Recherche et visualisation de mots sémantiquement liés

Alexander Panchenko[1, 2] Hubert Naets[1] Laetitia Brouwers[1]
Pavel Romanov[2] Cédrick Fairon[1]
(1) CENTAL, Université catholique de Louvain, Belgique
{prénom.nom}@uclouvain.be
(2) Bauman Moscow State Technical University, Russie
aromanov@it-claim.ru

RÉSUMÉ

Nous présentons *PatternSim*, une nouvelle mesure de similarité sémantique qui repose d'une part sur des patrons lexico-syntaxiques appliqués à de très vastes corpus et d'autre part sur une formule de réordonnancement des candidats extraits. Le système, initialement développé pour l'anglais, a été adapté au français. Nous rendons compte de cette adaptation, nous en proposons une évaluation et décrivons l'usage de ce nouveau modèle dans la plateforme de consultation en ligne *Serelex*.

ABSTRACT

Search and Visualization of Semantically Related Words

We present *PatternSim*, a new semantic similarity measure that relies on morpho-syntactic patterns applied to very large corpora and on a re-ranking formula that reorder extracted candidates. The system, originally developed for English, was adapted to French. We explain this adaptation, propose a first evaluation of it and we describe how this new model was used to build the *Serelex* online search platform.

MOTS-CLÉS : Mesure de similarité sémantique, relations sémantiques.

KEYWORDS: Semantic similarity measure, semantic relations.

1 Introduction

Les mesures de similarité sémantique permettent d'identifier des mots entretenant différents types de relations sémantiques entre eux (synonymes, hyper/hyponymes, méronymes, etc.) et d'en calculer le degré de similarité. Elles servent à automatiser la construction de ressources sémantiques utiles pour les applications de TAL telles que l'expansion de requêtes, la classification de documents, la désambiguïsation sémantique, etc.

Trois approches computationnelles principales coexistent :

Les mesures basées sur WordNet. Elles obtiennent d'excellents résultats, mais sont limitées par la couverture lexicale de WordNet — voir Wu et Palmer (1994), Leacock et Chodorow (1998) ou encore Resnik (1995).

Les méthodes basées sur dictionnaires (de type explicatif). Elles rencontrent les mêmes difficultés dans la mesure où elles dépendent de ressources préexistantes réalisées manuel-

lement — voir *ExtendedLesk* (Banerjee et Pedersen, 2003), *GlossVectors* (Patwardhan et Pedersen, 2006), *WiktionaryOverlap* (Zesch *et al.*, 2008) ou *Q-Ech* (Fairon et Ho, 2004).

Les approches basées sur corpus. Elles permettent d'obtenir une couverture plus large, car elles calculent les scores de similarité sur des corpus qui peuvent être facilement étendus. Malheureusement, ces dernières offrent généralement une précision plus faible, car elles reposent souvent sur des modèles relativement simples (du type *vector space models*) [1] — voir *ContextWindow* (Van de Cruys, 2010), *SyntacticContext* (Lin, 1998) ou *LSA* (Landauer *et al.*, 1998).

À côté de ces approches computationnelles, le recours au *crowdsourcing* et la mise en place de « jeux sérieux » ont également démontré leur intérêt pour le développement de ressources pour le TAL (Chamberlain *et al.*, 2013). En français, l'expérience la plus importante dans le domaine de la sémantique lexicale est celle de *JeuxDeMots* [2].

Dans cet article, nous décrirons *PatternSim* [3] , une système d'extraction de relations basé sur l'utilisation de corpus et de patrons lexico-syntaxiques. Bien que des techniques existent pour calculer des relations sémantiques à partir de patrons automatiquement appris sur corpus (Bollegala *et al.*, 2007), nous avons fait le choix d'utiliser une bibliothèque de patrons explicitement définis, nous rapprochant ainsi de la méthode classique de Hearst (1992) que nous étendons. Cette approche nous permet de contrôler les motifs extraits et d'éviter ainsi une partie du bruit inhérent à une méthode par apprentissage automatique.

L'originalité de notre approche réside dans le fait que les patrons lexico-sémantiques sont utilisés pour mesurer les similarités sémantiques et non simplement pour extraire les relations et que ces relations sont réordonnées à l'aide d'une heuristique permettant de les classer par ordre de pertinence. En outre, les données extraites et les logiciels réalisés sont disponibles sous license open source.

