Détermination de la valence affective de termes dans de grands corpus de textes

Yves Bestgen

Fonds national de la recherche scientifique
Université catholique de Louvain
Place du Cardinal Mercier, 10 B1347 Louvain-la-Neuve Belgique
yves.bestgen@psp.ucl.ac.be

RÉSUMÉ. *L'objectif de la recherche présentée ici est de développer une méthode d'extraction d'information à partir de grands corpus de textes permettant d'estimer la valence affective associée à n'importe quel terme. L'approche proposée combine deux techniques qui ont été développées indépendamment l'une de l'autre : l'analyse sémantique latente (ASL) et l'évaluation du contenu affectif d'un texte sur la base d'une partie des mots qui le composent. Une étude préliminaire visant à évaluer cette approche a été menée sur un corpus de plusieurs milliers d'articles publiés pendant un an dans un journal belge francophone. Une première analyse montre qu'en combinant l'ASL et un dictionnaire de valence affective de 3000 mots, il est possible d'approximer efficacement la valence affective de mots en se basant exclusivement sur les mots qui leur sont associés dans l'espace sémantique. Une seconde analyse, beaucoup plus exploratoire, applique la technique à des noms de firme. Dans la conclusion, on propose quelques pistes de développement.*

ABSTRACT. *The aim of this research is to develop a method of information extraction from large corpora of texts to estimate the affective valence associated to any term. Our approach combines two techniques : latent semantic analysis (LSA) and the determination of the emotional content of a text based on the words that compose it. A preliminary study designed to evaluate this approach has been conducted on a corpus of several thousands of articles published in a Belgian newspaper. A first analysis showed that, by combining LSA and a dictionary of 3000 words, it is possible to approximate efficiently the affective valence of words on the base of the words that are associated to them in the semantic space. A second analysis applied the technique to firm names. We conclude by proposing some improvements of the technique.*

MOTS-CLÉS : *Méthode d'extraction d'information, Valence affective de terme, Analyse sémantique latente.*

KEYWORDS : *Method of information extraction, Affective valence of terms, Latent semantic analysis.*

CIFT 2002, pages 1 à 14

1. Introduction

L'objectif de la recherche rapportée ici est de développer une technique d'extraction d'information permettant d'évaluer comment des termes sont présentés dans des textes. Par *terme*, nous entendons les mots de la langue qu'ils s'agissent de verbes, de noms communs, d'adjectifs ou d'adverbes, mais aussi et surtout les noms propres, les noms de firmes, les marques commerciales,…

Nous nous intéressons spécifiquement à la manière positive ou négative avec laquelle un terme est présenté dans des textes. Tel produit, institution, entreprise, parti politique, personne publique est-il présenté d'une manière plutôt positive, favorable ou bien à l'inverse d'une manière plutôt négative, défavorable. Cette dimension évaluative retient depuis longtemps l'attention de nombreux chercheurs en psychologie. Elle correspond à la composante affective de toute "attitude" par rapport à un objet (Osgood *et al.*, 1957) . C'est également la dimension la plus importante pour structurer le champ sémantique des mots d'émotions (Russel, 1980). Plus généralement, elle est une des composantes de la signification connotative d'un mot où elle correspond à la dimension affective (ou émotionnelle), mais aussi et surtout à la dimension axiologique en reflétant "*un jugement d'appréciation, ou de dépréciation, porté sur l'objet dénoté* (Kerbrat-Orechioni, 1977, p.110)" (voir Kerbrat-Orechioni (1977) pour une discussion approfondie de la notion de connotation). La notion de connotation étant beaucoup plus riche que ce que nous étudions ici (et elle-même émotionnellement fort chargée en linguistique et en psychologie), nous utiliserons dans la suite l'expression "valence affective" pour référer à cette dimension évaluative. *Connotation* présente aussi l'inconvénient d'être souvent associé à l'idée de dimension idiosyncrasique de la signification alors que ce que nous désignons par "valence affective" est très largement partagé par les membres d'une communauté linguistique (Kerbrat-Orechioni, 1977 ; Osgood, 1962 ; Osgood *et al.*, 1957 ; Vandendorpe, 1993).

Il s'agit donc d'extraire du corpus le sens qui est donné à ces termes et plus particulièrement la valence affective qui leur est associée. Ce genre d'information, qu'on appelle souvent "analyse de l'image" est classiquement recueilli au moyen d'enquêtes de type sondage d'opinion en posant à un échantillon représentatif diverses questions destinées à évaluer leur attitude par rapport à la cible de l'étude. Si cette approche est très informative, elle est coûteuse en ressources et en temps et elle doit être complètement recommencée pour chaque nouvelle cible à évaluer. De plus, elle donne une image globale de la cible, mais ne permet pas de cerner comment certains médias présentent cette cible.

