Une procédure automatique pour étendre des normes lexicales par l'analyse des cooccurrences dans des textes

Nadja Vincze * — Yves Bestgen **

** Université catholique de Louvain – CENTAL*
1, Place Blaise-Pascal – 1348 Louvain-la-Neuve
nadja.vincze@uclouvain.be

*** Chercheur qualifié du F.R.S. – Université catholique de Louvain – CECL*
10, Place Cardinal-Mercier – 1348 Louvain-la-Neuve
yves.bestgen@psp.ucl.ac.be

RÉSUMÉ. *Tant dans le domaine de la psychologie que dans celui du traitement automatique des langues, les normes portant sur des propriétés sémantiques, comme le caractère concret ou abstrait, la polarité ou le caractère émotionnel, constituent des ressources importantes. La construction manuelle de ces normes, par l'intermédiaire d'évaluateurs, est coûteuse, d'où l'intérêt de développer des méthodes de construction ou d'extension automatique. Plusieurs méthodes ont été proposées, mais elles portent sur une seule dimension : la polarité. Nous proposons de voir dans quelle mesure l'une d'entre elles peut être étendue à six autres normes, et ce pour le français et l'espagnol. Les expérimentations confirment l'efficacité de la technique non seulement pour étendre une norme, mais également pour mettre en évidence des mots pour lesquels les valeurs attribuées par les évaluateurs sont sujettes à caution.*

ABSTRACT. *Both in the field of psychology and in natural language processing, norms related to semantic properties, such as concreteness, polarity or emotionality, are important resources. The manual construction of these norms, by asking participants to rate the words, is expensive, hence the need to develop automatic methods of construction or extension. Several methods have been proposed, but they focus on only one dimension: polarity. We propose to determine whether one of these methods can be extended to six other norms, for French and Spanish. The experiments confirm the effectiveness of the technique, not only to extend a norm, but also to highlight the words for which the values that were assigned by the raters seem unreliable.*

MOTS-CLÉS : *normes lexicales, psycholinguistique, fouille d'opinion, analyse sémantique latente, corpus et collection de textes.*

KEYWORDS: *lexical norms, psycholinguistic, opinion mining, latent semantic analysis, corpus, text collections.*

TAL. Volume 52 – n° 3/2011, pages 191 à 216

1. Introduction

Depuis plus de 40 ans, de nombreuses études ont été menées pour récolter des normes à propos de propriétés formelles et sémantiques de mots, comme la fréquence d'emploi, la régularité orthographique, la familiarité, l'âge d'acquisition, le degré d'abstraction ou encore le caractère émotionnel. Certaines de ces propriétés peuvent être aisément récoltées au travers de procédures automatiques de comptage appliquées soit à des ressources linguistiques, comme des dictionnaires, soit à des corpus. D'autres propriétés, comme la familiarité ou le caractère plus ou moins agréable ou positif d'un mot[1], sont obtenues en demandant à des participants, souvent plusieurs dizaines, d'évaluer les mots par exemple sur une échelle en 7 points allant de *ce mot évoque une idée très désagréable* à *ce mot évoque une idée très agréable*. Dans ce genre d'études, chaque participant juge plusieurs centaines de mots, présentés le plus souvent sous la forme d'un questionnaire papier/crayon, sur lequel chaque mot est accompagné d'une échelle pour recueillir leurs réponses. Plusieurs ordres de présentation des mots sont employés afin de contrôler le sentiment de routine et la fatigue qu'une telle tâche ne peut que générer. L'importance de ces normes se marque dans la publication d'articles dont le seul objectif est de les recenser (Proctor et Vu, 1999) et dans la constitution de larges bases de données qui rassemblent le plus grand nombre possible de propriétés issues de différentes études comme le *MRC Psycholinguistic Database* (Wilson, 1988) pour l'anglais ou la base *Lexique 3*[2] pour le français (New, 2006).

Le tableau 1 présente de telles normes pour 6 mots choisis de manière à couvrir la diversité des scores sur ces normes. La norme de fréquence est issue de la base *Lexique 3* et correspond à la fréquence pour un million de mots dans un corpus de textes littéraires. Toutes les autres normes ont été recueillies en demandant l'avis de personnes et les valeurs reproduites dans le tableau correspondent à l'avis moyen de celles-ci. Pour la norme d'imagerie (IMA), Hogenraad et Orianne (1981) ont employé une échelle à 7 degrés d'intensité allant de *évoquant peu d'images* (1) à *évoquant beaucoup d'images* (7) pour obtenir l'avis de 24 personnes à propos de 1 130 des noms communs. La norme de polarité (POL) est composée de 3 252 mots, chacun évalué par une trentaine d'évaluateurs sur une échelle à 7 points selon que le mot évoque une idée allant de *très désagréable* (1) à *très agréable* (7) (Hogenraad *et al.*, 1995). La familiarité (FAM) et l'âge d'acquisition (AGE) ont été recueillis par Lachaud (2007). Pour la familiarité une échelle allant de – 3 à + 3 a été employée sur laquelle le participant devait indiquer dans quelle mesure le mot présenté lui était familier. Enfin, l'âge d'acquisition a été obtenu en demandant à la personne d'indiquer l'âge en années auquel elle pense avoir appris ce mot. Ces deux dernières

1. Cette dimension est appelée *valence* ou *évaluation* en psycholinguistique, *polarité* ou *orientation sémantique* en traitement automatique des langues.
2. Disponible à l'adresse http://www.lexique.org (consulté le 11 octobre 2011).

normes sont disponibles pour 1 225 mots dont des noms, mais aussi des verbes (*perdre*), des adjectifs (*gras*), des adverbes (*très*) et même des noms propres (*Inde*), chacun évalué par une vingtaine de personnes.

MOTS	**FREQ.**	**IMA**	**POL**	**FAM**	**AGE**
chance	136	3,1	6,4	2,8	5,5
chat	131	6,4	4,2	2,9	3,5
danse	35	6,1	5,6	2,5	5,0
flux	6	3,0	3,7	2,5	11,1
heure	924	4,0	3,9	3,0	5,8
honte	83	2,8	2,2	2,8	6,9

Tableau 1. *Extrait de normes lexicales : Freq = fréquence (pour un million de mots), IMA = imagerie, POL = polarité, FAM = familiarité, AGE = âge d'acquisition (en années)*

Suivant la discipline scientifique et la nature de la dimension étudiée, ces normes lexicales ont des usages très différents. En traitement automatique des langues, de telles normes sont tout particulièrement employées dans le champ de la fouille d'opinion. En effet, les domaines d'applications qui peuvent tirer profit d'une détection de la subjectivité ou de la polarité sont nombreux et variés, allant des systèmes de recommandations (Sun et *al.*, 2010), à la veille de produits (Popescu et Etzioni, 2005), en passant par les systèmes de questions-réponses (Somasundaran et *al.*, 2007). Les informations à propos du degré d'abstraction d'un mot sont utilisées pour filtrer des listes de candidats mots-clés (Da Sylva, 2010) ou pour identifier les différents sens d'un mot afin de les représenter de manière visuelle (Saenko et Darrell, 2009). En psycholinguistique, le champ où elles sont le plus employées, elles servent principalement lors de la sélection d'un matériel expérimental ou dans des études corrélationnelles visant à identifier les propriétés des mots qui affectent l'efficacité des processus cognitifs à l'œuvre lors de leur traitement (Bonin *et al.*, 2004 ; Kousta *et al.*, 2011). Strain *et al.* (1995) ont ainsi comparé le temps nécessaire pour prononcer des mots classés en huit catégories selon les trois critères dichotomiques suivant : mots fréquents ou rares, mots fortement ou faiblement imagés et mots dont la prononciation répond aux règles standard de conversion graphème/phonème ou non (ex. *oiseau* et *oignon*). Ils ont montré que le degré d'imagerie d'un mot affectait la vitesse de prononciation, mais seulement pour les mots rares et phonologiquement irréguliers. Cette observation, ultérieurement reproduite et étendue aux effets d'une quatrième variable, l'âge d'acquisition, a permis de mieux spécifier les contraintes qui pèsent sur les modèles cognitifs de l'accès lexical (Shibahara *et al.*, 2003). De telles normes sont également pertinentes pour le développement de tests de compétence en lecture (Desrochers et Saint-

Aubin, 2008) ainsi que pour l'analyse de textes d'apprenants d'une langue étrangère (Dewaele et Pavlenko, 2002). Dans des approches plus cliniques, elles servent à caractériser le déficit de personnes présentant des troubles de la personnalité (Pennebaker *et al.*, 2003) ou victimes d'atteintes neurologiques. Par exemple, Franklin *et al.* (1994) ont souligné les difficultés de compréhension orale spécifiques aux mots abstraits et peu imaginables de leur patient DRB victime d'un accident vasculaire cérébral. Dans ce genre d'études, il importe de disposer de normes qui couvrent autant les adjectifs que les noms ou encore les verbes puisque l'effet d'une propriété sémantique peut varier selon la catégorie grammaticale.

