de Raymond Domenech entre juillet 2008 et juillet 2009. Les conclusions de ces différents tests montrent des résultats de bonne qualité (la précision varie entre 0,80 et 0,90). Cependant ils sont de moyenne qualité sur l'aspect quantitatif (le rappel avoisine 0,50). La couverture des ressources lexicales est le principal facteur expliquant les nombreuses évaluations non détectées. Dans cet article, nous nous intéressons précisément à ce problème en cherchant à acquérir automatiquement de nouveaux mots ou expressions liés au langage évaluatif, habituellement manquants dans les ressources existantes, afin d'augmenter le rappel. Nous attachons cependant une importance particulière à ne pas faire chuter la qualité de la ressource initiale. Notre approche s'inspire des tests sémantiques décrits par Anscombre et Ducrot (1983), puis repris dans les travaux informatiques de Hatzivassiloglou et McKeown (1997) pour prédire l'orientation sémantique des adjectifs. Nous développons cette idée en proposant de nouveaux patrons de tests sémantiques permettant d'induire le degré de subjectivité d'adjectifs, d'expressions nominales ou verbales à partir des usages linguistiques des internautes (section 4).

Dans la section 2, nous définissons le concept de subjectivité et la modélisation du langage évaluatif que nous adoptons. Nous présentons les ressources lexico-sémantiques existantes pour la fouille d'opinions, le principe des méthodes automatiques d'enrichissement lexicale du domaine et leurs limites en comparaison avec la richesse du langage évaluatif définie précédemment. Dans la section 3, nous présentons la version initiale du lexique de l'évaluation développée manuellement, pour traiter le français. Cette version comporte 982 entrées et a été notamment utilisée par l'outil Apopsis pour DEFT'09. La méthode d'enrichissement automatique de ce lexique initial constitue le cœur de notre travail (section 4). Elle repose sur la modélisation de contraintes linguistiques et sur l'interprétation des résultats de requêtes envoyées à un moteur de recherche pour catégoriser de nouveaux candidats. Pour les besoins du protocole expérimental, nous cherchons à valider cette méthode, en contexte, en comparant la détection des évaluations dans un corpus de blogs avant et après l'enrichissement automatique du lexique de l'évaluation (section 5). Néanmoins, comme il a été évoqué par les organisateurs et les participants du défi DEFT'09, l'évaluation des méthodes de détection d'opinions à granularité fine reste une problématique importante du domaine. Elle est, dans la pratique, non résolue idéalement. Nous discutons plus précisément de cet aspect dans les sections 5 et 6.

2. Considérations récentes en fouille d'opinions

L'expression *sentiment analysis*, traduite de façon erronée en français par *analyse des sentiments*, est apparue *via* les travaux de Turney (2002) et Pang *et al.* (2002). Elle est reprise dans un très grand nombre de travaux sur la classification de critiques (de films, de livres, de voitures, de téléphones), ce qui vaut à cette expression d'être

fréquemment associée à cette tâche. Le terme *opinion mining* (fouille d'opinions) est utilisé de façon plus générale et englobe différentes problématiques de recherche liées à l'opinion : la construction de ressources lexicales, la classification de textes en deux classes objectif/subjectif, la détection de passages d'opinions et leurs catégorisations sémantiques, le résumé automatique de textes d'opinions, etc. En réalité, ces deux expressions réfèrent aux mêmes tâches et témoignent de la variété importante des problématiques et des approches de ce domaine au détriment parfois d'une bonne lisibilité terminologique. L'influence des objectifs applicatifs est déterminante dans le choix des méthodes. Pour suivre l'intérêt porté autour des textes issus de la blogosphère et des réseaux sociaux, un des courants actuels du domaine résumé par Liu (2009) consiste à adapter des modèles linguistiques du langage évaluatif vers des modèles informatiques pour détecter les évaluations à une granularité plus fine, améliorer leur catégorisation sémantique (en dépassant le simple aspect positif/négatif) et détecter le sujet cible qu'elles évaluent. Nous présentons dans les sous-sections suivantes, les modèles linguistiques du langage évaluatif à la base de nos travaux et précisons en quoi les ressources lexico-sémantiques existantes dans le domaine ne répondent pas précisément à ces modèles et aux attentes applicatives.

2.1. *Modélisation du langage évaluatif*

L'évaluation est définie par Lavelle (1950) comme *l'acte de rupture de l'indifférence par laquelle nous mettons toutes les choses sur le même plan et considérons toutes les actions comme équivalentes*. Tout acte de langage révélant une rupture d'indifférence relève donc du phénomène évaluatif. Ces actes mettent en jeu des mécanismes sémantiques, pragmatiques ou énonciatifs complexes faisant l'objet de nombreuses études (Benveniste, 1974 ; Anscombre et Ducrot, 1983). Selon Kerbrat-Orecchioni (1997), lorsqu'un sujet d'énonciation se trouve confronté au problème de la verbalisation d'un objet référentiel, réel ou imaginaire, et que pour ce faire il doit sélectionner certaines unités de son stock lexical et syntaxique, il a le choix entre deux types de formulations :

– le discours objectif, qui s'efforce de gommer toute trace de l'existence d'un énonciateur individuel ;

– le discours subjectif, dans lequel l'énonciateur s'avoue explicitement (*je trouve ça moche*) ou se pose implicitement (*c'est moche*) comme la source évaluative de l'assertion.

La classification sémantique de la subjectivité dans le lexique proposée par Kerbrat-Orecchioni (1997) oppose les lexèmes objectifs, relativement stables en langue (par exemple les adjectifs de couleur), aux lexèmes subjectifs associés à des échelles de valeurs propres à chaque locuteur. Tels sont, parmi les adjectifs, les affectifs (*poignant, drôle*), les évaluatifs comportant des traits axiologiques[6](*beau*, *imbécile*, etc.) ou

6. Associés à une échelle bien/mal.

modalisateurs[7] (*véritable*, *marginal*, etc.). Ces classes lexicales ne sont pas stables puisque renvoyant à des systèmes individuels. De plus, tout élément est susceptible de se charger de traits subjectifs qu'il n'a pas initialement. Une démarche analogue permet de classer les noms et les verbes. Dans le même courant d'idée, Charaudeau (1992) montre qu'il existe cinq modalités permettant à un locuteur d'exprimer une évaluation (l'opinion, l'accord ou le désaccord, l'acceptation ou le refus, le jugement et l'appréciation). Chacune de ces modalités révèle une attitude particulière du locuteur : sa croyance plus ou moins certaine par rapport à l'évaluation qu'il exprime, le champ d'expérience dans lequel il se positionne (éthique, moral, intellectuel, esthétique, etc), sa position par rapport à son énoncé (présence ou absence du « je »). Selon Charaudeau, il existe des marqueurs lexicaux et des structures linguistiques spécifiques à ces modalités (Tab. 1).

Marqueurs	**Modalité**
être sceptique, douter, à mes yeux, penser, être convaincu	Opinion (conviction faible à forte)
effectivement, d'accord, non, être faux	Accord ou désaccord
courageux, héros, lâche, mentir	Jugement (éthique, moral, intellectuel)
adorer, haïr, plaisir, triste	Appréciation (affect, hédonique)

Tableau 1. *Exemples de marqueurs lexicaux pour les modalités de l'évaluation*

La théorie de Galatanu (2000) sur l'évaluation complète le modèle de Charaudeau en hiérarchisant les modalités sur une échelle de subjectivité. Lorsqu'un locuteur organise son énoncé, il peut choisir d'objectiver ou de subjectiver son discours en activant certaines modalités ou en adoptant une configuration énonciative explicite ou implicite (présence ou absence du pronom *je*, inclusion du locuteur dans le pronom *nous*, etc.). Dans les exemples du tableau 2, la valeur mise en jeu « mentir » (modalité de jugement) intervient dans des stratégies argumentatives différentes. L'évaluation *Je n'aime pas qu'il mente* paraîtra ainsi plus personnelle (ou plus subjective) que *nous condamnons ses mensonges* ou *Oui, c'est un menteur*, alors qu'elles mettent pourtant toute la même valeur évaluative en jeu : *mentir*.

