# Exploration de corpus pour l'analyse de sentiments

Sigrid Maurel[(1)] et Luca Dini[(1)]

[(1)] CELI France, SAS
12-14, rue Claude Genin
38000 Grenoble
{maurel, dini}@celi-france.com
http://www.celi-france.com

## Résumé – Abstract

Dans cet article nous présentons l'amélioration de notre système d'extraction de sentiments et opinions SYBILLE. Par rapport à la première version nous avons changé la méthode statistique (initialement basée sur l'apprentissage automatique) vers une exploration ontologique de corpus. Cette nouvelle méthode qui s'insère avant la méthode symbolique permet de survoler les textes et d'en avoir très rapidement un premier aperçu. Elle extrait les concepts des domaines présents dans les textes et en fournissant une ontologie elle facilite le développement de la grammaire de la méthode symbolique.

In this article we present the improvement of our sentiment and opinion mining system SYBILLE. Compared to the first version we changed the statistic method (formerly based on machine learning) to an ontologic corpus discovery. This new method which comes before the symbolic method allows to skim the texts and to get very quickly a first glance. It extracts the concepts of the present domains of the texts and by giving an ontology facilitates the development of the grammar of the symbolic method.

## Mots-clefs – Keywords

extraction de sentiments et opinions, exploration de corpus

sentiment and opinion mining, corpus discovery

# 1 Introduction

## 1.1 Motivation

Cet article s'intéresse à la classification de textes d'opinion en langue française. Dans ce cas précis, la classification a pour objectif l'analyse de sentiments exprimés dans différents types de textes comme par exemple dans des forums de discussion sur Internet où les internautes échangent des avis et s'entraident. Les textes issus de forums sur Internet constituent des sources d'informations spontanées et récentes, incontournables pour acquérir, au jour le jour, des connaissances sur les consommateurs, pour anticiper leurs besoins et leurs attentes afin de tenter d'améliorer la relation client/fournisseur. En analysant ces textes d'opinion le fournisseur d'un produit ou d'un service peut mieux réagir aux desiderata de ses clients, le client peut de son côté s'inspirer des sentiments et opinions d'autres clients sur le produit auquel il s'intéresse et profiter ainsi d'une aide à la décision (acquérir ou ne pas acquérir le produit, choisir plutôt le produit A ou le produit B, etc.).

Comme le montrent de nombreux travaux de socio- et psycho-linguistique (Sproull & Kiesler, 1991), la communication médiée par ordinateur favorise l'expression des émotions, sentiments et opinions souvent contrôlés ou réprimés dans des cadres de communication plus traditionnels visant à étudier le point de vue des consommateurs (interviews face à face, enquêtes fermées, enquêtes ouvertes, etc.). De là, naît l'intérêt des analystes pour ces sources d'informations.

Les corpus utilisés pour le développement des systèmes de classification sont composés de textes (ou *threads*, fils de discussion) provenant de forums sur Internet qui parlent entre autres de tourisme, de jeux vidéo et d'imprimantes. Un texte (ou message) dans un forum contient un jugement argumenté de l'auteur du message, positif, négatif ou parfois mitigé, sur un sujet donné. Mais il contient aussi des parties exemptes de sentiments, comme c'est le cas par exemple dans la description du jeu vidéo sur lequel porte la critique. L'objectif de l'analyse est donc d'identifier avec précision les parties pertinentes pour la classification automatique du texte dans son entier.

Une des difficultés de la classification en *positif* et *négatif* réside dans la nécessité d'une bonne analyse syntaxique du texte, analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase, d'anaphore ou de coréférence (la reprise d'un argument présent plus tôt dans le document). Une autre difficulté du langage naturel pour l'analyse automatique de sentiments réside dans les contextes intentionnels, pour lesquels l'expression d'opinion n'est pas un vrai sentiment. C'est le cas dans une phrase comme:

> « Je croyais que la France était un beau pays. »

(Dini & Mazzini, 2002) ont montré le lien qui existe entre les structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion qu'elle véhicule. Ainsi l'analyse de la phrase par *paquets de mots* donne des résultats peu satisfaisants alors qu'une analyse syntaxique du texte peut aider à trouver les expressions qui contiennent des opinions. Les deux phrases suivantes contiennent les mêmes *paquets de mots* sans pour autant exprimer les mêmes sentiments. En effet, la première phrase contient un sentiment positif alors que la deuxième est négative:

> « Je l'ai apprécié pas seulement à cause de ... »
> « Je l'ai pas apprécié seulement à cause de ... »

Dans le cadre de sa participation à la campagne d'évaluation DEFT'07 (c.f. section 6 pour plus de détails), CELI France a mis au point trois méthodes pour classer les textes des différents corpus. La première est une méthode symbolique qui inclut un système d'extraction d'information adapté aux corpus. Elle est basée sur des règles d'un analyseur syntaxico-sémantique. Cet analyseur contient un lexique de mots qui véhiculent des sentiments sur lesquels réagissent les règles de la grammaire. La deuxième est une méthode statistique basée sur des techniques d'apprentissage automatique. Enfin, la dernière, SYBILLE, est une méthode hybride qui combine les techniques des deux précédentes pour aboutir à des résultats très précis.

Depuis cette première évaluation CELI France a amélioré son système SYBILLE. Nous avons changé la méthode statistique initialement basée sur un apprentissage automatique (Maurel *et al.*, 2009) par une exploration ontologique de corpus qui donne très vite une première idée sur les concepts véhiculés dans les textes. De ce fait elle nous permet d'orienter de manière très précise le développement de la méthode symbolique qui passe à la deuxième place lors du processus. La méthode hybride reste inchangée et combine les résultats des deux premières méthodes.

L'analyse des textes se fait au niveau de la phrase, les sentiments d'un document sont extraits phrase par phrase, et c'est seulement ensuite qu'une valeur globale est attribuée au message entier. Ceci permet d'extraire une information contextuelle qui est donc très précise.

Les sections suivantes présentent brièvement l'état de l'art et les corpus utilisés pour s'attarder ensuite sur les trois méthodes développées et en fournir une première évaluation. Une section sera dédiée à l'interface graphique SYBILLE pour présenter les possibilités que celle-ci offre aux utilisateurs, avant de conclure cet article.

## 1.2 État de l'art

L'analyse de sentiments se concentre aujourd'hui sur l'attribution d'une polarité à des expressions subjectives (les mots et les phrases qui expriment des opinions, des émotions, des sentiments, etc.) afin de décider de l'orientation d'un document (Turney, 2002), (Wilson *et al.*, 2004) ou de la valeur positive/négative/neutre d'une opinion dans un document (Hatzivassiloglou & McKeown, 1997), (Yu & Hatzivassiloglou, 2003), (Kim & Hovy, 2004).