Le développement de telles normes, surtout lorsqu'il nécessite le recours à des évaluations subjectives par plusieurs dizaines de personnes de chaque mot, est extrêmement coûteux en temps et en ressources, ce qui réduit fortement le nombre de mots qui les composent. Or, des normes couvrant de très nombreux mots sont indispensables. Dans la recherche citée ci-dessus, Strain *et al.* (1995, p. 1141) ont ainsi dû étendre les normes disponibles parce qu'ils ne disposaient pas d'assez de mots dans chacune des 8 classes nécessaires à leur expérience. Le tableau 1 peut être employé pour illustrer le problème. On y remarque que la familiarité, dont Gernsbacher (1984) prétend qu'elle est une meilleure mesure de la disponibilité d'un mot que la fréquence dans des corpus, y est toujours très élevée. La raison n'est pas que les normes de Lachaud (2007) ne contiennent pas de mots peu familiers puisqu'on y trouve *dais* (– 1,09) ou *yole* (– 0,21), mais que les autres normes ne contiennent que des mots familiers. Il s'ensuit qu'étudier l'impact de la familiarité en contrôlant celui de la polarité est impossible sur la base de ces normes. Pouvoir les étendre automatiquement ouvrirait donc de nouvelles perspectives.

L'objectif de la recherche présentée ici est de déterminer si une méthode que nous avons développée récemment pour étendre une norme de polarité en français (Vincze et Bestgen, 2011) peut être appliquée à d'autres normes, comme le degré d'abstraction des mots, leur caractère imagé ou le fait qu'ils évoquent plus ou moins fortement une réaction émotionnelle, ainsi qu'à d'autres langues. Pour cela, nous avons choisi, en plus du français, l'espagnol. Dans l'esprit de ce numéro spécial sur les ressources linguistiques libres, la technique proposée présente l'intérêt d'être totalement automatisée et assez simple à mettre en pratique. Elle ne nécessite qu'une collection de textes de quelques millions de mots, qui peut être facilement constituée à partir de sites Web, et une norme disponible pour quelques centaines de mots. Sur cette base, il est possible d'assigner une valeur sur cette norme à n'importe quel mot apparaissant une dizaine de fois au moins dans la collection de textes. De plus, la mise à disposition de la communauté scientifique de ressources ainsi constituées évite un gaspillage de coût et de temps, facilite la reproduction (et l'extension) de recherches fondées sur celles-ci et peut même engendrer une dynamique participative, chacun pouvant améliorer la qualité de la ressource ou en enrichir la couverture. Pour ces raisons, les normes issues de cette recherche sont mises à disposition des chercheurs (voir fichiers annexés à la présente soumission).

Après une présentation des travaux en fouille d'opinion qui ont ouvert la voie au développement automatique de normes lexicales, la section 3 décrit la méthode employée dans le cadre de cette étude. Le matériel utilisé pour évaluer son efficacité et l'ensemble des résultats pour les deux langues étudiées sont présentés dans la section 4. La conclusion souligne les différents emplois possibles de la méthode et présente quelques développements envisagés.

2. Travaux antérieurs

Depuis une dizaine d'années, la catégorisation automatique de textes, un champ très dynamique du traitement automatique des langues, s'est ouverte à la classification de documents selon leur caractère subjectif et objectif (Wiebe *et al.*, 2004) et selon leur orientation sémantique (Pang et Lee, 2008), avec une classification binaire positif/négatif ou multiclasse, selon le degré de polarité (Pang *et al.*, 2002). De nombreuses méthodes proposées pour réaliser ces tâches nécessitent des lexiques où à chaque entrée sont associés une polarité ou un degré de polarité. Ces méthodes calculent l'orientation sémantique d'un document selon les orientations de mots ou de groupes de mots qui le composent, éventuellement modifiées par des structures syntaxiques (comme la négation, les modificateurs adverbiaux, etc.). Bien entendu, plus la couverture du lexique est large, plus le nombre de mots identifiés dans les textes est élevé, ce qui augmente la précision de la détermination de l'orientation sémantique.

La constitution manuelle de ces ressources étant lente et surtout coûteuse, des méthodes de construction automatique ont vu le jour, parmi lesquelles on peut distinguer deux approches : celle fondée sur des ressources linguistiques et celle fondée sur des corpus de textes. Les approches qui s'appuient sur des ressources linguistiques, comme des dictionnaires ou des thésaurus, calculent généralement la similarité entre les mots à partir de leur relation de synonymie. Elles procèdent en partant de quelques mots dont la polarité est connue et en lançant un algorithme d'amorçage (*bootstrapping*) qui parcourt les liens synonymiques et antonymiques de la base, en attribuant la même orientation aux mots synonymes et *vice versa* (Kamps et Marx, 2002 ; Kim et Hovy, 2004). Kamps et Marx (2002) ont probablement été les premiers à proposer une telle procédure en dérivant de WordNet (Miller *et al.*, 1990) un graphe dans lequel chaque nœud représente un terme et un lien est présent entre deux nœuds s'ils sont synonymes. À partir de ce graphe, ils mesurent la distance minimale relative entre les nœuds dont la polarité est à déterminer et ceux correspondant à *good* et *bad*. La limitation principale de cette méthode est qu'elle ne s'applique qu'aux adjectifs. Esuli et Sebastiani (2006) ont étendu cette approche pour développer SentiWordNet et ont proposé une méthode fondée également sur WordNet, qui assigne à chaque synset trois valeurs (positive, négative et neutre) sur la base d'une procédure d'apprentissage semi-supervisée. Celle-ci consiste tout d'abord à étendre par synonymie deux ensembles de mots germes et à entraîner dans

un second temps un modèle de classification sur la base de leurs définitions, converties en formes vectorielles.

Les approches qui s'appuient sur des corpus de textes calculent les similarités entre les mots différemment, ne disposant pas d'informations sur leurs liens synonymiques. Hatzivassiloglou et McKeown (1997) ont proposé un algorithme capable de déterminer l'orientation sémantique d'adjectifs à partir de l'analyse de leurs cooccurrences avec des conjonctions. Plus récemment, de Marneffe *et al.* (2010) ont proposé d'apprendre automatiquement les échelles sous-jacentes à des ensembles d'adjectifs graduables par l'entremise de l'analyse de la fréquence de ces adjectifs dans un corpus de critiques étiquetées selon leur polarité, obtenant des résultats au moins équivalents à ceux d'approches fondées sur des ressources linguistiques. D'autres chercheurs emploient des modèles d'espaces vectoriels qui représentent le contenu sémantique des mots. Ces espaces sont construits à partir de l'analyse statistique des contextes de cooccurrences des mots ; la similarité entre les mots du corpus est alors représentée par leur proximité. Il existe deux grands types de modèles (Baroni et Lenci, 2010) : les modèles structurés et les modèles non structurés. Ces derniers représentent les relations de cooccurrences entre un élément et un contexte. Le contexte est défini soit par des documents (Deerwester *et al.*, 1990) soit par des fenêtres d'un certain nombre de mots autour de la cible (Schütze, 1997). Les modèles structurés prennent en compte les structures syntaxiques entre la cible et les éléments du contexte (Padó et Lapata, 2007 ; Baroni et Lenci, 2008). Un mot ne va être considéré comme contexte que s'il est lié à la cible par une relation lexico-syntactique pertinente. Ce sont surtout les méthodes non structurées qui ont été utilisées pour la détection de la polarité en fouille d'opinion, et notamment l'analyse sémantique latente (ASL, Latente Semantic Analysis, Deerwester *et al.*, 1990). Turney et Littman (2003) l'emploient pour estimer la proximité sémantique entre le mot auquel ils veulent attribuer une valeur de polarité et 14 mots germes, 7 positifs (*good*, *nice*, …) et 7 négatifs (*bad*, *nasty*, …). Un mot est d'autant plus positif qu'il est plus proche des germes positifs et plus éloigné des germes négatifs. De son côté, Bestgen (2002) utilise l'ASL pour identifier les mots fréquemment associés aux mots dont il veut déterminer la polarité. Contrairement à Turney et Littman qui se fondent sur 14 mots amorces, Bestgen utilise un dictionnaire de 3 000 mots dont la polarité a été jugée par des évaluateurs et attribue à chaque mot la polarité moyenne de ses plus proches voisins dont la polarité est connue. On notera que les similarités peuvent être calculées sans passer par l'analyse sémantique latente, mais que, dans ce cas, des corpus de très grande taille sont nécessaires (Turney, Littman, 2003). Velikovich *et al.* (2010), par exemple, obtiennent les cooccurrences entre des mots de différentes catégories grammaticales, mais aussi des unités polylexicales sur la base de 4 milliards de pages Web et construisent ensuite un graphe pondéré. La polarité est alors calculée à partir de deux ensembles de points de repère, positifs et négatifs, par propagation : à chaque nœud du graphe est assigné un degré de polarité (positive et négative) calculé à partir du poids total du plus lourd chemin entre le nœud et un élément de

l'ensemble positif, puis négatif, de départ. La polarité finale correspond à la différence entre la positive et la négative.