Initialement développé pour l'anglais, *PatternSim* a été adapté au français. Nous rendrons compte de ce travail d'adaptation et proposerons deux évaluations. Nous présenterons ensuite *Serelex* [4], un outil de consultation en ligne qui permet de naviguer dans le graphe des relations sémantiques calculées par *PatternSim* (Figure 1) et d'expérimenter les différentes mesures sémantiques implémentées.

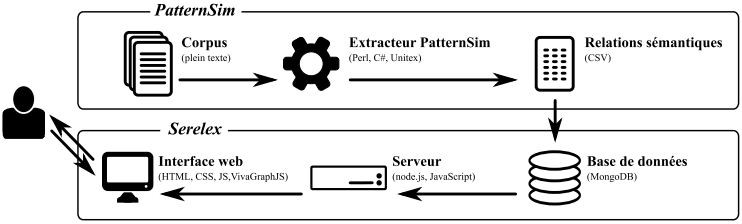

FIGURE 1 – Architecture de *PatternSim* et de *Serelex*

1. Pour un état de l'art plus complet, voir Panchenko (2013).
2. `http://www.jeuxdemots.org`
3. Le code source de *PatternSim* peut être téléchargé sur `https://github.com/cental/patternsim` ; les relations extraites automatiquement sont accessibles sur `http://patternsim.cental.be`.
4. `http://serelex.cental.be`

corpus	*nb de documents*	*nb de tokens*	*nb de lemmes*	*taille*	*concordances extraites*
Anglais					
Wikipedia	2 694 815	$2,026 \cdot 10^9$	3 368 147	5,88 Go	1 196 468
ukWaC	2 694 643	$0,889 \cdot 10^9$	5 469 313	11,76 Go	2 227 025
Wikipedia + ukWaC	5 387 431	$2,915 \cdot 10^9$	7 585 989	17,64 Go	3 423 493
Français					
frWaC	2 268 304	$1,597 \cdot 10^9$	7 047 431	8,00 Go	936 035
Wikipedia long abstracts	734 848	$0,053 \cdot 10^9$	966 789	283 Mo	22 363
frWaC + Wikipedia	3 003 152	$1,65 \cdot 10^9$	7 523 201	8,28 Go	958 398

TABLE 1 – Corpus utilisés dans *PatternSim* et concordances extraites.

2 *PatternSim* : système d'extraction de relations sémantiques

L'approche que nous allons décrire associe un système d'extraction de relations sémantiques (*PatternSim*) opérant sur de vastes corpus et une formule de réordonnancement (*Efreq-Rnum-Cfrq-Pnum*)(Panchenko *et al.*, 2012) qui classe les candidats par ordre de pertinence estimée.

2.1 Corpus & patrons d'extraction

L'identification de relations sémantiques dans un corpus repose sur l'exploitation de patrons d'extraction lexico-syntaxiques construits à la main dans le but d'identifier des cooccurrences significatives de mots telles que, pour l'anglais :

- `such NP as NP, NP[,] and/or NP;`
- `NP, including NP, NP [,] and/or NP;`
- `NP, i. e.[,] NP.`

Dans les contextes décrits par ces structures, les relations entre termes (NP) sont clairement établies, comme dans l'exemple suivant qui atteste du lien entre *foods* et *sandwiches* ou *burgers* :

- `such {non-alcoholic [sodas]} as {[root beer]} and {[cream soda]}[PATTERN=1]`
- `{traditional[foods]}, such as {[sandwiches]},{[burgers]}, and {[fries]}[PATTERN=2]`

Ces patrons, ainsi que de nombreuses variantes qu'il n'est pas possible de lister ici (insertions, permutations, variantes lexicales, etc.), sont décrits sous forme de graphes Unitex [5]. Ils sont utilisés pour la recherche et l'étiquetage de ces structures dans les corpus (comme on le voit dans l'exemple qui précède, les noms qui sont liés sémantiquement sont encadrés de crochets durant la phase d'étiquetage).

Pour nos expérimentations sur l'anglais,nous avons utilisé deux grands corpus représentant un volume total de 17,64 Go de données textuelles : Wikipedia et ukWaC [6] (Table 1).