Des travaux allant au-delà ont mis l'accent sur la force d'une opinion exprimée où chaque proposition dans une phrase peut avoir un fond neutre, faible, moyen ou élevé (Wilson *et al.*, 2004). Des catégories grammaticales ont été utilisées pour l'analyse de sentiments dans (Bethard *et al.*, 2004) où des syntagmes adjectivaux comme *trop riche* ont été utilisés afin d'extraire des opinions véhiculant des sentiments. (Bethard *et al.*, 2004) utilisent une évaluation basée sur la somme des scores des adjectifs et des adverbes classés manuellement, tandis que

(Chklovski, 2006) utilise des méthodes fondées sur un modèle pour représenter des expressions adverbiales de degré telles que *parfois*, *beaucoup*, *assez* ou *très fort*.

L'approche que nous avons adoptée pour la classification de textes d'opinion est caractérisée par une utilisation mixte d'une technologie symbolique fondée sur des règles et d'une technologie statistique reposant sur l'extraction d'une ontologie du domaine, approche dans laquelle la méthode symbolique a un poids plus important (Dini, 2002), (Dini & Mazzini, 2002), (Maurel *et al.*, 2007), (Maurel *et al.*, 2008), (Bosca & Dini, 2009). La technologie symbolique fait d'abord une analyse du texte phrase par phrase et en extrait ensuite les relations qui véhiculent des sentiments, tandis que la technologie statistique traite les textes en une seule phase et fournit une liste de concepts et termes spécifiques du texte.

Il convient de remarquer que, contrairement à d'autres approches actuelles, la technologie de l'analyse de sentiments développée à CELI France (SYBILLE) ne se limite pas à une analyse lexicale (c'est-à-dire identification et pondération de mots positifs et négatifs), mais s'étend à une analyse syntaxique et sémantique. L'analyse syntaxique est effectuée par le biais d'une analyse robuste de surface telle que celles décrites par (Aït-Mokhtar & Chanod, 1997), (Basili *et al.*, 1999), (Aït-Mokhtar *et al.*, 2001), donnant ainsi un résultat très proche de celui produit par des grammaires de dépendance.

## 2 Les corpus

Les données de type forums de discussion sur Internet s'articulent comme un flux d'interactions, comme par exemple: demande-réponse, argument-contre argument, commentaire-désaccord, etc. Ce flux est distribué sur une dimension temporelle qui nécessite un traitement chronologique du fil de discussion. Contrairement aux corpus utilisés par (Wilson *et al.*, 2004), il n'est pas nécessaire ici d'identifier la personne à qui est associé un sentiment, car dans 95 % des cas, les discours analysés sont des discours à la première personne. Un exemple de flux d'interactions est donné en figure 1.

Les corpus utilisés sont assez différents les uns des autres, que ce soit par la taille des corpus eux-mêmes que par la taille de chaque *thread* (fil de discussion). Nous avons utilisé les corpus de DEFT'07 auxquels nous avons ajouté des corpus collectés sur Internet. Ces corpus nous ont permis d'augmenter la diversité des sujets et de répondre aux exigences de nos clients, selon les domaines demandés. Nous avons donc des textes de domaines très différents, entre autres du domaine de la restauration rapide, du nucléaire, de l'alimentation infantile, etc.[1] Certains corpus sont structurés, d'autres contiennent beaucoup de messages en style *texto*, et le nombre de fautes d'orthographe présentes dans les messages varie aussi beaucoup.

Le comité d'organisation de DEFT'07 a pris soin de nettoyer ses corpus (Grouin *et al.*, 2007). Ainsi, les fins de ligne ont été normalisées, les caractères encodés en ISO-Latin, et les textes ont été annotés manuellement.[2] Les corpus de DEFT'07 contiennent des critiques de films, de livres et de spectacles, des tests de jeux vidéo, des relectures d'articles scientifiques (de différentes conférences sur l'intelligence artificielle) et des notes de débats parlementaires (sur la loi de l'énergie).

Les textes de nos corpus portent essentiellement sur le tourisme (en France et ailleurs dans le monde), les jeux vidéo (critiques et problèmes) et les imprimantes (conseils d'achats). Ils comprennent d'un côté des aides à la solution de problèmes, mais aussi des avis sur des lieux visités et des produits achetés. Chaque *thread* contient les messages des auteurs participant aux forums sur un sujet donné.

Les fautes d'orthographe[3] dans les textes des corpus posent parfois des problèmes d'analyse. Heureusement, les règles syntaxiques de la grammaire (voir la section 4 sur la méthode symbolique) sont dans la plupart des cas assez tolérantes pour permettre l'accord entre un nom et un adjectif, ou un nom et un verbe même si le *e* ou le *s* manque. Mais malheureusement, il y a aussi des messages tellement mal écrits dans les corpus (par exemple en style *texto*, c'est-à-dire avec beaucoup d'abréviations) que l'analyse peut échouer.[4]

[1] Ces derniers ne seront pas abordés plus profondément dans cet article, mais les ressources sont disponibles dans notre système SYBILLE.

[2] En ce qui concerne nos propres corpus, ils sont encodés en UTF-8 et nous n'avons effectué aucun nettoyage. Tous les corpus sont disponibles au format XML.

[3] Nous avons fait le choix de garder les textes tels quels, donc de ne pas appliquer un correcteur automatique d'orthographe ou un lexique d'abréviations. Ce choix s'explique par la volonté de garder toutes les caractéristiques stylistiques présentes dans les textes. Nous considérons qu'une uniformisation des entrées à ce moment-là du processus nous ferait perdre des informations utiles.