Exemple	**Sur-modalité**	**Modalité**
Je doute qu'il mente	Opinion faible explicite	Jugement implicite
Il est évident qu'il ment	Opinion forte implicite	Jugement implicite
Oui, c'est un menteur	Accord	Jugement implicite
Il ment		Jugement implicite
Je n'aime pas qu'il mente	Appréciation explicite	Jugement implicite
Nous condamnons ses mensonges	Jugement explicite	Jugement implicite

Tableau 2. *Exemple de discours évaluatif différent pour la même valeur* mentir

7. Associés à une échelle vrai/faux ou réel/irréel.

Pour adapter ces théories vers des modèles d'analyse automatique, la première étape consiste à construire des ressources pour stocker ces marqueurs lexicaux évaluatifs. Puis, il s'agit pour chaque entrée lexical, d'ajouter des traits sémantiques correspondant aux catégories décrites dans ces modèles (appréciation, jugemement, accord, opinion - explicite, implicite - positif, négatif). Les bons résultats obtenus par Apopsis durant le DÉfi Fouille de textes 2009 pour détecter les passages subjectifs[8] à partir de modèles informatiques tenant compte de ces théories montrent leur intérêt pour améliorer la qualité des analyses et résoudre des problèmes applicatifs réels.

2.2. *Ressources lexicales existantes sur le langage évaluatif*

Dans le domaine de la fouille d'opinions, plusieurs travaux importants ont abouti à la création de ressources lexicales pour améliorer les systèmes automatiques, en particulier pour l'anglais. Nous présentons ces ressources et discutons de leurs caractéristiques : le nombre d'entrées de la ressource, la nature des entrées (quelles catégories grammaticales ? présence de mots simples uniquement ?), les traits sémantiques représentés, la méthode utilisée pour constituer la ressource.

2.2.1. *WordNet-Affect*

Wordnet-Affect (Strapparava et Valitutti, 2004) est une ressource linguistique pour la représentation lexicale de connaissances sur les affects pour l'anglais. Un sous-ensemble de synsets[9] de WordNet appropriés est choisi pour représenter des concepts affectifs. Des informations additionnelles sont ajoutées aux synsets affectifs, en leur associant une ou plusieurs étiquettes qui précisent une signification affective. WordNet-Affect est constitué de 1309 concepts (539 noms, 517 adjectifs, 238 verbes et 15 adverbes) directement ou indirectement liés à un état mental ou émotionnel. Par exemple, les concepts affectifs représentant un état émotif sont représentés par des synsets marqués par l'étiquette Emotion (*Anger, Fear,* etc.). WordNet-Affect a été développé semi-manuellement en deux étapes : l'identification manuelle d'un premier « noyau » de synsets affectifs, l'utilisation des relations de synonymie/antonymie présentes dans WordNet afin de propager les informations de ce noyau à son voisinage. Les entrées lexicales de Wordnet-Affect sont très majoritairement des mots simples.

2.2.2. *SentiWordNet*

De façon complémentaire, SentiWordNet (Esuli et Sebastiani, 2006) est une ressource dédiée aux systèmes de classification de textes d'opinions. SentiWordNet assigne à chaque synset de WordNet trois valeurs : Positivité, Négativité, Objectivité (en respectant l'égalité : Positivité + Négativité + Objectivité = 1). Chacune des valeurs a été déterminée par apprentissage supervisé sur des corpus dont les textes

8. Précision : 80,8 % , Rappel : 92,6 % . Voir *Actes du cinquième DÉfi Fouille de Textes 2009*.
9. Un synset correspond à un groupe de mots interchangeables, dénotant un sens ou un usage particulier.

sont classés *positif*, *négatif* ou *objectif* en exploitant également les relations de synonymie/antonymie de WordNet par la suite.

L'approche statistique pour constituer cette ressource laisse toutefois présupposer des limites applicatives, l'adjectif *interesting* (intéressant) est ainsi classé comme étant *objectif*. Tout comme Wordnet-Affect, SentiWordnet ne contient que des mots simples et est donc loin de couvrir l'ensemble des segments évaluatifs des textes. Les passages évaluatifs dépassent très fréquemment le cadre du mot (*coup de foudre, chasse à l'homme, cri d'alarme*, etc.) et sont souvent constitués de plusieurs mots *objectifs* lorsqu'ils sont pris séparément. D'un point de vue quantitatif, bien qu'étant elles-mêmes incomplètes, ces deux ressources n'ont pas d'équivalent pour le français.

2.2.3. *Lexique des sentiments*

Développé manuellement pour le français, le lexique des sentiments (Mathieu, 2005), comporte un millier de mots - exclusivement des mots simples - exprimant des sentiments, des émotions et des états psychologiques. Ces mots sont répartis en 38 classes sémantiquement homogènes : 22 classes négatives (Peur, Tristesse, Irritation, etc.), 14 classes positives (Amour, Intérêt, Passion, etc.), 2 classes sans polarité (Étonnement et Indifférence). Chaque classe est nommée par le sentiment ou l'état psychologique décrit, comme la classe Peur qui contient les mots relatifs à un sentiment de peur (*peur, crainte, frayeur, effrayer, effrayant,* etc). Les classes sémantiques sont liées entre elles par des relations de sens, d'intensité et d'antonymie. Bien qu'étant une ressource intéressante pour le français, ce lexique des sentiments est limité par sa taille et la nature de ses entrées. L'auteur relève l'importance du niveau pragmatique et énonciatif du texte pour mesurer la subjectivité d'un mot et précise que « *la reconnaissance et l'interprétation du simple vocabulaire du domaine ne sont pas suffisantes le plus souvent, il faut prendre en compte d'autres éléments comme les expressions idiomatiques ou figées telles que être la prunelle de ses yeux* », p. 320.

2.3. *Enrichir automatiquement des ressources lexicales*

Pour tenter de contourner le coût de la création manuelle de ces ressources, certaines approches consistent à déterminer automatiquement quelques caractéristiques sémantiques de termes inconnus (est-il *objectif* ou *subjectif* ? *positif* ou *négatif* ?). Traditionnellement, les travaux du domaine se concentrent très majoritairement sur les adjectifs. Hatzivassiloglou et McKeown (1997) ont proposé une méthode pour déterminer la polarité de mots. Leur algorithme a pour objectif de déterminer l'orientation sémantique d'adjectifs à partir de l'analyse de leurs cooccurrences avec des conjonctions (*and*, *but*, *neither*, etc.). Cet algorithme est néanmoins limité spécifiquement aux adjectifs et la question de son application à d'autres catégories grammaticales n'est pas résolue.

La technique proposée par Turney (2002) se fonde sur un calcul de proximité sémantique (PMI-IR) entre un mot inconnu et quatorze mots connus servant de points

de repère : 7 adjectifs à polarité positive (*good*, *excellent*, etc.) et 7 à polarité négative (*bad*, *poor*, etc.). À partir d'un corpus de 10 000 000 mots, un mot inconnu est d'autant plus positif qu'il est plus proche des points de repère positifs et plus éloigné des points de repère négatifs. Turney a évalué l'efficacité de sa technique en comparant l'orientation prédite à celle définie dans le *General Inquirer Lexicon* (Stone, 1966) qui contient une liste de 3 596 mots étiquetés comme positifs ou négatifs. Cette technique nécessite donc une ressource de seulement 14 mots, elle étiquette correctement 65 % des mots. Cette méthode a été souvent adaptée depuis, notamment par Banea *et al.* (2008) pour construire des ressources de termes subjectifs pour les langues peu dotées. Bestgen (2002) propose une adaption de la méthode de Turney. La principale différence est qu'il sélectionne un ensemble spécifique de points de repère pour chaque mot à évaluer parmi plusieurs milliers de mots dont la polarité est connue alors que Turney emploie comme points de repère quelques mots sélectionnés *a priori*. Plus précisément, Bestgen s'appuie sur un dictionnaire de 3 000 mots dont la polarité a été évaluée par une trentaine de juges. La polarité inconnue d'un mot correspond à la polarité moyenne (non pondérée par la proximité) de ses 30 plus proches voisins dont la polarité est connue. Ici aussi, les plus proches voisins sont identifiés sur la base d'une analyse sémantique latente et correspondent aux 30 mots ayant le plus grand cosinus avec le mot inconnu. Pour évaluer cet indice, Bestgen a comparé les valeurs prédites pour 60 mots du dictionnaire aux valeurs réelles et a obtenu des corrélations comprises entre 0,56 et 0,70.