Le développement récent de techniques d'extraction d'information (Deerwester *et al.*, 1990 ; Kodratoff, 1999 ; Wilks, 1997) conjugué à la mise à disposition de grands corpus de textes sous forme électronique rend possible une autre voie : analyser comment une cible est présentée dans la presse. Bien sûr, cette deuxième approche n'est pas susceptible de remplacer les sondages, mais plutôt de les

"compléter" en apportant un deuxième point de vue sur l'image d'une cible : celui qui est transmis par les médias.

Le problème majeur que rencontre une telle approche réside dans la détermination de l'attitude de l'auteur du texte par rapport à la cible. Comme le soulignent Wilks (1997) et Das *et al.* (2001), les questions qui portent sur la manière dont un élément est présenté et évalué dans un texte (comme de décider si la critique d'un film ou un commentaire boursier est positif ou négatif) sont particulièrement complexes pour les techniques d'extraction d'information. Notre objectif n'est pas d'apporter une solution "exacte" à ce problème, mais, plus modestement, d'évaluer l'efficacité d'une approche partielle basée sur la combinaison de deux techniques : l'analyse sémantique latente et l'évaluation du contenu affectif d'un texte sur la base d'une partie des mots qui le composent.

2. Méthodologie proposée

La thèse centrale de ce travail est qu'il devrait être possible de déterminer la valence affective d'un terme en se basant sur les mots auxquels ce terme est fréquemment associé dans le corpus en question. La justification de cette approche se trouve dans l'analyse de la façon dont un signe (comme un mot) acquiert sa valence affective (Osgood *et al.*, 1957). Celle-ci peut-être le résultat des expériences de chacun avec l'objet dénoté, mais aussi des associations fréquentes avec d'autres mots dont la valence est connue. Ce mécanisme d'association est tout particulièrement actif pour les noms propres, les noms de sociétés et de marques commerciales (Kerbrat-Orechioni, 1977 ; Osgood *et al.*, 1957 ; Vandendorpe, 1993). On a ainsi pu montrer que le simple fait d'associer des mots négatifs comme "sale" ou "méchant" à des noms propres rend ceux-ci désagréables (Osgood *et al.*, 1957 ; Staats *et al.*, 1957).

Pour mettre en pratique cette approche, il est nécessaire de pouvoir déterminer les mots qui sont fréquemment associés aux cibles qu'on souhaite évaluer, mais aussi de connaître la valence affective de ces mots associés. Dans la suite de ce point, nous détaillerons comment ces deux conditions peuvent être remplies en commençant par la seconde parce qu'elle est la plus problématique.

2.1. *Intensité affective des voisins*

En effet, celle-ci semble nous entraîner dans un raisonnement circulaire : "*Pour déterminer la valence d'un terme, ... il est nécessaire de pouvoir déterminer la valence affective d'autres termes*" ? Il existe toutefois une solution proposée par Heise (1965) qui consiste à se baser, non sur tous les mots de la langue, mais sur un échantillon dont la valence affective a été déterminée par le recours à des juges. On sait en effet que la valence affective d'un mot peut être déterminée d'une manière très

fiable en demandant à des juges d'évaluer ce mot sur les échelles du différenciateur sémantique ou même plus simplement sur la seule dimension "agréable—désagréable" (Bestgen, 1994 ; Heise, 1965 ; Whissell *et al.*, 1986 ; voir aussi le point 3.1).

Une série de listes de mots évalués sur cette dimension ont été établies (Heise, 1965 ; Hogenraad *et al.*, 1995 ; Whissell *et al.*, 1986). Ces *dictionnaires de normes* (Heise, 1965) ont été principalement employés pour évaluer la valence affective de textes, mais aussi d'unités plus petites comme des phrases (Bestgen, 1994). La procédure proposée par Heise (1965) et ultérieurement développée par Anderson *et al.* (1982), Bestgen (1994), Hogenraad *et al.* (1989) et Whissell *et al.* (1986) est très simple. Dans un premier temps, on dresse la liste des mots différents et de leur fréquence dans l'unité textuelle. Cette liste est comparée à un dictionnaire qui contient un ensemble de mots dont on connaît la valence affective. A chaque fois qu'un mot se trouve dans les deux listes, on affecte la valeur indiquée dans le dictionnaire au mot du texte. Enfin, on calcule la moyenne des valeurs connues.

Malgré le caractère rudimentaire de cette technique, surtout marqué dans le fait qu'elle ne se base que sur les mots qui composent un texte, des arguments en faveur de sa validité ont pu être apportés (Anderson *et al.*, 1982 ; Bestgen, 1994 ; Whissell *et al.*, 1986). On a ainsi montré que les valeurs qu'elle attribue à des textes, ou à des passages de ceux-ci, permettent de prédire significativement la valence affective rapportée par des lecteurs lorsqu'on leur demande d'évaluer si la situation décrite par le texte est plutôt agréable ou désagréable. On a aussi montré l'efficacité de la technique pour comparer le style d'écrivains, pour déterminer l'impact de réécritures successives sur la qualité d'une oeuvre ou encore pour analyser des textes terroristes (Anderson *et al.*, 1986, 1989 ; Bestgen, 1994 ; Hogenraad *et al.*, 1995 ; Whissell *et al.*, 1986).

