Mise au jour semi-automatique de nuances sémantiques entre mots de sens proches

Mathias Rossignol[1] Pascale Sébillot[2]
(1) Centre de Recherche International MICA
UMI-2954 CNRS - HUT - INP Grenoble
1 Đai Cồ Việt, Hà Nội, Vietnam
(2) IRISA, Campus de Beaulieu, 35042 Rennes, France
Mathias.Rossignol@gmail.com, Pascale.Sebilot@irisa.fr

Résumé L'acquisition automatique sur corpus d'informations lexicales sémantiques donne une place importante à la constitution de classes sémantiques rassemblant des mots de sens proches. Or, l'intérêt pratique de celles-ci reste limité en l'absence d'information sur les distinctions individualisant les sens des mots qu'elles rassemblent. Nous présentons dans cet article un premier système permettant de mettre au jour, de manière semi-automatique et à partir des seules données textuelles rassemblées dans un corpus, des éléments de distinction sémantique fine entre mots appartenant à une même classe, atteignant ainsi un degré de définition du sens encore inédit en acquisition automatique d'informations sémantiques lexicales. La technique mise au point regroupe, en s'appuyant sur l'étude de grands voisinages autour des occurrences des mots comparés, des paires de mots distingués par des nuances similaires. Cette approche présente la faiblesse de ne permettre qu'une représentation implicite des nuances découvertes : les listes de paires de mots rapprochées doivent être interprétées afin de « comprendre » l'élément de distinction commun. En revanche, elle permet une automatisation importante du processus de recherche de nuances, suffisante pour assurer que le travail humain de validation des résultats n'introduise dans ceux-ci de biais interprétatif trop important.

Abstract The automatic corpus-based acquisition of lexical semantic information has given rise to numerous studies on the constitution of semantic classes, gathering words of similar meanings. However, the practical interest of those classes remains limited in the absence of knowledge about the nuances distinguishing word meanings inside them. We present in this paper a first system to put to light in a semi-automatic way such semantic nuances between words of a same class, using data from a textual corpus, thus reaching a degree of word meaning definition never before attained, to our knowledge, by automatic means. The developed technique exploits large contexts around word occurrences to bring together pairs of words distinguished by a similar nuance. The weakness of that approach is that it only provides an implicit representation of the discovered distinctions : a human interpretation is still needed to name them. However, it makes it possible to reach an important level of automation, so that the human validation work cannot introduce much bias in the results.

Mots-clefs : Classes sémantiques, nuances de sens, acquisition sur corpus

Keywords: Semantic classes, nuances, corpus-based acquisition

1 Introduction

L'acquisition automatique sur corpus d'informations lexicales sémantiques donne une place importante, depuis certains de ses travaux fondateurs (Hindle, 1990; Grefenstette, 1994), à la constitution de classes sémantiques. Cette tâche consiste à rassembler des mots de sens proches, interchangeables dans certains contextes sans affecter la cohésion du texte où la permutation a lieu ; linguistiquement, elle se place plus ou moins explicitement dans la lignée des théories sémantiques différentielles (Greimas, 1966; Rastier, 1996), mais aussi des travaux de Z. Harris (Harris, 1968), selon lesquels une équivalence d'usage des mots implique une équivalence sémantique entre eux.

En tant qu'outil, les classes sémantiques doivent permettre la géneralisation d'énoncés, la prise en compte de la variabilité lexicale, voire l'introduction de variations dans l'expression pour la synthèse de textes. On peut néanmoins s'interroger sur leur utilité pratique réelle pour les tâches abordées en traitement automatique des langues : en recherche d'information, par exemple, un utilisateur souhaitant faire l'acquisition d'une voiture ne souhaite sans doute pas que sa requête soit étendue à *camion*. En traduction, le titre de *The Jungle Book* de R. Kipling ne doit certainement pas être traduit par *le Livre de la Brousse*, à moins que l'on ne souhaite en délocaliser l'action en Afrique. En synthèse de texte, par exemple pour la rédaction de bulletins météorologiques, il semble capital de ne pas confondre *crachin*, *pluie* et *averse*même si l'on peut envisager de remplacer « averses » par « pluie intermittente ».

