L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ?

François Morlane-Hondère, Cécile Fabre

TALN 2010, Montréal, 19–23 juillet 2010

L'antonymie observée avec des méthodes de TAL : une relation à la fois syntagmatique et paradigmatique ?

François Morlane-Hondère Cécile Fabre
CLLE-ERSS, Université de Toulouse
francois.morlane@univ-tlse2.fr, cecile.fabre@univ-tlse2.fr

Résumé. Cette étude utilise des outils de TAL pour tester l'hypothèse avancée par plusieurs études linguistiques récentes selon laquelle la relation antonymique, classiquement décrite comme une relation paradigmatique, a la particularité de fonctionner également sur le plan syntagmatique, c'est-à-dire de réunir des mots qui sont non seulement substituables mais qui apparaissent également régulièrement dans des relations contextuelles. Nous utilisons deux méthodes – l'analyse distributionnelle pour le plan paradigmatique, la recherche par patrons antonymiques pour le plan syntagmatique. Les résultats montrent que le diagnostic d'antonymie n'est pas significativement meilleur lorsqu'on croise les deux méthodes, puisqu'une partie des antonymes identifiés ne répondent pas au test de substituabilité, ce qui semble confirmer la prépondérance du plan syntagmatique pour l'étude et l'acquisition de cette relation.

Abstract. In this paper, we use NLP methods to test the hypothesis, suggested by several linguistic studies, that antonymy is not only a paradigmatic but also a syntagmatic relation : antonym pairs, that have been classically described by their ability to be substituted for each other, also tend to frequently co-occur in texts. We use two methods – distributional analysis on the paradigmatic level, lexico-syntactic pattern recognition on the syntagmatic level. Results show that antonym detection is not significantly improved by combining the two methods : a set of antonyms do not satisfy the test for substitutability, which tends to confirm the predominance of the syntagmatic level for studying and identifying antonymy.

Mots-clés : sémantique lexicale, antonymie, analyse distributionnelle, patrons lexico-syntaxiques.

Keywords: lexical semantics, antonymy, distributional analysis, lexico-grammatical patterns.

1 Introduction

Nous nous intéressons à la relation d'antonymie et à son repérage automatique dans les textes, sur la base de travaux récents qui renouvellent les descriptions de cette relation lexicale. Les travaux de (Jones & Murphy, 2005), (Murphy, 2006), (Murphy *et al.*, 2009) se sont en effet intéressés à la façon dont les antonymes cooccurrent dans les textes, s'appuyant sur les observations antérieures de (Charles & Miller, 1989), (Justeson & Katz, 1991) et (Fellbaum, 1995), selon lesquelles les antonymes ont la particularité d'apparaître fréquemment ensemble – et même de marquer une opposition d'autant plus forte qu'ils cooccurrent significativement. L'ensemble de ces auteurs définissent ainsi l'antonymie comme une relation à la fois paradigmatique et syntagmatique. Paradigmatique, bien sûr, parce que deux antonymes s'opposent à l'intérieur d'un champ sémantique déterminé et se conforment donc au principe de substituabilité. Syntagmatique aussi parce que deux mots considérés comme antonymes tendent à apparaître ensemble dans des relations aussi bien inter qu'intrapropositionnelles et plus particulièrement dans des constructions contras-

tives étudiées par (Jones, 2002). Cette capacité des couples antonymiques à fonctionner simultanément sur ces deux plans est remarquable et, selon Murphy, ne serait pas partagée par les autres relations lexicales (synonymie, hyperonymie, co-hyponymie), pour lesquelles le plan paradigmatique prévaut.

Dans une expérience très récente, (Lobanova *et al.*, 2010) ont confirmé l'intérêt d'explorer cette dimension syntagmatique en montrant la possibilité d'acquérir des couples d'antonymes à partir d'une approche à base de patrons, sur le modèle de Hearst. Ce faisant, les auteurs signalent également les limites de cette approche : les patrons utilisés, pour être suffisamment précis, doivent être très spécifiques, ce qui limite le rappel. Et malgré cela, le taux de précision reste relativement réduit (entre 16% et 27% des paires de mots extraites sont jugées véritablement antonymiques). Notre objectif est donc de voir comment ce type de méthode peut se combiner avec une approche qui vise à vérifier l'appartenance du couple à un même paradigme distributionnel. Nous combinons l'approche par patrons, qui permet de détecter les cas de co-présence des antonymes, et l'approche distributionnelle, qui teste leur tendance à la substituabilité. Notre hypothèse est que la combinaison des deux méthodes devrait fournir la configuration optimale pour acquérir des couples d'antonymes. Ces deux méthodes de TAL offrent ainsi des moyens d'observer à grande échelle le comportement des couples antonymiques du point de vue paradigmatique et syntagmatique.