Ces dernières techniques sont particulièrement intéressantes pour créer automatiquement des ressources qui serviront à catégoriser des textes dans leur globalité de manière efficace (les erreurs étant globalement compensées par les bons indices) ; toutefois, à un niveau de granularité plus fin, ces techniques ne semblent pas applicables pour conserver une haute précision des résultats. Les auteurs ne font d'ailleurs pas état d'une évaluation en contexte pour vérifier la qualité de leur enrichissement. Enfin, ces méthodes ne cherchent pas à catégoriser les noms composés ou les expressions verbales du langage évaluatif.

3. Lexique de l'évaluation

Le lexique de l'évaluation que nous souhaitons enrichir automatiquement a été développé de manière semi-automatique à partir d'un corpus annoté.

3.1. *Du corpus annoté au lexique de l'évaluation*

Le lexique de l'évaluation se présente comme l'intégration structurée d'informations lexicales et sémantiques autour des valeurs prises par les termes lors de leurs apparitions en contexte. La typologie de l'évaluation retenue pour former ces entrées est fondée sur la théorie des modalités discursives de Charaudeau (1992) présentée dans la section 2. Afin d'extraire ces termes, un corpus annoté de 200 billets et leurs commentaires issus de blogs de la plate-forme Overblog sur des thèmes variés

(comme le cinéma, la politique, le développement durable, le sélectionneur de l'équipe de France de football, etc.) a été constitué (Dubreil *et al.*, 2008). Dans chaque billet de ce corpus, les différentes évaluations présentes et les concepts sur lesquels elles portent ont été annotés. L'étude des évaluations présentes dans ce corpus met en lumière quelques attentions à prendre en compte lors de la constitution du lexique. Des choix empiriques guidés par l'usage ont été mis en place, centrés sur l'ambiguïté d'interprétation et la désambiguïsation morphosyntaxique :

– lorsqu'un terme est présent dans une appréciation contenant une tournure négative (ex. : *pas sympa* : appréciation défavorable), la polarité de l'évaluation associée à ce terme est inversée (*sympa* - appréciation favorable) ;

– les collocations et les expressions figées sont prises en compte comme entrées lexicales (ex. : *hommage vibrant*, *tenir en haleine*) ;

– les ambiguïtés d'interprétation sont prises en compte, tant sur la polarité que sur le type d'évaluation où l'on peut retrouver l'entrée lexicale (ex. : *fou* peut être présent dans des appréciations favorables ou défavorables) ;

– les termes dont la sémantique est subjective mais non intrinsèquement positive ni négative (un vin *fruité*) ne sont à ce stade pas pris en compte dans le lexique.

```
Allez, <Appreciation type="AID" forme="transports">vous avez galéré
</Appreciation> dans les <CC id_c="C2">transports</CC> jusqu'à ce soir...
<Appreciation type="AED" forme="SNCF" ironie="oui">Merci</Appreciation>,
la <CC id_c="C1">SNCF</CC>!
```

Figure 1. *Texte extrait du corpus annoté manuellement. AID : appréciation implicite défavorable, AED : appréciation explicite défavorable*

3.2. *Définition d'une entrée lexicale évaluative*

Chaque entrée lexicale de notre lexique évaluatif se caractérise par des informations :

– **morphosyntaxiques** (*morpho*) : son lemme ou sa décomposition, sa catégorie grammaticale, ses éventuelles variantes orthographiques ;

– **sémantiques** (*evaluation*) : énumération des différentes valeurs évaluatives prises par les termes lors de leur apparition en contexte (type d'évaluation et catégorisation axiologique) ;

– sur les **contextes d'apparition** (*attestation*) : terme issu du corpus annoté, d'un ajout manuel ou d'un ajout automatique.

Pour chaque type d'évaluation (*evaluation*), on note le type de l'évaluation (*type*), son axiologie (*subtype*) et les différentes occurrences dans lesquelles le terme a été trouvé (*occurrence*) accompagnées du concept sur lequel elles portent. Ainsi le terme *grave* a été employé dans un jugement négatif (*type="Jugement" subtype="negatif"*) dont l'occurrence *pose de graves problèmes* porte sur le concept *réforme*.

```
<lexical_entry id="ADJ_grave" ambiguity="true"
    amb_type="onEvalType"
    amb_list="Appreciation#Jugement">
  <morpho> <name>grave</name> <category>ADJ</category></morpho>
  <evaluation type="Appreciation" subtype="defavorable">
    <occurrence><annotation subject="spidey">
      <form config="implicite" neg="true">
        ne lui arrivera rien de grave
      </form>
    </annotation></occurrence>
  </evaluation>
    <evaluation type="Jugement" subtype="defavorable">
    <occurrence><annotation subject="réforme">
      <form config="implicite" neg="true">
        pose de graves problèmes
      </form>
    </annotation></occurrence>
  </evaluation>
  <attestation>corpus annoté</attestation>
</lexical_entry>
```

Figure 2. *Entrée lexicale* grave*, dont l'ambiguïté porte sur le type d'évaluation*

3.3. *Constitution du lexique évaluatif*

Le lexique évaluatif est constitué de manière semi-automatique en quatre étapes :

– prétraitements : ajout d'informations lexicales et morphosyntaxiques ;

– extraction des évaluations annotées et du concept sur lequel elles portent ;

– validation des termes évaluatifs pertinents présents dans ces évaluations ;

– formatage du lexique de l'évaluation selon la DTD définie.

La validation des termes évaluatifs pertinents selon les règles décrites dans la section 3.1 (ambiguïtés, négation, etc.) est la plus délicate. Au final, le lexique l'évaluation contient 982 entrées lexicales dont 54 ambiguës. La majorité des termes évaluatifs sont majoritairement des adjectifs (Tab 3).

L'objectif de notre travail est d'enrichir automatiquement ce lexique évaluatif afin d'étendre sa couverture lexicale.

Catégorie	*Nb. d'entrées*	*Ambiguës*	*Catégorie*	*Nb. d'entrées*	*Ambiguës*
Adjectifs	493	26	Verbes	192	15
Noms	166	3	Syntagmes	24	0
Adverbes	60	9	Autres	47	1
			Total	982	54

Tableau 3. *Répartition par catégorie des entrées du lexique*

4. Apprentissage automatique de nouveaux termes subjectifs

Notre objectif consiste, en autres, à enrichir le lexique de l'évaluation en y ajoutant les termes porteurs de subjectivité non pris en compte jusqu'alors. Dans la langue, ces termes peuvent être des mots (*néfaste, zizanie, laminer*) ou des expressions (*rafler la mise, dialogue de sourd, vent de panique*). Il peut s'agir d'adjectifs, de noms (ou noms composés) ou de verbes (ou expressions verbales). D'autre part, il s'agit d'ajouter une information sémantique sur la polarité axiologique de nouveaux termes lorsque cela est pertinent. À ce stade, nous développons ce deuxième aspect uniquement sur les adjectifs. Pour réaliser ce double objectif, nous présentons le principe de *tests sémantiques* sur lequel se fonde notre méthode d'apprentissage automatique.

