# Une procédure pour identifier les modifieurs de la valence affective d'un mot dans des textes

Noémi Boubel[1] Yves Bestgen[2]

(1) UCLouvain, Cental, Place Blaise Pascal, 1, B-1348 Louvain-la-Neuve, Belgique
(2) UCLouvain, CECL, B-1348 Louvain-la-Neuve, Belgique
noemi.boubel@uclouvain.be, yves.bestgen@uclouvain.be

**Résumé :** Cette recherche s'inscrit dans le champ de la fouille d'opinion et, plus particulièrement, dans celui de l'analyse de la polarité d'une phrase ou d'un syntagme. Dans ce cadre, la prise en compte du contexte linguistique dans lequel apparaissent les mots porteurs de valence est particulièrement importante. Nous proposons une méthodologie pour extraire automatiquement de corpus de textes de telles expressions linguistiques. Cette approche s'appuie sur un corpus de textes, ou d'extraits de textes, dont la valence est connue, sur un lexique de valence construit à partir de ce corpus au moyen d'une procédure automatique et sur un analyseur syntaxique. Une étude exploratoire, limitée à la seule relation syntaxique associant un adverbe à un adjectif, laisse entrevoir les potentialités de l'approche.

**Abstract** This research is situated within the field of opinion mining and focuses more particularly on the analysis of the opinion expressed in a sentence or a syntagm. Within this frame of research, taking into account the linguistic context in which words which carry valence appear is particularly important. We propose a methodology to automatically extract such linguistic expressions from text corpora. This approach is based on (a) a corpus of texts, or text excerpts, the valence of which is known, (b) on a valence lexicon built from this corpus using an automatic procedure and (c) on a parser. An exploratory study, focusing on the syntactic relation associating an adverb to an adjective, shows the potential of the approach.

**Mots-clés :** modifieurs de valence, fouille d'opinion, lexique de valence
**Keywords:** contextual valence shifter, opinion mining, semantic orientation lexicon

## 1 Introduction

Depuis une dizaine d'années, la détection d'opinion et de sentiments dans les textes est devenue un sujet de recherche important en traitement automatique du langage, ainsi qu'un enjeu stratégique pour les entreprises et les institutions (Pang, Lee, 2008). La tâche classique de ce domaine consiste à déterminer automatiquement la polarité globale d'un texte. Progressivement, l'attention des chercheurs s'est également portée vers le niveau phrastique ou syntagmatique afin d'identifier les segments d'un texte qui expriment une opinion, la valence de celle-ci et sa cible (Hatzivassiloglou, Wiebe, 2000 ; Kessler, Nicolov, 2009 ; Vernier et al., 2009). Ce deuxième axe de recherche souligne l'importance de la prise en compte du contexte linguistique dans lequel apparaissent les mots porteurs de valence. Si certains travaux en fouille d'opinion commencent à prendre en compte ce genre de phénomènes (Kennedy, Inkpen, 2006 ; Musat, Trausan-Matu, 2010 ; Wilson et al., 2005), très rares sont ceux qui en ont fait leur objet central d'étude. Une recherche de Zaenen et Polanyi (2004) constitue cependant une exception notable. Leur hypothèse de travail est que la valence de termes polarisés peut être renforcée ou affaiblie par la présence d'autres items lexicaux, par la structure du discours et le type de texte, ou enfin par des facteurs culturels. Ces chercheuses proposent d'appeler *contextual valence shifters* les différents éléments ou procédés ayant un impact sur la valeur d'un mot comme la présence d'une négation, d'un adverbe de degré, d'un verbe modal ou d'une tournure ironique. Il faut toutefois noter que les arguments empiriques présentés reposent sur l'analyse de quelques exemples

le plus souvent construits à dessein. Plus récemment, Klenner, Petrakis, et Fahrni (2009) et Vernier et al. (2009) ont développé des ensembles de règles plus complexes et plus précises à partir d'une analyse linguistique approfondie d'un corpus et montré l'utilité de celles-ci pour détecter les évaluations localement exprimées dans un texte. Enfin, dans une autre optique, Vernier et Monceaux (2010) se servent de ces modifieurs de valence (là encore définis a priori) pour créer un lexique axiologique sur la base de tests sémantiques.