[4] D'après ce que nous avons pu observer, ces messages sont heureusement en minorité dans les corpus et ne modifient pas les résultats de façon significative.

```
    Avis sur les châteaux de la loire en France

angie-443*5, posté le 08-10-2006 à 16:18:50:
   J'ai besoin de vos conseil s.v.p. Je vais passer une ou deux
   journée dans la valée de la loire. Y-a-t-il un château en
   loire avec un jardin semblable à celui de versailles (en
   beauté et en superficie)? J'aime aussi l'aspect extérieurs
   des châteaux, plus que l'intérieur. Ce qui me plait d'une
   ville est tout d,abord ses rues piétonnes, animées et
   pittoresques, ses charmantes places et ses promenades.

[...]

BaLadeur, posté le 13-10-2006 à 11:23:43:
   Je partage l'avis d'Aston sur de nombreux points. Villandry
   est quelconque mais son jardin transformé en potager géant
   vaut le détour. Chenonceau est certainement le plus
   photogénique donc le plus connu et il le mérite largement.
   Si tu recherches la monumentalité comme à Versailles, la
   magnificence en plus, il faut absoluement voir Chambort.
   Enfin s'il faut ne visiter qu'une ville ce sera Tours.

[...]

zeus77, posté le 21-10-2006 à 21:59:33:
   A Amboise j'aime beaucoup le manoir du Clos-Lucé qui fut
   la dernière maison de Léonard de Vinci. Le parc est très
   agréable. Enfin un chateau où l'on pourrait vivre! Quel
   changement par rapport aux chateaux royaux. Un chateau que
   j'aime bien aussi c'est celui Du Moulin à Lassay sur Croisne
   entre Contres et Romorantin.

[...]
```

Figure 1: Exemple d'un flux d'interactions de messages, du domaine du *tourisme*. L'orthographe et la ponctuation n'ont pas été modifiées.

# 3 Méthode statistique distributionnelle

La méthode statistique distributionnelle (inspirée par (Bosca & Dini, 2009)) exploite le corpus afin d'identifier un échantillon de termes qui sont fortement distinctifs du domaine analysé. Les termes ainsi extraits sont ensuite mis en relation par moyens d'analyses statistiques d'occurences de mots à l'intérieur du corpus. L'issue résultante consiste en une représentation structurée (bien que pas une ontologie formelle) des concepts clés du domaine. Ce processus de découverte fonctionne en deux phases (détaillées ci-après), et est particulièrement efficace face à de grands corpus. Il permet d'avoir une idée du contenu et des mots-clés ou des concepts très rapidement.

Le processus qui analyse les textes fournit une moyenne d'exploration du corpus qui permet d'obtenir une sensation des sujets et des concepts discutés et quelles sont les relations entre ces concepts. Pour pouvoir configurer la grammaire de la méthode symbolique (c.f. la section 4) il est utile de savoir ne serait-ce que à peu près ce que l'on cherche dans les textes.

## 3.1 LOR

La méthode pour extraire une ontologie des textes a besoin de deux corpus. Le premier est le *corpus d'études* qui contient les textes d'un forum ou sous-forum avec les termes spécifiques qui nous intéressent particulièrement. Le deuxième est un corpus générique, un *corpus de référence* qui contient par exemple les textes de tous les sous-forums du forum en question, ou un ensemble de textes d'une source générale comme par exemple une encyclopédie. Dans le cas du *corpus d'études* de nos expériences il s'agit d'un corpus dynamique qui est généré à partir d'une requête de mots-clés sur une base de données contenant les textes.

| terme | score de pertinence | occurrences |
|---|---|---|
| laitage | 8.755508 | 912 |
| compote | 8.644531 | 2246 |
| biberon | 8.35303 | 9403 |
| blédina | 7.907746 | 354 |
| allaiter | 7.7495713 | 1992 |
| féculent | 7.651336 | 1227 |
| allaitement | 7.439835 | 1338 |
| bib | 7.280261 | 136 |
| candia | 7.0153956 | 1475 |
| diversification | 6.496531 | 1119 |

Table 1: LOR des termes pertinents du domaine de l'alimentation infantile.

Notre stratégie d'extraction de termes est basée sur la comparaison de fréquence entre le corpus du domaine (*corpus d'études*) et le corpus général (*corpus de référence*). Notre approche exploite comme mesure de termes spécifiques une version modifiée du bien connu *Log Odds Ratio* (LOR, c.f. (Everitt, 1992), (Baroni & Bisi, 2004)). La fonction de mesure de termes spécifiques adoptée dans nos expériences est une combinaison pondérée du LOR et de la mesure de fréquence de termes (TF). Elle peut être formalisée comme suit:

$$TermSpec = k * \frac{TermDF * GC_Docs}{TermGF * DC_Docs} + TermDF * (1 - k)$$

où $TermDF$ représente la fréquence d'un terme donné dans le corpus du domaine, $TermGF$ sa fréquence dans le corpus général, $DC_Docs$ le nombre de documents compris dans le corpus du domaine tandis que $GC_Docs$ est le nombre de documents dans le corpus général. Nous avons expérimenté l'extraction de terminologie avec trois valeurs différentes de $k$ (0, 0.5 et 1) ayant pour résultat donc trois fonctions de mesure différentes: une mesure de TF pure (avec $k = 0$), une mesure de LOR pure (avec $k = 1$) et une mesure équitablement ponderée de LOR/TF (avec $k = 0.5$); les paragraphes suivants décrivent en détail les différentes issues résultant de l'adoption de ces fonctions de mesure différentes.

Le tableau 1 donne un exemple du domaine de l'alimentation infantile. Il montre les dix premiers mots intéressants ou pertinents du corpus analysé. Le score attribué est une valeur de pertinence, ici le terme *laitage* est considéré plus pertinent que le terme *compote*, alors qu'il apparaît moins souvent.

## 3.2 RI

Les termes ainsi extraits (par LOR) du corpus du domaine sont enrichis avec une terminologie sémantiquement reliée par moyens d'un modèle distributionnel basé sur corpus. Une telle terminologie est basée sur l'hypothèse que le sens d'un terme donné émerge implicitement des contextes différents dans lesquels il apparaît (ici nous entendons par contexte l'unité de texte comme un paragraphe, un document ou une fenêtre textuelle). Puis, la deuxième phase du processus est une approche basée sur la co-ocurrence des mots, le sens d'un mot étant défini par son contexte. Cette méthode calcule donc un vecteur de sens pour chaque mot et plus les vecteurs de deux mots sont proches l'un de l'autre (plus l'angle entre eux est petit) plus leurs sens sont similaires.

L'indexage aléatoire *Random Indexing* (RI) exploite un modèle algébrique afin de représenter la sémantique des termes dans un espace à $N$ dimensions (un vecteur de $N$ coordonnées). L'approche RI crée une matrice *termes par contextes* où chaque ligne représente le degré d'appartenance d'un terme donné aux contextes différents. L'algorithme RI assigne une signature aléatoire à tous les contextes (un vecteur très épars de $N$ coordonnées, avec peu d'éléments non null, choisis aléatoirement) et génère ensuite le modèle de l'espace du vecteur en performant une analyse statistique des documents dans le corpus du domaine et en accumulant sur les lignes des termes toutes les signatures des contextes où les termes apparaissent.