Dans l'état actuel de la méthodologie, les relations extraites sont partiellement typées : on distingue les relations synonymiques des relations hiérarchiques d'hyperonymie et hyponymie en fonction des graphes qui ont permis d'extraire ces relations. Ce typage léger, n'a cependant pas été évalué et n'est pas utilisé à ce stade dans la mesure de similarité qui se veut globale, c'est-à-dire, toutes catégories confondues.

5. `http://www-igm.univ-mlv.fr/~unitex/`
6. `http://wacky.sslmit.unibo.it`

2.2 Mesure de similarité

(1) Dans un premier temps, les patrons d'extraction lexico-syntaxique sont appliqués sur le corpus. Cette opération permet d'extraire des concordances dans lesquelles les occurrences apparaissent étiquetées. En fonction de la complexité des grammaires d'extraction, cette étape peut être plus ou moins longue (de quelques secondes à plusieurs minutes par Mo). (2) Dans un second temps, les noms entre crochets (`[sodas]` dans `{non-alcoholic [sodas]}`, par exemple) sont lemmatisés à l'aide du dictionnaire DELA[7] ; les entités extraites sont combinées deux par deux. (3) Une matrice de similarité est remplie avec la fréquence des paires similaires. À ce stade, le score de similarité S_{ij} est égal au nombre de fois que les paires entre crochets ou entre accolades apparaissent dans le même contexte de concordance. (4) Pour finir, les paires de mots sont réordonnées à l'aide d'une formule optimisée à cet effet.

Plusieurs formules reposant sur différents paramètres ont été testées (Panchenko *et al.*, 2012). C'est la mesure qui combine l'ensemble de ces paramètres (*Efreq-Rnum-Cfreq-Pnum*) qui s'est révélée être la meilleure : $s_{ij} = \sqrt{p_{ij}} \cdot \frac{2 \cdot \mu_b}{b_{i*} + b_{*j}} \cdot \frac{P(c_i, c_j)}{P(c_i)P(c_j)}$.

Celle-ci prend en compte :
– la fréquence absolue des couples c_i, $c_j \in C$ qui apparaissent dans les concordances K (*Efreq*) ;
– le nombre de relations de c_i et c_j : les termes qui sont fortement liés à un grand nombre de mots sont pénalisés (*Rnum*) ;
– la fréquence de c_i et de c_j dans le corpus : les termes génériques (tels que « chose ») sont pénalisés (*Cfreq*) ;
– le nombre de patrons lexico-syntaxiques qui ont permis d'extraire la relation $c_i c_j$: les paires extraites par plusieurs patrons seront jugées plus robustes (*Pnum*).

2.3 Adaptation au français

Initialement développé pour l'anglais, le système a très récemment été adapté au français. Cette opération a nécessité l'utilisation d'un nouveau corpus ainsi que la traduction - et l'adaptation - des grammaires d'extraction. Ce travail de traduction a pris environ 25 heures. Le corpus français est composé de *frWaC*, un vaste corpus de 1,6 milliard de mots, collecté sur le Web dans le cadre du projet *WaCky*[6], et les résumés longs en français des pages de Wikipedia (Table 1).

Voici deux exemples de concordances extraites à l'aide des patrons d'extraction en français :

- `Cervera collecte aussi de nombreux{[insectes]=HYPER}, particulièrement des{[névroptères]=HYPO}.[PATTERN=4]`
- `L'{[acide] nicotinique=SYNO}, aussi connu sous le nom de{[ niacine]=SYNO},[PATTERN=11] est converti en nicotinamide in vivo.`

2.4 Évaluation

L'évaluation d'un système d'extraction de termes liés sémantiquement est une tâche peu aisée, en raison notamment du caractère relativement flou de la notion de « similarité sémantique » qui recouvre tous les types de relations sémantiques (synonymes, antonymes, hyperonymes, etc.). Il en résulte que la liste des termes « similaires » n'est pas une liste fermée, déterminée que l'on

7. `http://infolingu.univ-mlv.fr/`

pourrait consulter pour vérifier les résultats retournés par la mesure. C'est la raison pour laquelle nous avons choisi d'évaluer le système par rapport à plusieurs tâches.