À notre connaissance, aucune procédure systématique n'a été proposée pour identifier les expressions linguistiques qui modifient la valence d'un mot alors qu'une telle procédure semble indispensable en raison de la très grande diversité de ces expressions et de leur variabilité selon le genre de textes. L'objectif de cette recherche est d'essayer de répondre à cette lacune. Après avoir présenté la procédure que nous proposons et les différentes ressources qu'elle nécessite, nous rapportons une étude exploratoire qui illustre son fonctionnement.

## 2 Approche proposée

L'idée de départ consiste à employer une procédure automatique pour rechercher dans des textes classés selon leur polarité, comme des critiques de films ou de produits, les termes porteurs de valence les plus marqués et extraire des informations à propos du contexte dans lequel ils apparaissent. Notre hypothèse est que les expressions linguistiques qui ont pour fonction d'atténuer ou d'inverser la valence d'un terme seront tout particulièrement fréquentes dans le contexte d'un terme dont la valence est opposée à celle du texte dans lequel il apparaît. Par contre, celles qui ont pour fonction de renforcer la valence d'un terme s'observeront tout particulièrement dans le contexte d'un terme dont la valence est similaire à celle du texte dans lequel il apparaît. On peut aussi penser qu'un nombre plus important d'expressions ayant un effet atténuant pourrait être observé en relation avec de mots de polarité extrême présents dans des textes évaluatifs neutres.

Cette méthodologie nécessite un corpus de textes déjà classés selon la dimension évaluative, un lexique de mots extrêmes en termes de valence et un analyseur syntaxique. Idéalement, le corpus devrait être (comme toujours) le plus grand possible, mais, surtout, être prioritairement composé d'énoncés évaluatifs afin que l'emploi des constructions étudiées dans des fonctions non évaluatives n'altère pas trop la précision des analyses. Le lexique de valence, quant à lui, doit être spécifiquement adapté au corpus étudié afin de prendre en compte un maximum de mots très positifs et très négatifs (voir Vernier et al. (2009) pour un argument similaire). L'étude exploratoire présentée ci-dessous essaie de mettre en pratique ces contraintes.

## 3 Etude exploratoire

Afin d'évaluer les potentialités de la méthode proposée, nous avons opté, dans cette étude exploratoire, pour l'analyse de la relation syntaxique associant un adverbe et un adjectif ayant une valence très forte parce qu'on peut penser que cette construction, qui recouvre un grand nombre de mots différents, modifie fréquemment et fortement la valence de mots. Un autre intérêt de cette relation est que, si l'approche fonctionne, les résultats obtenus devraient largement correspondre aux intuitions linguistiques. Ceci aura néanmoins pour conséquence regrettable d'en réduire l'originalité.

### 3.1 Corpus

Un corpus de critiques de films nous a semblé idéal pour tester l'approche parce que celles-ci peuvent être trouvées en abondance sur internet. Le site allocine.fr, par exemple, propose, pour un grand nombre de films, de très brefs extraits d'articles de différents journaux auxquels est attribuée une note sur une échelle en 5 points allant d'un avis très négatif à un avis très positif[1] (voir exemple ci-dessous et dans la suite). Ces extraits mettent particulièrement en évidence l'opinion du journaliste et non le sujet du film ou l'intrigue, ce

[1] Les notes initialement attribuées par les auteurs des articles selon le système de notation en vigueur dans la source, et donc a priori fiable étant donné l'expertise des auteurs, sont remises par allocine.fr sur une seule et même échelle.

qui est une propriété intéressante pour la méthodologie proposée puisqu'on peut penser que la quasi-totalité des adjectifs de valence extrême présents dans ces extraits sera porteuse d'une information évaluative.

1. C'est gentiment amusant, pas trop crado bien qu'égrillard, mais ça n'offre pas le moindre intérêt sur le plan cinématographique. [« 1001 pattes » – L'Humanité – Note : 3]