La limitation principale de la plupart des méthodes présentées ci-dessus pour l'extension de normes autres que la polarité réside dans la nécessité de leur fournir une poignée de mots germes sélectionnés *a priori*. À notre connaissance, seuls Kamps et Marx (2002) ont proposé une paire de mots germes pour d'autres dimensions que la polarité : l'activité (actif et passif) et la puissance (fort et faible). Mais il faut noter que leur méthode ne fonctionne qu'avec des adjectifs, qu'elle repose sur le WordNet anglais et que son efficacité pour les dimensions d'activité et de puissance n'a pas été évaluée. De plus, ces paires de mots germes sont choisies de manière relativement arbitraire et une comparaison avec d'autres paires est rarement effectuée. Or des études en extraction de classes sémantiques (Pantel *et al.*, 2009 ; Kozareva et Hovy, 2010) ont montré l'importance du choix de ces germes, Pantel *et al.* (2009) ont par exemple observé une différence de 42 % de précision entre leur meilleur et leur plus mauvais ensemble de germes.

Tout récemment, Vincze et Bestgen (2011) ont développé une méthode spécifiquement conçue pour répondre à cette limitation puisqu'elle permet d'étendre une norme, disponible sur un nombre limité de mots, en identifiant automatiquement des mots germes (cf. section 3). Elle peut, théoriquement, être appliquée à n'importe quelle norme dans n'importe quelle langue pour autant que cette norme puisse être estimée sur la base des similarités sémantiques, c'est-à-dire pour autant que la valeur d'un mot sur cette norme puisse être estimée à partir de la similarité sémantique entre ce mot et d'autres mots dont la valeur est connue. Confirmer empiriquement la validité de cette assertion est l'objectif de l'étude rapportée ci-dessous.

3. ASG : une méthode pour étendre automatiquement des normes

ASG, pour *Apprentissage supervisé de mots germes*, emploie l'ASL pour construire un espace sémantique et calculer ensuite les proximités sémantiques avec des mots germes dont la valeur pour la dimension étudiée est connue. L'ASL a été préférée à d'autres méthodes non structurées parce qu'elle permet de capturer des informations du deuxième ordre[3] et aux méthodes structurées parce qu'elle est moins lourde dans les opérations de prétraitement (elle ne nécessite pas l'application d'analyseurs syntaxiques). De plus, Baroni et Lenci (2010) ont montré que les méthodes non structurées et structurées ont des performances comparables pour un même corpus. Les mots germes sont obtenus par une procédure d'apprentissage supervisé. Ils ne sont donc pas définis *a priori*, ce qui implique que la méthode peut

3. Le premier ordre définit deux mots comme étant similaires parce qu'ils apparaissent dans le même contexte, le deuxième ordre considère que deux mots sont similaires même s'ils n'apparaissent pas dans un même contexte parce que les contextes dans lesquels ils apparaissent sont similaires.

être appliquée à d'autres dimensions que la polarité. ASG requiert comme matériel d'apprentissage une norme lexicale pour la dimension en question et une collection de textes. Elle comporte les quatre étapes suivantes :

– sélectionner comme germes ou prédicteurs potentiels les mots qui sont présents dans la norme de référence ;

– effectuer ensuite une ASL sur une collection de textes afin d'obtenir un espace sémantique[4]. Concrètement, l'ASL part d'un tableau lexical qui contient le nombre d'occurrences de chaque mot dans chaque segment de textes, éventuellement modifié par une fonction (typiquement, la log entropie). Ce tableau fait ensuite l'objet d'une décomposition en valeurs singulières qui en extrait les dimensions orthogonales les plus importantes. Dans cet espace, le sens de chaque mot est représenté par un vecteur et l'on peut estimer la similarité sémantique entre deux mots par le cosinus entre leurs vecteurs. Dans la méthode ASG, on calcule les cosinus entre chacun des prédicteurs potentiels et tous les mots présents dans la norme. On obtient donc une matrice carrée avec en abscisse et en ordonnée les mots de la norme et en valeurs les cosinus entre leurs vecteurs dans l'espace sémantique. Le tableau 2 présente un petit extrait d'une des matrices utilisées dans les expériences décrites à la section 4 (corpus français et normes de polarité). On y trouve les cosinus entre quelques-uns des mots repris dans le tableau 1 et quelques prédicteurs potentiels. *Merveilleux* et *sage* sont les deux meilleurs prédicteurs de cette norme alors qu'*impressionnant* en est le moins bon, ce qui n'est pas nécessairement contre-intuitif si l'on pense que tant une déroute qu'un succès peuvent être impressionnants ;

MOTS	**POL**	**merveilleux**	**rage**	**impressionnant**
chance	6,4	0,197	0,162	0,039
chat	4,2	0,136	0,203	0,155
flux	3,7	0,119	0,148	– 0,008
honte	2,1	0,083	0,336	0,101

Tableau 2. *Exemple de matrices de cosinus résultant d'une ASL*

– utiliser ensuite une procédure de régression linéaire multiple afin de construire un modèle prédictif fondé sur les prédicteurs les plus efficaces pour prédire la norme. Concrètement, nous employons une procédure de régression linéaire mixte, par sélection et élimination (*stepwise*). Au début de la procédure, le modèle ne contient aucun prédicteur. À chaque étape, l'algorithme ajoute la variable qui

4. Un schéma synthétisant les différentes étapes d'une ASL est donné dans Piérard et Bestgen (TAL-47(2), 2006, p. 96).

améliore le plus la prédiction pour autant que sa contribution soit statistiquement significative pour un seuil de signification (*alpha*) donné. Mais comme l'ajout d'une variable dans le modèle modifie la contribution des autres variables à la prédiction, l'algorithme vérifie aussi à chaque étape que la variable présente dans le modèle qui contribue le moins à la prédiction apporte néanmoins toujours une contribution statistiquement significative à celle-ci. Si ce n'est pas le cas, cette variable est retirée du modèle. La procédure s'arrête lorsque aucune variable non encore présente dans le modèle ne peut y être ajoutée et qu'aucune variable présente dans le modèle ne doit en être retirée. Dans l'exemple du tableau 2, mais également dans les données réelles dont celui-ci est extrait, c'est le germe *merveilleux* qui est sélectionné en premier ; *rage* est sélectionné lors d'une étape ultérieure contrairement à *impressionnant*. Dans cette procédure, *alpha* est le seul paramètre qui doit être fixé. Classiquement, une valeur comprise entre 0,01 et 0,15 est employée. Plus cette valeur est élevée, plus facilement un prédicteur pourra entrer dans le modèle et plus celui-ci sera ajusté aux données ayant servi à l'apprentissage, faisant craindre une mauvaise généralisation à de nouvelles données. Un *alpha* trop petit limite fortement le nombre de prédicteurs et donc les capacités prédictives du modèle. Dans les analyses qui suivent, nous avons fait varier ce paramètre afin d'en déterminer l'impact ;

– employer le modèle construit à l'étape précédente pour estimer les valeurs – pour la dimension étudiée – des mots présents dans l'espace sémantique, mais non dans la norme initiale.

Il convient de mentionner que le critère de sélection des prédicteurs potentiels proposé à la première étape n'est pas unique. Il a l'avantage de limiter la sélection à des mots dont la valeur sur la dimension étudiée est connue, mais rien n'empêche de sélectionner les prédicteurs parmi l'ensemble des mots présents dans l'espace sémantique, que leur polarité soit connue ou non, puisque la seule condition est de pouvoir calculer la proximité entre chaque prédicteur potentiel et les mots repris dans la norme[5].

4. Expérimentations

Les expérimentations menées ont pour but de déterminer dans quelle mesure la procédure ASG, qui a été développée pour la norme de polarité, peut être utilisée pour prédire d'autres normes et ce dans différentes langues. Pour ce faire, nous avons utilisé plusieurs normes en français et en espagnol et nous avons extrait, pour chacune de ces langues, des espaces sémantiques à partir de textes littéraires.

5. Une analyse comparative de différentes manières de sélectionner les prédicteurs potentiels pour l'extension d'une norme de polarité est donnée dans (Vincze et Bestgen, 2011).

4.1 *Normes*

Cinq normes francophones ont été employées. Les normes d'abstraction et d'imagerie ont été recueillies par Hogenraad et Orianne (1981). La norme d'abstraction (F-ABS) a été obtenue en demandant à 24 évaluateurs belges francophones d'évaluer 450 noms communs de la langue française sur une échelle à 7 degrés d'intensité allant de *abstrait* (1) à *concret* (7) selon que le mot ne peut pas ou peut être expérimenté par les sens. La norme d'imagerie (F-IMA) a été obtenue lors de la même étude. Elle porte sur 1 130 noms évalués sur une échelle à 7 degrés d'intensité allant de *évoquant peu d'images* (1) à é*voquant beaucoup d'images* (7).

La norme de polarité (F-POL) est composée de 3 252 mots, chacun évalué par une trentaine de personnes sur une échelle à 7 points selon que le mot évoque une idée allant de *très désagréable* (1) à *très agréable* (7) (Hogenraad *et al.*, 1995). L'axe d'activation (F-ACT) renvoie au pouvoir des mots à suggérer ou à déclencher une activité (*non actif* : 1 ; *actif* : 7) et l'axe d'émotion (F-EMO) porte sur la capacité d'un mot à susciter une réaction émotionnelle plus ou moins forte (*non émotionnel* : 1 ; *émotionnel* : 7). Ces deux normes sont disponibles pour 3 000 mots et résultent également de jugements émis par une trentaine de personnes.

