Enrichissement de lexiques sémantiques approvisionnés par les foules : le système WISIGOTH appliqué à Wiktionary

Franck Sajous* — **Emmanuel Navarro**** — **Bruno Gaume***

* *CLLE-ERSS*
5 allées Antonio Machado – F-31058 Toulouse Cedex 9
{sajous,gaume}@univ-tlse2.fr

** *IRIT*
118 route de Narbonne – F-31500 Toulouse
navarro@irit.fr

RÉSUMÉ. *Bien que de nombreuses applications de TAL reposent sur des ressources lexicales sémantiques, celles-ci sont rarement simultanément de qualité satisfaisante et librement disponibles. Partant de la confrontation entre méthodes traditionnelles et tendances émergentes de construction et d'évaluation de ressources lexicales, nous présentons dans cet article une nouvelle méthode fondée sur Wiktionary, un dictionnaire multilingue libre, disponible en ligne et construit collaborativement, puis nous proposons un enrichissement semi-automatique de son réseau de synonymie utilisant des données endogènes et exogènes, recourant à une validation « par les foules ». Nous décrivons enfin une implémentation de ce système baptisée WISIGOTH.*

ABSTRACT. *Semantic lexical resources are a mainstay of various NLP applications. However, comprehensive and reliable resources rarely exist or are often not freely available. We discuss in this paper the context of lexical resources building and the problems of evaluation. We present Wiktionary, a freely available and collaboratively built multilingual dictionary and we propose a semi-automatic approach based on random walks for enriching its synonymy network, which uses endogenous and exogenous data. We then propose a validation "by crowds". Finally, we present an implementation of this system called WISIGOTH.*

MOTS-CLÉS : *réseaux de synonymie, similarité sémantique, ressources collaboratives, Wiktionary, enrichissement semi-automatique, marches aléatoires, réseaux petits mondes.*

KEYWORDS: *synonymy networks, semantic relatedness, collaboratively constructed resources, Wiktionary, semi-automatic enrichment, random walks, hierarchical small worlds.*

1. Introduction

Souligner l'importance des ressources lexicales sémantiques dans les applications de TAL tout en déplorant leur absence ou leur qualité insatisfaisante est désormais un lieu commun. Alors que le traitement de l'anglais est doté de WordNet (ci-après noté WN), de l'Université de Princeton (Fellbaum, 1998), ressource éprouvée depuis de nombreuses années, plusieurs langues telles que le français ne bénéficient encore d'aucune ressource de qualité satisfaisante. *« We desperately need linguistic resources ! »* écrit Sekine (2010), soulignant qu'il n'est pas réaliste de penser qu'une seule institution pourra développer des ressources à large échelle, qu'une collaboration est nécessaire et que partager les ressources est crucial. La première difficulté s'opposant à la construction de telles ressources découle des modalités de développement et du compromis coût/qualité qui en résulte. Recourir à des experts pour construire manuellement des ressources induit souvent un coût prohibitif. On ne peut par ailleurs préjuger de la qualité des ressources construites automatiquement (donc bruitées), qui devraient être validées par des experts, ce qui ramène au problème initial. Enfin, dans le cas où le recours à la validation par des experts est envisageable, la mise en place d'une mesure d'accord entre ces experts est problématique.

Après avoir décrit, en section 2, les tentatives antérieures de construction de ressources et dressé un inventaire des différentes méthodes d'extraction de relations lexico-sémantiques, nous nous intéresserons tout particulièrement à la synonymie et aux difficultés liées à sa modélisation et l'évaluation de celle-ci. Nous présentons, en section 3, de nouvelles tendances fondées sur l'édition collaborative qui constitue une piste intéressante pour la construction de ressources. Dans ce cadre, Wiktionary, un dictionnaire libre disponible en ligne, nous paraît être une clé pour régler simultanément le problème du coût de développement et, dans une certaine mesure, celui de l'évaluation. Nous présentons, en section 4, un processus d'enrichissement semi-automatique visant à densifier les réseaux de synonymie extraits de cette ressource. Nous mesurons l'impact de l'utilisation de différentes sources de données sur ce processus en section 5. Enfin, nous présentons une implémentation de notre système, nommée WISIGOTH, que tout internaute peut utiliser pour contribuer à son enrichissement.

2. Construction de ressources lexicales : contexte

WN est probablement le seul projet de construction d'une ressource lexicale sémantique à connaître un tel succès et à être aussi largement utilisé. D'autres projets tels qu'EuroWordNet (Vossen, 1998) et BalkaNet (Tufis, 2000), motivés par la réussite de WN, se sont inscrits dans la même lignée tout en visant une couverture moins ambitieuse. De plus, les ressources produites se sont figées dès la fin de leur développement initial alors que WN a continué d'évoluer. Jacquin *et al.* (2007) ont pointé les faiblesses de la partie française d'EuroWordNet et ont proposé des méthodes automatiques pour ajouter des relations manquantes. Ces méthodes, comme celles que nous énumérons en section 2.1, bien qu'intéressantes, nécessiteraient une validation

manuelle par des experts (donc coûteuse) pour produire une ressource fiable. Les problèmes de temps et coût de développement, ainsi que de disponibilité des ressources, sont de plus en plus pris en compte : en linguistique de corpus, par exemple, une méthode « AGILE », empruntée à l'informatique, a été proposée par Voormann et Gut (2008) pour permettre simultanément de maximiser la taille d'un corpus et de ses annotations tout en réduisant le temps et le coût de son développement. Brunello (2009) s'est intéressé au problème de disponibilité et propose une méthode pour construire des corpus libres en recourant au Web et aux métadonnées qui identifient des pages Web sous licence libre. Avant de montrer, en section 3, comment l'édition collaborative pourrait être un atout pour surmonter les obstacles récurrents dans le domaine de la construction de ressources lexicales, nous commencerons par décrire l'arrière-plan des méthodes classiques.

2.1. *Méthodes d'extraction de relations*

Proposées initialement par Hearst (1992) pour repérer en corpus les relations d'hyperonymie, les **approches par patrons lexico-syntaxiques** ont été affinées par Pantel et Pennacchiotti (2006) pour réduire l'intervention manuelle nécessaire. De tels patrons ne sont pas toujours aisés à mettre au point et sont par ailleurs spécifiques à la langue étudiée.

Des méthodes d'**enrichissement mutuel de ressources par liens interlangues** consistent à projeter des relations sémantiques en suivant des liens de traduction. Sagot et Fišer (2008) ont construit WOLF, un WordNet *libre* du français, en utilisant plusieurs ressources existantes pour amorcer d'une part une base de concepts (fondés sur la synonymie) français et anglais et pour construire un index interlangue à partir duquel les projections ont été opérées. De même, Soria *et al.* (2009) ont utilisé des liens de traduction pour effectuer une « fertilisation mutuelle automatique » de deux lexiques, italien et chinois, ayant une structure similaire à celle de WN. Ces expériences ont le mérite de prouver la faisabilité de la méthode, mais ne permettent pas de produire des ressources fiables en l'absence de validation manuelle.