Selon cette approche si deux termes différents ont un sens similaire ils devraient apparaître dans des contextes similaires (à l'intérieur d'un même document ou entourés des mêmes mots), en résultant caractérisés par des coordonnées proches dans l'espace sémantique ainsi généré. Dans nos études de cas nous avons appliqué la technique RI pour générer des clusters de termes en sélectionnant dans l'espace sémantique les termes avec la

| terme | score de pertinence | contexte |
|---|---|---|
| vache | 0.81468266 | lait de vache |
| tire | 0.7715641 | tire-lait |
| soja | 0.7615201 | lait de soja |
| poudre | 0.75552726 | lait en poudre |
| lactose | 0.75137496 | lait sans lactose |
| montée | 0.7460638 | montée de lait |
| chèvre | 0.6994075 | lait de chèvre |
| intolérance | 0.6894124 | intolérance au lait |
| biberon | 0.64999706 | biberon de lait |
| régurgiter | 0.63957196 | régurgiter le lait |

Table 2: RI des termes en contexte avec le mot *lait*.

distance minimale du mot analysé en exploitant la mesure de distance cosine.

Le tableau 2 donne un autre exemple du domaine de l'alimentation infantile. Il montre les dix premiers mots en contexte avec le mot *lait*.

### 3.3 Comparaison de LOR et RI sur corpus

Le premier pas de notre expérience découverte est de comprendre si nous pouvons produire une *photo instantanée* générale des contenus du corpus. Afin d'effectuer une telle tâche, les résultats de LOR appliqués sur la base de documents apparaissent quelque peu décevants. En effet, et la liste de termes comme LOR pure et la liste de LOR/TF pondérée (avec un poids de 0.5) semblent être plutôt orientées à mettre en évidence des termes imprévus que des termes descriptifs pertinents.

Nous notons bien sûr l'apparition de quelques termes qui sont probables à caractériser le domaine en question ou les opinions que les auteurs des textes peuvent en avoir, mais la tendance est occultée par les termes qui sont inattendus et probablement arrivés par des discussions hors-sujet. Afin de minimiser l'impact d'hors-sujet et de bruit venant d'analyses peu structurées de pages web, nous restreignons l'algorithme LOR uniquement à des phrases qui contiennent un mot-clé préalablement choisi.

Une fois que nous avons isolé un ensemble de concepts qui constitue le pivot de notre étude, nous pouvons enquêter sur les comportements différents des deux algorithmes. En même temps nous pouvons évaluer l'effet de phénomènes linguistiques comme l'ambiguïté sémantique et la partie syntaxique de notre méthodologie proposée.

## 4 Méthode symbolique

Une fois que le processus statistique est terminé et les connaissances du corpus acquises par l'exploration, le développement de la grammaire symbolique peut se baser dessus pour améliorer les règles symboliques.

Comme nous l'avons dit plus haut, la méthode symbolique se base sur une analyse syntaxique du texte faite par un analyseur fonctionnel et relationnel (c.f. les travaux sur l'analyse syntaxique et sémantique de (Basili *et al.*, 1999), (Aït-Mokhtar *et al.*, 2001), (Dini, 2002), (Dini & Mazzini, 2002), (Dini & Segond, 2007)). Cet analyseur traite, phrase par phrase, un texte donné en entrée et en extrait, pour chaque phrase, les relations syntaxiques présentes. Il s'agit de relations syntaxiques fonctionnelles de base, telles que le modifieur d'un nom, d'un verbe, sujet et objet d'une phrase, ainsi que de relations plus complexes telles que la coréférence entre deux syntagmes au sein d'une même phrase.

L'utilisateur a la possibilité d'élaborer une grammaire à sa guise et d'ajouter de nouvelles règles afin d'extraire les relations auxquelles il s'intéresse. Pour ce faire, il peut modifier les règles d'extraction de relations (par exemple ajouter des règles pour de nouvelles relations), augmenter/diminuer les traits sur les mots dans le lexique qui agissent sur les règles, enlever certaines parties du traitement, etc.

La polarité positive ou négative attribuée au message entier[5] dépend du rapport entre la quantité de relations

[5]L'attribution d'un sentiment global au message entier est utilisée dans des contextes spécifiques, comme par exemple pour l'évaluation

d'opinions positives et négatives. Une majorité de relations d'opinions positives détermine une polarité positive du message, tandis qu'une majorité de relations d'opinions négatives provoque une polarité négative.

## 4.1 Grammaire

La grammaire utilisée a été initialement développée afin d'extraire les relations de sentiments exprimés dans une phrase dans le cadre d'un projet sur le tourisme en France. Elle a été ensuite modifiée et améliorée en vue de la participation à DEFT'07 (c.f. section 6, (Maurel *et al.*, 2007)). Dans un deuxième temps, la grammaire a été divisée en deux parties: une première partie de base (la grammaire *générique*) s'appliquant à tous les textes qui contiennent des sentiments, et une deuxième partie pour chaque domaine différent, selon le sujet du corpus: tourisme, jeux vidéo, imprimantes, etc. Les différences se situent essentiellement dans les lexiques appliqués, chaque domaine ayant ses propres mots et expressions.

Ainsi les mots se rattachant à la vitesse (*lent*, *rapide*, etc.) ont des polarités différentes selon qu'ils qualifient une imprimante ou un voyage. De même, comme le montrent les phrases ci-dessous, l'adjectif *effrayant* est plutôt perçu comme positif dans une description romanesque alors qu'il est perçu comme négatif dans le domaine des assurances ou du tourisme:

> « Dans *Ghost*, les habitants du village sont vraiment effrayants! »
> « C'est effrayant de voir comment la côte est de plus en plus bétonnée. »

En général, une relation de sentiment a deux arguments: le premier est l'expression linguistique qui véhicule le sentiment en question, le deuxième est la cause ou l'objet du sentiment (si la cause est exprimée dans la phrase). Ceci donne pour la phrase

> « J'aime beaucoup Grenoble. »

la relation `SENTIMENT_POSITIF(aimer,Grenoble)`. L'attribut `POSITIF` de la relation, c'est-à-dire la valeur de sa classe, indique qu'il s'agit d'un sentiment positif dont l'objet est *Grenoble*. Dans le cas d'une phrase comme

> « Je déteste!!!!! »

la relation n'aura qu'un seul argument: `SENTIMENT_NEGATIF(détester)`, dans la mesure où l'objet du sentiment n'est pas exprimé dans la phrase.