Ces résultats encourageants laisse penser qu'il doit être possible d'adapter cette technique pour répondre à la question qui nous intéresse. Il s'agirait donc de dresser la liste des mots fréquemment associés à un terme dont on veut connaître la valence affective, d'identifier dans cette liste les mots qui sont dans le dictionnaire et d'attribuer au terme-cible la valence moyenne de ceux-ci.

2.2. *Identification des voisins*

La deuxième composante de la technique doit permettre d'identifier les mots qui sont associés aux termes dont on veut déterminer la valence affective. Pour atteindre cet objectif, il semble logique de s'appuyer sur une analyse des cooccurrences entre les termes présents dans le corpus. L'approche la plus évidente, qui consiste à rechercher les mots qui cooccurrent le plus fréquemment avec un terme-cible, se heurte toutefois à un problème majeur. Même dans un grand corpus de textes, la plus grande partie des mots sont relativement rares. Il s'ensuit que les cooccurrences le sont encore plus. Leur rareté les rend particulièrement sensibles à des variations

aléatoires (Burgess *et al.*, 1998 ; Kintsch, 2001 ; Rajman *et al.*, 1997). Par exemple, on trouve dans le corpus analysé ci-dessous (plus de 24 000 articles parus dans un journal francophone) que *patiemment* et *sensuel* sont parmi les sept mots qui sont le plus associés à *fenêtre*. Ils lui sont plus associés, selon cet indice, que *vitre*, *façade* ou même *châssis*[1].

Cette difficulté peut être contournée en analysant, non chaque terme indépendamment, mais l'ensemble de la matrice de cooccurrences au moyen d'une procédure statistique qui produit une lissage des associations de manière à éliminer les éléments anecdotiques. L'analyse sémantique latente, une technique issue de travaux sur l'indexation automatique de documents, est tout particulièrement indiquée pour effectuer ce traitement (Deerwester *et al.*, 1990 ; Landauer *et al.*, 1998). Cette technique vise à construire un espace sémantique de très grandes dimensions à partir de l'analyse statistique de l'ensemble des cooccurrences dans un corpus de textes[2]. Comme le souligne Landauer *et al.* (1998), elle peut être vue de deux manières. A un niveau théorique, elle peut servir de base pour développer des simulations des processus psycholinguistiques à l'oeuvre lors de la compréhension du langage (Landauer *et al.*, 1997), incluant, par exemple, un "modèle computationnel" du traitement des métaphores (Bestgen *et al.*, 2002 ; Kintsch, 2000 ; Lemaire *et al.*, 2001), mais aussi l'analyse de la cohérence dans des textes (Foltz *et al.*, 1998 ; Lories *et al.*, 1998). A un niveau plus appliqué, c'est une technique permettant d'inférer et de représenter le sens de mots sur la base de leur usage dans des textes. C'est ce second aspect qui nous intéresse ici.

Le point de départ de l'analyse est un tableau lexical (Lebart *et al.*, 1992) qui contient le nombre d'occurrences de chaque mot dans chacun des documents, un document pouvant être un texte, un paragraphe ou même une phrase. Pour un grand corpus de textes découpés en de nombreux segments, cette étape produit une immense matrice clairsemée (*sparse*). Pour extraire les dimensions sémantiques, le tableau de fréquence fait l'objet d'une décomposition en valeurs singulières, une sorte d'analyse factorielle, qui en extrait les dimensions orthogonales les plus importantes[3]. On peut considérer que cet espace sémantique était implicite aux données et donc latent, ce qui donne son nom à la méthode.

Tant les mots que les segments originaux sont positionnés dans cet espace sémantique, ce qui permet de mesurer leur proximité. Plus précisément, le sens de chaque mot y est représenté par un vecteur. Pour mesurer la similarité sémantique

---

[1] Ces calculs prennent bien sûr en compte la fréquence de chacun de ces mots dans le corpus.

[2] Une description détaillée de la technique peut être trouvée dans les nombreux articles téléchargeables à l'adresse *http://LSA.colorado.edu/* ainsi qu'à l'adresse *http://www.upmf-grenoble.fr/sciedu/blemaire/lsa.html*.