### 3.2 Lexique de valence

Pour déterminer l'orientation sémantique d'un maximum de termes présents dans le corpus, nous avons employé la méthode d'apprentissage supervisé de mots germes développée par Vincze et Bestgen (2011). À la suite des approches proposées par Bestgen (2002) et Turney et Littman (2003), cette méthode utilise une procédure de régression pour estimer la valence d'un terme sur la base de ses proximités thématiques dans un corpus avec d'autres mots dont la valence est connue. Dans la présente étude, les proximités thématiques entre les termes sont obtenues au moyen de l'analyse sémantique latente (Berry et al., 1993 ; Deerwester et al., 1990) du corpus de critiques et les mots, dont la valence est connue a priori et qui servent donc à construire le modèle prédictif, proviennent d'une norme lexicale de 3252 mots évalués sur une échelle à 7 points allant de *très désagréable* (1) à *très agréable* (7) par un minimum de 30 juges (Hogenraad et al., 1995). En sortie, la méthode attribue une valence à l'ensemble des termes contenus dans le corpus de critiques qui ont une fréquence supérieure ou égale à 10, soit 9368 mots. Pour la suite des analyses, nous avons retenu comme termes extrêmes négatifs les 10% de cette liste les plus extrêmes du côté négatif et procéder de manière similaire pour les termes extrêmes positifs. De ces deux listes, seuls les adjectifs ont été employés dans la suite des analyses. À titre d'exemple, les cinq adjectifs les plus négatifs et les plus positifs obtenus par cette procédure sont respectivement *insignifiant, sadique, prétentieux, sordide, glauque* et *merveilleux, délicieux, agréable, chaleureux, rafraîchissant*. Il est à noter que, lors de la détermination de la valence des termes, les notes attribuées aux critiques n'interviennent en aucune manière.

### 3.3 Analyse syntaxique

Afin de mettre en place la méthodologie proposée plus haut, nous avons développé une série de scripts perl permettant d'extraire les relations syntaxiques impliquant un terme d'une liste établie au préalable, grâce à un analyseur syntaxique robuste, XIP (Xerox Incremental Parser, Aït-Mokhtar et al., 2002). Le corpus est tout d'abord étiqueté par l'analyseur selon les relations de dépendances présentes. Puis les relations (ADJMOD) associant un adjectif présent dans notre lexique à un adverbe sont extraites du corpus et mises en relation avec la note de la critique dans laquelle ils apparaissent. Nous obtenons donc des triplets comme : *ADV="complètement", ADJ+="indigeste", NOTE="2"* ou *ADV="souvent", ADJ+="hilarant", NOTE="3"*.

## 4 Analyses

### 4.1 Traitements statistiques

La partie supérieure du Tableau 1 présente un résumé numérique du corpus analysé en termes de nombre de critiques, nombre de mots et nombre d'adjectifs extrêmes pour les différentes notes attribuées aux critiques. On observe que, non seulement, il y a moins de critiques extrêmes (1 ou 5), mais que les critiques les plus négatives sont aussi les plus brèves puisqu'elles ont une longueur moyenne de 21 mots alors que la longueur moyenne calculée sur l'ensemble des critiques est supérieure à 26 mots. De plus, s'il y a globalement plus d'adjectifs très positifs que d'adjectifs très négatifs, leur distribution en fonction de la note des critiques correspond à ce à quoi on pouvait s'attendre : une proportion plus grande d'adjectifs négatifs dans les critiques les plus négatives et l'inverse dans les critiques les plus positives.

Les données recueillies par l'approche proposée peuvent être présentées sous la forme de tables de contingence, une par adverbe analysé, construites sur la base des trois variables : la note des critiques, la valence de l'adjectif et l'adverbe. La partie inférieure du Tableau 1 donne cette table de contingence à trois entrées pour l'adverbe *jamais*. Les analyses les plus pertinentes pour nos objectifs visent à déterminer, séparément pour les adjectifs positifs et négatifs, si l'adverbe modifie plus fréquemment les adjectifs

présents dans les critiques qui ont reçu une note spécifique. Elle porte donc sur les tables de contingence formées par les variables Note et Adverbe (cellules en vert et en jaune).

Ces tables de contingence ont été analysées au moyen du test du Chi-carré qui évalue l'indépendance entre les deux variables formant la table. Lorsque ce test donne un résultat statistiquement significatif pour un alpha de 0.05, nous avons calculé la statistique du résidu ajusté qui indique les cellules qui contribuent le plus aux écarts à l'indépendance (Everitt, 1977) en employant le même alpha pour signaler les contributions potentiellement intéressantes. Ces analyses statistiques n'ont été effectuées que sur les adverbes ayant une fréquence totale dans le corpus au moins égale à 20.

| | | Note 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|---|
| Nbr. Critiques | | 3641 | 16135 | 23210 | 25904 | 8671 | 77561 |
| Nbr. Mots | | 77028 | 412990 | 619490 | 680796 | 239358 | 2029662 |
| Nbr. Adj. + | | 506 | 3356 | 7964 | 11088 | 3979 | 26893 |
| Nbr. Adj. - | | 1226 | 4871 | 4111 | 3783 | 1259 | 15150 |
| Adj+ | Jamais | 1 | 1 | 2 | 3 | 0 | 7 |
| | -------- | 505 | 3355 | 7962 | 11085 | 3979 | 26886 |
| Adj- | Jamais | 0 | 4 | 13 | 42 | 17 | 76 |
| | -------- | 1226 | 4867 | 4098 | 3741 | 1242 | 15174 |