Dans la section suivante, nous présentons d'abord les deux méthodes de repérage des couples – par projection de patrons (2.1) puis par analyse distributionnelle automatique (2.2). Nous présentons ensuite l'évaluation de la méthode, qui combine une comparaison avec une ressource lexicale (3.1) et le recours à des jugements de locuteurs (3.2). Nous analysons enfin (3.3) les résultats obtenus, qui montrent que le critère de la similarité distributionnelle ne semble pas améliorer le taux de détection de la relation d'antonymie obtenu par la technique par patrons, suggérant une prédominance du fonctionnement syntagmatique pour une partie des couples.

2 Combiner deux modes de repérage de l'antonymie

Cette expérience est menée sur l'antonymie adjectivale, à partir d'un grand corpus d'articles encyclopédiques en français[1] tirés de Wikipédia.

2.1 Plan syntagmatique : paires d'adjectifs liés par des patrons antonymiques

Nous avons sélectionné un ensemble de constructions antonymiques, décrites sous la forme de patrons lexico-syntaxiques[2], en nous appuyant sur des travaux antérieurs, en particulier (Jones, 2002). Notre approche consiste à projeter automatiquement ces patrons sur l'ensemble du corpus Wikipedia, et à extraire les couples qui occupent les positions X et Y (cf. tableau 1). Ainsi traduites en patrons de surface, ces constructions ramènent une certaine quantité de bruit, en particulier le patron *X ou Y*, X et Y pouvant également entretenir une relation d'équivalence. Du fait de la prédominance de ce patron ambigu *X ou Y* – quasiment 74% des couples rapportés le sont dans cette structure – et de la bonne précision des relations minoritaires, nous avons décidé de nous limiter à un sous-ensemble de couples : ceux qui sont reliés par au moins deux patrons de type distinct, soit 907 couples différents (sur les 13 295 couples adjectivaux extraits

[1] Le corpus constitué de l'ensemble des articles de la version francophone de Wikipedia (avril 2007), soit plus de 470 000 articles pour 194 millions de mots. Son traitement ainsi que la création de la base de voisins sont dus au travail de Franck Sajous (CLLE-ERSS).

[2] Nous avons imposé un certain nombre de contraintes aux patrons – en termes de distance, ou en nous appuyant sur des informations morpho-syntaxiques disponibles dans le corpus annoté, que nous ne détaillons pas ici.

au total).

Malgré la rigueur du filtrage que nous avons imposé, toutes les paires de mots qui apparaissent dans deux patrons différents ne peuvent pas être considérées comme des paires d'antonymes. Tout d'abord, la méthode produit des erreurs dues à l'analyse syntaxique : le système extrait la paire *ancien/international* à partir du segment *Le recrutement de nombreux internationaux ou anciens internationaux reconnus*, dans lequel *international* est analysé comme un adjectif. On peut également évoquer le cas des locutions adjectivales comme *bon marché*, qui ne sont pas traitées comme telles par Syntex. Par ailleurs, comme l'ont montré (Jones, 2002) et (Lobanova *et al.*, 2010), les patrons peuvent repérer des oppositions ponctuelles, non lexicalisées, entre deux mots. Dans l'exemple ci-dessous, l'opposition entre les deux adjectifs *administratif* et *humain* est induite discursivement par la construction contrastive :

*Le terme, néanmoins, reflète des connotations **plus** humaines **qu'**administratives.*

On connaît de fait la capacité du discours d'instaurer des relations non lexicalisées (Hoey, 1991). Nous introduisons donc avec la méthode distributionnelle une contrainte supplémentaire, qui vise à s'assurer de la proximité sémantique entre les deux adjectifs.