4.1. *Principe : des tests sémantiques*

4.1.1. *Mesurer le degré de subjectivité de termes*

L'incidence de certains adjectifs (*vrai*, *véritable*) ou adverbes (*littéralement*, etc.) sur l'énonciation a été étudiée en linguistique par Legallois (2005) et Suhamy (2006). On considère ainsi que l'adverbe *littéralement* ne doit pas être pris à la lettre, et qu'il a au contraire, de par l'usage courant, une fonction intensive qui révèle les représentations mentales et donc la subjectivité du locuteur. De là viennent des expressions comme *le contribuable est littéralement écrasé d'impôts*, *le pétrole qui flambe littéralement sur les marchés*, etc. Nous développons cette idée pour formuler une première hypothèse :

Principe A : Un terme neutre (adjectif, groupe nominal ou groupe verbal) est rarement intensifié par un adverbe.

Cela a du sens de dire :

– Il est *particulièrement* **dynamique**. C'est *véritablement* **une hérésie**.

– Il est *littéralement* **tombé sous le charme**. Il a *littéralement* **soulevé la foule**.

En revanche, les énoncés suivants semblent sémantiquement mal construits :

– C'est *terriblement* **scalaire**. C'est *particulièrement* **législatif**.

– C'est *littéralement* **un oiseau**. Il est *franchement* **allé à l'école**.

À partir de ce principe, nous définissons un ensemble S_1 de tests sémantiques construits selon l'algorithme 1 α est un terme donné dont le degré de subjectivité est inconnu et Δ_{int} est une liste de marqueurs d'intensité. Dans cette liste, nous conservons les adverbes étudiés par Legallois (2005) et Suhamy (2006). Nous y ajoutons, de façon empirique, les adverbes d'intensité les plus fréquents sur Yahoo! et les moins ambigus (exclusion de : *trop*, *vrai*). Les adverbes d'intensité étant beaucoup plus stables sémantiquement et d'un nombre assez réduit par rapport aux termes subjectifs, il ne nous paraît pas nécessaire de chercher à faire varier cette liste. En revanche, l'apparition au fil du temps de nouveaux termes autour de ces adverbes nous semble plus intéressante à détecter.

Algorithme 1 Construire les tests sémantiques pour α

$\Delta_{int} \leftarrow$ [Particulièrement, Terriblement, Parfaitement, Véritablement, Littéralement, Réellement, Franchement, Véritable]
$S_1 \leftarrow \emptyset$
pour tout δ_{int} élement de Δ_{int} **faire**
  $s \leftarrow \delta_{int}\ \alpha$ {ex. : *Véritable coup de trafalgar*}
  $S_1 \leftarrow S_1 + s$
fin pour
return S_1

Nous présentons le contexte d'utilisation de ces tests sémantiques dans le paragraphe 4.1.3.

4.1.2. *Mesurer la polarité axiologique des adjectifs*

Dans ce travail, nous empruntons l'idée de départ de Hatzivassiloglou et McKeown (1997) pour mesurer la polarité axiologique des adjectifs selon un second principe.

Principe B : Deux adjectifs ayant la même polarité axiologique sont rarement mis en opposition.

Cela a du sens de dire :

– C'est **joli** *mais* **inutile**.

– C'est **atroce** *mais* **efficace**.

Les énoncés suivants semblent moins pertinents :

– Elle est **jolie** *mais* **belle**.

– C'est **atroce** *mais* **douloureux**.

De la même façon nous construisons un ensemble S_2 de tests sémantiques construits selon l'algorithme 2 α est un adjectif subjectif dont la polarité axiologique est inconnue. Δ_{axiol} est un ensemble d'adjectifs dont la polarité axiologique est connue. Cet ensemble est composé de 100 adjectifs représentatifs (50 positifs,

50 négatifs) issus du lexique de l'évaluation présenté dans la section précédente. Les

Algorithme 2 Construire les tests sémantiques axiologiques pour l'adjectif α

$\Delta_{axiol} \leftarrow$ [liste de 100 adjectifs connus]
$S_2 \leftarrow \emptyset$
pour tout δ_{axiol} élement de Δ_{axiol} **faire**
  $s_a \leftarrow \delta_{axiol}$ mais α {ex. : *consciencieux mais introverti*}
  $s_b \leftarrow \alpha$ mais δ_{axiol} {ex. : *introverti mais consciencieux* }
  $s_c \leftarrow$ pas δ_{axiol} mais α {ex. : *pas introverti mais consciencieux* }
  $s_d \leftarrow$ pas α mais δ_{axiol} {ex. : *pas consciencieux mais introverti* }
  $S_2 \leftarrow S_2 + s_a + s_b + s_c + s_d$
fin pour
return S_2

tests s_c et s_d ont été ajoutés pour tenir compte de certains cas où un adjectif est modifié par une négation. D'autres constructions devraient idéalement être considérées : lorsque l'adjectif est modifié par un adverbe (*x mais **peu** consciencieux*, *x mais **trop** consciencieux*, etc.) et lorsque le connecteur relie un adjectif neutre et un adjectif axiologique (*un **court mais beau** poème*, etc.). Nous avons ici fait le choix de ne pas multiplier les différentes requêtes pour des raisons de coût de calcul. Nous supposons, *a priori*, que ces dernières constructions ont un nombre d'occurrences moins grand et perturberont l'apprentissage et la catégorisation automatique de façon marginale.

4.1.3. *Réaliser automatiquement des tests sémantiques*

Afin de réaliser ces tests sémantiques automatiquement et à grande échelle, nous nous appuyons sur les usages linguistiques des internautes et sur leur fréquence. L'hypothèse consiste à considérer que la pertinence d'un énoncé peut s'évaluer grâce au nombre d'occurrences de cet énoncé sur le Web. Par exemple, à partir du moteur de recherche Yahoo!, les requêtes exactes suivantes apportent un indice sur le degré de subjectivité des termes en gras :

– *véritablement* **scalaire** → 0 occurrence (**scalaire** → 650 000 occurrences ;[10])

– *véritablement* **mangé au restaurant** → 0 occ. (**mangé au restaurant** → 25 100 occ.) ;

– *véritablement* **une hérésie** → 13 occ. (**une hérésie** → 460 000 occ.) ;

– *véritablement* **soulevé la foule** → 15 occ. (**soulevé la foule** → 9 990 occ.).

Les termes *hérésie* et *soulever la foule* sont potentiellement subjectifs selon ces premiers résultats. De la même façon, pour le calcul de l'axiologie, l'adjectif négatif *introverti* apporte un indice sur la polarité de l'adjectif *consciencieux* :

10. Les valeurs numériques fournies par Yahoo! soulèvent évidemment des questions de fiabilités. À partir de certaines valeurs, Yahoo! ne retourne qu'un ordre de grandeur. Toutefois, ces valeurs, même approximatives, constituent *a priori* des indices suffisament pertinents pour valider ou non un candidat.

– **introverti** *mais* **consciencieux** → 26 occurrences.

D'un point de vue technique, Yahoo!Search BOSS[11] est un service permettant de construire et d'exécuter automatiquement des requêtes sur le moteur de recherche Yahoo!. L'API Java de Yahoo!Search BOSS permet, en réponse à une requête, d'obtenir un flux de réponses en XML. L'analyse de ce flux de textes format XML permet notamment de récupérer les champs d'informations suivants : nombre de résultats d'une requête, URL des sites proposés, résumés des sites proposés, etc. Nous utilisons le composant *fr.univ.nantes.lina.uima.YahooSearch*[12] développé pour l'occasion au sein de la plate-forme UIMA (Ferruci et Lally, 2004).

4.2. *Annotation humaine et modèle d'apprentissage*

Pour les besoins du protocole expérimental, cinq juges humains ont annoté des termes de la langue française : 500 adjectifs, 500 noms ou expressions nominales et 500 verbes ou expressions verbales. Ces termes sont extraits automatiquement d'environ 1 000 billets de blogs de la plate-forme Over-blog (Tab. 4), à partir du Treetagger (Schmid, 1994) et d'un algorithme de chunking d'expressions nominales et verbales inspiré de (Vergne et Giguet, 1998). Pour chacun des termes extraits, les juges humains doivent prendre la décision suivante : *le terme est-il subjectif ?* {oui, non, ambiguïté}. L'accord inter-annotateur est calculé à partir de l'indice Kappa de Fleiss[13] (Fleiss, 1971). Pour cette tâche, l'indice moyen obtenu entre juges humains est de 0,70, ce qui correspond à un bon taux d'accord d'après l'échelle de cet indice.