De nombreux travaux ont enfin porté sur la mise au point de **calculs de similarité sémantique** : différentes propriétés sont associées à des lexèmes et la propension à partager des propriétés communes est interprétée comme le signe d'une proximité sémantique. La plupart de ces méthodes reposent sur des modèles vectoriels : les lexèmes sont représentés par des vecteurs dont chaque coordonnée représente une propriété du lexème et la similarité sémantique des lexèmes est donnée par un calcul de similarité entre ces vecteurs. Les méthodes diffèrent ensuite par les propriétés qui constituent les vecteurs et les algorithmes de calcul de similarité. Un lexème peut être représenté, par exemple, par un vecteur contenant ses cooccurrents en corpus ou ses contextes syntaxiques. Heylen *et al.* (2008) utilisent la similarité entre vecteurs pour extraire des relations de synonymie et comparent les résultats obtenus en utilisant des sacs de mots et des contextes syntaxiques. Ils étudient également l'impact de certaines propriétés linguistiques (fréquence, spécificité et classes sémantiques) sur la similarité ainsi calculée : les contextes syntaxiques s'avèrent plus efficaces que les sacs de

mots et de meilleurs résultats sont obtenus avec les termes de haute fréquence et ceux appartenant à des classes sémantiques abstraites. L'effet de la spécificité sémantique n'est pas probant. Ils montrent également que les vecteurs rapprochés par leur méthode n'ayant pas trait à la synonymie reflètent souvent d'autres relations sémantiques (cohyponymie et hyperonymie/hyponymie). Des comparaisons de différentes mesures et pondérations de contextes syntaxiques sont données dans (Curran et Moens, 2002). Van der Plas et Bouma (2005), quant à eux, étudient quels contextes syntaxiques particuliers mènent aux meilleurs résultats, observant par exemple qu'une meilleure précision est obtenue avec la relation *objet* qu'avec la relation *modifieur*.
Ces méthodes peuvent également être utilisées avec des propriétés non linguistiques : certaines tirent parti d'architectures spécifiques telles que Wikipédia ou Wiktionary. Ollivier et Senellart (2007) et Weale *et al.* (2009) utilisent la structure hypertextuelle de ces ressources comme objet de leur calcul de similarité. Ces méthodes peuvent donner de bons résultats sur certaines tâches (78 à 93 % sur des tests de type TOEFL) mais ne sont bien sûr pas reproductibles en dehors de ces architectures spécifiques (*e.g.* réseaux classiques de type WordNet en cours de construction).
Enfin, les **marches aléatoires** sont des méthodes efficaces qui permettent de calculer la similarité entre sommets dans un graphe. Les graphes peuvent être construits à partir de différentes sources de données et modéliser, par exemple, des réseaux lexicaux dans lesquels un sommet représente un lexème et un arc, une relation sémantique[1]. Nous présentons, en section 4, une approche par marches aléatoires appliquées à des graphes bipartis contenant des données endogènes (liens de synonymie et de traduction, gloses de Wiktionary) et exogènes (contextes syntaxiques extraits de larges corpus).

2.2. *Sens et synonymie*

2.2.1. *Choix de modélisation*

Si l'hyperonymie est au cœur des ontologies, nombre d'applications de TAL exploitent des ressources reposant sur la relation de synonymie, sur laquelle reposent les *concepts*. De nombreux travaux ont été menés par les philosophes et les linguistes sur les questions du sens et de la synonymie. Du point de vue du TAL, il peut être intéressant, sans toutefois chercher à définir ce qu'*est* la synonymie, de se demander quelle synonymie est pertinente dans une visée applicative : là où le linguiste distingue, par exemple, synonymes *absolus*, *cognitifs* et *quasi-synonymes*[2], une tâche donnée peut nécessiter d'autres partitions. De plus, face à la difficulté de développer des ressources, il pourrait être tentant de chercher à développer une ressource idéalement générique. Kilgarriff (1997) écrit, à propos de la désambiguïsation lexicale (ci-après WSD), qu'il n'y a aucune raison de penser qu'un même ensemble de sens soit approprié pour plusieurs applications de TAL : différents buts et différents corpus conduisent à différentes représentations. Par exemple, Edmonds et Hirst (2002)

1. Les sommets peuvent également correspondre à des vecteurs tels que ceux présentés plus haut, les arcs étant pondérés par les distances calculées entre ces vecteurs.
2. Cruse (1986) emploie le terme de *plesionym*.

traitent du problème du choix lexical en traduction automatique, ce qui nécessite le recours à des nuances de sens subtiles, dépassant les synsets traditionnels. Ils proposent dans cette optique de développer un modèle fondé sur une ontologie à gros grain dans laquelle des clusters de quasi-synonymes représentent les sens de base. À un niveau plus fin, plusieurs types de contrastes répertoriés dans une liste finie (dénotationnel, stylistique, structurel, etc.) permettent de différencier les quasi-synonymes d'un même cluster. Le rôle central accordé à la granularité de représentation par ce modèle est extrêmement intéressant, cependant, construire une ressource de ce type à large couverture représente un travail considérable.
D'autres auteurs s'appuient sur des outils mathématiques pour modéliser la synonymie. Victorri et Fuchs (1996) et Ploux et Victorri (1998) utilisent par exemple les cliques maximales pour mettre en évidence les sens dans les réseaux lexicaux. Habert *et al.* (1996) écrivent : *« We argue that the various cliques in which a word appears represent different axes of similarity and help to identify the different senses of that word »*. Cependant, une large divergence oppose les ressources lexicales (*cf.* section 2.2.2) et la notion de clique maximale est très sensible : l'ajout ou la suppression de quelques liens conduit à des différences significatives de modélisation des sens. Pour pallier ce problème, recourir à des approches robustes est nécessaire, comme nous le proposons, à la suite de Gaume (2004), en section 4.

2.2.2. *Propriétés des réseaux de synonymie*

Afin de mieux comprendre la diversité qui caractérise les ressources lexicales, nous avons construit les graphes modélisant les réseaux de synonymie de sept dictionnaires de référence du français[3]. Nous présentons ci-dessous les propriétés structurelles de ces graphes (*cf.* tableau 1) puis les comparons entre eux.

Propriétés communes : la plupart des réseaux lexicaux sont des réseaux *Petits Mondes Hiérarchiques* (PMH) partageant des propriétés similaires (Watts et Strogatz, 1998 ; Albert et Barabasi, 2002 ; Newman, 2003 ; Gaume *et al.*, 2008 ; Gaume *et al.*, 2010) : une **faible densité** (nombre d'arcs faible), des **chemins courts** (la longueur moyenne L des chemins entre deux sommets est courte), un **taux de clustering** C élevé (ces graphes sont denses localement, bien que globalement peu denses) et **leur distribution des degrés d'incidence suit une loi de puissance** (une description plus complète peut être trouvée dans (Gaume, 2004)). Dans le tableau 1, n et m désignent respectivement le nombre de sommets et d'arcs, le degré moyen des sommets et λ le coefficient de la loi de puissance qui approxime la distribution des degrés d'incidence avec un coefficient de correlation de r^2. n_{lcc} et L_{lcc} sont respectivement le nombre de sommets et la longueur moyenne des chemins calculés sur la plus grande composante connexe. Si n et $\langle k \rangle$ varient selon les dictionnaires, L_{lcc} est faible, C est toujours élevé, et la distribution des degrés reste proche d'une loi de puissance ($r^2 > 0,9$) avec λ compris entre – 3,6 et – 2,2. Ces propriétés montrent que ces réseaux sont des PMH, tout comme ceux extraits de Wiktionary – voir (Navarro *et al.*, 2009).

3. Les relations de synonymie ont été extraites de chaque dictionnaire par l'INALF (aujourd'hui ATILF) puis corrigées au CRISCO.

Dictionnaire	n	m	$\langle k \rangle$	n_{lcc}	C	L_{lcc}	λ	r^2
Bailly	12 738	14 226	2, 38	560	0, 04	11, 11	− 2, 67	0, 94
Benac	21 206	33 005	3, 33	728	0, 02	9, 03	− 2, 68	0, 94
Bertaud-du-Chazaud	40 818	123 576	6, 16	259	0, 11	6, 13	− 2, 28	0, 92
Guizot	3 161	2 200	2, 08	1 018	0, 08	4, 69	− 3, 56	0, 95
Lafaye	3 120	2 502	2, 05	641	0, 01	9, 37	− 2, 58	0, 97
Larousse	25 505	79 612	7, 11	1 533	0, 18	6, 35	− 2, 46	0, 92
Robert	48 898	115 763	5, 44	3 340	0, 11	6, 43	− 2, 43	0, 94