Pour l'espagnol, nous avons employé l'adaptation par Redondo *et al.* (2007) des normes ANEW (*Affective Norms for English Words* (Bradley et Lang, 1999)), normes de référence en anglais pour les dimensions polarité, activation et dominance, non disponibles en français. Ces normes portent sur 1 034 mots évalués par un total de 720 personnes. Leur tâche était d'indiquer la réaction émotionnelle évoquée par des mots sur trois échelles à 9 points : la polarité (E-POL) (*négatif, triste* = 1 ; *positif, joyeux* = 9), l'activation (E-ACT) (*calme* = 1 ; *excité* = 9) et la dominance (E-DOM) (*se sentir dominé* = 1 ; *se sentir dominant* = 9). Contrairement aux autres normes employées, les normes ANEW sont systématiquement obtenues à l'aide d'un système de notation affective fondé sur des échelles non verbales, où les réactions émotionnelles sont représentées sous la forme d'images. Redondo *et al.* (2007) fournissent également des valeurs d'abstraction (E-ABS) et d'imagerie (E-IMA) pour une partie des 1 034 mots de la norme (612 pour E-ABS et 601 pour E-IMA). Ces valeurs ont été obtenues au moyen d'échelles à 7 points similaires à celles employées pour ces mêmes dimensions en français.

4.2 *Collection de textes*

Pour les deux langues étudiées, les espaces sémantiques, utilisés pour calculer les similarités entre les mots, ont été construits sur la base de textes littéraires disponibles sur le Web. Pour le français, la collection de textes a été constituée principalement à partir des bases ABU[6] et Frantext[7]. Elle contient

6. http://abu.cnam.fr/ (consulté le 6 octobre 2011).

approximativement 5 300 000 mots. Pour l'espagnol, une collection de textes similaires a été rassemblée sur la base des textes mis à disposition par le projet Gutenberg[8], qui offre librement accès à plus de 36 000 livres dans une cinquantaine de langues. La collection que nous avons constituée contient approximativement 4 900 000 mots.

Pour construire le tableau lexical nécessaire à l'ASL, chaque texte a été subdivisé en segments de 250 mots et les prétraitements suivants ont été effectués : lemmatisation par le logiciel TreeTagger (Schmid, 1994), suppression de mots-outils et suppression des mots de fréquence totale inférieure à 10. Les matrices de cooccurrences ont été soumises à une décomposition en valeurs singulières réalisée par le programme Svdpack (Berry, 1992) et les 300 premiers vecteurs propres ont été conservés, ce nombre étant généralement considéré comme un optimum (Landauer et *al*., 2004).

4.3 *Analyses et résultats*

Les analyses réalisées ont pour premier objectif de valider la méthode en évaluant l'efficacité avec laquelle ASG prédit des normes. Deux tâches ont été considérées : attribuer à un mot une valeur sur l'échelle sous-jacente à la norme et distinguer les mots occupant les positions les plus extrêmes sur cette même échelle. Le deuxième objectif est de montrer l'utilité de l'approche proposée pour étendre des normes.

4.3.1 *Efficacité de ASG dans la prédiction de normes*

Afin d'estimer l'efficacité de la méthode pour prédire de nouvelles données, nous avons employé la procédure de validation croisée *Leave-one-out* (LOOCV) dans laquelle chaque observation est prédite au moyen du modèle dérivé de l'analyse de toutes les autres observations disponibles, produisant ainsi une estimation non biaisée (Stone, 1974). Cette procédure n'est en fait qu'un cas particulier de la procédure de validation croisée *k-fold*, classique en TAL, dans laquelle le paramètre *k* est égal à *N*, le nombre total d'observations. Par rapport à une procédure *k-fold* avec *k* égal à 3, 5 ou 10, la procédure LOOCV (ou *N-fold*) évite les problèmes de variabilité des résultats liés à la ou aux quelques partitions sélectionnées aléatoirement et est particulièrement efficace lorsque le nombre total d'observations n'est pas très grand, comme c'est le cas dans certaines analyses rapportées ci-dessous (*N = 385*).

Toutefois, la combinaison des techniques *stepwise* et LOOCV pose un problème. Les procédures *stepwise* disponibles dans les logiciels de statistiques effectuent la

7. http://www.atilf.fr/Les-ressources/Ressources-informatisees/FRANTEXT/ (consulté le 6 octobre 2011).
8. http://www.gutenberg.org/browse/languages/es (consulté le 6 octobre 2011).

sélection des variables en s'appuyant sur l'ensemble des données. C'est seulement dans un second temps que la procédure LOOCV est appliquée en ne prenant pas en compte, à tour de rôle, chacune des observations. Il s'ensuit que les prédicteurs ont été sélectionnés sur la base de l'ensemble des données et que donc l'estimation LOOCV est favorablement biaisée (donne une estimation de l'efficacité de la prédiction supérieure à la valeur réelle). Pour éviter ce problème, nous avons employé la procédure de double validation croisée qui consiste à retirer, à tour de rôle, chacune des observations *avant* d'effectuer la sélection des variables par la procédure *stepwise* et l'estimation du modèle qui est employé pour prédire cette observation (Stone, 1974). En d'autres mots, autant d'analyses *stepwise* qu'il y a de données dans le matériel d'apprentissage sont effectuées, laissant à chaque fois une observation de côté, et les variables sélectionnées sont employées pour construire le meilleur modèle pour ces observations. Ensuite, on emploie ce modèle pour prédire l'observation laissée de côté. De cette façon, l'estimation LOOCV est toujours non biaisée malgré l'emploi d'une procédure *stepwise*. Ces analyses ont été effectuées au moyen du logiciel SAS (*Statistical Analysis System V9*) par l'entremise de macro-instructions qui tirent profit de ses fonctionnalités ODS (*Output Delivery System*).

L'ensemble des analyses rapportées ci-dessous a été mené sur les mots des normes présents dans les espaces sémantiques, les seuls pour lesquels il est possible de comparer les valeurs prédites aux valeurs observées. Le nombre de mots remplissant cette condition pour chaque norme est donné dans les tableaux 3 et 4.

4.3.1.1 Corrélations entre les normes et les valeurs prédites par ASG

La méthode ASG est ici utilisée pour prédire la valeur des mots sur la dimension étudiée. La qualité de cette prédiction a été évaluée au moyen de la corrélation (r de Bravais-Pearson) entre les valeurs estimées par la procédure LOOCV et les valeurs réelles fournies par les normes.

Le tableau 3 pour le français et le tableau 4 pour l'espagnol reprennent les corrélations moyennes des LOOCV pour différents seuils de probabilités pour entrer ou sortir du modèle de régression.

Normes	**F-ABS**	**F-IMA**	**F-POL**	**F-ACT**	**F-EMO**
N	385	924	2 687	2 420	2 420
0,01	0,75	0,69	0,62	0,44	0,58
0,05	0,75	0,70	0,61	0,45	0,58
0,1	0,73	0,68	0,60	0,46	0,58
0,15	0,73	0,68	0,60	0,47	0,59

Tableau 3. *Corrélations pour le français*

Normes	**E-ABS**	**E-IMA**	**E-POL**	**E-ACT**	**E-DOM**
N	551	540	792	792	792
0,01	0,63	0,68	0,63	0,51	0,59
0,05	0,67	0,65	0,65	0,53	0,60
0,1	0,64	0,66	0,65	0,50	0,59
0,15	0,62	0,62	0,63	0,52	0,58

Tableau 4. *Corrélations pour l'espagnol*

Pour les deux langues et pour toutes les normes, on observe très peu de différences entre les seuils de probabilité employés pour la sélection de prédicteurs. En revanche, on observe pour une même langue des différences relativement importantes entre les normes (ex. différence de 0,14 entre la dominance et le caractère concret pour l'espagnol). Il est intéressant de constater que l'on a des profils similaires dans les deux langues à ce niveau : les prédictions pour la dimension concrète sont systématiquement meilleures que pour les autres dimensions. Viennent ensuite l'imagerie et la polarité. Deux explications peuvent être avancées : certaines dimensions peuvent être plus difficiles à estimer par la méthode ou par les évaluateurs humains, ce qui a un impact au niveau d'une prédiction automatique. La première explication signifierait que, pour certaines dimensions, on ne peut pas prédire de valeurs sur la base des similarités sémantiques entre les mots. La seconde explication renverrait à des difficultés plus importantes éprouvées par les évaluateurs lors de l'utilisation de certaines échelles. Il en va sans doute un peu des deux, cependant nous manquons de données cohérentes sur les variances des accords entre les évaluateurs pour les normes utilisées.

Une autre constatation intéressante porte sur la taille des normes. En effet, il semble qu'elle ne soit pas un facteur prépondérant pour la qualité des prédictions. La norme abstraite, la mieux prédite pour les deux langues, contient moins de mots que d'autres normes moins bien estimées, par exemple 450 mots contre plus de 3 000 pour la polarité en français.

Plus généralement, même si les corrélations entre les valeurs estimées et réelles sont assez élevées, elles sont loin d'être parfaites. Il faut toutefois noter que des niveaux de corrélations semblables (approximativement 0,70) sont considérés comme plus que suffisants pour employer des jugements subjectifs d'âge d'acquisition par des adultes à la place d'un indice objectif comme l'âge à partir duquel des enfants sont capables de dénommer une image (Bonin *et al.*, 2004). De même, les corrélations typiquement obtenues entre les estimations subjectives de fréquence et les mesures objectives de fréquence, fondées sur des corpus, dépassent rarement 0,70 (Tanaka-Ishii et Terada, 2011).