PATRON	EXEMPLE
X ou Y	Cette connexion peut être **temporaire** *ou* **définitive**.
à la fois X et Y	Sa production est *à la fois* **fermière** *et* **industrielle**.
entre X et Y	les différences *entre* les vins **blancs** *et* les vins **rouges**
plus/plutôt/moins/autant/aussi (bien) X que Y	Il se déguste *aussi bien* **chaud** *que* **froid**.
X plutôt que Y	Il décrit une Terre **sphérique** *plutôt que* **plate**.
soit X soit Y	Les coups francs sont *soit* **directs** *soit* **indirects**.
ni X ni Y	Il n'est donc *ni* **explicite** *ni* **implicite**.

TAB. 1 – Patrons retenus pour la projection sur corpus.

2.2 Plan paradigmatique : paires d'adjectifs voisins sur le plan distributionnel

Nous utilisons une base de voisins distributionnels générée par le programme développé par (Bourigault, 2002), en aval de l'analyseur Syntex (Bourigault, 2007), à partir du corpus Wikipédia. L'analyse distributionnelle automatique consiste à rapprocher les mots qui partagent les mêmes contextes syntaxiques, conformément à l'hypothèse selon laquelle la proximité distributionnelle est un bon indice de proximité sémantique (voir en particulier (Baroni & Lenci, 2009) pour une présentation récente des principes de l'analyse distributionnelle automatique). Le processus d'extraction des voisins distributionnels prend en entrée les triplets <gouverneur, relation, dépendant> qui sont générés par Syntex (par exemple <voiture, ADJ, rapide>). Ces triplets fournissent les données à partir desquelles sont rapprochés les couples de mots, en utilisant la mesure de similarité de Lin, qui permet d'ordonner les 4 millions de couples générés. Il a été montré, notamment par (Geffet & Dagan, 2005), que les paires de mots constituées à l'aide de cette méthode entretiennent des relations sémantiques de nature très diverse. Les relations lexicales (synonymie, hyperonymie, antonymie) se mêlent à des relations de nature associative au sens large.

À ce stade, nous disposons donc de deux méthodes qui détectent l'une et l'autre des relations antonymiques mais sont toutes les deux, bien qu'à des degrés divers, bruitées. L'évaluation qui suit cherche à mesurer leur complémentarité, et plus particulièrement à déterminer si les couples repérés conjointement par les deux méthodes présentent un degré d'antonymie plus marqué. Sur les 907 couples retenus par la technique

des patrons, 612 sont également des voisins (environ 67%). L'hypothèse est que le fait que deux adjectifs soient proches distributionnellement permettrait de filtrer les cas de contrastes occasionnels produits en discours.

3 Évaluation des résultats

Nous proposons deux méthodes complémentaires pour juger du caractère antonymique des paires d'adjectifs extraites : la première passe par l'utilisation d'une ressource de référence, la deuxième consiste à solliciter des locuteurs en leur demandant d'identifier la relation sémantique entre les paires d'adjectifs.

3.1 Comparaison à une ressource de référence

Nous utilisons le dictionnaire d'antonymes et de synonymes du CRISCO, Dicosyn, disponible en ligne[3]. Nous montrons dans le tableau 2 la répartition des 907 paires d'adjectifs en trois catégories selon qu'ils sont recensés comme antonymes, comme synonymes, ou absents du dictionnaire. Le taux d'antonymes détecté est plus élevé [4] lorsque les deux adjectifs sont détectés conjointement par les deux techniques. Néanmoins, le calcul du chi^2 montre que cette différence n'est pas significative[5]. En d'autres termes, la propension des adjectifs à la substituabilité n'est pas une propriété décisive pour catégoriser les paires comme antonymes.

	Antonyme	Synonyme	Absent	
Voisins	32%	4%	64%	100%
Non-voisins	27%	6%	67%	100%

TAB. 2 – Proportion des voisins et non voisins présents ou absents du dictionnaire.