Il apparaît dans ces exemples que l'utilité des classes sémantiques, et du principe associé de permutabilité des mots en contexte, reste limité sans information sur ce qui, au-delà de leur sens commun, fait la spécificité des sens des mots qu'elles rassemblent. C'est la connaissance de ces nuances qui permet de définir plus précisément les conditions dans lesquelles une permutation est en effet sémantiquement pertinente. De même qu'il est aujourd'hui possible d'automatiser dans une large mesure l'apprentissage de classes sémantiques à partir de corpus, on souhaiterait donc pouvoir mettre au jour de manière tout aussi automatique les distinctions sémantiques structurant ces classes — ces « distinctions sémantiques » désignant des nuances beaucoup plus fines que les relations lexicales classiques comme l'hyperonymie (Caraballo & Charniak, 1999). Le travail présenté dans cet article constitue un premier ensemble d'expériences et de propositions pour répondre à cette problématique encore peu explorée.

Quoique aucune étude n'ait jusqu'ici, à notre connaissance, proposé de réelle automatisation de la recherche des nuances de sens entre mots, plusieurs auteurs ont montré qu'il est possible de mettre au jour, par étude des contextes d'usages des mots en corpus, des indices de ces nuances. Est ainsi présentée dans (Fabre *et al.*, 1997) une expérience d'utilisation du système ZELLIG (Habert *et al.*, 1996) afin d'évaluer les similarités sémantiques entre mots sur un corpus non technique — un recueil d'allocutions du président F. Mitterand. Le graphe de proximité sémantique obtenu permet d'identifier pour certains mots diverses facettes de leur sens, en fonction de leurs usages et de leurs rapprochements avec d'autres mots du corpus, ce qui participe à une définition plus précise de leur sens. Les auteurs de ce travail proposent également de préciser les sens des quasi-synonymes par un retour au texte et une étude directe de leurs distinctions d'usage.

Cette seconde approche est reprise et formalisée de manière plus approfondie dans (Pichon & Sébillot, 1999), dont l'objectif est la désignation explicite des nuances distinguant les noms d'une même classe. La technique employée consiste à associer à chacun de ces noms la liste des adjectifs apparaissant le plus fréquemment au voisinage de ses occurrences, puis à construire les listes d'adjectifs communs à tous les noms d'une classe, ou au contraire spécifiques à l'un d'eux, partagés par un sous-ensemble de mots, *etc*. Ces listes sont ensuite interprétées manuellement afin d'en extraire un ou plusieurs indices révélateurs d'un sens particulier. Ainsi, dans le cadre d'un corpus traitant de négociations territoriales, le fait que seul *autorité* soit associé à des adjectifs comme « temporaire » ou « transitoire » indique que ce nom désigne une réalité plus instable que *pouvoir* ou *gouvernement*, qui appartiennent à la même classe. Quoique s'appuyant beaucoup sur une analyse manuelle des données, cette première étude montre le bien-fondé de l'hypothèse selon laquelle les contextes d'usage des mots sont porteurs d'information concernant les nuances sémantiques autant que les points communs.

Le travail que nous présentons s'appuie donc sur ce même principe, mais en tentant d'automatiser autant que possible la mise au jour des nuances. Afin de rendre possible cette automatisation, nous ne pouvons toutefois aborder une problématique identique à celle de (Pichon & Sébillot, 1999), la désignation explicite des nuances supposant un degré élevé de « compréhension » des phénomènes étudiés. Notre objectif est donc d'effectuer des rapprochements entre paires de mots distingués par une même nuance. Ainsi, nous ne pourrons pas exprimer que *montagne* est « plus gros que » *colline*, mais pourrons en revanche indiquer que la nuance qui les distingue est similaire à celle séparant *rocher* de *caillou*, ou *rivière* de *ruisseau*, le nommage ultérieur par un humain de cette distinction commune restant bien entendu possible. Ces regroupements peuvent être rapprochés de la notion d'analogie (« a est à b ce que c est à d ») (Lepage, 2004), celle-ci étant toutefois principalement cantonnée en TAL à la morphologie et à la syntaxe.