Tableau 1. *Propriétés structurelles des graphes de synonymie*

Divergences : bien que tous ces graphes soient des PMH, le tableau 1 montre que les couvertures lexicales (n) et le nombre de liens de synonymie (m) varient significativement d'un graphe à l'autre. Étant donnés $G_1 = (V_1, E_1)$ et $G_2 = (V_2, E_2)$ deux des sept graphes extraits des dictionnaires, nous calculons le rappel, la précision et la F-mesure de la couverture lexicale de G_1 par rapport à celle de G_2 (colonne de gauche du tableau 2). Notons que $R_\bullet(G_1, G_2) = P_\bullet(G_2, G_1)$ et $F_\bullet(G_1, G_2) = F_\bullet(G_2, G_1)$. Ces mesures renseignent sur les couvertures lexicales relatives de G_1 et G_2 et non sur leur accord en terme de synonymie. Pour mesurer cet accord, nous projetons les réseaux de synonymie sur leurs sommets communs $V_1 \cap V_2$ puis construisons le sous-graphe $G_{1_{\langle V_1 \cap V_2 \rangle}}$ de $G_1 : G_{1_{\langle V_1 \cap V_2 \rangle}} = (V_{1_{\langle V_1 \cap V_2 \rangle}}, E_{1_{\langle V_1 \cap V_2 \rangle}})$ avec $V_{1_{\langle V_1 \cap V_2 \rangle}} = V_1 \cap V_2$ et $E_{1_{\langle V_1 \cap V_2 \rangle}} = E_1 \cap ((V_1 \cap V_2) \times (V_1 \cap V_2))$. Nous construisons $G_{2_{\langle V_1 \cap V_2 \rangle}}$ de la même manière. Nous calculons enfin le rappel, la précision et la F-mesure des arcs de $G_{1_{\langle V_1 \cap V_2 \rangle}}$ par rapport à ceux de $G_{2_{\langle V_1 \cap V_2 \rangle}}$ (colonne de droite du tableau 2).
Le tableau 3 montre la valeur de ces mesures pour chaque paire de graphes. La comparaison des couvertures lexicales est donnée dans les colonnes (•) et l'accord sur les réseaux de synonymie restreints aux sommets communs dans les colonnes (↕). Comprise entre 0,27 et 0,69, et avec une score moyen de 0,46, la F-mesure (en gras) varie significativement selon les paires de dictionnaires considérés.
Ces résultats montrent qu'une divergence notable oppose les réseaux de synonymie de référence construits par les lexicographes. Outre la subjectivité des experts, cela s'explique notamment par des choix éditoriaux (*e.g.* public visé, contrainte de taille en version imprimée, etc.). À la suite de ces observations, nous avons compilé les sept dictionnaires dans une même ressource, que nous avons ensuite divisée selon les catégories syntaxiques[4] pour obtenir trois ressources : DicoSyn.Noms, DicoSyn.Verbes et DicoSyn.Adjs. Cet ensemble servira d'étalon pour évaluer Wiktionary et notre méthode d'enrichissement (*cf.* sections 3.2 et 4).

2.3. *La question récurrente de l'évaluation*

Bien qu'ayant donné lieu à de nombreux travaux, le problème de l'évaluation des ressources construites automatiquement reste ouvert. Une approche traditionnelle est la **comparaison à une ressource étalon**. En supposant l'existence et la disponibilité

4. Ce travail de catégorisation automatique et de validation manuelle a été effectué à CLLE-ERSS avec l'aide de Lydia-Mai Ho-Dac.

Couverture lexicale	**Synonymie**
$R_{\bullet}(G_1, G_2) = \frac{\lvert V_1 \cap V_2 \rvert}{\lvert V_2 \rvert}$	$R_{\updownarrow}(G_1, G_2) = \frac{\lvert E_{1\langle V_1 \cap V_2 \rangle} \cap E_{2\langle V_1 \cap V_2 \rangle} \rvert}{\lvert E_{2\langle V_1 \cap V_2 \rangle} \rvert}$
$P_{\bullet}(G_1, G_2) = \frac{\lvert V_1 \cap V_2 \rvert}{\lvert V_1 \rvert}$	$P_{\updownarrow}(G_1, G_2) = \frac{\lvert E_{1\langle V_1 \cap V_2 \rangle} \cap E_{2\langle V_1 \cap V_2 \rangle} \rvert}{\lvert E_{1\langle V_1 \cap V_2 \rangle} \rvert}$
$F_{\bullet}(G_1, G_2) = 2.\frac{R_{\bullet}(G_1,G_2).P_{\bullet}(G_1,G_2)}{R_{\bullet}(G_1,G_2)+P_{\bullet}(G_1,G_2)}$	$F_{\updownarrow}(G_1, G_2) = 2.\frac{R_{\updownarrow}(G_1,G_2).P_{\updownarrow}(G_1,G_2)}{R_{\updownarrow}(G_1,G_2)+P_{\updownarrow}(G_1,G_2)}$

Tableau 2. *Comparaison de la couverture lexicale et de la synonymie entre étalons*

		Benac		**Bertaud**		**Guizot**		**Lafaye**		**Larousse**		**Robert**	
		(•)	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**	**(•)**	**(↕)**
Bail.	**R**	0,50	0,56	0,29	0,20	0,84	0,60	0,90	0,61	0,40	0,18	0,24	0,20
	P	0,82	0,60	0,93	0,78	0,21	0,49	0,22	0,52	0,81	0,62	0,91	0,71
	F	0,62	**0,58**	0,44	**0,32**	0,34	**0,54**	0,36	**0,56**	0,54	**0,28**	0,37	**0,31**
Ben.	**R**			0,47	0,31	0,85	0,58	0,90	0,68	0,52	0,18	0,30	0,18
	P			0,90	0,76	0,13	0,42	0,13	0,51	0,63	0,60	0,70	0,64
	F			0,62	**0,44**	0,22	**0,49**	0,23	**0,58**	0,57	**0,27**	0,42	**0,28**
Bert.	**R**					0,93	0,78	0,96	0,81	0,76	0,44	0,52	0,54
	P					0,07	0,16	0,07	0,17	0,47	0,38	0,63	0,49
	F					0,13	**0,27**	0,14	**0,29**	0,58	**0,41**	0,57	**0,51**
Guiz.	**R**							0,79	0,68	0,11	0,19	0,06	0,18
	P							0,78	0,69	0,88	0,72	0,91	0,82
	F							0,78	**0,69**	0,19	**0,29**	0,11	**0,29**
Laf.	**R**									0,11	0,18	0,06	0,17
	P									0,93	0,65	0,95	0,77
	F									0,20	**0,28**	0,11	**0,28**
Lar.	**R**											0,44	0,50
	P											0,85	0,54
	F											0,58	**0,52**

Tableau 3. *Accord entre couples de dictionnaires : rappel (R), précision (P) et F-mesure (F) (pour le dictionnaire de la ligne par rapport au dictionnaire de la colonne)*

d'une telle ressource, cette dernière a ses limites et donc l'évaluation qui en découle également. L'étalon choisi peut en effet avoir été conçu dans une optique particulière (comme c'est le cas pour les dictionnaires) et ne refléter qu'un aspect de la réalité. Il n'est donc nécessairement ni générique, ni approprié à toute évaluation. Un étalon doit être lui-même évalué ou caractérisé avant d'être choisi pour une évaluation donnée. Nous avons montré en section 2.2.2 qu'il n'y a pas de consensus entre les ressources de référence. Ainsi, le choix d'un étalon plutôt que d'un autre influence largement une évaluation et, partant, les résultats obtenus ne permettent pas toujours de tirer des conclusions définitives : si, par exemple, un système propose deux lexèmes comme synonymes et qu'ils n'apparaissent pas comme tels dans l'étalon, le système est-il en défaut ou est-ce dû au caractère non exhaustif de l'étalon ? Dans (Navarro *et al.*, 2009), nous expliquons comment nous avons dû adapter notre matériel expérimental pour le comparer aux étalons choisis (symétriser les arcs du graphe lorsque l'étalon était WN et fusionner les sous-sens pour comparer à DicoSyn), et la perte d'information potentielle qui en résulte, pouvant ainsi biaiser les résultats.

Une **évaluation manuelle** peut également consister à extraire un échantillon de relations et à les étiqueter manuellement comme correctes ou incorrectes. Cependant,

un tel jugement est entaché de la subjectivité liée notamment à une représentation du monde différente pour chaque annotateur et mène souvent à un accord interjuges faible. En cas d'accord satisfaisant, celui-ci est régulièrement utilisé comme seul critère de qualité alors que rien ne garantit la pertinence des jugements rendus. Murray et Green (2004) ont analysé les facteurs corrélés à des taux d'accord faibles sur une tâche de désambiguïsation lexicale et ont montré que les accords élevés sont corrélés à des compétences lexicales *comparables* entre les annotateurs et non à des compétences *élevées*. Deux juges « naïfs » sont susceptibles d'obtenir le même taux d'accord que deux experts ; ajouter un expert à un groupe de naïfs tend à diminuer ce taux. Ils concluent que la mesure de l'accord interannotateurs seule ne constitue pas un critère de qualité d'une annotation et qu'elle doit être couplée à une mesure de compétence par rapport à l'annotation envisagée.