[3] Contrairement à une analyse factorielle classique, les dimensions extraites sont très nombreuses (plusieurs centaines) et non interprétables. Elles peuvent toutefois être vues comme analogues aux traits sémantiques fréquemment postulés pour décrire le sens des mots (Landauer *et al.*, 1998).

entre deux mots, on calcule le cosinus entre les vecteurs qui les représentent. Plus deux mots sont sémantiquement proches, plus les deux vecteurs qui les représentent pointent dans la même direction et donc plus leur cosinus se rapproche de 1. Un cosinus de 0 indique une absence de similarité puisque les vecteurs correspondants sont orthogonaux. L'analyse des cosinus entre un terme et tous les autres termes nous permet donc de rechercher ses plus proches voisins.

3. Evaluation de la technique

Dans cette section, nous rapportons une étude préliminaire consistant en l'application de la méthode proposée à un corpus de plusieurs milliers d'articles publiés pendant un an dans un journal belge francophone. Cette présentation nous permettra d'indiquer concrètement les étapes et les traitements que la méthode nécessite et les décisions méthodologiques qui doivent être prises. Elle nous permettra aussi d'effectuer deux analyses afin de recueillir de premières indications à propos de son efficacité.

3.1. *Dictionnaire de valence*

Le dictionnaire employé pour cette recherche est composé de 3000 mots évalués sur la dimension agréable — désagréable (Hogenraad *et al.*, 1995). Un minimum de 30 juges (étudiants dans des établissements d'enseignement supérieur ou des écoles techniques) ont procédé à l'évaluation des termes sur une échelle à 7 points allant de *très désagréable* (1) à *très agréable* (7).

Le tableau 1 donne quelques exemples de mots, sélectionnés aléatoirement dans le dictionnaire, qui font partie de l'échantillon qui sera utilisé pour la validation de l'approche.

Mot	Valence	Mot	Valence
détresse	1.4	contrôlable	3.5
imbécile	1.4	outil	4.3
tristesse	1.6	risquer	4.5
hostilité	2.2	entier	4.9
impassible	2.6	revenir	5.0
superstitieux	2.8	admiratif	5.7
hâte	3.1	doux	6.0
ambigu	3.2	sincérité	6.1

Tableau 1. *Valences affectives de quelques mots sur une échelle allant de* très désagréable *(1.0) à* très agréable *(7.0).*

La fidélité de ce genre de normes a été mesurée en obtenant 4 années plus tard des évaluations pour un sous-échantillon de 413 mots (Bestgen, 1994). L'accord inter-juges, mesuré par le coefficient thêta (Armor, 1974), est de 0.94. De plus, la corrélation entre les valeurs obtenues dans cette seconde étude et les normes originales est de 0.93. Enfin, la comparaison de ces normes obtenues auprès d'étudiants à des normes établies sur la base de jugements formulés par des personnes âgées de 17 à 90 ans souligne leur stabilité (Messina *et al.*, 1989).

3.2. *Constitution de l'espace sémantique*

Le corpus de textes utilisés pour construire l'espace sémantique est constitué de tous les articles parus dans le journal *Le Soir*, quotidien belge francophone à grande diffusion et à visée généraliste, durant l'année 2000 et disponible sur un CD-Rom, soit plus de 20 000 000 de mots.

Dans un premier temps, les mots ont été lemmatisés dans le but de réduire le nombre de formes graphiques différentes. Cette étape a été effectuée en comparant les formes graphiques présentes dans le corpus avec une liste de formes fléchies et des lemmes correspondants. La procédure actuellement employée est totalement automatisée et ne prend pas en compte le contexte pour lever certaines ambiguïtés. Ainsi, dans le cas de formes graphiques qui peuvent correspondre à plusieurs lemmes, on procède en affectant systématiquement la forme graphique au lemme le plus fréquent selon la base de données Brulex (Content *et al.*, 1990).

Dans un second temps, les formes fonctionnelles (pronoms, articles,…) et une trentaine de termes introduits dans la base de données pour indexer les articles (comme *éditorial*, *entretien*, *interview*, *légende*) ont été supprimés. Cette opération a réduit la taille du corpus à un peu moins de 12 000 000 de mots.

Le corpus a été segmenté en fonction des articles. La taille de ceux-ci variait très fortement : entre 29 et 4 429 mots. Afin d'homogénéiser un peu le corpus, les articles très courts (moins de 120 mots) et les articles très longs (plus de 570 mots), qui présentent le risque de ne pas être homogènes au niveau du contenu, ont été supprimés. Ces limites ont été définies sur la base d'une analyse de la distribution de fréquence des longueurs d'articles afin d'éliminer les 10% d'articles les plus courts et les 10% les plus longs. On a ainsi obtenu 24 776 documents.