En prélude à la description des techniques que nous avons développées, nous précisons à la section 2 les prérequis à nos travaux : d'une part, les données employées pour l'expérimentation, d'autre part, la méthode de construction des classes sémantiques que nous avons pour ambition de structurer. Nous présentons ensuite à la section 3 les considérations théoriques ayant guidé nos choix méthodologiques, avant d'exposer à la section 4 la technique développée pour mettre au jour les nuances entre mots de classes sémantiques. Celle-ci permet une première organisation des données, qui requiert toutefois encore un filtrage manuel avant d'être totalement exploitable. Nous présentons à la section 5 la procédure que nous avons mise au point afin de « cadrer » cette intervention humaine et de restreindre les possibilités d'introduction de connaissances *a priori* par l'utilisateur-validateur. La section 6, enfin, est consacrée à un rapide examen des résultats obtenus.

2 Conditions expérimentales

Nous présentons dans cette section le corpus d'expérimentation adopté ainsi que son mode de préparation, puis les étapes d'apprentissage automatique aboutissant à la constitution des classes sémantiques que nous nous proposons de structurer.

2.1 Corpus étudié

Le corpus employé a fins d'expérimentation au cours de notre recherche est constitué de 14 ans d'archive du mensuel *le Monde Diplomatique*, rassemblant quelque 11 millions de mots (environ 6 000 articles, 100 000 paragraphes). Les textes collectés abordent des sujets très variés — nous ne travaillons pas sur un corpus spécialisé —, et la langue qui y est développée est très riche, et fréquemment complexe. Ce corpus constitue donc un cas « difficile » pour l'acquisition automatique d'informations lexicales sémantiques. Il a été étiqueté et lemmatisé grâce aux outils du projet MULTEXT (Armstrong, 1996), mais n'a pas subi d'analyse syntaxique par souci de fiabilité, de simplicité et de généricité.

2.2 Étapes antérieures d'acquisition d'informations sémantiques

Découpage thématique du corpus Afin de réduire l'impact sur notre étude des phénomènes de polysémie des mots étudiés, nous avons dans un premier temps découpé le corpus d'étude en sous-corpus thématiques grâce au système FAESTOS, présenté dans (Rossignol & Sébillot, 2003). Celui-ci nous permet de caractériser et détecter de manière totalement automatique les occurrences des principaux thèmes abordés dans un corpus, sans aucune donnée extérieure ni connaissance *a priori* de ces thèmes. Nous disposons à l'issue de ce premier traitement d'une quarantaine de sous-corpus rassemblant chacun l'ensemble des pararaphes du corpus d'origine abordant un thème donné — de quelques dizaines à quelques centaines de milliers de mots.

Constitution de classes sémantiques Chacun de ces sous-corpus est ensuite exploité indépendamment afin de construire un ensemble de classes sémantiques « thématisées », selon une méthode présentée dans (Rossignol & Sébillot, 2006) et dont nous rappelons ici très rapidement le principe. La petite taille relative des sous-corpus nous oblige à réaliser cette opération en deux étapes : dans un premier temps, une « proximité sémantique » approximative entre mots p est calculée en employant l'intégralité des données du corpus d'origine, par une méthode classique de comparaison des ensembles de mots des voisinages d'occurrences. Cette proximité reste très imparfaite pour la classification sémantique, notamment du fait du « bruit » introduit dans l'analyse par les nombreux mots polysémiques, mais elle constitue une première source d'informations permettant ensuite de réaliser sur chaque sous-corpus une étude et comparaison plus approfondie des usages de mots, et construire des arbres de classification hiérarchique rapprochant les mots de sens similaires. Si la première des deux étapes de calcul de similarité sémantique entre mots est générique est aboutit à la définition d'une mesure p pour les noms, verbes, adjectifs, *etc.*, la seconde est pour l'heure limitée dans son applicabilité aux noms. C'est donc la structuration de classes sémantiques de noms qui retient ici notre attention.