2.4.1 Évaluation pour l'anglais

Une évaluation détaillée du système anglais a été présentée dans Panchenko *et al.* (2012). Celle-ci a montré (1) que la formule *Efreq-Rnum-Cfreq-Pnum* permettait de construire la variante la plus efficace du mécanisme de réordonnancement des relations sémantiques, (2) que la précision moyenne sur l'anglais (c'est-à-dire le nombre de relations jugées correctes) varie entre 0.736 (quand on ne considère que la meilleure relation) et 0.599 (quand on considère les 20 meilleures relations) et (3) que ce système permettait souvent d'égaler ou de dépasser des méthodes reposant sur des ressources linguistiques beaucoup plus complexes.

2.4.2 Évaluation pour le français

Deux expériences ont été réalisées : dans la première, nous avons confié à des juges humains une tâche d'évaluation et dans la seconde, nous avons comparé nos résultats aux relations sémantiques disponibles dans *JeuxDeMots*.

Cinquante mots ont été sélectionnés aléatoirement depuis cinq rubriques du journal *Le Monde* du 28 mars 2013. Les 30 meilleures relations sémantiques pour chacun de ces mots ont été sélectionnées en utilisant deux mesures : *Efreq* (c'est-à-dire les relations triées par simple fréquence) et *Efreq-Rnum-Cfreq-Pnum* (c'est-à-dire les relations réordonnées à l'aide de la mesure du même nom). En tout, ce sont deux ensembles de 1348 paires liées à 47 mots [8], que nous avons extraits et que nous avons demandé d'évaluer à respectivement quatre et trois annotateurs humains. La tâche de ceux-ci consistait à indiquer si les mots de chaque paire étaient ou non sémantiquement liés.

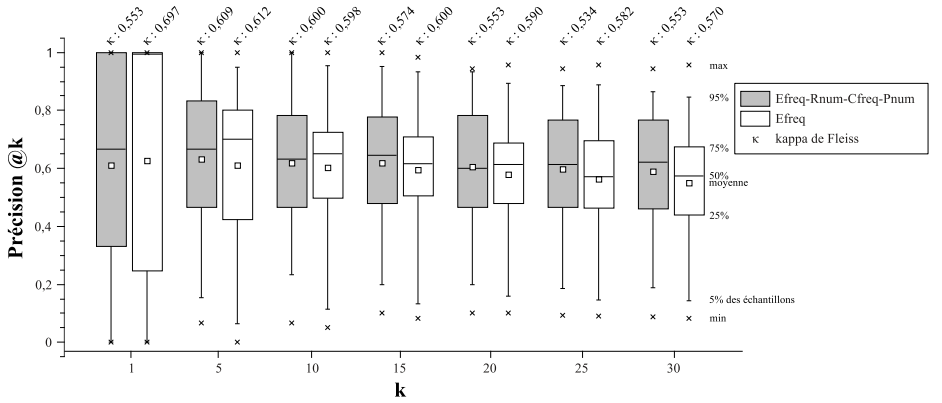

FIGURE 2 – Extraction des relations sémantiques : précision pour les *k* premières relations.

8. 1348 paires et non 1500, car trois mots très techniques ne figurent pas dans la base de données et car certains des 47 mots restants n'ont qu'un petit nombre de relations sémantiques (« kératine », par exemple).

Nous avons calculé pour chacun des 47 mots la précision moyenne pour ses k premières relations. La figure 2 montre le résultat de l'évaluation. Pour la mesure *Efreq*, la précision moyenne, indiquée par le petit carré blanc, varie entre 0,628 (première relation de chacun des 57 mots) et 0,550 (30 premières relations), pour un accord inter-annotateurs (*kappa* de Fleiss) allant de « substantiel » (0,6-0,8) à « modéré » (0,4-0,6). Pour la mesure *Efreq-Rnum-Cfreq-Pnum*, la précision moyenne varie de 0,633 (5 premières relations) à 0,592 (30 premières relations), pour un kappa allant de 0,609 à 0,524. Un test t pour échantillons appariés révèle que la mesure *Efreq-Rnum-Cfreq-Pnum* améliore de façon significative les résultats pour $k = 25$ (t : 2,131 ; P : 0,0192) et pour $k = 30$ (t : 2,9 ; P : 0,0028) ; l'amélioration n'est pas significative pour les autres valeurs de k. Parmi les 47 mots issus du Monde, les mots les plus concrets (« baleine », « laboratoire », « kératine ») ont le plus de relations estimées correctes, tandis que les mots les plus abstraits (« ampleur », « conclusion ») n'ont que très peu de relations jugées exactes.

