L'analyse sémantique latente et l'identification des métaphores

Yves Bestgen et Anne-Françoise Cabiaux

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

Résumé

Après avoir présenté le modèle computationnel de l'interprétation de métaphores proposé par Kintsch (2000), nous rapportons une étude préliminaire qui évalue son efficacité dans le traitement de métaphores littéraires et la possibilité de l'employer pour leur identification.
Having introduced Kintsch's computational model of metaphor comprehension (2000), we report a preliminary study aiming at determining its efficiency in modeling the processing of literary metaphors and its ability to detect them.

Mots Clés

Métaphore - Analyse sémantique latente - interprétation et détection - textes littéraires
Metaphor - Latent semantic analysis - interpretation and detection - literary texts

1 Introduction

De nombreux auteurs ont souligné combien les énoncés que nous produisons quotidiennement contiennent un grand nombre d'expressions figuratives qui posent des problèmes tant aux modèles psychologiques de la compréhension du langage qu'aux approches linguistiques en traitement automatique de la langue (Gibbs, 1994; Martin, 1992). Durant ces dix dernières années, ces deux disciplines ont fait d'importants progrès tout particulièrement au niveau du traitement de la métaphore. De plus en plus de données empiriques sont venues étayer les hypothèses des chercheurs quant à la manière dont nous comprenons une métaphore (Cacciari, Glucksberg, 1994; Kintsch, 2000). Simultanément, des méthodes pour détecter et interpréter automatiquement les métaphores ont été proposées et implémentées (Fass, 1991; Ferrari, 1996; Ferrari et al., 2000; Martin, 1992). Si ces disciplines sont loin de s'ignorer, leurs visées très différentes n'ont jusqu'à présent pas permis de réel rapprochement. Récemment toutefois, Kintsch (2000) a entrouvert la voie à un tel rapprochement en proposant un "modèle computationnel" de l'interprétation des énoncés métaphoriques. Après avoir décrit l'algorithme qu'il propose, nous présentons une étude préliminaire qui confronte cet algorithme à des métaphores variées d'origine littéraire et qui teste la possibilité de l'appliquer à leur détection.

2 Un modèle computationnel de l'interprétation des métaphores basé sur l'analyse sémantique latente

Kintsch (2000) propose un modèle computationnel qui s'appuie sur la conception interactive de l'interprétation des métaphores. Selon celle-ci, tant le véhicule[1] que la topique contribuent au sens de la métaphore puisque le véhicule propose des propriétés parmi lesquelles la topique sélectionne celles qui sont acceptables et les adaptent en fonction de ses propriétés sémantiques. Par exemple, dans la métaphore reprise par Kintsch à Glucksberg "*Mon avocat est un requin*", *avocat* (topique) sélectionne les traits de *requin* (véhicule) qui peuvent lui être attribués. Ce seront par exemple *sanguinaire* ou *vicieux*. Une caractéristique importante de cette conception est qu'elle peut être étendue à n'importe quelle prédication, la topique y jouant le rôle de l'argument que le véhicule-prédicat enrichit de certaines de ses propriétés (Glucksberg, McGlone, 1999; Kintsch, 2001). Ceci rejoint la thèse, actuellement privilégiée en psycholinguistique, selon laquelle une métaphore est comprise par les mêmes processus mentaux que ceux qui s'appliquent aux énoncés dont le sens littéral est pertinent (Gibbs, 1994; Glucksberg et al., 1997; voir Martin (1992) pour un autre emploi de cette même thèse).

Pour implémenter cette conception, il est nécessaire de pouvoir identifier les traits sémantiques qui participeront au sens de la métaphore et de proposer un algorithme capable d'effectuer la sélection. La première composante est fournie par *l'analyse sémantique latente* (ASL). Issue de travaux sur l'indexation automatique de documents, cette technique vise à construire un espace sémantique de très grandes dimensions à partir de l'analyse statistique des cooccurrences dans un corpus de textes[2]. Le sens de chaque mot y est représenté par un vecteur. Pour mesurer la similarité sémantique entre deux mots, on calcule le cosinus entre les vecteurs qui les représentent. Plus deux mots sont sémantiquement proches, plus les deux vecteurs qui les représentent pointent dans la même direction et donc plus leur cosinus se rapproche de 1. Un cosinus de 0 indique une absence de similarité puisque les vecteurs correspondants sont orthogonaux. L'algorithme employé pour déterminer le sens d'une prédication vise à sélectionner parmi les "traits" du prédicat ceux qui sont proches de l'argument. On procède en recherchant parmi les n plus proches voisins du prédicat les k plus proches voisins de l'argument. Afin de garantir que les termes sélectionnés sont suffisamment liés aux deux éléments de la prédication, un seuil de proximité minimal est imposé aux termes sélectionnés. Le sens de la prédication est alors déterminé en prenant le centroïde du prédicat, de l'argument et des k termes qui viennent d'être sélectionnés, c'est-à-dire en additionnant les vecteurs correspondants. L'adéquation du "sens" attribué par cette procédure à une prédication, et donc aussi à une métaphore, est évaluée en déterminant la proximité entre ce nouveau vecteur et des points de repères considérés comme proches du sens de la métaphore (p.e., *vicieux, sanguinaire* pour la métaphore *mon avocat est un requin*).