Enfin, on peut comparer plusieurs ressources *via* une **évaluation par la tâche** en mesurant les performances d'un système utilisant les différentes ressources pour réaliser une tâche donnée. Des ressources lexicales sémantiques peuvent être évaluées, par exemple, dans des tâches de recherche d'information, de traduction automatique, de WSD, etc. Pour estimer les performances de ce système, le processus d'évaluation doit pouvoir déterminer, pour une entrée donnée, quelle sortie doit être produite. Ce problème nécessite alors la construction d'un étalon et soulève les mêmes problèmes que ceux mentionnés plus haut. Par exemple, Kilgarriff (1998) a montré les difficultés rencontrées dans la préparation d'un étalon pour la campagne SENSEVAL.

Nous avons évoqué dans la section 2.2.1 le rôle central de la granularité dans la modélisation de la synonymie. Cette notion joue également un rôle crucial dans les processus d'évaluation. Dans une tâche de WSD, Palmer *et al.* (2007) ont montré qu'un regroupement des sens d'un dictionnaire électronique permettait de régler des désaccords « à la marge » entre annotateurs. De manière générale, l'accord augmentait de 10 à 20 % lorsqu'il était mesuré sur une annotation du système utilisant une granularité plus grosse. Cependant, ils soulignent qu'atteindre des taux d'accord très élevés avec des mots fortement polysémiques est un objectif non réaliste. Ils notent également qu'accroître cet accord n'est pertinent que si cela n'affecte pas ou peu la qualité résultante des applications de TAL.

3. Édition collaborative « par les foules »

Depuis la création de Wikipédia, l'exactitude des ressources collaboratives construites « par les foules » a été mise en doute. Wikipédia était alors la seule ressource de ce type et la question de son bien-fondé a mené à une polémique : Giles (2005) a prétendu que l'encyclopédie en ligne était aussi précise que l'encyclopédie Britanica, qui a pour sa part réfuté les critères d'évaluation utilisés. Depuis, les wikis se sont multipliés. Plus modérés que Giles, Zesch et Gurevych (2010) ont montré, à travers une tâche de mesure de similarité sémantique, que les ressources fondées sur « la sagesse des foules » ne sont pas meilleures que celles fondées sur celle des linguistes, mais sont sérieusement compétitives. Elles dépassent même les ressources construites par les experts dans certains cas, notamment en terme de couverture.

3.1. *Plusieurs modèles*

Au problème de la qualité du contenu s'ajoute celui de la participation « des foules ». En effet, les approches collaboratives n'impliquent pas la seule participation de collègues et d'étudiants mais doivent mobiliser une plus large population ne partageant pas nécessairement les motivations des chercheurs. Plusieurs tendances récentes refaçonnent le domaine de la création de ressources. Parmi elles, quelques chercheurs ont mis au point des **jeux en ligne** destinés à attirer des internautes « ordinaires » qui jouent pour le plaisir, indépendamment de la finalité première du jeu. C'est le cas de JeuDeMots, développé par Lafourcade (2007), qui a réussi à collecter un nombre important de relations sémantiques (essentiellement non typées, mais également des relations telles qu'hyperonymie et méronymie). Cependant, réaliser un jeu suffisament distrayant est une tâche difficile en soi et il est probable que certaines tentatives échouent si elles ne produisent pas des jeux suffisament attractifs.
Le système « **Mechanical Turk** », créé par Amazon (AMT), consiste à définir des micro-tâches que des travailleurs (*turkers*) accomplissent contre une rétribution minimale. Ces tâches, ou HITs (*human intelligence tasks*), permettent d'effectuer ce dont un système informatique est incapable : étiqueter une image, exprimer sa préférence pour une couleur, etc. Utilisé généralement dans le commerce électronique, ce système a également servi à juger la qualité des articles de Wikipédia en recueillant les appréciation d'un groupe d'utilisateurs. Des chercheurs ont également eu recours à ce modèle pour mettre en place des systèmes d'évaluation. Snow *et al.* (2008), par exemple, ont évalué les performances d'annotateurs naïfs utilisés comme *turkers* dans des tâches telles que désambiguisation lexicale, jugement de similarité sémantique, annotation de sentiments dans les documents textuels, etc. Ils ont obtenu, dans cinq tâches sur sept, le résultat contre-intuitif qu'un système entraîné sur les annotations des *turkers* surpasse celui entraîné par un expert unique. Ils expliquent ce résultat par le fait que le jugement de plusieurs naïfs permet de corriger le biais introduit par l'unicité d'un annotateur, fût-il expert. Le modèle d'Amazon a montré qu'il pouvait être approprié à des tâches d'annotation. Il faut néanmoins accueillir ce constat avec mesure et garder à l'esprit qu'AMT a été conçu pour accomplir des tâches courtes et ne doit être utilisé que pour ces tâches-là. Dans le cas contraire, les *turkers* peu scrupuleux peuvent être tentés d'abuser le système en passant un minimum de temps sur chaque tâche et en donnant des réponses au hasard. Par ailleurs, même avec des « *turkers* honnêtes », vérifier la compétence d'un annotateur pour une tâche donnée peut être nécessaire (*cf.* section 2.3).
La mise en place d'infrastructures sophistiquées et coûteuses dans la perspective d'une édition « par les foules » présente le risque d'aboutir à des coquilles vides attendant indéfiniment d'être remplies. En effet, dans le contexte actuel de compétition pour attirer les internautes, les plates-formes dépourvues de contenu, aussi prometteuses soient-elles, n'attirent personne. Toute solution envisagée qui sous-estime la contrainte d'attractivité est vouée à l'échec. Seules quelques initiatives collaboratives et quelques réseaux sociaux ont réussi à subsister et concentrent la majorité des internautes. Le simple fait de s'adosser à l'une de ces « *success stories* » du Web laisse espérer une multitude de visiteurs. Wikipédia et Wiktionary sont certainement

parmi les meilleurs exemples de réussite et la communauté du TAL peut apporter ses compétences et outils à leurs utilisateurs, bénéficiant en retour d'un nombre important de contributeurs. Des jalons ont été mis en place dans une approche que nous avons nommée « **piggybacking** » (Sajous *et al.*, 2010) et que nous poursuivons ici : nous appuyant sur l'architecture de Wiktionary et son contenu déjà existant, nous proposons aux contributeurs des synonymes calculés automatiquement afin qu'ils les valident. Nous espérons ainsi aider ces contributeurs à accélérer l'enrichissement du dictionnaire tout en profitant de son infrastructure pour mettre au point des méthodes également applicables à d'autres types de ressources.

3.2. *Wiktionary*

Wiktionary, le *« compagnon lexical de Wikipédia »*, est un dictionnaire multilingue libre et accessible en ligne. Comme les autres satellites de la Wikimedia Foundation, c'est un projet collaboratif : n'importe quel internaute peut contribuer et ses modifications sont publiées immédiatement. Chaque article peut mentionner des informations sur l'étymologie, contenir des gloses, exemples, traductions et des relations sémantiques (voir (Navarro *et al.*, 2009 ; Sajous *et al.*, 2010) pour une description plus détaillée). Du point de vue du TAL, Wiktionary peut sembler un terrain de jeu idéal et la couverture lexicale apparente renforce cette impression. Nous présentons ci-après un examen plus approfondi qui nuance des propos souvent (trop ?) enthousiastes.