À titre illustratif, nous nous intéressons dans la suite de cet article à l'étude des classes sémantiques construites pour le thème « nouvelles technologies de l'information ». Celui-ci donne lieu à la constitution d'un arbre de classification dont nous extrayons manuellement une soixantaine de classes rassemblant de deux à huit noms, comme par exemple {*ordinateur*, *micro-ordinateur*}, {*système*, *programme*, *dispositif*}, {*groupe*, *atelier*, *corporation*, *câblo-opérateur*, *entreprise*, *opérateur*, *firme*, *compagnie*}, {*mondialisation*, *déréglementation*, *globalisation*, *dérégulation*}, {*intelligence*, *cerveau*}, ou {*technologie*, *technique*}.

Les données employées par la méthode de recherche des nuances décrite dans cet article sont la liste de ces classes, les données textuelles du sous-corpus thématique associé, ainsi que la mesure de proximité sémantique générale p, que nous sommes amenés à utiliser de nouveau.

Les fondations techniques de notre recherche étant ainsi posées, nous introduisons à la section suivante les motivations théoriques de notre approche afin de préciser celle-ci.

3 Motivations théoriques

Notre objectif est de rassembler des paires de mots telles que leurs éléments sont séparés par des distinctions sémantiques similaires. Ces distinctions doivent, comme il est inévitable de le faire en acquisition automatique d'informations lexicales sémantiques, être inférées à partir de l'observation des usages des mots dans le corpus étudié — et en particulier, naturellement, de leurs différences. Nous nous dirigeons donc vers une comparaison de caractérisations d'usages de mots, la principale question posée étant alors celle de sélection des données textuelles les plus pertinentes pour rendre manifestes les nuances de sens entre mots.

Les auteurs de (Pichon & Sébillot, 1999) font usage des adjectifs apparaissant à proximité immédiate des mots étudiés, mais il semble que cela ne constitue pas un choix optimal : en effet, une étude menée sur l'ensemble des paires nom-adjectif d'un extrait de cent paragraphes de notre corpus d'étude nous a montré que seuls 7 % des paires N-A ainsi extraites correspondaient à des cas où l'adjectif est révélateur d'une spécificité du nom. Il semble en fait que les relations sémantiques entre mots proches (en position) dans le texte soient le lieu de l'expression du « commun », propre à assurer l'impression de cohésion à la lecture du texte, alors que les nuances et distinctions jouent un rôle plus important dans la structuration des énoncés au niveau du raisonnement, de l'articulation des idées, de la « dialectique ».

C'est pourquoi nous avons choisi de retenir pour caractériser les usage des mots étudiés non pas leurs voisins immédiats, mais un contexte plus large, de l'échelle de la phrase entière. Nous présentons à la section suivante la technique de comparaison entre paires de mots fondée sur ce principe que nous avons mise en œuvre.

4 Méthode développée

Afin de maximiser les chances de voir s'opérer comme nous le souhaitons des rapprochement entre paires de mots dont les deux éléments sont distingués par une même nuance, il est naturellement nécessaire de ne pas s'intéresser à la structuration d'une unique classe sémantique : plus les paires de mots étudiées sont nombreuses, plus les chances de découvrir parmi elles des ensembles de paires représentatifs de nuances spécifiques augmentent. C'est pourquoi nous considérons simultanément toutes les classes sémantiques construites pour un thème donné grâce aux traitements rapidement introduits à la section 2.2, et cherchons des points communs entre toutes les paires de mots telles que leurs deux éléments appartiennent à une même classe.

Nous présentons ici dans un premier temps le mode de sélection des indices textuels permettant de caractériser ces paires, puis la mesure de similarité que nous avons développée afin de rapprocher celles correspondant à des différences d'usage similaires.

4.1 Représentation des informations contextuelles pour une paire de mots

Conformément au principe adopté de travail « à longue portée », nous considérons pour chaque occurrence des noms à caractériser un voisinage constitué par la totalité des mots de la phrase où elle apparaît, à l'exception de ses voisins les plus proches (dans une fenêtre d'exclusion de trois mots à droite et à gauche de chaque occurrence). Les catégories de mots retenues pour prendre part à la caractérisation sont les noms, verbes et adjectifs.