Tableau 1 : Statistiques descriptives et tables de contingence pour *jamais*

### 4.2 Analyse des résultats

Le Tableau 2 met en évidence une série d'adverbes qui fonctionnent comme des inverseurs ou des atténuateurs de valence. Par exemple, l'adverbe *jamais* utilisé avec un adjectif négatif est significativement trop rare dans des critiques très négatives ou négatives, et trop fréquent dans les critiques positives et très positives. On notera que cet adverbe est rarement employé avec un terme positif, un résultat auquel nous ne nous attendions pas. Parmi les autres adverbes trop fréquemment associés à des adjectifs de la polarité inverse à celle de la critique ou trop peu fréquemment associés à des adjectifs de même polarité que la critique, on peut mentionner *pas*, *parfois, un peu, trop, moins, peu, assez, plutôt*. Parmi les adverbes apparaissant trop souvent dans des critiques de même polarité que l'adjectif qui lui est associé, et qui donc présentent le profil d'un intensifieur de valence, on peut citer *franchement, complètement, totalement, bien, profondément, particulièrement*.

D'autres résultats semblent plus étonnants. C'est le cas, par exemple, de *tellement* [2] ou de *particulièrement*, (a priori intensifiant), trop souvent associés à un adjectif positif dans une critique négative ou de *très* trop rarement associé à un adjectif positif dans des critiques très positives, ou encore de *faussement* [3] dont l'impact suggéré par l'analyse statistique serait l'intensification. Des exemples de ces cas permettent de se rendre compte de la diversité et de la complexité des phénomènes jouant sur la polarité canonique d'un terme.

2. Une comédie tellement sotte qu'on l'oublie à la seconde où l'on a quitté la salle. Tellement inoffensive, inefficace, et ratée que même les ados risquent cette fois de passer la main.
3. L'ancienne version, crépitante, visait à tout dire sur les hommes, celle-ci passe les femmes au gril, sur le ton de la presse people et avec une connivence faussement méchante.

Plus généralement, les adverbes repris dans le Tableau 2 sont de nature très diverse : adverbes de comparaison (*aussi, à la fois*), adverbes de fréquence (*souvent, parfois, peu*), adverbes mettant peut-être plus en avant le locuteur (*franchement, totalement*) ou adverbes porteurs eux-mêmes d'une certaine valence

(*faussement* ou *trop*). De plus, le contexte plus large joue également souvent un rôle important ; l'inversion de la valence d'un adjectif, par exemple, peut se faire à un autre niveau (tournures négatives diverses, combinaison de termes de polarité différente, utilisation de l'ironie par divers procédés [4] [5]).

4. La compétition pour le navet le plus farfelu de l'année est lancée sur des très hautes bases
5. (…) le spectateur rendra facilement les armes devant un film aride, peu loquace, aussi chaleureux que l'Allemagne de l'Est de la grande époque.

| Valence | Fréquence | Note 1 | Note 2 | Note 3 | Note 4 | Note 5 |
|---|---|---|---|---|---|---|
| **Positif** | **Trop rare** | | | profondément, particulièrement | moins, trop, peu, plutôt, parfois, un peu | moins, pas, plutôt, assez, un peu, trop, très, parfois |
| | **Trop fréquent** | particulière-ment, pas, tellement | peu, pas | moins, plutôt, assez, trop, un peu, parfois, toujours, fort | particulièrement, très | plus, à la fois, absolument, profondément |
| **Négatif** | **Trop rare** | parfois, un peu, jamais, trop, presque | jamais, parfois, un peu, moins | complètement | trop, assez, aussi, franchement, vite, souvent, totalement, lourdement, plutôt, bien, de plus | un peu, assez, trop, aussi |
| | **Trop fréquent** | complètemen, faussement, aussi, de plus, proprement, franchement | lourdemen, vite, souvent, totalement, assez, franche-ment, bien, aussi | un peu, parfois, trop, pas, moins, assez | jamais | jamais, à la fois, presque |

Tableau 2 : Adverbes identifiés par l'analyse statistique[2]

Les analyses qui précèdent portent uniquement sur les adverbes suffisamment fréquents pour établir des conclusions statistiques. La procédure a également permis d'extraire 183 adverbes associés une seule fois à un des adjectifs retenus dans notre lexique et 376 adverbes n'apparaissent qu'un maximum de 10 fois. En se basant sur la note de la critique dans laquelle ils apparaissent ainsi que sur la polarité de l'adjectif associé, une étude similaire à celle qui vient de l'être peut être menée. À titre d'exemple, on peut citer *abyssalement vide, artificiellement glauque, authentiquement laids, américainement naïf*.