3.2.1. *Encodage et structuration des données*

Les projets de la Wikimedia Foundation utilisent un système de gestion de contenu appelé MediaWiki. Un langage à balises tel que HTML ayant été jugé trop complexe pour les contributeurs (que l'on souhaite nombreux), le contenu des articles est enregistré dans un langage nommé *code wiki*. Malheureusement, ce langage, n'ayant pas de syntaxe formelle, est sous-spécifié et des déviances par rapport au langage standard – supposé – sont fréquentes. En analysant manuellement les modifications des contributeurs, nous avons remarqué un nombre significatif d'erreurs dans les articles dû à une incompréhension ou à un non-respect de la syntaxe. Nous pressentons également l'effet, peut-être plus dommageable, que nombre d'utilisateurs ne deviendront jamais contributeurs seulement parce qu'ils resteront rebutés par cette syntaxe.

Un article contient potentiellement plusieurs sections de langues, la première étant celle de l'édition de Wiktionary consultée. Une section de langue peut contenir plusieurs sections liées à une catégorie syntaxique donnée. Dans ces sections, on peut trouver des gloses et des exemples, répartis selon plusieurs sens. Ensuite viennent les traductions et les relations sémantiques. Une grande variation existe cependant par rapport à ce cas prototypique. Chaque langue a ses propres conventions et au sein d'une langue donnée, les conventions écrites ne sont pas toujours respectées. La notion de *flexibilité* étant même revendiquée comme une propriété intrinsèque des projets de la Wikimedia Foundation, l'analyse automatique du contenu de Wiktionary n'est pas tâche aisée. Elle l'est encore moins lors de l'analyse des fichiers historiques re-

traçant la chronologie des articles (*i.e.* contenant toutes leurs versions datées), lorsque à la fois syntaxe et conventions évoluent avec le temps. Les implications de choix discutables sur les ressources extraites sont données dans (Navarro *et al.*, 2009). Le débat, par exemple, sur la pertinence de la représentation des sous-sens des mots en TAL n'aura pas lieu ici : leur format d'encodage trop lâche empêche leur utilisation.

3.2.2. *Lexèmes et relations sémantiques : une croissance à deux vitesses*

Afin d'étudier comment évolue une ressource collaborative telle que Wiktionary, nous avons analysé son « *dump* historique[5] ». Ces fichiers contiennent les versions intégrales de tous les articles après chaque édition. Comme on le voit dans la figure 1, l'édition anglaise a connu une croissance constante en terme de lexèmes alors que l'édition française a connu deux sauts. Le premier, début 2006, est dû à un import automatique d'articles du Dictionnaire de l'Académie Française (DAF). Des imports d'articles du Littré ont également été effectués, de manière plus étalée. Le nombre d'imports automatiques pour l'anglais est non significatif. On observe pour le français un second saut en 2008, plus important, qui concerne les noms et les adjectifs. Ce saut est dû à un import automatique de 76 347 gentilés (il n'affecte donc pas les verbes) extraits d'un site spécialisé. Aucun import automatique n'a été réalisé pour ajouter de relations sémantiques. Les contributeurs étant plus enclins à ajouter de nouvelles entrées, la croissance de ces relations est plus lente que celle des lexèmes. Parmi elles figurent essentiellement des synonymes et quelques antonymes. La figure 2 montre l'évolution du nombre de relations sémantiques et de liens de traduction pour l'anglais et le français. Aucun import automatisé concernant les traductions n'est explicitement mentionné dans Wiktionary. Cependant, nous avons remarqué dans l'édition française un ajout massif de traductions effectué début 2006 par un automate, sans que cet import ne soit documenté. Après examen, nous avons trouvé une note succinte dans la page de discussion de l'auteur de cet automate mentionnant l'import depuis un site en ligne dont ni l'origine du contenu ni la licence n'étaient clarifiées. Finalement, malgré une augmentation importante du nombre de relations sémantiques et de traduction, l'écart par rapport à la croissance du nombre de lexèmes continue de se creuser (*cf.* tableau 4). Or ces relations sont essentielles aux applications de TAL.

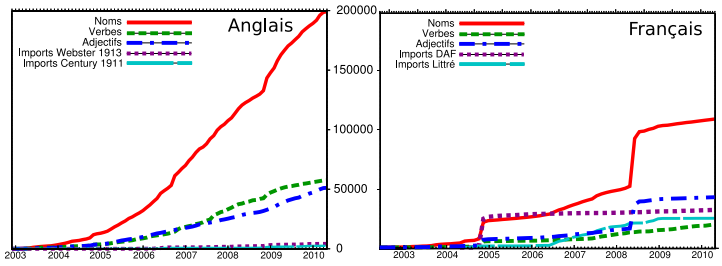


Figure 1. *Évolution du nombre de lexèmes et imports automatisés dans Wiktionary*

5. Ces *dumps* sont disponibles à l'adresse : `http://download.wikipedia.org/`

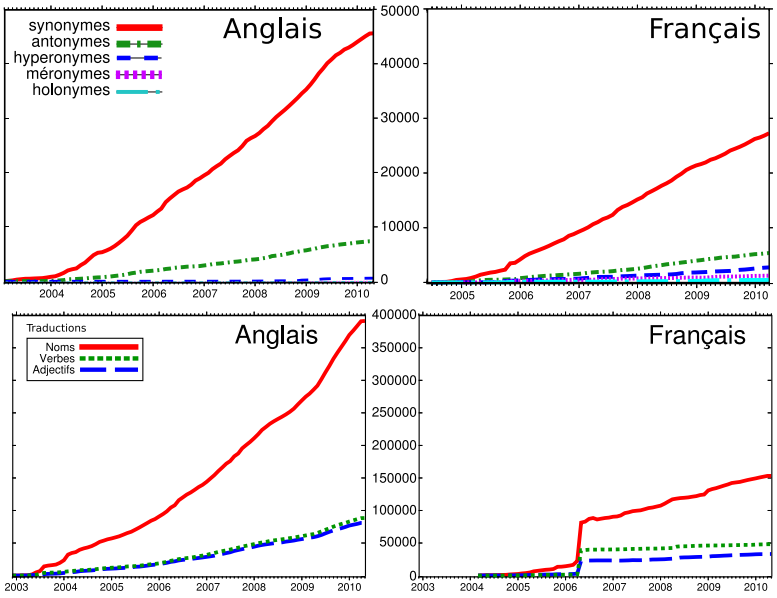


Figure 2. *Wiktionary : évolution des relations sémantiques et des liens de traduction*

		2007			2010		
		Noms	**Verbes**	**Adjectifs**	**Noms**	**Verbes**	**Adjectifs**
FR	Lexèmes	38 973	6 968	11 787	106 068 (× **2,7**)	17 782 (× **2,6**)	41 725 (× **3,5**)
	Synonymes	9 670	1 793	2 522	17 054 (× **1,8**)	3 158 (× **1,8**)	4 111 (× **1,6**)
	Traductions	106 061	43 319	25 066	153 060 (× **1,4**)	49 859 (× **1,2**)	32 949 (× **1,3**)
ANG	Lexèmes	65 078	10 453	17 340	196 790 (× **3,0**)	67 649 (× **6,5**)	48 930 (× **2,8**)
	Synonymes	12 271	3 621	4 483	28 193 (× **2,3**)	8 602 (× **2,4**)	9 574 (× **2,1**)
	Traductions	172 158	37 405	34 338	277 453 (× **1,6**)	70 271 (× **1,9**)	54 789 (× **1,6**)

Tableau 4. *Wiktionary : croissance des éditions française et anglaise de 2007 à 2010*

3.2.3. *Nomenclature et couverture lexicale*

Le Wiktionary anglais fait état de *« 2 038 436 entries with English definitions from over 400 languages »* et la version française de *« 1 821 097 articles [qui] décrivent en français les mots de plus de 700 langues »*. Ces chiffres impressionants sont néanmoins à tempérer : les méta-articles (pages d'aide, de discussion, etc.) sont comptés comme entrées et nombre de mots étrangers (à la langue du Wiktionnaire considéré) sont curieusement inclus dans la nomenclature (*e.g. lexicon* apparaît comme une entrée anglaise dans le Wiktionnaire français). Les formes fléchies font l'objet d'articles quand on pourrait les attendre dans les articles consacrés à leur forme citationnelle. Alors que dans l'édition française, les locutions sont catégorisées comme telles, l'édition anglaise les classe sous la catégorie de leur tête (voire une autre catégorie), ce qui gonfle artificiellement le nombre de lexèmes annoncé (*e.g. « caught on the hop »* apparaît comme un verbe standard).