Tous les mots dont la fréquence dans le corpus était inférieure à 5 ont été supprimés, faisant ainsi passer le nombre de mots différents de 168 305 (dont un grand nombre de noms propres) à 34 113. La matrice de cooccurrences des 34 113 termes dans les 24 776 documents a été soumise à une décomposition en valeurs singulières réalisée par le programme Svdpack (Berry, 1992) et les 295 premiers vecteurs propres ont été conservés, c'est-à-dire tous ceux que le programme a trouvés en 333 itérations.

3.3. *Analyses et résultats*

Deux analyses ont été menées afin d'estimer l'efficacité de la méthode proposée. La première est une épreuve de validation visant à vérifier qu'il est bien possible d'évaluer la valence affective de termes en se basant sur les mots qui leur sont associés dans des textes. La seconde analyse se veut illustratrice. Elle consiste en l'application de la technique à quelques termes potentiellement intéressants afin d'illustrer la faisabilité et l'intérêt de l'approche. Nous avons sélectionné huit noms de banques actives en Belgique afin de déterminer leur valence affective dans les articles du *Soir*. L'une de celle-ci étant au centre d'une importante affaire de fraude fiscale, il s'agira de s'assurer qu'elle reçoit bien le score d'évaluation le plus négatif.

3.3.1. ***Epreuve de validation***

La logique du test de validation est la suivante. Si la méthode est efficace, il devrait être possible d'estimer la valence affective d'un mot sur la base de ses proches voisins dont on connaît la valence affective. Nous avons donc sélectionné aléatoirement un échantillon de 30 mots dans le dictionnaire de normes et nous avons calculé le cosinus entre chacun de ces mots et les 34 113 mots inclus dans l'espace sémantique. Nous avons ensuite recherché, toujours pour chaque mot-cible, les plus proches voisins dont nous connaissions le score d'évaluation (en ne prenant bien sûr pas en compte le mot cible lui-même). Enfin, la moyenne des scores d'évaluation des voisins a été calculée pour servir d'estimateur de l'évaluation du mot. Ces estimations ont été comparées aux valeurs données dans le dictionnaire au moyen du coefficient de corrélation de Pearson (Howell, 1998).

Pour mettre en pratique cette procédure, il est nécessaire de fixer un paramètre : le nombre de plus proches voisins (*ppv*) utilisés pour estimer le score d'un mot. Deux analyses ont été réalisées en fixant ce paramètre à 5 et à 30. Les corrélations entre les estimations et les scores réels sont très élevées : $r = 0.69$ pour $ppv = 5$ et $r = 0.70$ pour $ppv = 30$. Ces valeurs sont statistiquement très significatives ($p < 0.0001$).

Afin de confirmer ces résultats, un deuxième échantillon de 30 mots a été sélectionné aléatoirement dans le dictionnaire de norme et a été traité par la même procédure. La corrélation entre les estimations et les scores réels pour $ppv = 30$ est équivalente à celle observée pour le premier puisqu'elle est égale à 0.67. Elle est plus faible pour $ppv = 5$ ($r = 0.56$), mais toujours très significative ($p < 0.002$).

La figure 1 présente un diagramme de dispersion (Howell, 1998) pour les deux échantillons et *ppv* égale à 30. On y a présenté les valeurs pour le dictionnaire en ordonnée et les valeurs prédites par la technique en abscisse. Dans le graphe idéal, correspondant à une corrélation de 1, tous les points tomberaient sur la droite. Dans la pire des cas, les points se présenteraient sous la forme d'un cercle plein. La présence d'un ovale bien allongé confirme la corrélation très élevée, mais inférieure à 1.

L'efficacité globale de la procédure ne doit pas occulter la présence d'estimations inadéquates qui se marquent par l'existence de points fort éloignés de la droite. Ces points soulignent pour une part les limitations de l'implémentation effectuée ici. C'est par exemple le cas du mot *ennuyer* qui selon le dictionnaire doit avoir un score de 2.2, soit désagréable, alors que la procédure lui attribue un score de 5.0, soit plutôt agréable. Ce score excessif est provoqué par la présence parmi les plus proches voisins d'*ennuyer* de mots comme *aimer, envie*, *plaire*, *heureux, bonheur,* et *adorer* dont l'origine vient pour une part de ce que le terme *ennuyer* est souvent nier dans le journal : *on ne s'est pas ennuyé*. L'analyse sémantique latente n'ayant accès à aucune information syntaxique, cette utilisation négative n'est pas prise en compte.