## 5 Conclusion

Nous avons proposé une méthodologie pour extraire automatiquement de corpus de textes des expressions linguistiques susceptibles de modifier la valence de mots. Cette approche s'appuie sur un corpus de textes, ou d'extraits de textes, dont la valence est connue, sur un lexique de valence construit à partir de ce corpus au moyen d'une procédure automatique et sur un analyseur syntaxique. L'étude exploratoire, limitée à la seule relation syntaxique associant un adverbe à un adjectif, laisse entrevoir les potentialités de l'approche. Elle laisse toutefois de nombreuses questions en suspens. Les principales nous semblent porter sur le degré d'efficacité qui pourra être obtenu avec d'autres structures et la possibilité de généraliser les constructions identifiées dans un corpus donné à d'autres genres de textes.

[2] Les adverbes sont ordonnés dans chaque cellule en fonction de la taille des résidus ajustés (en valeur absolue). Il s'ensuit que les premiers mentionnés sont plus typiques que les derniers.

## Références

AÏT-MOKHTAR S., CHANOD J.-P., ROUX C. (2002). Robustness beyond shallowness : incremental deep parsing. *Natural Language Engineering* 8, 121-144.

BERRY, M.. DO, T., O'BRIEN, G., KRISHNA, V., VARADHAN, S. (1993). SVDPACKC: Version 1.0 User's Guide, *Tech. Rep. CS-93-194*, University of Tennessee, Knoxville, TN.

BESTGEN, Y. (2002). Détermination de la valence affective de termes dans de grands corpus de textes. Actes de *CIFT'02*, 81-94.

DEERWESTER, S., DUMAIS, S.T., FURNAS, G.W., LANDAUER, T.K., HARSHMAN, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science* 41, 391-407.

EVERITT, B. (1977) *The Analysis of Contingency Tables*. Chapman & Hall.

HATZIVASSILOGLOU, V., WIEBE, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity. Proceedings of the *18th Conference on Computational Linguistics*, 299-305.

HOGENRAAD, R., BESTGEN, Y., NYSTEN, J.L. (1995). Terrorist Rhetoric : Texture and Architecture, In Nissan et Schmidt (Eds.), *From Information to Knowledge*, 48-59, Intellect Book.

KENNEDY, A., D. INKPEN. (2006). Sentiment Classification of Movie Reviews Using Contextual Valence Shifters. *Computational Intelligence* 22, 110–125.

KESSLER J.S., NICOLOV, N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. Proceedings of *ICWSM 2009*.

KLENNER, M., S. PETRAKIS, A. FAHRNI. Robust Compositional Polarity Classification. Proceedings of *RANLP 2009*.

MUSAT, C., TRAUSAN-MATU, S. (2010). The Impact of Valence Shifters on Mining Implicit Economic Opinions. Proceedings of *AIMSA 2010*, 131–140.

PANG, B., LEE L. (2008) Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–135

TURNEY P. (2002). Thumbs up or Thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of *ACL'02*.

TURNEY, P.D., LITTMAN, M. (2003). Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems* 21, 315--346.

VERNIER, M., MONCEAUX, L. (2010). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement automatique des langues* 51, 125-149.

VERNIER, M., MONCEAUX, L., DAILLE, B. (2009). Catégorisation des évaluations dans un corpus de blogs multi-domaine. *Revue des Nouvelles Technologies de l'Information* (RNTI-E-17), 45–70.

VINCZE, N., BESTGEN, Y. (2011). Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée. Actes de *TALN2011*.

WILSON, T., J. WIEBE, P. HOFFMANN. « Recognizing Contextual Polarity in Phrase-level Sentiment Analysis ». Proceedings of the *Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 347–354.

ZAENEN, A., POLANYI L. (2004). Contextual Valence Shifters. Proceedings of *AAAI Spring Symposium on Exploring Attitude and Affect in Text*, 106–111.