Selon ce modèle, le seul facteur qui change lorsqu'on analyse, non un énoncé littéral, mais une métaphore, est le paramètre n, c'est-à-dire le nombre de voisins du prédicat parmi lesquels on cherche les plus proches voisins de l'argument. Pour un énoncé littéral, on se limite aux 20 plus proches voisins, alors que pour un énoncé métaphorique, il faut aller jusqu'à 200 voire

---

[1] Sur la base d'une définition très générale de la métaphore comme *une manière d'expliquer quelque chose en utilisant les termes d'autre chose*, on définira la topique comme la cible de la métaphore (ce dont on veut parler) et le véhicule comme la source de la métaphore (ce qui permet d'en parler).

[2] Une description détaillée de la technique ASL (Latent semantic analysis — LSA) peut être trouvée dans les nombreux articles téléchargeables à l'adresse *http://LSA.colorado.edu/* dont Landauer et al. (1998)

500.

3 Notre recherche

Les arguments empiriques avancés par Kintsch (2000, 2001) pour soutenir son modèle computationnel sont très réduits. Il n'a étudié qu'un tout petit nombre de métaphores (7), toutes issues d'un matériel expérimental et de type attributif. Par ailleurs, la possibilité d'appliquer cette technique à l'identification d'énoncés métaphoriques n'a pas été envisagée. La présente recherche est un premier pas vers la prise en compte de ces limitations. Tout d'abord, nous analysons l'efficacité de l'algorithme lorsqu'il est appliqué à des métaphores littéraires de différents types. Ensuite, nous prenons en compte l'idée centrale d'une continuité entre énoncés littéraux et métaphoriques en vérifiant que l'algorithme fonctionne avec des expressions perçues par des lecteurs comme très métaphoriques ou peu métaphoriques. Enfin, nous rapportons les résultats d'une première tentative pour dériver un indice de l'intensité figurative d'une prédication sur la base de l'algorithme.

3.1 Matériel

3.1.1 Sélection des métaphores

Vingt phrases contenant des expressions métaphoriques ont été sélectionnées dans neuf contes de Maupassant. Dix phrases exprimaient une métaphore vive et 10 une métaphore morte. Ont été considérées comme *mortes* les métaphores employant des mots dans un sens que le dictionnaire *Petit Robert* qualifie de figuratif alors que le sens des mots employés dans les métaphores considérées comme vives n'était pas mentionné dans ce même dictionnaire. La liste des expressions métaphoriques est donnée dans le Tableau 1.

3.1.2 Estimation de l'intensité figurative

Pour cette tâche de jugement, le matériel principal était composé des 20 expressions métaphoriques. Toutefois, afin que les juges n'évaluent pas que des énoncés a priori métaphoriques, une version littérale de chaque énoncé métaphorique a été écrite. Par exemple, la version littérale de "*la voix sourde et profonde du torrent* " était "*le bruit sourd et profond du torrent*". Au total, 40 personnes ont indiqué s'ils considéraient "*que la formulation de l'énoncé est plutôt littérale ou plutôt figurée*" en utilisant une échelle graduée allant de *tout à fait littéral* (1) à *tout à fait figuré* (7). Chaque juge a évalué 20 expressions : 10 en version métaphorique et les 10 autres en version littéralisée. Aucun juge n'a vu les deux versions d'un même énoncé.