Adjectifs	Noms	Verbes
aborigène **(O)**	république **(O)**	prendre la grosse tête **(S)**
téléphonique **(O)**	république bananière **(S)**	tricoter **(O)**
néo-nazi **(A)**	vie de chien **(S)**	échapper des griffes **(S)**
populiste **(S)**	vie de famille **(S)**	voler la vedette **(S)**
télégénique **(S)**	souffle de fraîcheur **(S)**	glandouiller **(S)**

Tableau 4. *Exemples de termes devant être catégorisés par les juges humains en trois classes : subjectif **(S)**, non subjectif **(O)**, ambigu **(A)***

Afin de constituer un modèle d'apprentissage, chaque terme α catégorisé par les juges humains est ensuite automatiquement utilisé dans les tests sémantiques décrits précédemment. Après une étape de normalisation[14], le résultat de chaque

11. http://developer.yahoo.com/search/boss/
12. Tutoriel et sources disponibles ici : http://www.uima-fr.org
13. Par contraste avec l'indice kappa de Cohen, celui-ci permet de calculer l'accord interannotateurs pour plus de deux juges humains.
14. $\delta_i normalisation = \frac{\delta_i}{\sum_{j=1}^{n} \delta_j}$

test sémantique est ainsi un descripteur de α dans le modèle d'apprentissage ($\delta_i = \frac{Occ(s_i,\alpha)}{Occ(\alpha)}$). Nous ajoutons à α un dernier descripteur qui correspond au nombre d'occurrences de ce terme indépendamment de tous tests sémantiques. Chaque instance d'apprentissage est donc décrite par 10 attributs[15] (410 pour les adjectifs[16]).

Pour donner un aperçu en deux dimensions, la figure 3 représente la répartition des termes de type noms ou noms composés catégorisés manuellement selon deux axes :

– **Y** : le nombre d'occurrences du terme associé à un test sémantique de subjectivité (ex. : *littéralement* **pété les plombs** (1 330 occurrences)) ;

– **X** : le nombre d'occurrences du terme seul (ex. : **péter les plombs** (78 700)).

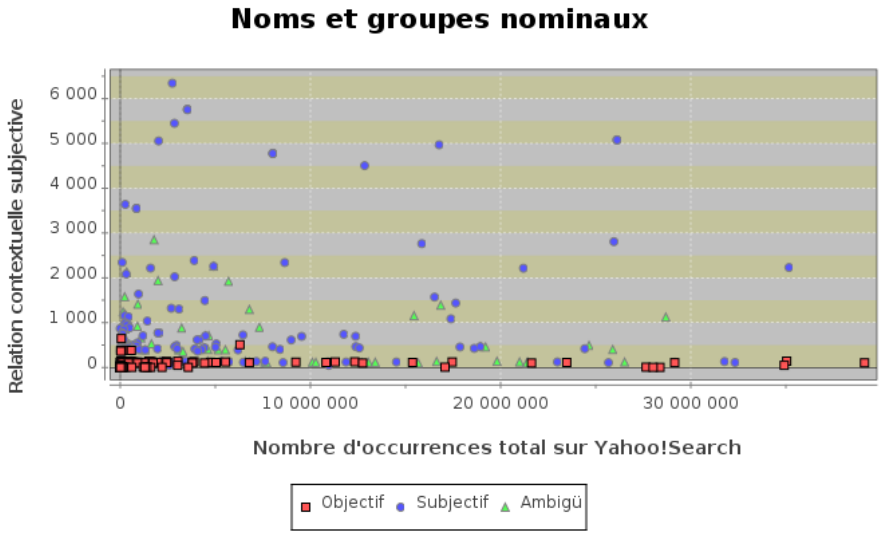

Figure 3. *Répartition des termes catégorisés par les juges humains selon deux axes : nombre d'occurrences du termes (X) et nombre d'occurrences du termes dans des tests sémantiques de subjectivité (Y) dans l'index de Yahoo!Search*

Malgré un espace des descripteurs réduit à deux dimensions, il est visuellement possible de discriminer la zone des termes catégorisés par les humains comme étant objectifs et la zone des termes subjectifs. Les graphiques obtenus pour les verbes/groupes verbaux et les adjectifs sont d'allures équivalentes (Fig. 3) . À partir de ces données et pour chaque catégorie grammaticale, nous entraînons un classifieur SVM (Joachims, 1997) pour déterminer un hyperplan optimal discriminant les termes subjectifs des autres termes et obtenir ainsi une fonction de classification.

15. Un descripteur pour chaque adverbe d'intensité considéré.
16. Un descripteur pour chaque adjectif connu considéré (quatres constructions possibles).

4.3. *Mise en œuvre informatique*

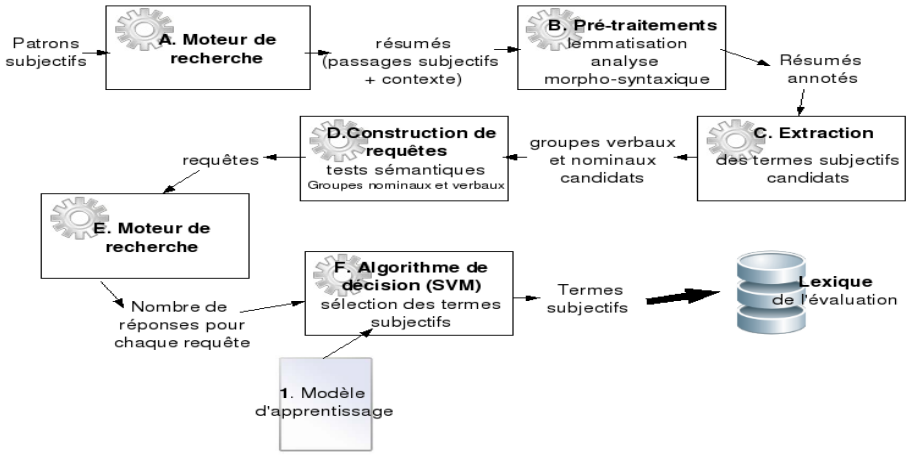

Figure 4. *Chaîne de traitement pour collecter des termes candidats (groupes nominaux et groupes verbaux) sur le Web, mesurer leur degré de subjectivité et enrichir automatiquement le lexique de l'évaluation*

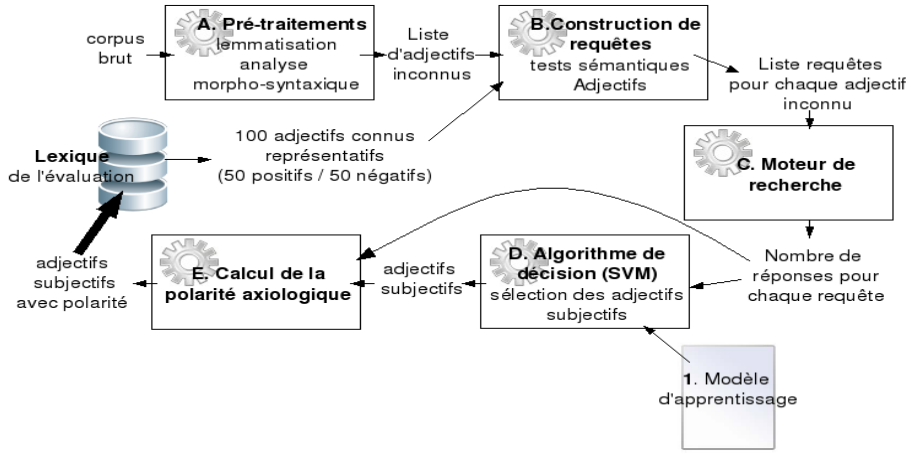

Figure 5. *Chaîne de traitement pour collecter des adjectifs candidats sur le Web, mesurer leur degré de subjectivité, calculer leur polarité axiologique et enrichir automatiquement le lexique de l'évaluation*

Les figures 4 et 5 présentent les chaînes de traitements pour enrichir automatiquement le lexique de l'évaluation. Les deux chaînes sont légèrement différentes selon la catégorie grammaticale des termes candidats.