La limitation principale des analyses présentées ci-dessus réside dans l'absence de comparaison entre les performances d'ASG et celles d'autres techniques

automatiques capables d'estimer ce genre de normes en français ou en espagnol. Plusieurs raisons rendent de telles comparaisons difficiles à réaliser. Tout d'abord, aucune norme estimée par ce genre de procédures n'est, à notre connaissance, disponible dans ces langues. Un des objectifs majeurs de la présente recherche est justement de mettre à disposition de telles ressources. La deuxième raison est que les techniques similaires proposées en TAL (cf. section 2) portent sur la polarité. De plus, ces techniques ont le plus souvent été développées pour l'anglais et s'appuient sur des ressources comme WordNet qui ne sont encore que partiellement disponibles dans d'autres langues. La technique proposée par Turney et Littman (2003) pour la polarité, décrite dans la section 2 et fréquemment considérée comme technique état de l'art (Hassan *et al.*, 2011), fait toutefois exception puisqu'elle peut être aisément implémentée dans d'autres langues pour autant que les 14 points de repère sur lesquels elle se fonde soient traduits dans celles-ci. Nous l'avons donc appliquée aux normes de polarité francophones et espagnoles sur la base des mêmes espaces sémantiques[9]. La corrélation entre la valeur estimée par cette technique et la valeur réelle pour le français est de 0,38 ; elle est de 0,49 pour l'espagnol. Dans les deux cas, elle est nettement plus faible que la valeur obtenue par ASG (égale ou supérieure à 0,60 dans les deux langues).

4.3.1.2 Efficacité de ASG pour sélectionner des mots selon une norme

Comme les normes lexicales sont fréquemment employées en psychologie pour sélectionner un matériel expérimental composé de mots appartenant à chacun des deux pôles de l'échelle (les mots concrets *vs* les mots abstraits, les mots positifs *vs* les mots négatifs) tout en contrôlant plusieurs autres propriétés des mots comme la fréquence, l'âge d'acquisition ou la régularité phonologique, la deuxième analyse vise à évaluer l'efficacité de la procédure dans l'assignation de mots dans une des deux catégories que constituent ces deux pôles. Pour ce faire, chacune des normes a été divisée en trois parts approximativement égales, de façon à ne garder que les deux tiers extrêmes de chaque dimension et éliminer les mots perçus par les évaluateurs comme les plus neutres et donc ceux à propos desquels l'appartenance à un des deux pôles est la moins bien établie. Appliquée par exemple à la norme F-ABS, cette étape produit trois ensembles de mots, les 127 mots abstraits qui ont tous une valeur sur la norme strictement inférieure[10] à 3,7 , les 124 mots concrets qui ont une valeur supérieure à 5,6 et un dernier ensemble, qui ne sera pas employé dans la

9. Les points de repère ont été traduits de la manière suivante : *good = bon, bueno ; nice = gentil, agradable ; excellent = excellent, excelente ; positive = positif, positivo ; fortunate = heureux, afortunado ; correct = correct, correcto ; superior = supérieur, superior ; bad = mauvais, malo ; nasty = méchant, feo ; poor = médiocre, pobre ; negative = négatif, negativo ; unfortunate = malheureux, desgraciado ; wrong = faux, falso ; inferior = inférieur, inferior.*

10. Le fait de ne prendre dans les ensembles extrêmes que les valeurs strictement inférieures à la valeur limite explique les inégalités de taille des trois ensembles.

suite des analyses, composé des 134 mots dont la valeur est comprise entre 3,7 et 5,6.

Ensuite, on détermine pour chacun des mots analysés dans quel pôle de l'échelle la procédure automatique le classe. Comme pour la norme, la valeur prédite par ASG peut être comprise comme une mesure de la confiance avec laquelle le mot peut être classé dans un des deux pôles, les mots recevant un score proche du milieu de l'échelle étant les plus neutres et donc ceux à propos desquels ASG est le moins informatif. Nous avons donc effectué ces calculs en faisant varier le pourcentage de mots du milieu de l'échelle pour lesquels le score attribué par ASG n'est pas pris en compte : élimination de 0 % (tous les mots sont donc employés), des 33,3 % du milieu comme pour la norme, des 50 % et des 80 %[11].

Les résultats de ces analyses se présentent sous la forme de tables de contingence 2 × 2, une pour chaque norme et chaque seuil d'élimination. À titre d'exemple, deux de ces tables sont données dans le tableau 5. Elles résultent de l'analyse de la norme F-ABS pour un pourcentage d'élimination des scores attribués par ASG de 0 et de 25 %.

		ASG 0 %		
		Abs.	***Conc.***	***Total***
Norme	***Abs.***	112	15	127
	Conc.	13	111	124
	Total	125	126	251

		ASG 25 %		
		Abs.	***Conc.***	***Total***
Norme	***Abs.***	102	6	108
	Conc.	4	93	97
	Total	106	99	205

Tableau 5. *Tables de contingence pour le classement des mots abstraits et concrets par ASG*

Précision : (112 + 111) / 251 = 0,89 Kappa : Ao = 0,89 Aa = (125 / 251 × 127 / 251) + (126 / 251 × 124 / 251) = 0,50 k = (0,89 – 0,50) / (1 – 0,50) = 0,78 Rappel : (112 + 111) / 251 = 0,89	Précision : (102 + 93) / 205 = 0,95 Kappa : (0,95 – 0,50) / (1 – 0,50) = 0,90 Rappel : (102 + 93) / 251 = 0,78

Tableau 6. *Calculs des trois indices d'efficacité sur les données du tableau 5*

11. Il est à noter que l'élimination d'une partie des mots pour ASG est effectuée indépendamment de l'élimination d'un tiers des mots pour la norme.

Sur ces tables, trois indices résumant l'efficacité de ASG ont été calculés. Les deux plus importants sont l'exactitude (ou proportion d'accord ou précision) qui correspond au rapport entre le nombre de mots bien classés par ASG et le nombre de mots à classer et le coefficient Kappa (k) de Cohen, qui est également une mesure de l'accord, mais corrigée pour le hasard. En employant la notation d'Arstein et Poesio (2008), la formule du kappa est

$$k = \frac{(A_o - A_a)}{(1 - A_a)} \quad [1]$$

dans laquelle A_o représente la proportion d'accord observée et A_a la proportion d'accord attendue par le seul fait du hasard, calculée sur la base des totaux marginaux. Un kappa au moins égal à 0,80 est considéré comme démontrant une bonne ou même une excellente fiabilité (Arstein et Poesio, 2008 ; Landis et Koch, 1977).

Dans le cas présent, l'exactitude est égale à la précision (qui correspond au rapport entre le nombre de mots bien classés par ASG et le nombre de mots classés par cette procédure). Le troisième indice fourni est le rappel, le rapport entre le nombre de mots bien classés par ASG et le nombre de mots qu'il fallait *initialement* classer selon la norme. Ce dernier est évidemment fortement influencé par le pourcentage de mots qui ont été éliminés des prédictions effectuées par ASG. Étant donné que l'objectif de la procédure d'assignation automatique est d'identifier avec la plus grande certitude possible des mots appartenant aux deux pôles de l'échelle, la précision doit être privilégiée par rapport au rappel.

	F-ABS		**F-IMA**		**F-POL**		**F-ACT**		**F-EMO**	
	Pr	*k*	*Pr*	*k*	*Pr*	*k*	*Pr*	*k*	*Pr*	*k*
	N	*Ra.*	*N*	*Ra.*	*N*	*Ra.*	*N*	*Ra.*	*N*	*Ra.*
0,00 %	0,89	0,78	0,88	0,76	0,79	0,57	0,72	0,43	0,79	0,59
	251	0,89	574	0,88	1 385	0,79	1 544	0,72	1 493	0,79
33,30 %	0,96	0,92	0,95	0,91	0,86	0,72	0,77	0,55	0,87	0,73
	190	0,73	412	0,68	1 085	0,62	1 089	0,55	1 062	0,62
50,00 %	0,98	0,96	0,97	0,93	0,90	0,80	0,81	0,62	0,89	0,79
	150	0,59	330	0,56	899	0,51	831	0,44	827	0,50
80,00 %	1,00	1,00	0,99	0,97	0,96	0,92	0,88	0,76	0,96	0,91
	63	0,25	153	0,26	416	0,24	359	0,20	386	0,25

Tableau 7. *Catégorisation pour le français*

	E-ABS		**E-IMA**		**E-POL**		**E-ACT**		**E-DOM**	
	Pr *N*	*k* *Ra.*	*Pr* *N*	*k* *Ra.*	*Pr* *N*	*k* *Ra.*	*Pr* *N*	*k* *Ra.*	*Pr* *N*	*k* *Ra.*
0,00 %	0,85 363	0,69 0,85	0,83 358	0,66 0,83	0,80 525	0,60 0,80	0,75 526	0,51 0,75	0,77 526	0,55 0,77
33,30 %	0,91 263	0,82 0,66	0,90 258	0,80 0,65	0,90 377	0,81 0,65	0,85 362	0,71 0,59	0,85 372	0,70 0,60
50,00 %	0,92 208	0,85 0,53	0,92 204	0,83 0,52	0,94 303	0,87 0,54	0,90 282	0,81 0,48	0,90 290	0,80 0,50
80,00 %	0,96 91	0,91 0,24	0,92 87	0,84 0,22	0,96 138	0,93 0,25	0,96 127	0,92 0,23	0,98 130	0,95 0,24

Tableau 8. *Catégorisation pour l'espagnol*

Le tableau 7 pour le français et le tableau 8 pour l'espagnol donnent, pour chaque dimension étudiée, le nombre de mots de la norme pris en compte dans l'analyse (après élimination du tiers du milieu pour la norme et du pourcentage indiqué en tête de ligne pour ASG), et les trois indices décrits ci-dessus : les scores de précision (exactitude), de rappel et le coefficient Kappa (k). Ces résultats correspondent aux valeurs prédites par ASG pour un seuil de probabilité de 0,05.