Comme on pouvait s'y attendre, les juges ont été très sensibles à l'opposition entre les expressions métaphoriques (score minimum de 4,4) et les expressions littéralisées (score maximum de 3,8). L'accord inter-juges, mesuré par l'alpha de Cronbach, pour l'ensemble des expressions est excellent (0,97) et reste acceptable lorsqu'on analyse les seuls énoncés métaphoriques (0,76). Comme on peut le voir dans le Tableau 1, les métaphores vives ont généralement été évaluées comme plus figuratives que les métaphores mortes. La différence entre les valeurs moyennes pour les 10 métaphores vives (5,8) et celles pour les 10 métaphores mortes (5,2) est très significative selon un test *t* pour échantillons indépendants ($t(1,18) = 3,07$; $p<,01$). On observe aussi que certaines expressions ont été jugées comme

nettement plus figuratives que d'autres.

3.2 Constitution de l'espace sémantique

Le corpus de textes utilisés pour constituer l'espace sémantique est constitué de 206 contes de Maupassant, soit tous les contes que nous avons pu trouver en version électronique, à l'exception des neuf contes dans lesquels les métaphores ont été reprises. Au total, ce corpus contient 600 000 mots. Chaque conte a été segmenté en fonction des paragraphes. Toutefois, un paragraphe comptant moins de 50 mots était réuni avec le ou les suivants jusqu'à que cette taille minimale soit atteinte. On a ainsi obtenu 6567 segments. Les mots ont été lemmatisés par comparaison avec une liste de formes fléchies et des lemmes correspondants. Tous les mots dont la fréquence dans le corpus était inférieure à 3 ont été supprimés. Après ces différents traitements, le corpus contenait 6067 mots différents. La matrice de cooccurrence des 6067 mots dans les 6567 segments a été soumise à une décomposition en valeurs singulières réalisée par le programme SVDPACK (Berry, 1992) et les 150 premiers vecteurs propres ont été conservés.

	Expressions Métaphoriques	Mots décrivant la métaphore	Fig	Top	ASL	Diff
V	un orage de *sottises*	violent inattendu excessif énorme	6,3	0,07	0,12	0,05
V	des fusées de *gaieté*	explosion brusque éclater violence	6,2	0,25	0,31	0,06
V	le *palais de justice* est l'égout des infamies	vil écœurement aboutir échouer	6,2	-0,03	0,04	0,07
V	l'abbaye est un feu d'artifice de *pierre*	léger superbe étonnant	6,1	0,10	0,25	0,15
V	la *lune* verse une pluie de lumière	généreux répandre distribuer	6,0	0,26	0,29	0,03
V	la voix sourde et profonde du *torrent*	humain communication parole vibrer	6,0	0,07	0,25	0,18
V	les *arbres* bataillent contre le vent	résister volonté difficile	5,9	0,07	0,25	0,18
V	cueillir les *mots* sur la bouche	rechercher attraper délicat doux	5,9	0,22	0,29	0,07
M	le *jour* mourant	fin tristesse agonie	5,6	0,20	0,32	0,12
M	voler *l'œil* tendu vers quelque chose	volonté désir attrait	5,6	0,20	0,22	0,02
M	vider le sac des *arguments*	quantité paquet désordre	5,5	0,13	0,23	0,10
M	être cloué à son *fauteuil*	impuissance souffrance pénible	5,5	0,16	0,23	0,07
V	la *mer* est perfide	dangereux traître méchant sournois	5,4	0,18	0,22	0,04
M	s'écrier d'une *voix* de tonnerre	peur bruyant violent	5,2	0,17	0,26	0,09
M	être soulevé *d'admiration*	intensité irrésistible émotion force	5,0	0,19	0,33	0,14
M	porter son *chagrin*	lourd fardeau pénible	5,0	0,15	0,34	0,19
M	la *voix* s'envole sans écho	léger libre monter	4,8	0,15	0,21	0,06
M	la reine des *gibiers*	solennel important premier majesté	4,8	0,11	0,15	0,04
M	jeter un *aboiement* furieux	lancer brutal violent force	4,7	0,11	0,41	0,30
V	la nappe transparente de *l'eau*	lisse vaste plat couvrir	4,4	0,33	0,37	0,04

Tableau 1 : Liste des métaphores vives (V) et mortes (M) analysées et résultats obtenus : intensité figurative estimée par les juges (Fig), cosinus entre la topique et le "sens" (Top), cosinus entre le centroïde pour l'ASL et le "sens" (ASL), différence entre ces deux cosinus (Diff). La topique est en italique et le véhicule est souligné. Les métaphores sont rangées par ordre décroissant d'intensité figurative.