Dans un second temps, nous avons extrait de la dernière version des données lexicales de *JeuxDeMots* [9] 1 318 479 paires de mots liés sémantiquement. Nous avons comparé ces paires aux 5 849 497 paires extraites à l'aide de *PatternSim*. 86 283 paires sont communes aux deux ensembles (ce qui représente 6,54% des paires de *JeuxDeMots* et 12,04% si on ne conserve que les 54% de paires qui possèdent une entrée commune avec *PatternSim*) ; 18 099 paires communes figurent parmi les 20 premières relations réordonnées à l'aide de la formule *Efreq-Rnum-Cfreq-Pnum*. Dans la tâche précédente, si *JeuxDeMots* avait été un juge, il aurait validé 138 relations, soit 10% de celles-ci, pour 38 des 47 mots extraits du *Monde*.

3 *Serelex* : consultation en ligne des relations sémantiques

Serelex est un moteur de recherche lexical qui, pour un mot donné, propose automatiquement une liste de candidats sémantiquement proches. Pour ce faire, *Serelex* exploite les relations sémantiques extraites à l'aide de *PatternSim* (cf. Figure 1). Ainsi, à la différence des dictionnaires de synonymes et autres thésaurus (Thesaurus.com, VisualSynonyms.com), *Serelex* se base uniquement sur l'information extraite de corpus textuels.

Les requêtes de l'utilisateur sont lemmatisées à l'aide du dictionnaire DELA [10] et une recherche approximative est lancée dans le cas où aucune forme correspondant à la requête de l'utilisateur n'est trouvée. Les résultats de chaque requête sont triés en fonction des scores de similarité enregistrés dans la base de données. Le classement des termes suggérés tient compte notamment de la fréquence des termes dans le corpus et de la fréquence des termes dans les requêtes des utilisateurs.

Le système est accessible au travers d'une interface graphique ou via un web service RESTful. Dans l'interface graphique (cf. Figure 3), un simple champ de texte permet à l'utilisateur d'entrer une requête sous forme de mot-clé (par exemple, un mot simple comme *mathématique*, *Stanford* ou encore une expression polylexicale comme *tour d'ivoire*). La liste de mots proposée affiche les 20 mots les plus liés sémantiquement à la requête. Une représentation sous forme de graphe [11] offre simultanément une représentation visuelle de ces 20 suggestions et de tous les liens sémantiques existant entre ces mots, ce qui permet de grouper visuellement les résultats par sens. Ainsi, dans

9. Version du 24 février 2013 de `http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR/`
10. `http://infolingu.univ-mlv.fr/`
11. Représenté grâce à un algorithme de type Barnes-Hut (Barnes et Hut, 1986) basé sur les forces.

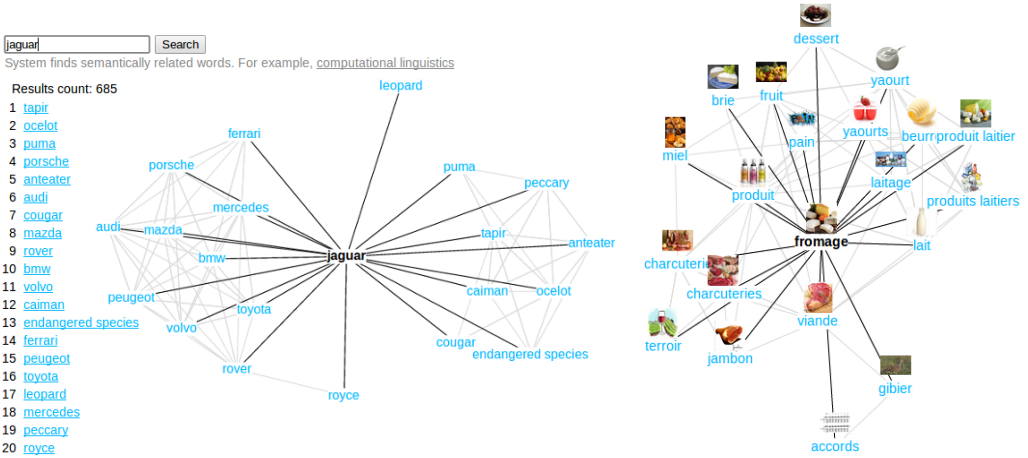

FIGURE 3 – Interface graphique de *Serelex* : résultats de la requête « jaguar » en anglais et de la requête « fromage » en français (affichage avec images).