Malgré des catégorisations surprenantes et des imports automatiques massifs (la moitié des noms sont des gentilés, *cf.* section 3.2.2), le nombre de lexèmes contenus dans les éditions française et anglaise de Wiktionary reste élevé. Il nous a alors paru pertinent d'évaluer sa couverture lexicale effective. Nous avons pour cela comparé sa nomenclature à celle du Trésor de la Langue Française informatisé (TLFi) et examiné le recouvrement entre les noms, verbes et adjectifs (formes canoniques) du dictionnaire collaboratif par rapport à ceux présents dans Morphalou[6]. Nous voyons dans le tableau 5 que le Wiktionnaire comprend trois quarts des noms du TLFi, quasiment tous ses verbes, et deux tiers de ses adjectifs. Nous avons également évalué la proportion couverte par chaque dictionnaire du vocabulaire des trois corpus suivants : 30 millions de mots issus de 515 romans du XX^{e} siècle de la base Frantext, 200 millions de mots issus des articles du quotidien *Le Monde* (période 1991 à 2000) et 260 millions de mots extraits de l'encyclopédie Wikipédia (version 06/2008). Nous avons lemmatisé et catégorisé ces corpus avec TreeTagger[7], puis gardé pour chacun les lemmes catégorisés dont la fréquence est supérieure ou égale à 5. Nous observons que les deux lexiques suivent la même tendance : le vocabulaire de Frantext est mieux couvert que celui du *Monde*, lui-même mieux couvert que celui de Wikipédia. La moindre couverture pour Wikipédia peut s'expliquer notamment par la diversité des domaines couverts et des rédacteurs, ainsi que par la présence des mots étrangers. Le faible pourcentage de noms couverts peut s'expliquer par le fait que les articles de l'encyclopédie comportent de nombreux mots isolés inconnus de TreeTagger qu'il étiquette souvent comme des noms communs. On observe sur les trois corpus que le Wiktionnaire couvre plus de noms et de verbes (2 à 7 %) et que Morphalou couvre mieux les adjectifs (1 à 4 %). Si l'on construit l'union des deux nomenclatures (noté M∪W), on augmente clairement la couverture pour les noms (5 %) et les adjectifs (10 %). Ces observations nous amènent à formuler le constat que, bien que bruité et souffrant d'un manque de formalisme, Wiktionary, par sa couverture et sa disponibilité, mérite d'être envisagé comme une ressource lexicale intéressante pour le TAL. De plus, les résultats présentés dans le tableau 5 confirment les observations de Zesch (2010) : les ressources créées par des experts et celles construites collaborativement par les foules ne se confondent pas mais peuvent contenir des connaissances complémentaires. En effet, s'il est vrai que l'on trouve dans les entrées du Wiktionnaire absentes du TLFi des néologismes liés notamment au domaine d'Internet tels que *googler* et *wikifier*, ou des variations diatopiques régionales (*dracher*) et nationales (*diplomation*, *hommerie*), le Wiktionnaire compte également des termes spécialisés comme *clitique* et l'adjectif *métier* (*e.g.* une application métier) et d'autres, passés dans l'usage courant comme *sinogramme, homophobie, sociétal, fractal, ergonomique, médicaliser, étanchéifier, désactiver, décélérer, paramétrer*…

6. Lexique distribué par l'ATILF issu de la nomenclature du TLFi, disponible à l'adresse : `http://www.cnrtl.fr/lexiques/morphalou/`

7. `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

	Taille de la nomenclature			Couverture des vocabulaires des corpus (%)								
				Frantext			*Le Monde*			Wikipédia		
	Morphalou	Wiktionnaire	Intersection	Morph.	Wikt.	M∪W	Morph.	Wikt.	M∪W	Morph.	Wikt.	M∪W
N.	41 005	134 203	29 604	76,4	80,6	84,4	47,3	54,1	58,1	23,5	26,7	31,6
V.	7 384	18 830	6 964	84,2	86,5	87,1	75,1	80,0	80,8	66,3	71,5	72,2
Adj.	15 208	42 263	10 014	88,9	84,6	94,0	78,9	76,8	88,1	73,9	72,4	84,7

Tableau 5. *Couverture lexicale du Wiktionnaire (2011) et de Morphalou (1.0)*

4. Enrichissement semi-automatique

Pour combler l'écart entre le nombre de lexèmes et celui des relations sémantiques (*cf.* section 3.2.2), nous proposons de développer un système d'enrichissement de réseaux sémantiques, à commencer par les relations de synonymie. Dans une expérience précédente (Navarro *et al.*, 2009), nous avons mis au point un système de densification automatique des réseaux de synonymie, que nous avons appliqué au Wiktionary français et anglais et évalué en utilisant DicoSyn et WN. Les résultats obtenus étaient prévisibles : en ajoutant peu de liens, nous augmentions le rappel et diminuions la précision faiblement ; en ajoutant un nombre important de liens, le rappel augmentait aussi significativement que la précision diminuait. Cette évaluation souffrait au demeurant du biais introduit par l'utilisation d'étalons (*cf.* section 2.3). De plus, se posait la question du nombre approprié de synonymes devant être ajoutés pour un lexème donné. Ce nombre pourrait être fonction du degré d'incidence (nombre de voisins) du lexème, mais dans une ressource en cours de construction telle que Wiktionary, on ne saurait juger si le nombre actuel de synonymes d'un lexème reflète sa polysémie ou le degré d'achèvement de la rédaction de l'entrée correspondante. Une validation manuelle permettrait de pouvoir non seulement juger de la pertinence des propositions, mais également de décider du nombre d'ajouts à effectuer.
Nous décrivons ci-après une méthode d'enrichissement de ressources déjà amorcées par calcul de relations de synonymie candidates dans une perspective de validation « par les foules » et étudions l'impact des différentes sources de données utilisées.

4.1. *Le modèle de graphes pondérés bipartis*

Nous nous appuyons dans les expériences qui suivent sur des *graphes bipartis symétriques et pondérés*. Afin d'homogénéiser la présentation, toutes les sources de données que nous utiliserons seront modélisées par un graphe $G = (V, V', E, w)$ constitué de deux sous-ensembles de sommets. V fera toujours référence à l'ensemble des lexèmes de Wiktionary pour la langue et la catégorie syntaxique considérée. V' désignera un second sous-ensemble de sommets dont le contenu sera constitué de différentes sources de données. L'ensemble E des arcs est tel que $E \subseteq (V \times V')$ et modélise les relations qu'entretiennent les sommets de V (lexèmes) avec ceux de V' (propriétés). Un poids est donné à chaque arc par la fonction $w : E \rightarrow \mathbb{R}^+$.

• **Graphe de traduction** $G_{Wt} = (V, V_{Wt}, E_{Wt}, w_{Wt})$
$V' = V_{Wt}$ est ici constitué des traductions des lexèmes du Wiktionnaire étudié. E_{Wt} représente les liens de traduction : un arc relie $v \in V$ à $t \in V_{Wt}$ si t est une traduction de v dans Wiktionary. Tous les liens ont la même pondération : $\forall e \in E,\ w_{Wt}(e) = 1$.

• **Graphe de synonymie** $G_{Ws} = (V, V_{Ws}, E_{Ws}, w_{Ws})$
$V' = V_{Ws}$ est ici une duplication de V. Un arc relie $v \in V$ à $u \in V_{Ws}$ si $v = u$ ou si u (resp. v) est mentionné comme synonyme dans l'entrée de v (resp. u). Les arcs ont ici aussi tous la même pondération : $\forall e \in E,\ w_{Ws}(e) = 1$.

• **Graphe de gloses** $G_{Wg} = (V, V_{Wg}, E_{Wg}, w_{Wg})$
$V' = V_{Wg}$ correspond ici à l'ensemble des lexèmes lemmatisés trouvés dans les gloses de toutes les entrées. Un arc relie $v \in V$ à $g \in V_{Wg}$ si g apparaît dans une des gloses de v. Pour chaque lexème, ses gloses ont été concaténées, lemmatisées et étiquetées avec TreeTagger, puis les mots vides ont été supprimés. Parmi plusieurs pondérations envisageables, nous avons utilisé la fréquence : le poids de l'arc liant $u \in V$ et $g \in V_{Wg}$ est le nombre d'occurrences de g dans la glose de u. Nous prévoyons dans une expérience future d'utiliser la position du terme dans la glose.