À chaque nom m étudié, nous associons l'ensemble I_m de tous les noms, verbes et adjectifs apparaissant dans les mêmes phrases que m mais éloignés de lui par une distance minimale de 3 mots. Si plusieurs mots d'une même classe sémantique sont simultanément présents dans une même phrase, nous n'associons pas tous les mots de la phrase à chacun d'eux, puisque notre but est de faire apparaître autant que possible des différences d'usage. Chaque mot-indice est donc rattaché à la caractérisation du mot qui est le plus proche de lui, en nombre de positions, dans la phrase.

Les objets de cette étude sont les paires de mots (m_1, m_2) telles que m_1 et m_2 appartiennent à une même classe sémantique. Notre objectif est de définir une mesure de similarité dont la valeur soit d'autant plus élevée que, pour deux paires (m_1, m_2) et (m'_1, m'_2), les différences d'usages distinguant m_1 de m_2 sont similaires à celles distinguant m'_1 de m'_2. Afin de caractériser cette différence d'usage, nous associons à chaque paire (m_1, m_2) une paire d'ensembles (E_1, E_2) définis par $E_1 = I_{m_1} \setminus I_{m_2}$ et $E_2 = I_{m_2} \setminus I_{m_1}$. E_1 contient donc l'ensemble des indices contextuels exclusivement associés à m_1, et E_2 ceux exclusivement associés à m_2.

4.2 Calcul d'une « similarité de distinction » entre paires de mots

La similarité entre deux paires de mots est calculée par un indice de Jaccard modifié afin de prendre en compte l'information portée par la mesure de « proximité sémantique générale » p entre mots calculée au cours de la constitution des classes sémantiques et déjà évoquée à la section 2.2. La modification consiste à comparer non pas directement les ensembles de voisins, mais des « ensembles flous » dérivés de ceux-ci grâce à p.

Si nous travaillons sur deux paires de mots (m_1, m_2) et (m'_1, m'_2), respectivement caractérisées par les paires d'ensembles (E_1, E_2) et (E'_1, E'_2), nous définissons tout d'abord un « cardinal d'intersection floue » cif par :

$$cif\left(E_1, E'_1\right) = \frac{1}{2}\left[\sum_{m_1 \in E_1} \max_{m_2 \in E'_1}\left(p\left(m_1, m_2\right)\right) + \sum_{m_2 \in E'_1} \max_{m_1 \in E_1}\left(p\left(m_2, m_1\right)\right)\right] \tag{1}$$

Les valeurs de p étant normalisées pour appartenir à l'intervalle $[0, 1]$, cette formule peut être comprise comme le cardinal de l'intersection de deux ensembles flous F_1 et F'_1 tels que le degré d'appartenance d'un mot à F_1 (resp. F'_1) soit égale à sa proximité maximale avec l'un des mots de E_1 (resp. E'_1). Nous l'employons pour calculer la similarité σ entre deux ensembles de mots-indices E_1 et E'_1 par la formule :

$$\sigma\left(E_1, E'_1\right) = \frac{cif\left(E_1, E'_1\right)}{\text{Card}\left(E_1\right) + \text{Card}\left(E'_1\right) - cif\left(E_1, E'_1\right)} \tag{2}$$

σ n'est autre qu'un indice de Jaccard adapté afin de prendre en compte l'information fournie par p. Nous calculons enfin la similarité s entre les deux paires de mots considérées en réalisant la moyenne des similarités observées entre leurs ensembles d'indices E_1 et E'_1 d'une part, E_2 et E'_2 d'autre part :

$$s\left(\left(m_1, m_2\right), \left(m'_1, m'_2\right)\right) = \frac{\sigma\left(E_1, E'_1\right) + \sigma\left(E_2, E'_2\right)}{2} \tag{3}$$

Les mesures de similarité entre paires de mots ainsi calculées sont statistiquement normalisées afin que chaque ligne et colonne de la matrice de similarité ait une moyenne nulle et un écart-type unitaire selon une méthode présentée dans (Rossignol & Sébillot, 2006). Une exploration manuelle des rapprochements entre paires suggérés par cette mesure montre l'existence de regroupements « intéressants », faisant sens par rapport à notre objectif. Néanmoins, toutes nos tentatives d'exploitation directe de ces valeurs de similarité par des méthodes de classification classiques mènent à une perte d'information importante par rapport aux rapprochement pertinents qu'une analyse manuelle des similarités permet de mettre au jour. Nous avons donc défini une procédure d'exploitation manuelle de ces résultats évitant autant que possible l'« enrichissement » de ceux-ci par des connaissances non intrinsèques aux textes considérés.