L'objectif de la grammaire est d'extraire le plus d'informations possible dans le *thread*, en particulier les sentiments positifs et négatifs, les lieux et produits. Pour ceci, les *threads* sont analysés phrase par phrase. Chaque phrase peut contenir zéro, une ou plusieurs relations de sentiment. Il est tout à fait possible d'avoir des relations de sentiments positifs et négatifs dans une même phrase:

> « En qualité d'impression, la Epson est meilleure, en texte comme en photo, malheureusement c'est aussi la plus chère. »
> ⇒ `SENTIMENT_POSITIF(meilleur,Epson)`
> ⇒ `SENTIMENT_NEGATIF(cher,ce)`

Les parties de la grammaire qui varient selon le corpus se distinguent essentiellement par le lexique de mots qui reçoivent les traits `positif` et `négatif` correspondant aux valeurs des classes des textes. Par exemple, le lexique de la grammaire du *tourisme* contient les mots *joli* et *beau*:

> « Ce monument est vraiment *beau*. »

Pourtant, dans un corpus qui porte sur le cinéma, les livres ou les jeux vidéo, ces mêmes mots n'expriment pas toujours des sentiments. Ils ont donc été supprimés du lexique de la grammaire des *jeuxvidéo* parce qu'ils produisent trop de relations éronnées:

DEFT'07 (c.f. section 6). Sinon nous n'attribuons pas de sentiment global mais gardons les sentiments attribuées à chaque phrase.

> « Cela dépendra moins de vous que de l'imbécillité contagieuse des ennemis qui attendent sagement derrière un petit muret, leur *beau* visage buriné dépassant allègrement. »

Comme on le voit dans la phrase précédente, dans ce contexte, les mots de type *joli* ou *beau* sont utilisés pour décrire une action ou un personnage, mais pas un sentiment. La difficulté réside dans le fait de pouvoir distinguer les parties subjectives des parties objectives d'un texte. La description d'une action peut contenir des phrases avec des sentiments, donc subjectives, qui se réfèrent au déroulement de l'histoire. Cependant ces phrases devront être considérées comme étant objectives pour l'évaluation.

## 4.2 Lexique de sentiments

L'analyse du texte se base sur les mots du lexique qui ont reçu des traits spécifiques marquant le sentiment positif ou négatif. Il s'agit pour la plupart de verbes (*aimer*, *apprécier*, *détester*, ...) et d'adjectifs (*magnifique*, *superbe*, *insupportable*, ...), mais aussi de quelques noms communs (*plaisir*, ...) et d'adverbes (*malheureusement*, ...). Par exemple, quand une relation de modifieur du nom est extraite (*paysage magnifique*) et que le modifieur (*magnifique*) porte le trait `sents`, la relation de sentiment ( ⇒ `SENTIMENT_POSITIF(magnifique,paysage)`) est extraite ensuite entre le nom et son modifieur. Après cette phase d'analyse, il y a évidemment des règles plus complexes pour extraire les relations des phrases plus compliquées.

Le lexique a été défini par un linguiste au fur et à mesure de l'avancé de chaque projet. A chaque fois qu'un mot intéressant est apparu dans les textes qui n'était pas encore dans le lexique il a été ajouté à ce dernier, selon le domaine du texte. Pour chaque domaine il y a le même lexique de base et ensuite un lexique spécifique qui contient les mots du domaine en question.

L'attribut de la relation (`positif` ou `négatif`) d'un sentiment sera inversé quand une négation est présente dans la phrase, comme par exemple:

> « J'aime pas du tout les randonnées en montagne! »
> ⇒ `SENTIMENT_NEGATIF(aimer,randonnée)`
> « Ce n'est pas un mauvais restaurant. »
> ⇒ `SENTIMENT_POSITIF(mauvais,restaurant)`

Quand cela est possible, les pronoms *qui* et *que* se rapportant à une entité présente ailleurs dans la même phrase, seront remplacés par cette même entité:

> « Grenoble est une ville qui vaut vraiment le détour hiver comme été. »
> ⇒ `SENTIMENT_POSITIF(valoir,ville)`

Certains noms communs ainsi que des verbes de type interrogatif ont reçu un trait (`no-sents`) pour empêcher l'extraction de relations. Dans *Je cherche un bon hôtel.*, *Bon voyage!* ou *Bonne journée!* il ne s'agit pas de sentiments proprement dit exprimés par l'auteur du texte, mais plutôt de souhaits comme on peut les trouver surtout au début ou à la fin de messages. C'est pour cette raison que nous essayons d'éviter d'extraire ces relations.

Les noms de lieu et de produit ont également des traits spéciaux pour pouvoir extraire d'autres relations qui seront potentiellement intéressantes dans le futur. Voici un extrait du lexique où les mots reçoivent des traits en plus de ceux qu'ils portent déjà (la valeur 1 ajoute ce trait au mot, la valeur 0 l'enlève).

Chaque mot qui peut véhiculer un sentiment reçoit le trait `sents`, puis le trait `positif` ou `négatif` selon sa polarité. D'après la taxonomie d'(Ogorek, 2005) (c.f. la section suivante 4.3) sont ajoutées des valeurs de sentiment plus fines comme `à l'aise`, `détendu`, etc. Les mots qui ne doivent pas entrer en relation de sentiment reçoivent le trait `no-sents`. Les traits `genre` et `plateforme` servent à extraire d'autres relations intéressantes dans le domaine des *jeuxvidéo*.

```
Lexique:

agréable = {sents=1, positif=1, à l'aise=1}
sympathique = {sents=1, positif=1, détendu=1}
aimer = {sents=1, positif=1, enchanté=1}
conseiller = {sents=1, positif=1, conseil=1}
```

```
plaisir = {sents=1, positif=1, enchanté=1}
décevant = {sents=1, negatif=1, triste=1}
cher = {sents=1, negatif=1, cher=1}
regretter = {sents=1, negatif=1, triste=1}
malheureusement = {sents=1, negatif=1, triste=1}
appétit = {no-sents=1}
vacance = {no-sents=1}
chercher = {no-sents=1}
aventure = {genre=1}
PC = {plateforme=1}
```

La taille du lexique varie selon le domaine d'application. Le lexique de la grammaire de base des sentiments contient environ 250 mots (noms, verbes, adjectifs, etc.) avec des traits de sentiment (`positif` et `négatif`). À ce lexique de base, s'ajoutent environ 150 mots dans le domaine du *tourisme*, et environ 250 mots dans le domaine des *jeuxvidéo*.