Dans le cas des groupes nominaux et verbaux (Fig. 4) , il s'agit dans un premier temps d'extraire des termes candidats. Pour cela, le composant **A** recherche dans l'index de Yahoo!Search les passages des documents contenant un des marqueurs d'intensité de l'ensemble Δ_{int} décrit précédemment. Chaque passage est analysé morpho-syntaxiquement (**B**) ce qui permet d'extraire le terme (nom, groupe nominal, verbe, groupe verbal) qui suit le marqueur d'intensité. On obtient ainsi une première liste brute de termes candidats (**C**). Chacun des termes candidats est ensuite associé à une liste de tests sémantiques envoyés automatiquement à Yahoo!Search (**D**) sous forme de requêtes. Le nombre d'occurrences de chaque requête (**E**) sert de descripteur pour catégoriser le terme candidat avec l'algorithme de décision SVM (**F**). Les termes classés *subjectifs* sont insérés dans le lexique de l'évaluation en ajoutant une information indiquant qu'ils ont été ajoutés automatiquement.

Pour les adjectifs, la démarche est analogue à deux nuances près : la première liste brute d'adjectifs est extraite à partir d'un corpus de 5 000 billets. 13 094 adjectifs inconnus sont ainsi extraits. Chacun de ces adjectifs est ensuite associé à son ensemble de tests sémantiques envoyés sous forme de requêtes à Yahoo!Search. De la même façon, le nombre d'occurrences de chaque requête sert de descripteur pour catégoriser le terme candidat. Pour chaque adjectif α inconnu, le calcul de la polarité axiologique est fondé sur le nombre d'occurrences de α comme terme négatif (neg) et sur le nombre d'occurrences de α comme terme positif (pos) : $axiol(\alpha) = \frac{pos}{total} - \frac{neg}{total}$. Une visualisation graphique sous forme de nuage de mots est générée pour chaque adjectif inconnu (Fig. 6 et Fig. 7) et permet d'observer les termes connus du lexique qui s'opposent à α et ceux qui s'associent à α selon les usages linguistiques.

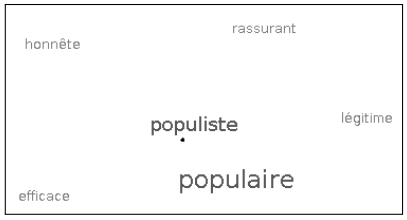

Figure 6. *Nuage de mots représentant les adjectifs connus du lexique qui s'**opposent** (polarité inverse) le plus fréquemment à l'adjectif* populiste *selon les tests sémantiques s_a et s_b*

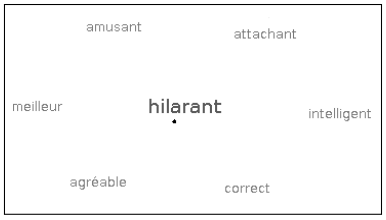

Figure 7. *Nuage de mots représentant les adjectifs connus du lexique qui s'**associent** (polarité identique) le plus fréquemment à l'adjectif* hilarant *selon les tests sémantiques s_c et s_d*

D'un point de vue technique, ces deux chaînes de traitements sont réalisées entièrement dans la plate-forme UIMA (Ferruci et Lally, 2004) dédiée à l'analyse de

textes par annotations successives. Elle permet de poser et de réutiliser des annotations sur un objet linguistique et de les échanger entre plusieurs composants dans un format normalisé (XMI). L'analyse morphosyntaxique dans UIMA est réalisée à partir d'une encapsulation du TreeTagger (Schmid, 1994). L'algorithme de décision SVM est implémenté à partir de la librairie Weka LibSVM[17].

5. Résultats

	Nombre	**Exemples** (sélectionnés aléatoirement)
Adjectifs	596	larmoyant, exhorbitant, opiniâtre, lunatique, incestueux, cocace, famélique, infantile, subversif, polluant
Noms ou noms composés	1 390	régal, fléau, plébiscite, camouflet, marée humaine, descente aux enfers, gain de temps, cacaphonie, bouffée d'air frais, capharnaüm
Verbes ou expressions verbales	488	jouer un rôle décisif, faire basculer le match, subjuguer, voler la vedette, marquer un tournant, toucher le fond, porter son équipe, trouver ses marques, inonder le marché, ovationner

Tableau 5. *Termes ajoutés au lexique (exemples aléatoires).*

La méthode exposée précédemment permet d'extraire 2 474 nouveaux termes (Tab. 5) et de les considérer comme des candidats suffisament pertinents pour être ajoutés au lexique de l'évaluation initiallement constitué de 982 entrées, essentiellement des termes simples. Seuls les 596 adjectifs ont été à ce stade catégorisés sur leur polarité axiologique (24,1% des nouveaux termes subjectifs). Nous cherchons à évaluer la méthode présentée sur deux aspects :

– la catégorisation de nouveaux termes *subjectifs* ;

– la catégorisation de la *polarité axiologique* des nouveaux adjectifs.

Évaluer directement l'enrichissement du lexique nécessiterait une ressource de référence, or c'est l'absence d'une telle ressource qui fait l'objet de ce travail. De plus, nous avons précédemment souligné l'importance du contexte pour déterminer si un terme est oui ou non subjectif. Ceci nous amène à proposer un protocole d'évaluation différent qui permet d'observer l'impact de l'enrichissement lexical sur un problème applicatif concret. Nous nous démarquons donc des travaux de référence du domaine (Turney, 2002) pour lesquels l'enrichissement du lexique est évalué hors contexte à partir de la ressource General Inquirer. Ce dernier protocole d'évaluation nous semble comporter un biais puisqu'il considère toujours correct les termes subjectifs qui peuvent être employés dans un contexte objectif non évaluatif.

17. http://www.cs.iastate.edu/ yasser/wlsvm/

Nous avons ainsi extrait 5 000 nouveaux billets et commentaires de blogs de la plate-forme Over-blog sans contrainte sur leur catégorie thématique ou sur leur taille. Nous utilisons l'outil Apopsis (Vernier *et al.*, 2009b) pour annoter les passages évaluatifs . En sortie de la chaîne de traitements, deux fichiers (format CSV) sont générés pour lister d'une part, les évaluations détectées à partir du lexique initial et d'autre part, les évaluations détectées à partir de la partie enrichie du lexique. La polarité axiologique du passage évaluatif et son contexte phrastique sont également renseignés selon le format CSV suivant : **passage évaluatif** ; polarité axiol. du passage ; *contexte (phrase)*.

– **une politique énergétique ambitieuse** ; positif ; *Ségolène Royal a d'ailleurs proposé de mettre en place une politique énergétique ambitieuse [...].*

5.1. *Accord interannotateurs*

Deux juges humains se sont chargés d'évaluer la précision[18] des passages évaluatifs annotés grâce à la partie enrichie du lexique. Le fichier CSV correspondant est divisé en cinq échantillons (A, B, C, D et E) répartis entre le juge humain 1 (A, B, C) et le juge humain 2 (C, D et E).