5 Procédure de filtrage manuel des résultats

La tâche d'analyse des similarités entre toutes les paires de mots étudiées étant clairement insurmontable manuellement, nous limitons le volume des propositions soumises à validation : à chaque paire de mots est associée la liste des cinq paires présentant avec elle les similarités les plus fortes (que nous nommons « liste de connexions »). Un premier mode de validation possible consiste à proposer à l'utilisateur de juger chacune des associations ainsi retenues (selon le critère « existe-t-il un point commun entre la nuance distinguant les mots de la paire A et celle distinguant les mots de la paire B ? »), puis à extraire du réseau des liens ainsi validés des composantes fortement connexes, lesquelles constituent alors des classes révélatrices de nuances particulières. Outre sa lourdeur pour l'utilisateur chargé d'effectuer la validation (les associations restent nombreuses), cette approche peut donner lieu à de nombreuses « dérives » : liens validés en fonction de choix antérieurs mais de manière non systématique, *etc*.

C'est pourquoi nous réalisons la validation des rapprochements de paires en deux temps. Tout d'abord, en parcourant rapidement les listes de connexions, nous avons relevé les associations entre paires les plus immédiatement marquantes selon un critère d'analogie (« a_1 est à b_1 ce que a_2 est à b_2 »), et associé à chacune de ces associations une dénomination reflétant la relation apparente entre les extrémités des deux paires ainsi rassemblées. Dans le domaine des « nouvelles technologies », par exemple, on peut noter le rapprochement des paires (*imprimerie*/*photographie*) et (*télécopieur*/*téléviseur*) et donner à la nuance commune le nom « texte / image ».

La validation à proprement parler des résultats est ensuite assistée par un système automatisé au fonctionnement assez simple : pour chaque paire de mot étudiée p, il construit l'ensemble des « noms de nuances » définis à l'étape précédente qui ont déjà été validés pour une au moins des paires présentes dans la liste de connexions de p, puis propose au « valideur » de confirmer ou infirmer la pertinence de ces nuances pour caractériser la différence de sens distinguant les éléments de p. Par exemple, si la paire (*imprimerie*/*photographie*) fait partie de la liste de connexions de p, le valideur devra évaluer la validité de la nuance « texte / image » pour distinguer les deux mots de p. Les associations validées sont prises en compte par le système afin de générer d'éventuelles nouvelles propositions impliquées par ces nouvelles connaissances. Le travail de validation se termine lorsque toutes les possibilités ont été épuisées.

L'ensemble des deux étapes décrites correspond à une heure de travail humain environ pour étudier les propositions faites sur un ensemble d'une soixantaine de classes sémantiques. Nous présentons maintenant les résultats auxquels elles permettent d'aboutir.

concret / abstrait

donnée	–	information
autoroute	–	réseau
dimension	–	échelle
courrier	–	messagerie
cerveau	–	intelligence
expérimentation	–	recherche
machine	–	logiciel
dispositif programme	–	système
matériel équipement	–	infrastructure
navigation transmission transport	–	communication

texte / image

imprimerie	–	photographie
tirage	–	audience
télécopieur	–	téléviseur

FIG. 1 – Paires de mots du domaine « nouvelles technologies » distinguées par les nuances « concret / abstrait » et « texte / image »

6 Résultats

Les résultats présentés ici ont principalement une valeur indicative et illustrative de l'intérêt des résultats que la méthode — très expérimentale — proposée peut permettre de faire apparaître.