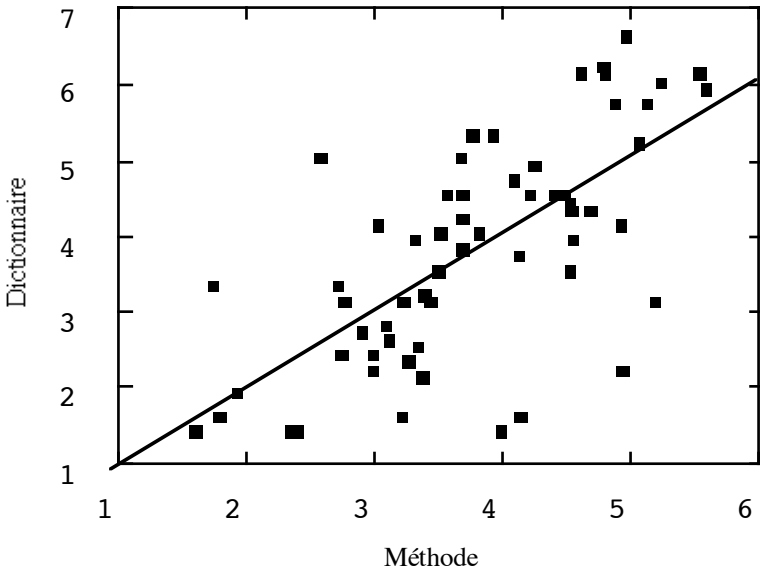

Figure 1. *Diagramme de dispersion pour les deux échantillons (60 mots). La valence selon le dictionnaire est en ordonnée et celle prédite par l'analyse est en abscisse.*

Toutefois, les écarts peuvent aussi provenir de ce que l'usage d'un terme dans la base de textes ne correspond pas à celui qui a été pris en compte par les évaluateurs lors de la construction du dictionnaire. Par exemple, le mot *justice* est évalué dans le dictionnaire comme très positif, sa valeur étant de 5.4. L'estimation à partir du corpus, calculée avec *ppv* = 30, n'est que de 3.0. On peut penser que ceci résulte du fait que les personnes ont évalué la *justice* en tant que *le fait d'être juste*, alors que, dans le journal, ce terme est aussi employé pour désigner *la justice en tant*

qu'institution et que les difficultés que celle-ci rencontre pour être juste y sont largement discutées. Dans ce genre de cas, il est loin d'être évident que ce soit l'estimation par la technique qui est incorrecte pour l'objectif que nous visons.

3.3.2. ***Analyse exploratoire***

La seconde analyse se veut illustratrice. Elle consiste en l'application de la technique à quelques termes potentiellement intéressants afin d'illustrer la faisabilité de l'approche. Nous avons sélectionné huit noms de banques actives en Belgique afin de déterminer leur valence affective dans les articles du *Soir* de l'année 2000. L'une d'entre elle, la *Kb-Lux* était durant cette année au centre d'une importante affaire de fraude fiscale. On peut donc s'attendre à ce que sa valence affective soit nettement plus faible que celle des autres banques. Comme dans l'analyse précédente, deux valeurs pour *ppv* ont été employées : 5 et 30.

Comme on peut le voir dans le tableau 2, la *Kb-Lux* reçoit bien un score nettement inférieur à celui des autres banques et ce dans les deux solutions.

	ppv = 5	ppv = 30
Kb-Lux	2.8	3.3
Argenta	3.7	4.2
Artesia	4.2	4.2
Bbl	3.9	4.3
Dexia	3.8	4.1
Fortis	3.4	4.2
Kbc	3.9	4.2
Paribas	3.8	3.9

Tableau 2. *Valences affectives de huit noms de banques sur une échelle allant de très désagréable (1.0) à très agréable (7.0).*

Lorsqu'on compare les scores pour les deux valeurs du paramètres *ppv*, on observe que la valence d'une des huit banques change très nettement : celle pour *Fortis* qui passe de 3.4 à 4.2. Une analyse des mots associés qui ont provoqué cette modification montre que le faible score de *Fortis* pour *ppv* = 5 est produit par le mot *déprécier* (valence = 2.5) qui est associé à tous les noms de banques lorsqu'on prend les 30 plus proches voisins, mais ne l'est qu'à *Fortis* lorsqu'on prend les 5 plus proches voisins.

Une autre remarque s'impose : le fait que la procédure ne détecte que peu de différences entre les banques, à l'exception du cas spécifique de *Kb-Lux* Il est impossible de dire s'il s'agit là d'un défaut de la technique ou si simplement ces banques sont présentées par le journal d'une manière très similaire. Toutefois, on peut penser que cette équivalence est largement provoquée par le fait que les termes

utilisés sont très liés (ce sont tous des noms de banques) et qu'ils sont donc proches des mêmes voisins. Il pourrait donc être intéressant d'envisager d'autres manières de calculer les scores de valence par exemple en privilégiant les mots qui sont spécifiquement associés à chacun des termes que l'on veut comparer.

4. Discussion et conclusion

Comme nous l'avons indiqué, la présente recherche ne se veut qu'un premier pas dans le développement d'une technique permettant d'estimer la valence affective de termes employés dans un corpus de textes. Les résultats sont encourageants d'autant plus que de nombreuses améliorations sont possibles et mériteraient d'être testées. En guise de conclusion, nous en indiquerons quelques unes.