Dans tous les cas, peu importe la langue, le coefficient Kappa et la précision augmentent en parallèle à l'emploi d'un seuil d'élimination de plus en plus strict et atteignent des valeurs très élevées. Il s'ensuit que, pour toutes les normes, le maximum d'efficacité est obtenu pour un seuil de 80 %. Ceci indique que les erreurs produites par la méthode automatique se concentrent sur les mots les plus neutres. Comme on pouvait s'y attendre, les différences entre les normes sont similaires à celles observées dans les analyses corrélationnelles (tableaux 3 et 4).

4.3.1.3 Analyse qualitative des erreurs

Les analyses qui précèdent soulignent l'efficacité de la méthode ASG, mais aussi ses imperfections. Afin d'essayer de comprendre l'origine des désaccords entre les normes et les estimations, nous avons examiné en contexte les mots classés différemment lorsque seuls 50 % des mots extrêmes pour ASG sont conservés. Deux facteurs ressortent de cette analyse.

Une première explication peut être trouvée dans la polysémie : certains mots possèdent plusieurs sens ne se trouvant pas tous du même côté des échelles de normes. Le sens le plus fréquent dans les textes, qui n'est pas toujours le sens premier, sans doute pris en compte par les évaluateurs, conditionne alors le contexte et donc la prédiction de ASG. Pour la dimension concrète, on peut citer *revuelta*, un mot qui signifie à la fois *la révolte* (*motín*) et *le coin* (*esquina*). Il a été classé

comme concret par ASG, alors que les évaluateurs l'ont considéré comme abstrait. Ceux-ci avaient sans doute en tête le premier sens du mot, plus courant. Cependant, on trouve dans le corpus une série d'occurrences du deuxième sens : *revuelta del camino*, *de la calle*, *del sendero*, etc. Pour l'imagerie, le mot *esfera* a été classé par ASG comme peu imagé, au lieu d'imagé. Cela s'explique sans doute par le fait que son premier sens, celui de *sphère*, renvoie beaucoup plus vite une image mentale que son deuxième, celui de milieu (*ambiente*), très peu imagé et relativement présent dans les textes : *en la esfera social en que ella vivía.* On peut aussi citer *pueblo*, qui renvoie à la fois à la ville et au peuple, moins imagé. En français, le mot *nature* dans son sens premier renvoie à quelque chose de très imagé, mais peut également renvoyer au tempérament d'un être, qui l'est beaucoup moins : *sa nature brutale ne se prêtait à aucune nuance de sentiment.* Le phénomène a également été rencontré pour la polarité : *heroína* a été classé par ASG comme positif au lieu de négatif probablement parce que ses deux sens sont radicalement opposés : la drogue est négative tandis que le féminin de héros est très positif et très fréquent dans des textes littéraires : *una heroína de leyendas y de cuentos fantásticos.* Pour le caractère émotionnel, on peut citer le nom *favori* en français, qui peut renvoyer à une entité préférée ou, au pluriel, à des touffes de barbe qui descendent sur les joues. Ce deuxième sens est beaucoup moins émotionnel et assez présent dans des descriptions littéraires : *le contraste de ses favoris blancs et de sa face cramoisie*. *Dépression* est un autre exemple, renvoyant à un état de souffrance mentale fort émotionnel et à une diminution de la pression atmosphérique, qui l'est beaucoup moins. Enfin, nous avons rencontré un cas pour la norme d'activité, *granada*, qui comme en français peut renvoyer à une arme ou à un fruit, beaucoup moins actif et très présent dans les textes.

Des travaux relativement récents en fouille d'opinion ont tenté de parer à ce problème en déterminant la polarité non plus des mots, mais de leurs sens. Ils se fondent pour la plupart sur WordNet, qui offre un classement par sens (synsets). Dès 1966, Stone *et al.* ont construit manuellement un lexique, portant entre autres sur la polarité, à partir des différents sens de mots mentionnés dans deux dictionnaires. On peut également citer le SentiWordNet d'Esuli et Sebastiani (2006, cf. section 2) ainsi que le Micro-WNOp de Cerini *et al.* (2007) composé de 1 105 synsets annotés manuellement de façon similaire au SentiWordNet. Wiebe et Mihalcea (2006) ont également annoté les synsets de WordNet, mais en leur attribuant une étiquette objective ou subjective. Quelques études ont montré que le recours à ce genre de ressources en parallèle à une désambiguïsation sémantique améliore la classification d'opinions (Akkaya et *al.*, 2009; Martín-Wanton *et al.*, 2010).

À un niveau plus subtil, les cas d'emplois figurés peuvent également renverser les valeurs pour plusieurs dimensions. Par exemple, *pasta* en espagnol, qui renvoie à la pâte, très concrète, peut également caractériser un trait de caractère, une façon d'être, beaucoup plus abstraite : *ser de buena pasta*. En français, on peut citer le mot *ivresse* (négatif pour les évaluateurs, mais classé positif par ASG), qui dans son sens premier renvoie à un état provoqué par l'alcool, plus négatif que son emploi figuré :

la tranquille ivresse de l'amour. Il en est de même pour *fièvre* ou *transport*, qui au sens figuré peuvent renvoyer à des sentiments, pour la plupart positifs (joie, amour, etc.) et qui ont été mal classés pour le caractère émotionnel.

Deuxièmement, le mot mal classé peut apparaître majoritairement dans des contextes opposés sur l'échelle de la norme. Tout d'abord, certains mots s'emploient fréquemment avec un complément de valeur opposée. Par exemple, *rapto*, mot abstrait en espagnol (*enlèvement*) qui a été classé concret, a généralement un complément du nom concret : *rapto de su hija, de la novia, de personas*. De même, *morceau* et *bout*, qui sont des mots considérés comme peu imagés par les évaluateurs, ont été classés comme imagés, probablement par l'influence de leurs compléments, la plupart du temps imagés : *morceau de pain, de sucre*, etc. On trouve également des cas de ce type pour la norme de polarité, comme *renoncement*, étant donné qu'on renonce généralement à quelque chose de positif, et, pour la norme d'émotion, on est généralement *insensible* à quelque chose qui devrait nous toucher : *insensible à mes caresses, à mes charmes*. D'autres mots peuvent être la cause ou la conséquence de quelque chose de valeur opposée. *Calmer*, pour la dimension d'activation, est généralement lié à un état de forte excitation, ce qui peut expliquer qu'il ait été considéré comme très actif par ASG. Le cas a également été rencontré pour la norme de polarité : *pardon* a été classé comme négatif par ASG probablement parce qu'il est la conséquence d'actes ou de paroles négatives. De même, quand on en vient à devoir *maîtriser* une situation ou une personne, c'est qu'elles sont plutôt négatives : *par la violence de ses mouvements [...] qu'il eut peine à maîtriser*.

Des explications plus spécifiques peuvent également être avancées. Par exemple, plusieurs mots se rapportant à l'homme, considérés comme imagés par les évaluateurs dans les deux langues, ont été classés comme peu imagés par ASG : *hombre, humano, persona, homme*. À la lecture de passages de textes, il apparaît que ces mots sont souvent associés à des qualités ou à des concepts abstraits : la liberté, la justice, la beauté, l'intellectuel, la vie, etc. Notons ici que le même mot dans les deux langues a été classé identiquement pour une même raison.

En conclusion, si ces deux facteurs soulignent les limites de l'approche proposée, ils peuvent aussi être vus comme des apports de celle-ci en pointant des mots dont la valeur indiquée dans la norme est peut-être sujette à caution. Il ne s'agit pas ici de prétendre que les estimations obtenues à partir de textes sont meilleures que celles fournies par les évaluateurs, mais, plus modestement, que les mots donnant lieu à un désaccord important méritent une analyse approfondie avant d'être employés dans un matériel expérimental, par exemple.