Sur le thème des « nouvelles technologies », l'ensemble de traitements présentés permet la mise au jour des 7 nuances suivantes : « concret / abstrait », « simple / sophistiqué », « agressif / consensuel », « condition de », « ancien / moderne », « production / consommation » et « texte / image ». La figure 1 présente une sélection de quelques unes parmi les vingt paires de mots environ reconnues comme distinguées par la nuance « concret / abstrait », ainsi que les trois correspondant à la nuance « texte / image ».

On peut noter à la lecture des résultats partiels présentés que leur validité dépend de l'interprétation des mots considérés, et peut requérir un retour au texte afin d'être réllement avérée. C'est le cas par exemple de la paire (*autoroute*/*réseau*), où *autoroute* est compris dans le sens d'« autoroute de l'information », son seul usage dans le sous-corpus associé au thème « nouvelles technologies ». Il en va de même pour *navigation*, compris dans le sens de « navigation sur Internet ». Le fait que ces deux mots gardent, malgré leur usage exclusivement métaphorique, une connotation de « concret » montre que les auteurs se plaçant sur le registre de la métaphore au monde matériel pour évoquer les réseaux électroniques tendent à se maintenir sur ce registre, tissant ainsi dans leur texte un « réseau sémantique » concret.

Les résultats obtenus pour d'autres thèmes du corpus sont de volume variable : au moins trois nuances sont en général mises au jour, la moyenne s'établissant à cinq si l'on ne considère que les thèmes donnant naissance des sous-corpus d'au moins 100 000 mots — en-deçà de cette valeur, les résultats se dégradent par manque de données. On constate que la plupart des thèmes donnent lieu à l'émergence d'une nuance dominante, souvent fortement liée à un discours « stéréotypé » développé autour de ce thème : pour les « nouvelles technologies », il s'agit de l'opposition « concret / abstrait » présentée ; pour le thème de la criminalité, la gradation que l'on pourrait nommer « plus grave / moins grave » est très fréquente et distingue aussi bien les crimes que les peines associées ; autour du thème des « arts du spectacle », la distinction dominante est triple, séparant théâtre, musique et cinéma.

Quoique la technique proposée permette de mettre au jour de nombreuses nuances, certaines d'entre elles, dont on aurait pu attendre l'apparition restent absentes. Autour du thème de la criminalité, par exemple, les distinctions entre civils et militaires parmi les acteurs des événements relatés, ou entre forces légales (*police*, *armée*, *etc.*) et criminelles (*gang*, *mafia*, *etc.*) dans les classes sémantiques regroupant des groupes organisés, ne sont pas détectées par le système. Nous proposons à la section suivante, en conclusion de cette présentation, quelques directions de recherches pouvant permettre d'améliorer ces résultats.

7 Conclusion

Nous avons présenté dans cet article un premier système permettant de mettre au jour, de manière semi-automatique et à partir des seules données textuelles, des éléments de distinction sémantique fine entre mots appartenant à une même classe, atteignant ainsi un degré de définition du sens encore inédit en acquisition automatique sur corpus d'information sémantiques lexicales. La technique mise au point réalise en s'appuyant sur l'étude de grands voisinages autour des occurrences des mots comparés un regroupement de paires de mots distingués par des nuances similaires. Cette approche présente, par rapport à des travaux antérieurs (Pichon & Sébillot, 1999), la faiblesse de ne permettre qu'une représentation implicite des nuances découvertes : les listes de paires de mots rapprochées doivent être interprétées afin de « comprendre » l'élément de distinction commun. En revanche, elle permet une automatisation bien plus importante du processus de recherche de sèmes spécifiques, suffisante pour assurer que le travail humain de validation des résultats ne puisse introduire dans ceux-ci de biais interprétatif trop important.