### 4.3 Annotation manuelle de textes

La configuration de la grammaire générique a été faite sur la base d'un travail d'annotation manuelle (à l'aide du logiciel Protégé 3.2[6] avec le plugin Knowtator[7]) de *threads* venant du domaine du tourisme. Ce corpus du *tourisme* contient une centaine de *threads* annotés (avec comme sujet différentes régions et destinations en France). Chaque *thread* est composé de messages des utilisateurs du forum; la longueur varie entre dix et 55 messages par document. Un message peut ne contenir qu'une phrase ou plusieurs paragraphes. L'annotation de ce corpus avec Protégé et Knowtator a été faite dans la lignée des travaux de (Riloff *et al.*, 2005), (Riloff *et al.*, 2006), (Wiebe & Mihalcea, 2006).

L'annotation inclut les informations de cause/objet, d'intensité et de l'émetteur du sentiment. Dans

> « J'aime énormément Grenoble. »

*aimer* véhicule le sentiment, *Grenoble* est l'objet du sentiment et *je* est l'émetteur du sentiment. L'adverbe *énormément* exprime l'intensité, le sentiment ici est plus intense que dans la phrase

> « J'aime bien Grenoble. »

L'annotation pour le *tourisme* ne contient pas seulement les deux valeurs *positif* et *négatif* pour classer les sentiments, mais est détaillée beaucoup plus finement (c.f. par exemple les travaux de (Mathieu, 2000), (Mathieu, 2006)). Le schéma d'annotation choisi est même plus fin et on voit donc que la classification des sentiments que l'on propose permet un grand nombre de modalités et va au-delà de la simple opposition positif-négatif.

En effet, nous avons repris la taxonomie d'(Ogorek, 2005) qui propose 33 sentiments différents (17 positifs et 16 négatifs) auxquels nous avons ajouté les pseudo-sentiments comme *bon-marché*, *conseil*, *cher* et *avertissement*, car dans le domaine du *tourisme* il y a beaucoup de messages concernant les prix des prestations dont les auteurs des messages sont contents (ou pas).

Les sentiments de la taxonomie d'Ogorek sont classés en groupes[8] comme AMOUR-DÉSIR (*amour, envie, tendresse, désir*), JOIE (*enchanté, excité, heureux, joyeux*), TRISTESSE- DÉTRESSE (*découragé, bouleversé, démoralisé, triste*), COLÈRE-DÉGOÛT-MÉPRIS (*colère, mépris, désapprobation*), etc.

## 5 SYBILLE, la méthode hybride

La méthode hybride est une combinaison des deux méthodes précédentes (c.f. sections 3 et 4). La méthode statistique distributionnelle sert dans un premier temps à faciliter le développement de la méthode symbolique

[6] `http://protege.stanford.edu/`
[7] `http://bionlp.sourceforge.net/Knowtator/index.shtml`
[8] Sauf les pseudo-sentiments concernant les prix et conseils introduits par notre équipe comme *gratuit*, etc.

selon le domaine des textes. Pour chaque domaine d'application une ontologie de concepts propre est extraite. La création de lexique pour la grammaire symbolique est ainsi facilitée et accélérée.

La méthode statistique distributionnelle permet de faire une première fouille dans les textes pour obtenir les concepts du domaine des textes analysés. Ensuite, l'utilisateur qui a configuré la grammaire de la méthode symbolique peut modifier et améliorer celle-ci pour obtenir de meilleurs résultats. Le travail prend alors la forme d'un cycle où les résultats s'améliorent constamment.

L'analyse du *thread* se fait au niveau des phrases et permet d'améliorer le résultat en ajoutant ou supprimant par exemple des mots au lexique. Ceci a l'avantage de montrer exactement quelles phrases du document expriment un sentiment, les phrases objectives n'étant pas pris en compte.

C'est une approche qui permet de pouvoir extraire rapidement les concepts intéressants du domaine d'application et d'améliorer en même temps le développement de la grammaire de la méthode symbolique. Ceci permet d'intégrer les spécificités du cahier des charges, c'est-à-dire les particularités de chaque corpus (à l'aide de lexiques différents selon le domaine d'application).

La méthode hybride a été évaluée dans sa première version (c'est-à-dire avec la méthode statistique basée sur l'apprentissage automatique, avant l'intégration de l'exploration de corpus), notamment au moment du concours DEFT'07 (c.f. la section 6), avec la mesure du F-score[9]. Elle a été utilisée pour trois des quatre corpus DEFT'07 et a donné les meilleurs résultats pour les corpus *jeuxvidéo* avec un F-score de 0,71, contre 0,54 (méthode symbolique) et 0,70 (méthode statistique) et *relectures* avec un F-score de 0,54, contre 0,48 (méthode symbolique) et 0,51 (méthode statistique).[10] Pour le corpus *débats politiques* seule l'ancienne méthode statistique a été utilisée.

Une évaluation avec la nouvelle méthode statistique est prévue dès que le système a été mis au point.

# 6 Première évaluation

La section suivante décrit la première évaluation du système SYBILLE, faite en 2007. Depuis le changement des méthodes nous n'avons malheureusement pas encore eu l'occasion de refaire une nouvelle évaluation empirique avec des objectifs bien définis et des résultats satisfaisants. Ce sera l'objet d'une future publication.

DEFT (le DÉfi Fouille de Texte) est une campagne d'évaluation dont le thème était en 2007[11] la classification de textes d'opinion, présents dans différents types de textes. Plusieurs groupes de recherche (laboratoires universitaires ou entreprises privées) ont pu tester leurs systèmes de classification sur les mêmes textes. Dans la phase initiale, chaque groupe inscrit a reçu les deux tiers de chacun des quatre corpus différents qui avaient comme sujet des critiques de films et de livres, des tests de jeux vidéo, des relectures d'articles scientifiques et des notes de débats parlementaires. Pour les trois premiers corpus, une note à trois valeurs (positif, moyen ou négatif) a été attribuée à chaque texte par le comité des organisateurs, une note à deux valeurs seulement (positif ou négatif) pour le dernier corpus. Après un certain temps pendant lequel chaque groupe a mis au point son ou ses systèmes de classification, un troisième tiers de chaque corpus a été envoyé pour faire les tests dont les résultats ont dû être soumis quelques jours plus tard.

La grammaire de l'analyseur a été paramétrée pour répondre aux besoins des différents corpus DEFT'07, du point de vue lexical mais aussi pour résister aux fautes d'orthographes répétitives. Le point le plus important à modifier a été la classification du message entier qui peut contenir plusieurs sentiments avec une seule valeur globale, et en particulier l'introduction de la notion de sentiment moyen. Dans notre approche standard, au niveau des phrases, les sentiments sont positifs ou négatifs. Il n'est pas nécessaire d'utiliser des sentiments moyens dans le domaine du tourisme, dans la mesure où la taxonomie utilisée (c.f. section 4.3) permet de nuancer suffisamment.