• **Graphe de contextes syntaxiques** $G_{Wpc} = (V, V_{Wpc}, E_{Wpc}, w_{Wpc})$
Nous avons extrait de l'encyclopédie française Wikipédia un corpus de 260 millions de mots que nous avons analysé avec Syntex (Bourigault, 2007). Cet analyseur produit des relations de dépendances syntaxiques que nous avons utilisées pour construire une table de fréquences de couples `<lexème, contexte>`, un contexte dénotant ici un autre lexème et une relation syntaxique. V_{Wpc} désigne l'ensemble de ces contextes et un arc $e = (v, c) \in E_{Wpc}$ signifie qu'un lexème v apparaît dans le contexte c. Nous utilisons l'information mutuelle comme pondération de ces arcs :

$$\forall (v, c) \in E, w_{Wpc}(< v, c >) = \log(\frac{f(v, c)f(*, *)}{f(v, *)f(*, c)})$$

où $f(v, c)$ est la fréquence du lexème v dans le contexte c, $f(v, *)$, $f(*, x)$ et $f(*, *)$ sont respectivement la fréquence totale de v (tous contextes confondus), la fréquence totale de c (avec tout lexème) et le nombre total de couples.

Les tailles des graphes ainsi construits et des graphes combinant les différentes sources de données sont récapitulées dans le tableau 6. Par exemple, $s + t + g$ est le graphe résultant de l'agglomération des sommets et des arcs des graphes de synonymie, de traduction et de gloses :
$G = \big(V, \quad V' = V_{Ws} \cup V_{Wt} \cup V_{Wg}, \quad E = E_{Wt} \cup E_{Wt} \cup E_{Wg}, \quad w\big)$.
Deux sommets provenant d'ensembles différents (de différents V', *e.g.* un de V_{Wt} et un de V_{Wg}) restent dupliqués même s'ils correspondent au même lexème. Nous pondérons les arcs du graphe combiné en multipliant le poids des arcs initiaux par un coefficient dépendant du type d'arc. Un graphe noté $\alpha_s.s + \alpha_t.t + \alpha_g.g + \alpha_c.c$ aura la fonction de pondération suivante :

$$w(e) = \begin{cases} \alpha_s.w_{Ws}(e) & \text{si} \quad e \in E_{Ws}, \\ \alpha_t.w_{Wt}(e) & \text{si} \quad e \in E_{Wt}, \\ \alpha_g.w_{Wg}(e) & \text{si} \quad e \in E_{Wg}, \\ \alpha_c.w_{Wc}(e) & \text{si} \quad e \in E_{Wc}. \end{cases}$$

D'autres manières d'agréger les données et de pondérer les arcs peuvent évidemment être envisagées. À ce stade, les configurations que nous proposons permettent déjà d'augmenter sensiblement le nombre de candidats pertinents proposés (*cf.* section 5.3).

		Anglais			**Français**		
		n	n'	m	n	n'	m
Adjectifs	trads	8 178	43 976	54 840	5 335	23 976	32 944
	syns	8 723	8 723	27 257	4 482	4 482	12 754
	gloses	45 703	39 409	218 993	41 620	42 455	263 281
	contextes	–	–	–	6 262	129 199	934 969
	s + t	13 650	52 699	82 097	7 849	28 458	45 698
	s + t + g	47 280	92 108	301 090	42 507	70 913	308 979
	s + t + g + c	–	–	–	42 517	200 761	1 248 779
Verbes	trads	7 473	52 862	70 432	3 174	30 162	49 866
	syn	7 341	7 341	23 927	3 190	3 190	9 510
	gloses	42 901	36 051	222 004	17 743	16 942	101 458
	contextes	–	–	–	4 273	2 312 096	5 499 611
	s + t	11 423	60 203	94 359	5 054	33 352	59 376
	s + t + g	44 295	96 254	316 363	18 226	50 294	160 834
	s + t + g + c	–	–	–	18 229	2 374 679	5 700 602
Noms	trads	29 489	235 233	277 897	18 468	129 426	153 033
	syns	31 227	31 227	86 195	19 407	19 407	53 869
	gloses	194 694	127 198	1 218 414	105 760	69 994	844 805
	contextes	–	–	–	22 711	1 671 655	8 719 464
	s + t	50 305	266 460	36 4092	30 810	148 833	206 902
	s + t + g	202 920	393 658	1 582 506	111 228	218 827	1 051 707
	s + t + g + c	–	–	–	111 290	1 898 564	9 818 553

Tableau 6. *Ordre et taille des graphes bipartis utilisés pour le calcul de similarité. n et n' : nombre de sommets de V et V' ayant au moins 1 voisin. m : nombre d'arcs. s, t, g et s : graphes de synonymie, traduction, gloses et contextes syntaxiques.*

4.2. *Calcul de similarité par marches aléatoires*

Nous calculons pour chaque sommet s_i sa similarité avec l'ensemble des autres sommets de manière à proposer comme synonymes candidats les sommets ayant les scores de similarité les plus élevés avec s_i. Cette similarité est calculée en effectuant des marches aléatoires dans les graphes présentés en section 4.1. Ce type d'approche proposé notamment par (Gaume *et al.*, 2005 ; Gaume et Mathieu, 2008) pour mesurer la *ressemblance topologique* dans les graphes et par (Hughes et Ramage, 2007) pour calculer la similarité sémantique dans les réseaux lexicaux est en effet adapté ici : nous avons vu en section 2.2.2 que les réseaux de synonymie, bien que globalement peu denses, sont denses localement, la plupart des arêtes d'un réseau de synonymie se trouvant dans ces zones denses. On peut emettre l'hypothèse que la plupart des liens encore absents d'un réseau en cours de construction manquent entre paires de

sommets d'une même zone dense. Or, un marcheur aléatoire entrant dans une zone dense est capturé par cette zone (il y reste longtemps avant d'en sortir). Un marcheur débutant sur un sommet s_i a donc une forte probabilité de passer au début de sa balade sur les sommets des zones denses auxquelles s_i appartient (Gaume *et al.*, 2010). Nous considérons un marcheur parcourant aléatoirement un des graphes $G = (V \cup V', E, w)$ décrit plus haut, partant de v. À chaque pas, la probabilité qu'il se déplace des sommets i à j est donnée par la cellule (i, j) de la matrice de transition P, définie comme suit :

$$[P]_{ij} = \begin{cases} \frac{w((i,j))}{\sum_{k \in \mathcal{N}(i)} w((i,k))} & \text{si} \quad (i,j) \in E, \\ 0 & \text{sinon.} \end{cases} \quad [1]$$

où $\mathcal{N}(i)$ est l'ensemble des voisins du sommet i : $\mathcal{N}(i) = \{j/(i,j) \in E\}$. Ainsi, la position du marcheur après t pas est donnée par la distribution de probabilités $X_t(v) = \delta_v P^t$, où δ_v est un vecteur-ligne de dimension $|V \cup V'|$ contenant la valeur 0 à toutes ses coordonnées, excepté celle correspondant au sommet v qui vaut 1. On notera $X_t(v, u)$ la valeur de la coordonnée u de ce vecteur, qui représente la probabilité pour le marcheur d'atteindre u en partant de v après t pas. Plusieurs mesures de similarité fondées sur cette probabilité ont été testées dans (Sajous *et al.*, 2010) et donnent des résultats équivalents. Nous effectuerons la suite des expériences avec $X_2(v, u)$, qui correspond à une marche aléatoire de longueur 2.