3.2 Questionnaires

Nous avons extrait aléatoirement trois jeux de 100 paires dans notre ensemble de départ. La seule contrainte a porté sur le fait que chacun de ces jeux devait être composé pour moitié de couples apparaissant parmi les voisins. Afin de fournir des éléments contextuels aux sujets (pour la plupart étudiants en linguistique), chaque paire était accompagnée d'une série de noms modifiés par les adjectifs en question sélectionnés aléatoirement dans le corpus (ex. : *formation*, *études*, *licence* pour la paire *général/professionnel*). Chaque jeu de données a été soumis à deux sujets différents qui avaient pour consigne de classer chaque paire selon les relations sémantiques proposées, à savoir : *opposition forte*, *opposition faible*, *synonymie*, *autre relation*, *pas de relation*, *ne sais pas*. Le fait de diviser la relation d'opposition en deux (opposition forte et faible) vise à tenir compte de la distinction faite en linguistique cognitive entre des antonymes dits *canoniques* ou *directs* – associations binaires et conventionnelles (*gai/triste*) – et des paires de mots ayant des sens opposés, mais ne présentant pas ce caractère conventionnel (*gai/déprimé*) (Murphy, 2006).

Le recours à deux sujets par jeu de données nous a permis d'évaluer le taux d'accord. Sur les trois jeux, le kappa moyen pour la relation d'opposition forte est de 0.64 alors qu'il est de 0.27 pour l'opposition

[3]http ://www.crisco.unicaen.fr/cgi-bin/cherches.cgi

[4]Nos résultats sont nettement meilleurs que ceux de (Lobanova *et al.*, 2010), qui ne retrouvent dans leurs résultats que 1% à 3% d'antonymes recensés dans les ressources lexicales qu'ils utilisent pour le néerlandais. Mais cela s'explique avant tout par le fait que nous disposons, avec Dicosyn, d'une ressource bien plus complète.

[5]chi^2 = 2.24 ; ddl = 1 ; $p < 0.05$

faible. Par conséquent, nous ne retenons que les résultats obtenus pour le jugement d'opposition forte. Le tableau 3 rapporte, pour chaque jeu, la proportion des paires sur lesquelles les deux sujets ont tous les deux indiqué une opposition forte. Les résultats montrent qu'il n'y a pas de différence significative de jugement chez les locuteurs, ce qui corrobore le constat que nous avons fait lors de l'utilisation du dictionnaire d'antonymes : le fait que deux adjectifs trouvés dans des patrons antonymiques soient des voisins distributionnels ne semble pas influer sur leurs chances de porter une relation d'antonymie.

Jeu 1		Jeu 2		Jeu 3	
Patrons seuls	Patrons + voisins	Patrons seuls	Patrons + voisins	Patrons seuls	Patrons + voisins
12 %	11 %	16 %	16 %	14 %	16 %

TAB. 3 – Résultats obtenus pour la catégorie *opposition forte* après dépouillement des questionnaires.

3.3 Analyse des résultats

Les résultats obtenus semblent contre-intuitifs : ils indiquent que le critère de substituabilité ne renforce pas, comme on aurait pu s'y attendre, la stabilité du contraste observé par le biais des patrons lexico-syntaxiques. Nous pouvons à ce stade proposer plusieurs éléments d'explication à ce constat.

Tout d'abord, une partie des paires d'adjectifs ne sont pas détectés par l'analyse distributionnelle parce qu'il s'agit de mots rares dans le corpus, qui ne passent pas le seuil de similarité[6] que nous avons imposé aux paires de voisins extraites. C'est ainsi le cas de paires comme *dioïque/monoïque* (resp. 50 et 62 occurrences dans le corpus). Cette remarque vaut également pour les couples dont l'un des membres a une fréquence trop faible pour que le score de la paire ne dépasse ce seuil. Ce constat signale une limite évidente de l'analyse distributionnelle : l'expérience ne peut être concluante que pour les mots ayant une fréquence importante dans le corpus.