	Juge 2 : Correct	Juge 2 : Erreur	Total
Juge 1 : Correct	1 329	110	1 439
Juge 1 : Erreur	361	164	525
Total	1 690	274	1 964

Tableau 6. *Tableau de contingence récapitulatif de l'accord observé (0,76 %)*

Nous cherchons à évaluer l'accord inter-annotateur à partir de 1 964 passages évaluatifs de l'échantillon C vérifiés par les juges humains 1 et 2, ce qui représente un peu plus de 10 % du nombre total des passages évaluatifs (Tab. 7) . Nous constatons (Tab. 6) que les deux juges détectent 8,1 [19] fois plus d'annotations de passages subjectifs correctes que d'annotations erronées. De ce fait, comme l'ont montré (Feinstein et Cicchetti, 1990), le déséquilibre entre les deux catégories de données (prévalence) rend inadéquat le test Kappa de Cohen traditionnellement utilisé pour mesurer l'accord interannotateurs. D'après les données, nous pouvons uniquement conclure que les juges humains ont un taux d'accord observé de 76 % et jugent correctes, ensemble, 67 % des annotations. Le tableau 7 récapitule les évaluations des juges humains, il permet d'observer que les erreurs et les désaccords portent essentiellement sur les noms et noms composés. Ceux-ci sont plus enclins à déclencher des stéréotypes culturels différents chez les juges humains. Ainsi, beaucoup d'exemples contenant les noms *crise économique*, *politique écologique*, *terrorisme* ou *pandémie* entraînent des désaccords chez les juges humains.

18. Nombre d'annotations correctes divisé par le nombre total d'annotations.
19. (1 329 corrects divisés par 164 erreurs).

– La pandémie de grippe, réelle ou inventée, permet de mettre en scène le final [...]

– Le mot pandémie est d'actualité, nous l'entendons même depuis des mois.

Les cas de passages subjectifs erronés sont causés par des termes (en particulier des noms) dont l'usage subjectif existe bien mais sont dépendants de leur contexte (*bijou, farce, daube, rafale, sanction, pollution*).

Échantillon	A	B	C	D	E	Total
Verbes/GV	329	600	317	270	616	2 132
Erreur	90	82	29	42	87	330
Correct	235	515	288	226	529	1 793
Non évalué	4	3	0	2	0	9
PRÉCISION	**72,3 %**	**86,3 %**	**90,8 %**	**83,7 %**	**85,8 %**	**84,5 %**
Noms/GN	1 548	3 131	1 637	1 307	3 612	11 235
Erreur	515	1 161	261	218	869	3 024
Correct	858	1 785	1 376	1 044	2 594	7 657
Non évalué	175	185	52	45	149	606
PRÉCISION	**62,5 %**	**60,6 %**	**87,6 %**	**82,7 %**	**74,9 %**	**72,0 %**
Adjectifs	611	1 157	610	499	1 373	4 250
Erreur	18	77	28	34	73	230
Correct	588	1 069	582	464	1 297	4 000
Non évalué	5	11	0	1	3	20
PRÉCISION	**97,0 %**	**93,3 %**	**95,4 %**	**93,2 %**	**94,7 %**	**94,6 %**
Polarité des adjectifs	588	1 069	582	464	1 297	4 000
Erreur	48	100	19	32	54	253
Correct	540	969	563	432	1 243	3 747
PRÉCISION	**91,8 %**	**90,7 %**	**96,8 %**	**93,1 %**	**95,8 %**	**93,7 %**

Tableau 7. *Précision obtenue pour la détection des passages évaluatifs à l'aide du lexique enrichi selon deux juges humains*

5.2. *Protocole d'évaluation quantitative et qualitative*

Comme évoqué par les organisateurs de DEFT'09, l'absence de corpus de référence pour évaluer des tâches telles que la détection locale de la subjectivité est un problème non résolu actuellement. Il empêche la mesure de rappel[20] puisqu'elle nécessiterait une lecture exhaustive de l'ensemble du corpus par les juges humains. Néanmoins, nous proposons un aperçu de l'amélioration quantitative des passages annotés en comparant le nombre de passages annotés à partir de la ressource initiale, le nombre de passages annotés par la partie enrichie automatiquement et le nombre de passages annotés catégorisés comme corrects par les juges humains.

20. Nombre d'annotations correctes divisé par le nombre d'annotations attendues.

Nombre de passages annotés	Valeur absolue	Pourcentage
avec ressource initiale	68 536	78,8 %
avec ressource enrichie	17 669	21,2 %
avec ressource enrichie (correcte)	13 450	15,6 %
Total	83 204	100 %

Tableau 8. *Évaluation quantitative de l'enrichissement du lexique de l'évaluation*

D'un point de vue qualitatif, sur les 17 669 évaluations détectées grâce au lexique enrichi, 13 450 ont été considérées correctes. La précision est donc de 76,12 %, ce qui correspond, avec une légère baisse, aux précisions obtenues dans les tests effectués avec le lexique initial : (Vernier *et al.*, 2009a ; Vernier *et al.*, 2009b).

6. Discussions et perspectives

D'un point de vue quantitatif, la taille du lexique de l'évaluation est passée de 982 entrées à 3 456 entrées (+ 252 %). Une première remarque concerne l'augmentation de la détection des évaluations de seulement 20 % en comparaison. Toutefois, cet enrichissement est loin de ne pas être significatif pour la raison suivante : par rapport aux termes du lexique de l'évaluation constitué manuellement (*beau, inquiétude, aimer*), les termes appris (*blasphématoire, la politique de l'autruche, faire tordre de rire*) ont une fréquence d'apparition plus faible, expliquant d'ailleurs qu'ils soient oubliés lors de la constitution manuelle de ressources pour la fouille d'opinions. En revanche, leur rareté rend leur usage particulièrement porteur d'informations et leur détection automatique est de ce fait intéressante et nécessaire pour des problèmes applicatifs réels. Dans l'exemple ci-dessous, l'enrichissement lexical permet ainsi de saisir l'opinion qui fait ici véritablement sens dans le discours (*la spoliation du peuple*) à partir de termes complexes (*spoilier le peuple*, *payer le prix fort*), là où une ressource *classique* n'aurait cerné que le *bonheur* (avec éventuellement sa tournure négative).

> *La guerre actuelle du Congo a* ***spolié le peuple***[+lexique enrichi] *congolais de ses moyens de subsistance. Pour une guerre dont la finalité est* ***loin de faire son bonheur***[+lexique initial]*, le peuple congolais* ***paie le prix fort***[+lexique enrichi].

Qualitativement, les résultats obtenus sont assez proches des résultats des méthodes d'apprentissage de termes subjectifs de Turney (2002) et Bestgen (2002). Notre approche se distingue par l'étape de présélection des termes candidats par des requêtes linguistiquement motivées diminuant l'ajout de bruits dans la ressource. Cette présélection nous permet d'obtenir un ensemble de mots, mais également des expressions qui ne sont habituellement pas traitées par les approches du domaine. En revanche, mis à part les adjectifs, nous ne mesurons pas à ce stade leur polarité axiologique. De ce point de vue, les méthodes de (Turney, 2002) et (Bestgen, 2002)

sont complémentaires avec celle que nous proposons et sont une de nos perspectives pour finaliser l'enrichissement du lexique. Les adjectifs ajoutés, et à un degré moindre les groupes verbaux, permettent en contexte une détection des évaluations précise (la précision varie de 71,5 % à 97,0 %). L'ambiguïté et les sources d'erreurs proviennent essentiellement des noms et groupes nominaux en raison du caractère davantage polysémique des noms. Ainsi, les noms *farce* ou *daube* ont bien un usage subjectif fréquent (*c'est une farce cette assemblée de politiciens*, *front page une vrai daube*) mais *occurrent* également dans des contextes complètement objectifs (*bien mélanger la farce*, *voici ma recette de daube de poisson*). Comme le précise Kerbrat-Orecchioni (1997) les termes subjectifs ne sont pas stables dans le langage (en témoigne les termes *collaboration* ou *collaborer* qui peuvent dénoter un stéréotype culturel négatif dans le contexte de la seconde guerre mondiale mais qui ont aujourd'hui un sens beaucoup plus neutre voire positif). En se fondant sur les usages linguistiques des internautes, la méthode d'apprentissage que nous proposons acquiert les stéréotypes linguistiques les plus socialement admis (*pantouflard*, *négationniste*, *escroquerie*), mais aussi ceux qui sont majoritaires au moment de l'apprentissage (les termes *écologie, écologique, pollution* sont ainsi classés comme étant subjectifs de par leur intensité dans le contexte et les discours actuels). Ceci parfois au prix d'erreurs malheureuses (l'apprentissage automatique détecte notamment l'adjectif *arabe* comme étant statistiquement subjectif et porteur d'un stéréotype culturel négatif d'après l'usage linguistique des internautes) qui sont elles-mêmes sources d'erreurs de catégorisation lors d'une évaluation en contexte.