Les sentiments moyens pour DEFT'07 n'ont pas été extraits à l'aide de mots dans le lexique avec un trait moyen, mais d'après des structures de phrase. Par exemple à une phrase qui contient un sentiment positif et un sentiment négatif coordonnés par *mais* est attribué un sentiment moyen à la place:

> « Ce jeu est *amusant* au début **mais** *ennuyant* la deuxième semaine. »

[9] Le F-score utilisé dans nos expériences est calculé de la manière suivante: $F_{score}(\beta) = \frac{(\beta^2+1)*Précision*Rappel}{\beta^2*Précision+Rappel}$ avec $\beta = 1$.

[10] Pour le corpus *aVoiraLire* le meilleur résultat a été obtenu par la méthode statistique avec un F-score de 0,52, contre 0,51 (méthode hybride) et 0,42 (méthode symbolique). Ce corpus n'est probablement pas assez uniforme (il parle de livres, films actuels au cinéma, disques, films plus anciens enregistrés, ...) pour pouvoir faire une liste de termes plus performante.

[11] `http://deft07.limsi.fr/`

Quelques mots clés (surtout des adverbes comme *malgré, pourtant, ...*) sont utilisés pour aider à classifier un texte qui contient des phrases avec des sentiments positifs et négatifs (c.f. les travaux de (Sándor, 2005)). Le texte entier est alors classé comme moyen.

# 7 L'interface graphique SYBILLE

Pour conclure cet article voici quelques figures qui présentent notre interface graphique SYBILLE, une interface qui aide l'analyste et le client à naviguer parmi les messages analysés pour en prendre connaissance. L'interface que nous utilisons est une dérivation du navigateur *Longwell* du MIT[12].

Les figures 2 et 3 suivantes montrent l'interface graphique du système SYBILLE, dans le domaine des *imprimantes*. Sur la figure 2 en haut à droite il y a un champ dans lequel l'utilisateur peut faire une recherche **(1)** de messages qui contiennent un mot de son choix; sinon, il peut ne choisir que les messages positifs ou négatifs **(2)**. Une autre façon de faire une recherche serait de se limiter aux messages qui n'évoquent qu'une marque précise **(3)**, ou en dessous un domaine d'application plus spécialisé **(4)**, ou encore un mot précis d'un domaine. On offre aussi l'option de sélectionner un forum donné parmi tous ceux qui ont été analysés. Les options de recherche peuvent être combinées à volonté pour limiter le nombre de réponses souhaitées.

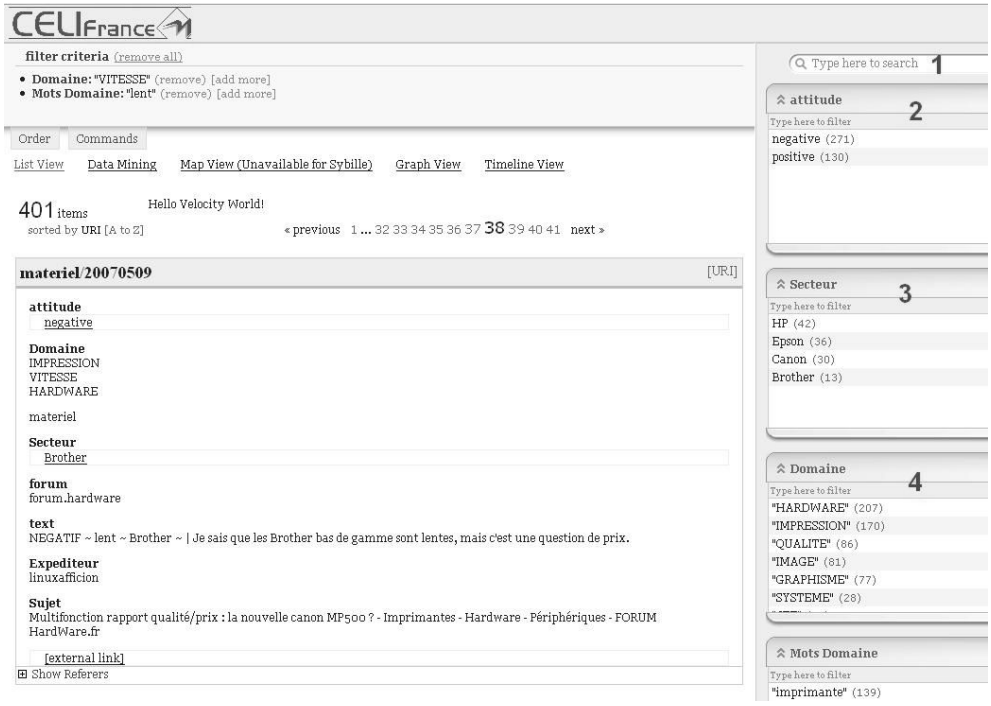


Figure 2: L'interface graphique SYBILLE, ici pour le domaine des imprimantes. Plusieurs moyens différents permettent de naviguer dans les résultats des messages analysés.

La figure 3 montre en détail la relation de sentiment qui est indiquée avec ses arguments **(5)**, la phrase qui contient le sentiment et un lien (*external link* **(6)**) vers le *thread* entier qui permet de visualiser le contexte.

[12] `http://simile.mit.edu/wiki/Longwell`

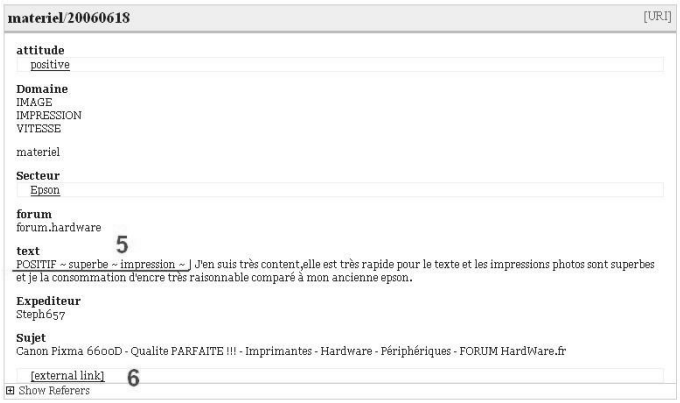


Figure 3: Exemple détaillé d'une relation de sentiment positif, toujours dans le domaine des imprimantes.

# 8 Conclusion

Nous avons présenté dans cet article comment l'utilisation de grammaires syntaxiques, un outil du traitement automatique du langage naturel, peut améliorer la qualité d'un système d'extraction de sentiments. Nous avons décrit une méthode symbolique avec une grammaire adaptée au domaine des textes, et une méthode statistique distributionnelle pour la création d'une ontologie des concepts du domaine des textes. Cette dernière nous offre la possibilité d'une exploration rapide d'un corpus d'un domaine donné qui souligne en même temps les attitudes différentes des différents groupes d'opinions par rapport à l'objet de l'étude.