Deux autres constats intéressent par contre directement l'étude de l'antonymie. Deux antonymes peuvent s'opposer lorsqu'ils qualifient un ensemble très limité de noms. C'est notamment le cas de *annuel* et *vivace*, qui ne s'opposent que lorsqu'ils portent sur les noms *plante* ou *espèce*, ou de *ras* et *long*, qui modifient *poil*. Ils partagent dès lors très peu de contextes et ne sont pas considérés comme des voisins distributionnels. Un dernier cas de figure nous place au cœur de l'hypothèse de Murphy, Jones et leurs collègues, à savoir le comportement syntagmatique de la relation d'antonymie : en effet, le fait que certaines paires d'antonymes ne soient pas repérées par l'analyse distributionnelle permet de dégager un sous-ensemble de couples qui privilégient la relation de cooccurrence à la relation de substituabilité. On a alors affaire à des couples qui fonctionnent sur un mode quasi locutionnel. C'est en particulier le cas de paires comme *coupable/innocent*, *pur/impur*, *officiel/officieux*. Si la combinaison des deux méthodes ne permet pas de détecter plus efficacement les antonymes, cette méthode fournit des éléments d'observation qui permettent d'affiner la description de la relation d'antonymie, en distinguant des paires pour lesquelles le double plan paradigmatique et syntagmatique fonctionne à plein, et d'autres pour lesquelles c'est le plan syntagmatique qui prime.

4 Conclusion

Notre objectif était d'apporter de nouveaux éléments d'expérimentation issus de techniques de TAL pour tester l'hypothèse d'un double fonctionnement paradigmatique et syntagmatique de l'antonymie, formu-

[6] Score de Lin < 0.1

lée par une série de travaux récents en linguistique cognitive et linguistique de corpus. Les résultats que nous avons obtenus semblent en effet corroborer cette hypothèse, en montrant que des paires de mots trouvées dans des contextes contrastifs n'ont pas plus de chance d'être perçues comme antonymes lorsqu'elles appartiennent à une même classe distributionnelle. Cela va à l'encontre de l'intuition qui consisterait à penser que le test de substituabilité fournirait un filtre efficace pour séparer les vrais antonymes des paires de mots qui entretiennent une relation d'opposition très éphémère suggérée par des contextes discursifs particuliers. Cette méthode met ainsi au jour des paires d'antonymes qui ne sont pas voisins distributionnels, ce qui permet de dégager des cas prototypiques d'association purement syntagmatique, qui tranchent avec l'hypothèse classique d'un fonctionnement paradigmatique. Ces premiers résultats nous encouragent à poursuivre ce type d'analyse pour essayer de dégager une typologie au sein de la relation d'antonymie – ou peut-être un continuum entre fonctionnement paradigmatique et syntagmatique. Une autre perspective consisterait à vérifier l'hypothèse d'un comportement particulier de l'antonymie au sein des relations lexicales, et à s'intéresser cette fois à l'articulation entre les dimension syntagmatique et paradigmatique pour les relations de synonymie et d'hyperonymie.

Références

BARONI M. & LENCI A. (2009). One distributional memory, many semantic spaces. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, p. 1–8, Athènes.

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9e conférence sur le Traitement Automatique de la Langue Naturelle*, p. 75–84, Nancy.

BOURIGAULT D. (2007). *Un analyseur syntaxique opérationnel :* SYNTEX. Mémoire d'habilitation à diriger des recherches. Université Toulouse II – Le Mirail.

CHARLES W. & MILLER G. (1989). Context of antonymous adjectives. *Applied psycholinguistics*, **10**.

FELLBAUM C. (1995). Co-occurrence and antonymy. *International journal of lexicography*, **8**.

GEFFET M. & DAGAN I. (2005). The distributional inclusion hypotheses and lexical entailment. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, p. 107–114.

HOEY M. (1991). *Patterns of lexis in text*. Oxford University Press (Oxford).

JONES S. (2002). *Antonymy : a corpus-based perspective*. Routledge.

JONES S. & MURPHY L. (2005). Using corpora to investigate antonym acquisition. *International Journal of Corpus Linguistics*, **10**(3), 401–422.

JUSTESON J. & KATZ S. (1991). Co-occurrence of antonymous adjectives and their contexts. *Computational linguistics*, **17**.

LOBANOVA A., VAN DER KLEIJ T. & SPENADER J. (2010). Defining antonymy : a corpus-based study of opposites by lexico-syntactic patterns. *International Journal of Lexicography*, **23**(1), 19–53.

MURPHY L. (2006). Antonyms as lexical constructions : or, why paradigmatic construction is not an oxymoron. *Constructions all over : case studies and theoretical implications.* Special volume of *Constructions*, **SV1**(8).

MURPHY M. L., PARADIS C., WILLNERS C. & JONES S. (2009). Discourse functions of antonymy : a cross-linguistic investigation of Swedish and English. *Journal of Pragmatics*, **41**(11), 2159–2184.