La question du protocole d'évaluation pour mesurer la pertinence d'une ressource lexico-sémantique du langage de la subjectivité soulève un problème épistémologique. Par définition, il n'y a pas de référence stable à laquelle se comparer puisque les marqueurs de subjectivité sont instables et varient selon les configurations d'énonciations et les stéréotypes culturels. Nous pensons que, pour cette tâche, la seule façon d'approximer un protocole évaluatif correct est d'observer en contexte, sur des échantillons, le comportement de la ressource développée. La répétition des campagnes d'évaluations, en particulier en fouille d'opinions (à l'image des éditions de DEFT et TREC) est sans doute le meilleur cadre pour élaborer des ressources les plus optimales possibles (corpus annoté, lexique de l'évaluation générique, lexique de l'évaluation par domaine thématique, par type de corpus (blogs, journaux, forums), etc.) en confrontant les outils automatiques entre eux ainsi qu'avec les expertises humaines et en observant les cas de désaccords et d'erreurs. Pour les besoins de ce travail, nous avons mobilisé deux juges humains comme point de comparaison, mais il s'agit là d'un travail particulièrement coûteux et qui ne peut être répété que ponctuellement et collaborativement pour améliorer son efficacité.

En conclusion, nous avons présenté une méthode pour enrichir une ressource sur le langage évaluatif désormais constituée de 3 456 entrées (les ressources existantes pour le français jusqu'à présent contenaient environ 1 000 entrées). Elle permet de détecter davantage de passages évaluatifs sans faire chuter notablement la précision sur les échantillons évalués manuellement. Cette méthode s'appuie sur l'indexation des usages linguistiques des internautes par un moteur de recherche et par un grand

nombre de requêtes dont les résultats sont à la base d'un algorithme de classification standard (SVM). La classification permet de discriminer 2 474 nouveaux termes subjectifs (mots ou expressions). La nature des termes appris est intéressante de par leur rareté et l'information qu'ils révèlent et du fait qu'ils correspondent à des stéréotypes culturels dont la valeur subjective ne relève pas de leur noyau sémantique mais de leur usage courant. La méthode proposée se distingue des autres en permettant de saisir à un instant donné, par des contraintes linguistiques, le degré de subjectivité de mots et d'expressions, ce qui est une particularité importante de ce domaine puisque les marqueurs de subjectivité sont particulièrement instables.

7. Bibliographie

Anscombre J., Ducrot O., *L'argumentation dans la langue*, Pierre Mardag, 1983.

Aue A., Gamon M., « Customizing Sentiment Classifiers to New Domains: A Case Study », *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, 2005.

Banea C., Mihalcea R., Wiebe J., « A Bootstrapping Method for Building Subjectivity Lexicons for Languages with Scarce Resources », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, may, 2008. http://www.lrec-conf.org/proceedings/lrec2008/.

Benveniste E., *Problèmes de linguistique générale II*, Gallimard, 1974.

Bestgen Y., « Détermination de la valence affective de termes dans de grands corpus de textes », *Actes du Colloque International sur la Fouille de Texte*, p. 81-94, 2002.

Charaudeau P., *Grammaire du sens et de l'expression*, Hachette Education, COMMUNICATION, PARA UNIVERSITAIRE, 1992.

Conrad J. G., Schilder F., « Opinion mining in legal blogs », *ICAIL '07: Proceedings of the 11th international conference on Artificial intelligence and law*, ACM, New York, NY, USA, p. 231-236, 2007.

Dubreil E., Vernier M., Monceaux L., Daille B., « Annotating Opinion - Evaluation Of Blogs », *Workshop on LREC 2008 Conference, Sentiment Analysis: Metaphor, Ontology and Terminology (EMOT-08)*, 2008.

Esuli A., Sebastiani F., « SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining », *Proceedings of Language Resources and Evaluation (LREC)*, 2006.

Feinstein A., Cicchetti D., « High agreement but low kappa : I. The problems of Two Paradoxes », *J. Clin. Epidemiol.*, p. 543-548, 1990.

Ferruci D., Lally A., « Uima : an architectural approach to unstructured information processing in the corporate research environment. », *Natural Language Engineering,10(3-4)*, p. 327-348, 2004.

Fleiss J. L., « Measuring nominal scale agreement among many raters », *Psychological Bulletin, Vol. 76, No.5*, p. 378-382, 1971.

Galatanu O., « Signification, sens, formation », *Education et Formation, Biennales de l'éducation, (sous la direction de Jean-Marie Barbier, d'Olga Galatanu)*, PUF, Paris, 2000.

Hatzivassiloglou V., McKeown K., « Predicting the Semantic Orientation of Adjectives », *Proceedings of the Joint ACL/EACL Conference*, p. 174-181, 1997.

Joachims T., « Text Categorization With Support Vector Machines: Learning with many relevant features », *Rapport Interne Ls8-Report 23*, Universität Dortmund, 1997.

Josef Ruppenhofer S. S., Wiebe J., « Finding the Sources and Targets of Subjective Expressions », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, may, 2008. http://www.lrec-conf.org/proceedings/lrec2008/.

Kerbrat-Orecchioni C., *L'Énonciation, de la subjectivité dans le langage*, Colin (réédition 2002), 1997.

Lavelle L., *Traité des valeurs*, vol. tome 1, PUF, 1950.

Legallois D., « Pour une définition énonciative de l'enclosure *vrai* », *Les marqueurs linguistiques de la présence de l'auteur*, David Banks, L'Harmattan, 2005.

Liu B., « Sentiment Analysis », *Invited talk at the 5th Annual Text Analytics Summit*, 2009.

Martin J., White P., *The Language of Evaluation, Appraisal in English*, Palgrave Macmillan, London, New York, 2005.

Mathieu Y. Y., « Annotation of Emotions and Feelings in Texts », *ACII*, p. 350-357, 2005.

Mishne G., Glance N., « Predicting Movie Sales from Blogger Sentiment », *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, p. 155-158, 2006.

Pang B., Lee L., Vaithyanathan S., « Thumbs up? Sentiment Classification using Machine Learning Techniques », *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 79-86, 2002.

Schmid H., « Probabilistic Part-of-Speech Tagging Using Decision Trees », *Proceedings of the International Conference on New Methods in Language Processing*, p. 44-49, 1994.

Stone P. J., *The General Inquirer: A Computer Approach to Content Analysis*, The MIT Press, 1966.

Stoyanov V., Cardie C., « Annotating Topics of Opinions », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, may, 2008. http://www.lrec-conf.org/proceedings/lrec2008/.

Strapparava C., Valitutti A., « WordNet-Affect: an affective extension of WordNet », *Proceedings of LREC*, vol. 4, p. 1083-1086, 2004.

Suhamy H., « Métaphore et dualité », *Bulletin de la Société de stylistique anglaise (ISSN 0240-4273), numéro 28*, p. 9-24, 2006.

Turney P., « Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews », *Proceedings of the Association for Computational Linguistics (ACL)*, p. 417-424, 2002.

Vergne J., Giguet E., « Regards Théoriques sur le «Tagging» », *Actes de Traitement Automatique des Langues Naturelles (TALN'98)*, p. 22-31, 1998.

Vernier M., Monceaux L., Daille B., « DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique », *Actes de l'atelier de clôture du cinquième Défi Fouille de Textes (DEFT'09)*, Paris, France, p. 101-112, June, 2009a.

Vernier M., Monceaux L., Daille B., Dubreil E., « Catégorisation des évaluations dans un corpus de blogs multi-domaine », *Numéro spécial de la revue RNTI (Revue des Nouvelles Technologies de l'Information) - fouille de données d'opinion*, p. 45-70, 2009b.