Les données soumises à l'analyse sémantique latente consiste en une matrice de cooccurrences de chaînes graphiques (rarement lemmatisées) dans des segments de textes. Il s'ensuit que ni l'ordre des mots et donc la structure syntaxique, ni la polysémie de certaines chaînes graphiques ne sont pris en compte. On rappellera que, malgré ces limitations, l'analyse sémantique latente a permis de développer des simulations des processus psycholinguistiques à l'oeuvre lors de la compréhension du langage (Landauer *et al.*, 1997, 1998). Il semble néanmoins nécessaire de confronter cette technique à d'autres procédures qui prennent en compte soit indirectement, soit plus directement des informations syntaxiques. Dans le premier cas, il s'agirait par exemple d'inclure dans le calcul des cooccurrences un facteur reflétant la proximité des mots dans le texte (Burgess *et al.*, 1998). Le second cas offre, selon nous, une voie encore plus prometteuse. Procéder à une analyse syntaxique même partielle du texte permettrait d'appliquer directement les techniques développées pour estimer la valence affective de phrases (Bestgen, 1994) en dépassant enfin la limitation induite par le traitement d'une phrase ou d'un paragraphe comme un "sac" de mots (Bestgen, 1994 ; Hogenraad *et al.*, 1989 ; Lebart *et al.*, 1992).

Si on souhaite dépasser le caractère exploratoire des essais décrits ici, il semble également nécessaire de prétraiter linguistiquement le corpus avant d'effectuer les autres analyses. Par exemple, différencier automatiquement les noms propres des noms communs est nécessaire si on veut pouvoir étudier une série de noms de firmes ou de produits qui désignent aussi des objets courants ou des couleurs. De même, il serait intéressant de pouvoir identifier lors du prétraitement du corpus des termes constitués de plusieurs mots comme par exemple *banque centrale européenne* afin de pouvoir analyser la valence affective de ces entités spécifiques. Rechercher simplement les segments répétés (au sens de Lebart *et al.* (1992), c'est-à-dire n'importe quelle suite de plusieurs mots qui s'observent plusieurs fois dans le corpus) ne serait pas très efficace car cela générerait un trop grand nombre de segments peu pertinentes. Ceux-ci pourraient toutefois être filtrés par une procédure prenant en compte la catégorie morpho-syntaxique des constituants afin de

privilégier ceux qui ont le plus de chance de correspondre à un terme intéressant, par exemple les séquences *Nom—Préposition—Nom* ou *Nom—Adj.—Adj.* (Feldman *et al.*, 1998). Finalement, on devrait aussi considérer les bénéfices que pourrait apporter la participation à l'analyse d'un expert qui vérifierait la liste des plus proches voisins d'un terme afin, par exemple, de prendre en compte la polysémie de certains mots associés.

On n'insistera donc jamais assez sur le caractère exploratoire de la présente étude. D'autres recherches sont nécessaires afin de tester les développements proposés ci-dessus.

Remerciements

Yves Bestgen est chercheur qualifié du Fonds national belge de la recherche scientifique. Cette recherche a bénéficié du soutien du Fonds de la recherche fondamentale collective.

5. Bibliographie

Anderson C.W., McMaster G.E., « Computer assisted modeling of affective tone in written documents », *Computers and the Humanities*, Vol. 16, 1982, p. 1-9.

Anderson C.W., McMaster G.E., « Modeling emotional tone in stories using tension levels and categorical states », *Computers and the Humanities*, Vol. 20, 1986, p. 3-9.

Anderson C.W., McMaster G.E., « Quantification of rewriting by the brothers Grimm : A comparison of successive versions of three tales », *Computers and the Humanities*, Vol. 23, 1989, p. 341-346.

Armor D.J., « Theta reliability and factor scaling », In H.L. Costner (Ed.), *Sociological methodology, 1973-1974*, Jossey-Bass, 1974, p.17-50.

Berry M.W., « Large scale singular value computation », *International Journal of Supercomputer Application*, Vol. 6, 1992, p. 13-49.

Bestgen Y., « Can emotional valence in stories be determined from words ? », *Cognition and Emotion*, Vol. 8, 1994, p. 21-36.

Bestgen Y., Cabiaux A.F., « L'analyse sémantique latente et l'identification des métaphores », *Actes de la 9ème conférence sur le Traitement automatique des langues Naturelles*, TALN '02, Le Chesnay, Inria, 2002, p.331-337.

Burgess C., Livesay K., Lund K., « Explorations in Context Space : Words, Sentences, Discourse », *Discourse Processes*, Vol. 25, 1998, p. 211-257.

Content A., Mousty P., Radeau M., « Brulex : une base de données lexicales informatisée pour le français écrit et parlé », *L'Année Psychologique*, Vol. 90, 1990, p. 551-566.