la Figure 3, on voit clairement apparaître deux clusters correspondant à deux sens de « jaguar » (*voiture* vs *animal*). L'utilisateur peut poursuivre ses recherches en cliquant sur les nœuds du graphe qui permettent de naviguer facilement entre les différents éléments. Il est également possible d'afficher les résultats sous forme d'images, ainsi qu'on peut le voir à propos de l'exemple « fromage ».

4 Conclusion et perspectives

Nous avons présenté dans cet article un système d'extraction de relations sémantiques à base de patterns linguistiques (*PatternSim*) et une interface web de visualisation de ces relations (*Serelex*). Ce système d'extraction, initialement développé pour l'anglais, a pu être adapté au français uniquement en remplaçant les grammaires d'extraction et les corpus utilisés et sans aucun usage d'autres ressources. Les évaluations que nous avons menées en anglais montrent que la mesure de réordonnancement que nous utilisons fournit des résultats comparables à ceux obtenus à l'aide de techniques faisant un usage important de dictionnaires ou de ressources telles que WordNet. Les premiers résultats que nous avons obtenus pour le français se révèlent également positifs et réaffirment l'intérêt de la mesure de réordonnancement. Des modifications mineures dans les grammaires d'extraction et un corpus plus étoffé devraient nous permettre d'atteindre rapidement la même qualité que celle obtenue en anglais.

Remerciements

Cette recherche a été partiellement financée par Wallonie-Bruxelles International (WBI) et par la Région Wallonne (projet ELIS-IT).

Références

BANERJEE, S. et PEDERSEN, T. (2003). Extended gloss overlaps as a measure of semantic relatedness. *In IJCAI*, volume 18, pages 805–810.

BARNES, J. et HUT, P. (1986). A hierarchical 0 (n log iv) force-calculation algorithm. *nature*, 324:4.

BOLLEGALA, D., MATSUO, Y. et ISHIZUKA, M. (2007). Measuring semantic similarity between words using web search engines. *In WWW*, volume 766.

CHAMBERLAIN, J., FORT, K., KRUSCHWITZ, U., LAFOURCADE, M. et PEOSIO, M. (2013). Using games to create language resources : Successes and limitations of the approach. *Theory and Applications of Natural Language Processing*, page 42.

FAIRON, C. et HO, N.-D. (2004). Quantité d'information échangée : une nouvelle mesure de la similarité des mots. *In Le poids des mots. Actes des 7es journées d'analyse statistique des données textuelles*, pages 423–433.

HEARST, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. *In ACL*, pages 539–545.

LANDAUER, T. K., FOLTZ, P. W. et LAHAM, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

LEACOCK, C. et CHODOROW, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet*, pages 265–283.

LIN, D. (1998). Automatic retrieval and clustering of similar words. *In ACL*, pages 768–774.

PANCHENKO, A. (2013). Similarity measures for semantic relation extraction. *Thèse de doctorat en linguistique*.

PANCHENKO, A., MOROZOVA, O. et NAETS, H. (2012). A semantic similarity measure based on lexico-syntactic patterns. *In Proceedings of KONVENS 2012*, pages 174–178.

PATWARDHAN, S. et PEDERSEN, T. (2006). Using WordNet-based context vectors to estimate the semantic relatedness of concepts. *Making Sense of Sense : Bringing Psycholinguistics and Computational Linguistics Together*, pages 1–12.

RESNIK, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In IJCAI*, volume 1, pages 448–453.

Van de CRUYS, T. (2010). *Mining for Meaning : The Extraction of Lexico-Semantic Knowledge from Text*. Thèse de doctorat, University of Groningen.

WU, Z. et PALMER, M. (1994). Verbs semantics and lexical selection. *In ACL'1994*, pages 133–138.

ZESCH, T., MÜLLER, C. et GUREVYCH, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. *In LREC'08*, pages 1646–1652.