La première évaluation de notre système de classification SYBILLE en 2007 a montré que la combinaison de la méthode symbolique et l'ancienne méthode statistique donne des résultats plus précis que chacune des méthodes employée séparément. L'intérêt de la méthode hybride repose sur la prise en compte des contextes d'application de ses résultats. Il est bien connu que la méthode purement symbolique a souvent pour le client un coût d'entrée plutôt élevé. Cette considération est liée au temps de configuration, de repérage ou de création de lexiques spécifiques, de taxonomies etc.

L'utilisation d'une méthode hybride permet, au contraire, de minimiser les coûts de configuration, en réduisant une partie du travail à l'annotation de textes, une tâche qui dans la plupart des cas peut être réalisée par le client lui-même. Les algorithmes d'apprentissage automatique sont alors en mesure de donner des premiers jugements au niveau du texte entier.

Ce qui est le plus important, c'est qu'avec ce type de système on peut ajouter, selon la méthode exposée dans cet article, une couche *symbolique* au fur et à mesure, de plus en plus importante dès que les exigences d'une application deviennent plus précises. On peut par exemple superposer une couche d'identification de jugement, qui permet d'avoir une visibilité sur les jugements sans devoir lire le texte dans son entier. On peut identifier certains patrons sémantiques qui sont d'importance capitale pour une application donnée et qui doivent avoir la priorité sur les résultats statistiques (par exemple le souci de sécurité exprimé par les internautes sur un certain modèle de voiture).

Les exemples pourraient être multipliés. Ce qui apparaît avant tout intéressant, c'est que la démarche hybride est importante non seulement pour des raisons scientifiques de performance (le meilleur résultat entre les technologies que nous avons adoptées) mais, aussi et surtout pour des raisons de développement et d'acceptation par le marché.

# Références

Aït-Mokhtar S. et Chanod J.-P. (1997). Subject and object dependency extraction using finite-state transducers. In P. Vossen, G. Adriaens, N. Calzolari, A. Sanfilippo & Y. Wilks, Eds., *Automatic information extraction and building of lexical semantic resources for NLP applications*, p. 71–77. Association for Computational Linguistics.

Aït-Mokhtar S., Chanod J.-P. et Roux C. (2001). A multi-input dependency parser. In *Actes d' IWPT'01*.

Baroni M. et Bisi S. (2004). Using cooccurrence statistics and the web to discover synonyms in a technical language. In *Actes de LREC'04*, p. 1725–1728.

Basili R., Pazienza M. T. et Zanzotto F. M. (1999). Lexicalizing a shallow parser. In *Actes de TALN'99*.

Bethard S., Yu H., Thornton A., Hatzivassiloglou V. et Jurafsky D. (2004). Automatic extraction of opinion propositions and their holders. In *Actes d' AAAI'04*.

Bosca A. et Dini L. (2009). Ontology based law discovery. In S. Montemagni & D. Tiscornia, Eds., *Semantic processing of legal texts*. Springer, à paraître.

Chklovski T. (2006). Deriving quantitative overviews of free text assessments on the web. In *Actes d' IUI'06*, p. 155–162.

Dini L. (2002). Compréhension multilingue et extraction de l'information. In F. Segond, Ed., *Multilinguisme et traitement de l'information (Traité des sciences et techniques de l'information)*. Editions Hermes Science.

Dini L. et Mazzini G. (2002). Opinion classification through information extraction. In A. Zanasi, C. A. Brebbia, N. F. F. Ebecken & P. Melli, Eds., *Data Mining III*, p. 299–310. WIT Press.

Dini L. et Segond F. (2007). La linguistique informatique au service des sentiments. In *Revue de l'électricité et de l'électronique*, p. 66–77. Editions SEE.

Everitt B. (1992). *The Analysis of Contingency Tables*. Chapman and Hall, 2nd edition.

Grouin C., Berthelin J.-B., El Ayari S., Heitz T., Hurault-Plantet M., Jardino M., Khalis Z. et Lastes M. (2007). Présentation de DEFT'07 (DÉfi Fouille de Textes). In *Actes de DEFT'07*, p. 1–8.

Hatzivassiloglou V. et McKeown K. R. (1997). Predicting the semantic orientation of adjectives. In *Actes d' ACL'97*, p. 174–181.

Kim S.-M. et Hovy E. (2004). Determining the sentiment of opinions. In *Actes de COLING'04*, p. 1267–1373.

Mathieu Y. Y. (2000). *Les verbes de sentiment. De l'analyse linguistique au traitement automatique*. CNRS Editions.

Mathieu Y. Y. (2006). A computational semantic lexicon of french verbs of emotion. In J. G. Shanahan, Y. Qu & J. Wiebe, Eds., *Computing attitude and affect in text*: *Theorie and applications*, p. 109–124. Springer.

Maurel S., Curtoni P. et Dini L. (2007). Classification d'opinions par méthodes symbolique, statistique et hybride. In *Actes de DEFT'07*, p. 111–117.

Maurel S., Curtoni P. et Dini L. (2008). L'analyse des sentiments dans les forums. In *Actes de FODOP'08*, p. 9–22.

Maurel S., Curtoni P. et Dini L. (2009). Extraction de sentiments et d'opinions basée sur des règles. In *Fouille des données d'opinions*. RNTI, à paraître.

Ogorek J. R. (2005). Normative picture categorization: Defining affective space in response to pictorial stimuli. In *Actes de REU'05*.

Riloff E., Patwardhan S. et Wiebe J. (2006). Feature subsumption for opinion analysis. In *Actes d' EMNLP'06*, p. 440–448.

Riloff E., Wiebe J. et Phillips W. (2005). Exploiting subjectivity classification to improve information extraction. In *Actes d' AAAI'05*.

Sándor A. (2005). A framework for detecting contextual concepts in texts. In *Actes du Electra Workshop*.

Sproull L. et Kiesler S. (1991). *Connections*: *New ways of working in the networked organization*. Cambridge: MIT Press.

Turney P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Actes d' ACL'02*.

Wiebe J. et Mihalcea R. (2006). Word sense and subjectivity. In *Actes d' ACL'06*, p. 1065–1072.

Wilson T., Wiebe J. et Hwa R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Actes d' AAAI'04*.

Yu H. et Hatzivassiloglou V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Actes d' EMNLP'03*, p. 129–136.