3.3 Analyses et résultats

3.3.1 L'ASL permet-elle d'approximer le sens de la métaphore?

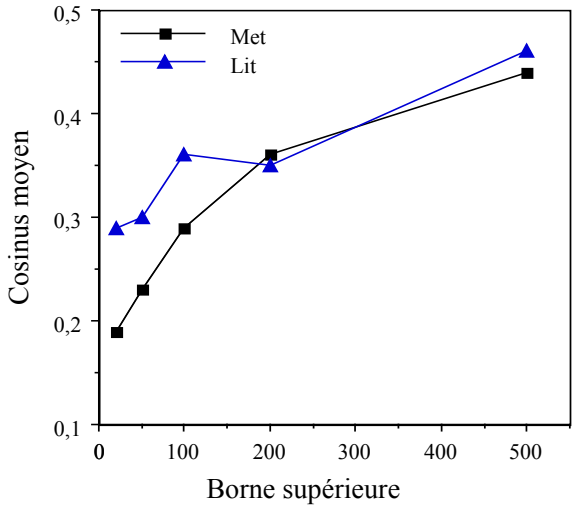
Pour répondre à cette question, nous avons employé la procédure proposée par Kintsch. Nous

avons calculé le cosinus entre le centroïde construit par l'ASL (paramètres : n=500, k=5 et un seuil minimum pour les cosinus de 0.216) et le centroïde représentant le sens de la métaphore déterminé sur la base de 3 à 4 mots. Nous avons aussi calculé le cosinus entre la topique et ce même centroïde représentant le sens de la métaphore. Le Tableau 1 donne pour chaque paire Véhicule—Topique, les mots qui représentent le sens de la métaphore. Si l'algorithme est capable d'approximer ce sens, le centroïde qu'il construit devrait être plus proche de celui-ci que ne l'est la topique. Le Tableau 1 donne les valeurs de ces deux cosinus et la différence obtenue en soustrayant le cosinus pour la topique du cosinus pour l'ASL. Toutes les différences sont positives, soulignant qu'à chaque fois le centroïde proposé est plus proche que la topique du sens de la métaphore. En moyenne, cette différence est de 0,10, valeur très significativement différente de 0 (t(19)=6,31; p>,0001). On notera toutefois que l'algorithme a été peu efficace lors de la recherche du sens de l'expression "*Le palais de justice est l'égout des infamies*" puisque le cosinus entre le centroïde pour l'ASL et le sens attendu n'est que de 0,04.

3.3.2 L'efficacité varie-t-elle selon l'intensité figurative des métaphores?

Nous avons essayé de répondre à cette question de deux manières. Tout d'abord, nous avons comparé l'efficacité de l'ASL pour les métaphores vives et mortes. La réponse à cette question est clairement négative. On n'observe aucune différence entre les cosinus moyens selon que les métaphores sont considérées comme mortes ou comme vives. La deuxième approche a consisté à comparer l'interprétation proposée par l'ASL avec les évaluations des juges. On observe une corrélation négative significative (r = -0,45; p<,05) entre l'efficacité de l'ASL et les évaluations moyennes des juges. L'ASL approxime donc d'autant mieux le sens d'une métaphore que les juges l'ont jugée peu figurative. Ce résultat semble à première vue problématique pour l'approche. Toutefois, on peut aussi le voir comme positif parce que plus un énoncé est figuratif, plus les lecteurs ont des difficultés pour le comprendre (Miall, Kuiken, 1994). L'ASL peinerait donc là où les lecteurs rencontrent aussi des difficultés.

3.3.3 Peut-on distinguer les versions métaphoriques des versions littérales?



Comme indiqué dans l'introduction, la procédure proposée par Kintsch s'applique aussi bien aux énoncés littéraux que métaphoriques. La seule différence est qu'il est nécessaire pour un énoncé métaphorique d'accroître le nombre de voisins du prédicat (véhicule) pris en compte pour trouver des concepts suffisamment associés à l'argument (topique). Dans le cas d'énoncés littéralement vrais, on devrait trouver parmi les mots les plus proches du prédicat des mots proches de l'argument. Par contre, dans le cas des énoncés métaphoriques, les mots proches du prédicat devraient être peu liés à l'argument. Ce n'est que lorsqu'on prend en compte des mots de moins en moins associés au prédicat qu'on peut espérer trouver pour la métaphore des mots suffisamment associés à l'argument. Sur la base