4.3.2 *Extension des normes*

Si la technique proposée peut être utilisée pour contrôler les valeurs attribuées par des évaluateurs comme expliqué ci-dessus, sa fonction principale est d'étendre une norme en estimant la valeur de nouveaux mots sans avoir recours à des

évaluateurs. Afin d'illustrer cet emploi, le tableau 9 présente les 200 mots qui ont obtenu les scores d'abstraction les plus extrêmes selon la procédure ASG, 100 pour le pôle abstrait et 100 pour le pôle concret, ordonnés à chaque fois du plus extrême au moins extrême. Lorsqu'un de ces mots fait partie de la norme F-ABS, il est suivi par son score entre parenthèses. Une analyse de ces deux listes permet de mettre en lumière les potentialités, mais également les limites de la procédure. Tout d'abord, on observe que nombre de mots parmi les plus extrêmes sur cette dimension ne sont pas dans la norme initiale et même ne sont pas des noms communs, alors que seule cette catégorie grammaticale est incluse dans la norme F-ABS. Ceci souligne la valeur heuristique de l'approche. Ensuite, la présence parmi les mots les plus concrets de *tulle*, *yole* ou encore *brodequin* traduit un effet très marqué du type de textes employés pour extraire l'espace sémantique : des œuvres littéraires du XIX^{e} et du XX^{e} siècle. Prendre en compte la fréquence d'emplois des mots dans des corpus récents comme celui qui repose sur les sous-titres (New *et al.*, 2007) permettrait d'identifier automatiquement les mots peu pertinents en raison de leur rareté.

Abstrait : pouvoir, raisonner, importance (2,0), idée (2,0), nécessité, personnalité (2,3), apprécier, comparer, motif, égard, différer, excès, chance (1,9), condition, bien, actuel, envisager, devoir, résumer, signification (1,7), concevoir, logique, incapable, admettre, essentiel, humilier, analogue, savoir, terme, réaliser, acquérir, combinaison, justifier, raison (1,7), constater, origine, résultat, exercer, exclure, conséquence, exaltation, analyser, réagir, différent, formule, rapprochement, genre (1,9), consister, certainement, inspirer, résulter, organisation, mal (2,0), tourmenter, dominant, égoïste, réflexion, exprimer, intellectuel, invincible, définitif, scrupule, pourvu, absoudre, réel, durable, identique, utile, croyance, caractère, naturel, déterminer, apparent, effet (2,5), conduite, convenir, correspondre, absolu, immédiat, incomplet, futur, autrement, préoccupation, vouer, accorder, considération, conclure, transformer, inquiétude, immuable, manifester, complexe, certitude, facilement, modification, égoïsme, pressentir, chose (3,1), dissiper, déraisonnable.
Concret : vernir, tulle, bouton (6,5), empeser, penché, châle, cirer, treillis, blouse, bague, jupe, pantalon, velours, mauve, épingle, collerette, satin, chausser, gland, baise, vif-argent, dais, soie, plissé, col, rigole, jupon, corsage, ruban, cloporte, nourrisson, feutre, damas, entrecroiser, dentelle, ombrelle, luisant, capitonner, rameur, bretelle, friper, aviron, manche, bonnet, bottine, yole, cuisse, chapeau (6,5), veste, escarpin, rame, moucheron, pantoufle, grenouillère, cigogne, plume (6,4), mouchoir, asperger, jarretière, sangle, retrousser, boîte, cornet, glauque, causeur, jarret, tricot, entrouvrir, ramer, agrafe, porte-fenêtre, casquette, brodequin, nu, tic-tac, sapin, flasque, autruche, crinière, cocarde, chaînette, velu, surplomber, touffe, chausson, somptuosité, bêche, baignoire, gazon, brun, cercler, parement, canotier, scarabée, plumet, muet, poignet, vêtir, guérite, casque.

Tableau 9. *Mots ayant le degré d'abstraction le plus extrême selon ASG*

Le tableau 10 donne les 100 mots les plus émotionnels selon ASG. La norme F-EMO contenant beaucoup plus de mots que la norme F-ABS, il est logique de retrouver dans cette liste plus de mots issus de celle-ci. On note néanmoins que

41 de ces 100 mots ne se trouvent pas dans la norme initiale et parmi ceux qui s'y trouvent un certain nombre y ont reçu une valeur peu élevée alors qu'ils semblent très émotionnels comme *amertume* (4,2), *délire* (4,7), *adorer* (5,0), *affreux* (4,8)… On observe aussi l'impact de la lemmatisation dans des mots comme *chérer* pour *chère*, *éperdre* pour *éperdu*. Le dernier mot *plessis* trouve également son origine dans la lemmatisation qui affecte le lemme *plessis* au nom *Plessis*. Ce dernier est presque exclusivement employé dans une seule des œuvres incluses dans la collection de textes, *Histoire de ma vie*, dans laquelle le château de Plessis Picard est associé à des moments heureux de la vie de George Sand. Ceci souligne encore une fois l'impact des textes à l'origine de l'espace sémantique, mais aussi la possibilité, à la suite de Bestgen (2002), d'employer ASG pour attribuer une valeur sur une norme à des noms propres.

Émotionnel : tendresse (6,1), aimer (6,7), bonheur (6,3), ressentir (5,1), caresse (6,4), cœur (5,5), tourment (5,0), bien-aimé, éperdre, jaloux (6,0), fou (4,9), torturer (5,6), amertume (4,2), déchirement (6,2), consoler, désespoir (6,3), égorger, délire (4,7), étreinte, pleurer (6,2), adorer (5,0), jalousie (6,0), pleurant, étreindre, surhumain, séparation, délirant, enlacer, affoler (5,8), épouser, effusion, isolement, grossesse (5,9), revivre, attendrissement, affreux (4,8), éperdument (4,6), chaste, joie (6,3), inoubliable, amour (6,8), consolation, insensé, ingrat, chérir (5,8), sensuel (5,9), imagination (5,7), filial, haïr (6,2), éperdu (4,6), sanglot (6,4), cruel (5,7), maternité, jalouser, pleurs, affolement (5,3), espérance (5,5), sangloter (5,8), tristesse (6,2), angoisse (5,7), tuer (5,9), chagrin (6,3), navrant (3,4), fiancer, parricide, forcené, déchirant (6,1), drame (5,8), inexprimable, bravade, amoureux (6,3), émotion (6,3), épouvantable (5,3), décevoir, chagriner (5,5), transe (4,8), amèrement, sort, chérer, bouleverser (5,3), enfant (5,7), délice (4,6), ranimer, unisson, fêter (5,3), séduire (5,6), pressentiment (4,7), volupté (4,8), lascif, orphelin (5,6), caresser (6,2), fille (5,3), poignant (5,0), détester (5,4), irrésistible, remords (5,5), choyer (5,3), indicible, préférence (5,0), plessis.

Tableau 10. *Mots ayant le degré d'émotion le plus extrême selon ASG*

5. Conclusion

Nous avons présenté une méthode automatique pour estimer des normes lexicales sur la base d'une analyse de collections de textes. Les analyses effectuées dans les deux langues que sont l'espagnol et le français soulignent leur efficacité pour les dimensions d'abstraction, d'imagerie, de polarité, de dominance et d'émotion. La procédure est moins efficace pour l'activation. Pour être mise en pratique, elle ne nécessite qu'un début de normes et une collection de textes de quelques millions de mots aisée à rassembler sur le Web pour de nombreuses langues. Elle se fonde exclusivement sur les cosinus, issus d'une analyse sémantique latente, entre le mot à estimer et les mots présents dans la norme et sur une procédure de régression multiple disponible dans de nombreux logiciels de statistiques.

En plus de l'extension de normes, ce type d'estimation peut être employé pour les contrôler afin d'identifier des mots pour lesquels les jugements sont fortement différents de l'estimation obtenue sur la base de textes. Une analyse de l'origine de ces différences s'impose alors. Cette méthode peut aussi être utilisée pour réduire le nombre d'évaluateurs lors d'une recherche visant à accroître une norme en présélectionnant des mots qui semblent plus particulièrement intéressants ou pour réduire le nombre d'évaluateurs nécessaires pour évaluer chaque mot.

Les expérimentations rapportées présentent un certain nombre de limites qui justifient des recherches complémentaires. Tout d'abord, la taille de la collection de textes nécessaire pour obtenir une efficacité suffisante n'a pas été évaluée dans les expérimentations. De même, l'impact du genre de textes employés pour extraire l'espace sémantique mériterait une analyse approfondie. Obtiendrait-on de meilleurs résultats en effectuant des estimations sur la base de plusieurs collections de textes et en combinant ces différentes estimations ? Une telle approche permettrait d'étudier la stabilité des estimations et d'attribuer un indice de confiance dans la valeur associée à chaque mot. Elle permettrait aussi de tirer avantage d'un des points forts de la méthode : produire des normes en fonction de l'usage prévu en sélectionnant une norme spécifique (obtenue auprès d'enfants ou d'adultes) et un corpus spécifique, par exemple composé de textes destinés à un public spécifique ou proche de la langue parlée. Une autre question à laquelle les expérimentations n'apportent qu'une réponse très partielle est la taille minimale de la norme nécessaire pour obtenir une bonne performance. Il apparaît que quelques centaines de mots sont amplement suffisants, mais peut-on en employer moins ? Enfin, l'origine de la faible efficacité de ASG pour la norme d'activation mériterait certainement d'être expliquée.

6. Bibliographie

Akkaya C., Wiebe J., Muhalcea R., « Subjectivity word sense disambiguation », *Conference on Empirical Methods in Natural Language Processing*, 2009, p. 190-199.

Artstein R., Poesio M., « Inter-Coder Agreement for Computational Linguistics », *Computational Linguistics*, vol. 34, 2008, p. 555-596.