Cette première recherche ne propose naturellement pas de réponse complète à la problématique infiniment complexe abordée, et de nombreux développement doivent encore être menés à bien. Une première nécessité est naturellement d'étendre les recherches aux autres catégories morphosyntaxiques « pleines » que le nom, sur lesquels se sont focalisés nos efforts. Les études préliminaires que nous avons réalisées dans ce sens suggèrent que, le phénomène linguistique considéré étant plus « fin » que la simple similarité sémantique permettant la constitution de classes de mots, des techniques différentes devront sans doute être développées et adaptées à chaque catégorie de mots. Loin d'être un handicap, cela constitue au contraire une source de richesse potentielle très importante : si par exemple un travail sur les adjectifs permet l'identification au sein de ceux-ci d'une opposition « mélioratif / dépréciatif » non identifiée par ailleurs pour les noms, il peut être possible de la projeter sur ceux-ci grâce aux associations noms-adjectifs.

Plus essentiellement, nous avons identifié le principe d'une détection des nuances par analyse des phrases d'apparition des mots à caractériser considérées dans leur totalité : c'est là à la fois trop et trop peu d'information. Trop car, naturellement, cette approche introduit une quantité de « bruit » très importante dans les données étudiées ; il est donc nécessaire dans un premier temps d'étudier empiriquement les motifs linguistiques pouvant être révélateurs de nuances sémantiques afin de tenter d'orienter plus efficacement la sélection des indices contextuels, l'« éparpillement » de ceux-ci justifiant d'ailleurs peut-être que l'approche purement statistique que nous avons adoptée soit abandonnée en faveur de techniques d'apprentissage symbolique. Trop peu, car le choix de l'échelle textuelle considérée elle-même demande également à être approfondi : si nous restons persuadé que le niveau du développement des idées est bien le plus pertinent pour la tâche entreprise, l'approximation de celui-ci par la phrase est un pis-aller discutable. Là encore, il paraît nécessaire afin de découvrir les indices pouvant permettre un « découpage » optimal de soumettre les données disponibles à une exploration linguistique.

Références

ARMSTRONG S. (1996). Multext: Multilingual Text Tools and Corpora. In *Lexikon und Text*, Niemeyer, Allemagne.

CARABALLO S. A. & CHARNIAK E. (1999). Determining the Specifity of Nouns from Text. In *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing (EMNLP) and Very Large Corpora (VLC)*, University of Maryland, College Park, MD, EU.

FABRE C., HABERT B. & LABBÉ D. (1997). La polysémie dans la langue générale et les discours spécialisés. *Sémiotiques*, **13**, 15–30.

GREFENSTETTE G. (1994). Corpus-derived First, Second and Third Order Word Affinities. In *6th Congress of the European Association for Lexicography (Euralex 94)*, Amsterdam, Pays-Bas.

GREIMAS A. J. (1966). *Sémantique Structurale*. Paris, France: Larousse.

HABERT B., NAULLEAU É. & NAZARENKO A. (1996). Symbolic Word Clustering for Medium-Size Corpora. In *16th International Conference on Computational Linguistics (COLING 96)*, Copenhague, Danemark.

HARRIS Z. (1968). *Mathematical Structures of Language*. New York, NJ, EU: John Wiley & Sons.

HINDLE D. (1990). Noun Classification from Predicate-Argument Structures. In *28st Annual Meeting of the Association for Computational Linguistics (ACL 90)*, Pittsburgh, PA, EU.

LEPAGE Y. (2004). Lower and Higher Estimates of the Number of True Analogies Contained in a Large Multilingual Corpus. In *20th International Conference on Computational Linguistics (COLING 04)*, Genève, Suisse.

PICHON R. & SÉBILLOT P. (1999). Différencier les sens des mots à l'aide du thème et du contexte de leurs occurrences : une expérience. In *6e conférence francophone internationale sur le Traitement Automatique des Langues Naturelles (TALN 99)*, Cargèse, France.

RASTIER F. (1996). *Sémantique Interprétative*. Paris, France: Presses Universitaires de France, 2e edition. 1ère édition en 1987.

ROSSIGNOL M. & SÉBILLOT P. (2003). Extraction statistique sur corpus de classes de mots-clés thématiques. *TAL (Traitement automatique des langues)*, **44**(3), 217–246.

ROSSIGNOL M. & SÉBILLOT P. (2006). Acquisition sur corpus non spécialisés de classes sémantiques thématisées. In *8èmes Journées internationales d'analyse statistiques des données textuelles (JADT 2006)*, Besançon, France.