Das S., Chen M., Yahoo! for Amazon : Opinion extraction from small talk on the web, Working Paper (under review), Décenbre 2001, Santa Clara University.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R., « Indexing by Latent Semantic Analysis », *Journal of the American Society for Information Science*, Vol. 41, 1990, p. 391-407.

Feldman R., Fresko M., Kinar Y., Lindell Y., Liphstat O., Rajman M., Schler Y., Zamir O., « Text Mining at the Term Level », *PKDD 1998*, 1998, p. 65-73.

Foltz P.W., Kintsch W., Landauer T.K., « The measurement of textual coherence with Latent Semantic Analysis », *Discourse Processes*, Vol. 25, 1998, p. 285-307.

Heise D.R., « Semantic differential profiles for 1000 most frequent english words », *Psychological Monographs*, Vol. 79, 1965, p. 1-31.

Hogenraad R., Bestgen Y., « On the thread of discourse : Homogeneity, trends, and rhytms in texts », *Empirical Studies of the Arts*, Vol. 7, 1989, p. 1-22.

Hogenraad R., Bestgen Y., Nysten J.L., « Terrorist Rhetoric : Texture and Architecture », In Nissan et Schmidt (Eds.), *From Information to Knowledge*, Intellect, 1995, p. 48-59.

Hogenraad R., Daubies C., Bestgen Y., Une théorie et une méthode générale d'analyse textuelle assistée par ordinateur. Le système PROTAN, Manuel, 1995, Université catholique de louvain.

Howell D., *Méthodes statistiques en sciences humaines*, Adaptation française de M. Roegiers, V. Yzerbyt et Y. Bestgen), Paris, De Boeck Université, 1998.

Kerbrat-Orecchioni C., *La connotation*, Lyon, Presses universitaires de Lyon, 1977.

Kintsch W., « Metaphor comprehension : A computational theory », *Psychonomic Bulletin and Review*, Vol. 7, 2000, p. 257-266.

Kintsch W., « Predication », *Cognitive Science*, Vol. 25, 2001, p. 173-202.

Kodratoff Y., « Knowledge Discovery in Texts : A Definition, and Applications », in Ras et Skowron (Eds.), *Foundation of Intelligent Systems*, LNAI 1609, Springer, 1999.

Landauer T.K., Dumais S.T., « A solution to Plato's problem : the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge », *Psychological Review*, Vol. 104, 1997, p. 211–240.

Landauer T.K., Foltz P.W., Laham D., « An introduction to Latent Semantic Analysis », *Discourse Processes*, Vol. 25, 1998, p. 259-284.

Lebart, L., Salem A., *Statistique textuelle*, Paris, Dunod, 1992.

Lemaire B., Bianco M., Sylvestre E., Noveck I., « Un modèle de compréhension de textes fondé sur l'analyse de la sémantique latente », In H. Paugam Moisy, V. Nyckees, J. Caron-Pargue (Eds.), *La cognition entre individu et société* (actes du colloque de l'ARCo), Hermès, 2001, p. 309-320.

Lories G., Bestgen Y., Possibilités et limites de l'analyse sémantique latente. Réunion du groupe de contact FNRS pour l'étude des processus cognitifs : From text to

representation, Novembre 1998, Université catholique de Louvain.

Messina D., Morais J., Cantraine F., « Valeur affective de 904 mots de la langue française », *C.P.C. : European Bulletin of Cognitive Psychology*, Vol. 9, 1989, p. 165-187.

Osgood C.E., « Studies on the generality of affective meaning systems », *American Psychologist*, Vol. 17, 1962, p. 10-28.

Osgood C.E., Suci G.J., Tannenbaum P.H., *The measurement of meaning*, Urbana, University of Illinois Press, 1957.

Rajman M., Besançon R., « Text Mining : Natural Language Techniques and Text Mining Applications », *Proceedings of the seventh IFIP 2.6 Working Conference on Database Semantics*, Chapam & Hall, 1997.

Russel J. A., « A circumplex model of affect », *Journal of Personality and Social Psychology*, Vol. 39, 1980, p. 1161-1178.

Staats C.K., Staats A.W., « Meaning established by classical conditioning », *Journal of Experimental Psychology*, Vol. 54, 1957, p. 74-80.

Vandendorpe C., « Quelques considérations sur le nom propre », *Langage et Société*, Vol. 66, 1993, p. 63-75.

Whissell C.M., Fournier M., Pelland R., Weir D., Makarec K., « A dictionary of affect in language : IV. Reliability, validity, and applications », *Perceptual and Motor Skills*, Vol. 62, 1986, p. 875-888.

Wilks Y., « Information Extraction as a Core Language Technology », in M. T. Paziensa (Ed.) *Information Extraction*, Springer, 1997, p. 1-9.