Baroni M., Lenci A. « Concepts and properties in word spaces », *Italian Journal of Linguistics*, vol. 20, 2008, p. 55-88.

Baroni M., Lenci A. « Distributional Memory: A General Framework for Corpus-Based Semantics », *Computational Linguistics*, vol. 36, 2010, p. 673-721.

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

Bestgen Y., « Détermination de la valence affective de termes dans de grands corpus de textes », *Actes de CIFT'*02, 2002, p. 81-94.

Bonin P., Barry C., Méot A., Chalard M., « The influence of age of acquisition in word reading and other tasks: A never ending story? », *Journal of Memory and Language*, vol. 50, 2004, p. 456-476.

Bradley M. M., Lang P. J., « Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings », Tech. Rep. No. C-1, Gainesville, FL: Center for Research in Psychophysiology, University of Florida, 1999.

Cerini S., Compagnoni V., Demontis A., Formentelli M., Gandini G., « Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining », *Language resources and linguistic theory: Typology, second language acquisition, English linguistics*, Franco Angeli Editore, Milano, 2007.

Da Sylva L., « Extraction semi-automatique d'un vocabulaire savant de base pour l'indexation automatique », *Actes de TALN 2010*, 2010.

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.

De Marneffe, M.-C., Manning C. D., Potts C., « Was It Good? {I}t Was Provocative. Learning the Meaning of Scalar Adjectives », *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, p. 167-176.

Desrochers A., Saint-Aubin J., « Sources de matériel en français pour l'élaboration d'épreuves de compétences en lecture et en écriture », *Canadian Journal of Education*, vol. 31, 2008, p. 305-326.

Dewaele J., Pavlenko A., « Emotion vocabulary in interlanguage », *Language learning*, vol. 52, 2002, p. 263-322.

Esuli A., Sebastiani F., « Sentiwordnet: A publicly available lexical resource for opinion mining », *Proceedings of LREC'06*, 2006, p. 417-422.

Franklin S., Howard D., Patterson, K., « Abstract word anomia », *Cognitive Neuropsychology*, vol. 11, 1995, p. 549-566.

Gernsbacher M. A., « Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy », *Journal of Experimental Psychology: General*, vol. 113, 1984, p. 256-281.

Hassan A., AbuJbara A., Jha R., Radev D., « Identifying the semantic orientation of foreign words », *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, p. 592--597.

Hatzivassiloglou V., McKeown K. R., « Predicting the semantic orientation of adjectives », *Proceedings of EACL 1997*, 1997, p. 174-181.

Hogenraad R., Bestgen Y., Nysten J. L., « Terrorist rhetoric: Texture and architecture », In E. Nissan & K.M. Schmidt (Eds.) *From information to knowledge. Conceptual and content analysis by computer*, Oxford, England: Intellect Books, 1995, p. 54-67.

Hogenraad R., Orianne E. « Valences d'imagerie de 1.130 noms de la langue française parlée », *Psychologica Belgica*, vol. 21, 1981, p. 21-30.

Kamps J., Marx M., « Words with Attitude », *Proceedings of the 1st Interational Conference on Global WordNet*, 2002, p. 332-341.

Kim S. M., Hovy E., « Determining the sentiment of opinions », *Proceedings of COLING*, 2004, p. 1367-1373.

Kousta S. T., Vigliocco G., Vinson D.P., Andrews M., Del Campo E., « The representation of abstract words: why emotion matters », *Journal of experimental psychology: General*. vol. 140, 2011, p. 14-34.

Kozareva Z., Hovy E., « Not All Seeds Are Equal: Measuring the Quality of Text Mining Seeds », *HLT-NAACL'2010*, 2010, p. 618-626.

Lachaud C.M., « CHACQFAM : une base de données renseignant l'âge d'acquisition estimé et la familiarité pour 1225 mots monosyllabiques et bisyllabiques du Français », *Année Psychologique*, vol. 107, 2007, p. 39-63.

Landauer T. K., Laham D., Derr M., « From paragraph to graph: Latent Semantic Analysis for information visualization », *Proceedings of the National Academy of Science* 101, 2004, p. 5214-5219.

Landis J. R., Koch G. G., « The measurement of observer agreement for categorical data », *Biometrics*, vol. 33, 1977, p. 159-174.

Martín-Wanton T., Balahur-Dobrescu A., Montoyo-Guijarro A., Pons-Porrata A., « Word Sense Disambiguation in Opinion Mining: Pros and Cons », *Research in Computing Science*, vol. 46, p. 119 - 129.

Miller G. A., Beckwith R., Fellbaum C., Gross D., Miller K., « WordNet: An on-line lexical database », *International Journal of Lexicography*, vol. 3, 1990, p. 235-244.

New B., « Lexique 3 : Une nouvelle base de données lexicales », *Actes de TALN 2006*, 2006.

New B., Brysbaert M., Veronis J., Pallier C. « The use of film subtitles to estimate word frequencies », *Applied Psycholinguistics*, vol. 28, 2007, p. 661-677.

Padó S., Lapata M., « Dependency-based construction of semantic space models », *Computational Linguistics*, vol. 33 (2), 2007, p. 161-199.

Pang B., Lee L., « Opinion Mining and Sentiment Analysis », *Foundations and Trends in Information Retrieval* , vol. 2, 2008, p. 1-135.

Pang B., Lee L., Vaithyanathan S., « Thumbs up? Sentiment classification using machine learning techniques », *Proceedings of the ACL-02*, 2002, 79-86.

Pantel P., Crestan E., Borkovsky A., Popescu M.Vyas V., « Web-scale distributional similarity and entity set expansion », *Proceedings of the 2009EMNLP*, 2009, 938-947.

Pennebaker J. W., Mehl M. R., Niederhoffer K. G., « Psychological aspects of natural language use: Our words, our selves ». *Annual Review of Psychology, 54*, 2003, p. 547-577.

Piérard S., Bestgen, Y., « Validation d'une méthodologie pour l'étude de deux types marqueurs de la segmentation dans un grand corpus de texte ». *Traitement Automatique des Langues*, vol. 47, 2006, p. 89-110.

Proctor R. W., Vu, K. L., « Index of norms and ratings published in the Psychonomic Society journals », *Behavior Research Methods, Instruments, & Computers*, vol. 31, 1999, p. 659-667.

Popescu A.-M., Etzioni O., « Extracting Product Features and Opinions from Reviews », *Proceedings of the HLT/EMNLP*, 2005, p. 339-346.

Redondo J., Fraga I., Padrón I., Comesaña M., « The Spanish adaptation of ANEW (Affective Norms for English Words) », *Behavior Research Methods*, vol. 39, 2007, p. 600-605.

Saenko K., Darrell T., « Filtering abstract senses from image search results », *Advances in Neural Information Processing Systems*, vol. 22, 2009, p. 1589-1597.

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

Schütze H., « Ambiguity Resolution in Natural Language Learning », *CSLI Publications*, 1997.

Shibahara N., Zorzi M., Hill M. P., Wydell T., Butterworth, B., « Semantic effects in word naming: Evidence from English and Japanese Kanji », *Quarterly Journal of Experimental Psychology*, vol. 56A, 2003, p. 263-286.

Stone M., « Cross-validatory choice and assessment of statistical predictions (with discussion) », *Journal of the Royal Statistical Society*, Series B, vol. 36, 1974, p. 111-147.

Stone P.J., Dunphy D.C., Smith M.S., Ogilvie D.M., « The General Inquirer: A Computer Approach to Content Analysis », *MIT Press*, 1966.

Strain E., Patterson K., Seidenberg M. S., « Semantic effects in single-word naming », *Journal of Experimental Psychology:Learning, Memory, & Cognition*, 21, 1995, p. 1140-1154.

Somasundaran S., Wilson T., Wiebe J., Stoyanov V., « QA with attitude: Exploiting opinion type analysis for improving question answering in on-line discussions and the news », *Proceedings of the International Conference on Weblogs and Social Media*, 2007.

Tanaka-Ishii K., Terada H., « Word familiarity and frequency », *Studia Linguistica*, vol. 65, 2011, p. 96-116.

Turney P. D., Littman M., « Measuring Praise and Criticism: Inference of Semantic Orientation from Association », *ACM Transactions on Information Systems*, vol. 21, 2003, p. 315-346.

Velikovich L., Blair-Goldensohn S., Hannan K., McDonald R., « The Viability of Web-derived Polarity Lexicons ». *Proceedings of NAACL*, 2010, p. 777-785.

Vincze N., Bestgen, Y., « Identification de mots germes pour la construction d'un lexique de valence au moyen d'une procédure supervisée », *Actes de TALN11*, vol. 1, 2011, p. 223-234.

Wiebe J., Mihalcea R., « Word sense and subjectivity », *Proceedings of the ACL-2006*, 2006, p. 1065-1072.

Wiebe J., Wilson T., Bruce R., Bell M., Martin M., « Learning subjective language », *Computational Linguistics*, vol. 30, 2004, p. 277-308.

Wilson, M.D., « The MRC Psycholinguistic Database: Machine Readable Dictionary, Version 2 », *Behavioral Research Methods, Instruments and Computers*, vol. 20, 1988, p. 6-11.