de ce raisonnement, nous avons construit un test afin de déterminer si l'ASL est capable de distinguer les énoncés métaphoriques des énoncés littéraux. Le matériel est composé des énoncés métaphoriques et des versions littérales construites pour la tâche de jugement. Pour chacun de ces énoncés, nous avons classé les mots de la base par ordre décroissant de leur cosinus avec le prédicat et nous avons analysé les 5 tranches suivantes : les mots occupant les places de 1 à 20, de 21 à 50, de 51 à 100, de 101 à200 et de 201 à 500. Dans chacune de ces tranches, nous avons recherché, suivant la procédure habituelle, les 5 mots les plus fortement associés à l'argument. La Figure 1 représente le cosinus moyen entre les 5 associés et l'argument pour les énoncés métaphoriques et littéraux. Les courbes sont en accord avec la prédiction. On observe une grande différence entre les énoncés métaphoriques et littéraux pour les premiers tronçons. Lorsque l'énoncé est métaphorique, les mots les plus associés au prédicat (1-20, 21-50, 51-100) sont moins liés à l'argument. Plus on accepte de prendre en compte des mots peu associés au prédicat (101-200, 201-500), plus cette différence se réduit.

4. Discussion et conclusion

Les résultats rapportés ci-dessus semblent encourageants. L'algorithme de Kintsch permet d'approximer le sens de métaphores littéraires de différents types et il a été possible d'en dériver un indice qui distingue les énoncés métaphoriques d'énoncés littéraux. Bien sûr, on n'insistera jamais assez sur le caractère exploratoire de la présente étude. Il se marque dans le matériel analysé (choix du véhicule et de la topique pour des expressions métaphoriques complexes) et dans une série de décisions arbitraires prises lors de la constitution du corpus pour l'espace sémantique (uniquement des contes de Maupassant segmentés en paragraphe d'au moins 50 mots et lemmatisés) et lors de la décomposition en valeurs singulières (150 dimensions). Même si les valeurs choisies correspondent aux recommandations issues d'études antérieures (voir note 1), chacune de ces décisions mériterait une étude empirique afin d'en déterminer l'impact sur l'efficacité de l'algorithme. La possibilité d'obtenir des résultats positifs avec un corpus d'une taille assez modeste est en soi très encourageant. Nous sommes néanmoins encore très loin de disposer, par la présente approche, d'une procédure automatique pour l'identification et l'interprétation de métaphores et de nombreuses autres recherches sont nécessaires afin de répliquer et d'étendre ces résultats.

Remerciements

Yves Bestgen est chercheur qualifié du Fonds National belge de la recherche scientifique. Cette recherche a bénéficié du soutien apporté par un crédit FRFC.

Références

Berry M.W. (1992), Large scale singular value computation. *International journal of Supercomputer Application*, Vol. 6, pp. 13-49.

Fass D. (1991), met*: A method for discriminating metonymy and metaphor by computer, *Computational Linguistics*, Vol. 17, pp. 49-90.

Ferrari S. (1996), A corpus-based approach for metaphor processing, Proceedings of *LEDAR Workshop on Language Engineering for Document Analysis and Recognition*, pp 114-121.

Ferrari S., Giguet E, Lucas N, Vergne J. (2000), Projet LINGUIX, recherche d'informations et traitements linguistiques : le cas des métaphores, Actes du *3ème colloque international sur le*

document électronique (CIDE 2000), pp. 279-293.
Cacciari C., Glucksberg S. (1994), Understanding figurative language. In Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 447-477), Academy Press.
Gibbs W.R. (1994), Figurative Thought and Figurative Language. In M.A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 411-446). Academy Press.
Glucksberg S., McGlone M.S. (1999), When love is not a journey: What metaphors mean, *Journal of Pragmatics*, Vol. 31, pp. 1541-1558.
Glucksberg S., McGlone M.S., Manfredi D. (1997), Property attribution in metaphor comprehension, *Journal of Memory and* Language, Vol. 36, pp.50-67.
Kintsch W. (2000), Metaphor comprehension: A computational theory, *Psychonomic Bulletin and Review*, Vol. 7, pp. 257-266.
Kintsch W. (2001), Predication, *Cognitive Science*, Vol. 25, pp. 173-202.
Landauer T.K., Foltz P.W., Laham D. (1998), An introduction to Latent Semantic Analysis. *Discourse Processes*, Vol. 25, pp. 259-284.
Martin J. (1992), Computer understanding of conventionnal metaphoric language, *Cognitive Science*, Vol. 16, pp. 233-270.
Miall D.S., Kuiken D. (1994), Foregrounding, Defamiliarization, and Affect Response to Literary Stories, *Poetics*, Vol. 22, pp. 389-407