5. Évaluation

5.1. *Pertinence des candidats proposés*

En vue d'un enrichissement semi-automatique dans lequel les contributeurs valident ou invalident les candidats proposés, nous considérons qu'une liste (courte) de suggestions est *acceptable* si elle contient au moins 1 candidat pertinent. Ainsi, notre évaluation consiste essentiellement à compter pour combien de lexèmes le système produit une liste acceptable. Nous examinons également pour combien de lexèmes sont proposées des listes contenant 2, 3 ou plusieurs candidats pertinents. Soit $G_{GS} = (V_{GS}, E_{GS})$ un réseau de synonymie étalon, V_{GS} ses lexèmes, et $E_{GS} \subseteq V_{GS} \times V_{GS}$ ses relations de synonymie. Nous évaluons ci-après la pertinence des relations proposées pour enrichir la ressource déficiente par rapport aux relations de l'étalon. Nous effectuons cette évaluation pour les lexèmes communs à la ressource à enrichir et l'étalon. Nous supprimons de la liste des candidats ceux qui ne figurent pas dans l'étalon[8] et limitons cette liste à $k \leq 5$ candidats. Pour chaque lexème $v \in V \cap V_{GS}$, on note $\Gamma_k(v)$ la liste des candidats *« évaluables »* :

$$\Gamma_k(v) = [c_1, c_2, \ldots, c_{k'}] \quad \text{avec} \quad \begin{cases} k' \leq k \\ \forall i,\ c_i \in C(v) \cap V_{GS} \\ \forall i, \text{sim}(v, c_i) \geq \text{sim}(v, c_{i+1}) \end{cases} \quad [2]$$

8. Il peut s'agir d'un néologisme ou d'un terme de spécialité, plus rarement d'une forme mal orthographiée de Wiktionary, ou être dû à une couverture non exhaustive de l'étalon.

$\Gamma_k(v)$ contient un maximum de k candidats (mais peut en contenir moins ou être vide). De plus, $\Gamma_k(v)$ dépend de la ressource étalon. On note $\Gamma_k^+(v)$ l'ensemble des candidats corrects de $\Gamma_k(v)$: $\Gamma_k^+(v) = \left\{c^+ \in \Gamma_k(v) / (v, c^+) \in E_{GS}\right\}$. On note N_k l'ensemble des lexèmes pour lesquels on propose k candidats et N_k^{+p} le sous-ensemble de lexèmes pour lesquels *au moins p candidats pertinents* sont proposés :

$$N_k = \left\{v \in V \cap V_{GS} / |\Gamma_k(v)| = k\right\} \ , \ N_k^{+p} = \left\{v \in N_k / |\Gamma_k^+(v)| \geq p\right\} \quad [3]$$

Pour mesurer l'impact des données utilisées sur le calcul des candidats, nous mesurons le ratio R_k entre le nombre de listes *suggérées*[9] et le nombre de lexèmes *évaluables*, et P_k, le ratio entre le nombre de listes *acceptables* et le nombre de listes *suggérées* :

$$R_k = \frac{|N_k|}{|V_{GS} \cap V|}, \qquad P_k = \frac{|N_k^{+1}|}{|N_k|} \quad [4]$$

Nous utilisons WN et DicoSyn comme ressources étalons respectivement pour l'anglais et le français. Pour les détails d'extraction des réseaux de synonymie de ces ressources, nous renvoyons à (Navarro *et al.*, 2009).

5.2. *Qualification des relations proposées*

Afin de mieux cerner le type de relations capturées par notre mesure de similarité, nous avons étudié la nature des synonymes candidats rejetés par l'étalon choisi. En utilisant WN, nous avons réparti les candidats anglais des listes $\Gamma_k(v)$ présentées plus haut ($k \leq 5$) suivant les relations *synonymie, hyperonymie, hyponymie, cohyponymie* pour les noms et *synonymie, hyperonymie, troponymie, cotroponymie* pour les verbes[10]. Nous avons gardé dans cette évaluation la finesse de granularité de WN en respectant ses définitions strictes : deux lexèmes sont synonymes s'ils apparaissent dans un même synset, et nous considérons seulement les relations strictes d'hyperonymie/hyponymie/cohyponymie (*e.g.* w_1 est en relation d'hyperonymie avec w_2 si w_2 apparaît dans un synset qui est un fils direct d'un synset dans lequel apparaît w_1).

5.3. *Résultats*

Le tableau 7 montre les résultats obtenus à partir des différentes sources de données. Le graphe de traduction donne une meilleure précision que celui de synonymie. Ce résultat était prévisible car les lexèmes, dans Wiktionary, ont plus de liens de traduction que de synonymie. De plus, ces liens sont répartis sur un nombre important de langues, ce qui rend l'information relativement fiable. Le meilleur rappel et la moindre précision sont obtenus avec les gloses et les contextes syntaxiques. Ce résultat était

9. *i.e.* le nombre de lexèmes pour lesquels au moins 1 candidat est proposé.

10. Nous avons également considéré, puis écarté les relations d'antonymie, holonymie/méronymie et causalité, qui concernaient entre 1 ‰ et 0,1 ‰ des cas.

			R_5	P_5	$\lvert N_5\rvert$	$\lvert N_5^{+1}\rvert$	$\lvert N_5^{+2}\rvert$	$\lvert N_5^{+3}\rvert$	$\lvert N_5^{+4}\rvert$	$\lvert N_5^{+5}\rvert$
ANG	A.	syns	17,4	49,1	2 456	1 207	439	**165**	**57**	**22**
		trads	9,2	**65,7**	1 299	853	406	144	27	3
		gloses non pondérées	**93,5**	25,9	**13 205**	3 421	774	154	34	2
		gloses pondérées	**93,5**	26,6	**13 205**	**3 510**	**794**	158	30	1
		$s+t$	21,9	**58,1**	3 096	1 800	805	283	91	29
		$\mathbf{10.s+10.t+g}$	**95,0**	35,9	**13 417**	**4 819**	**1 567**	**455**	**125**	**32**
		$\mathbf{10^2.s+10^2.t+g}$	**95,0**	35,9	**13 417**	**4 818**	**1 567**	**455**	**125**	**32**
	N.	syns	8,7	34,6	3 862	1 335	483	**200**	**95**	**54**
		trads	8,5	**51,0**	3 759	1 916	655	178	41	2
		glosses non pondérées	**95,6**	14,8	**42 337**	6 252	926	106	6	0
		gloses pondérées	**95,6**	15,3	**42 337**	**6 467**	**933**	114	5	1
		$s+t$	14,5	**47,6**	6 440	3 063	1 061	348	110	45
		$\mathbf{10.s+10.t+g}$	**96,4**	23,3	**42 688**	**9 944**	2 344	**561**	**142**	**43**
		$\mathbf{10^2.s+10^2.t+g}$	**96,4**	23,3	**42 688**	9 942	**2 345**	**561**	**143**	**43**
	V.	syns	23,9	44,5	2 153	959	431	**216**	**115**	**59**
		trads	24,7	**60,4**	2 223	1 342	**609**	187	43	1
		gloses non pondérées	**98,5**	27,0	**8 852**	2 389	518	98	10	2
		gloses pondérées	**98,5**	28,1	**8 852**	**2 490**	548	100	13	2
		$s+t$	37,6	**58,0**	3 380	1 962	918	358	119	**43**
		$\mathbf{10.s+10.t+g}$	**99,2**	41,0	8 916	**3 655**	**1 352**	**448**	**136**	34
		$10^2.s+10^2.t+g$	**99,2**	40,9	**8 917**	3 644	1 351	**448**	**136**	34
FR	A.	syns	11,9	75,2	480	361	224	**139**	55	**16**
		trads	6,0	**91,4**	243	222	184	117	**56**	11
		gloses non pondérées	**90,2**	32,2	**3 627**	1 167	309	91	12	1
		gloses pondérées	**90,2**	33,6	**3 627**	**1 220**	**337**	100	17	0
		contextes	86,2	20,7	3 468	719	157	40	11	1
		$s+t$	15,7	**81,3**	631	513	375	243	105	28
		$10.s+10.t+g$	89,5	44,3	3 602	1 594	728	371	154	38
		$10^2.s+10^2.t+g$	91,2	43,6	3 668	1 600	729	370	155	38
		$10^2.s+10^2.t+10.g+c$	**97,3**	41,9	3 913	1 640	680	347	143	32
		$\mathbf{10^3.s+10^3.t+10^2.g+c}$	**97,3**	45,3	**3 915**	**1 774**	**791**	**408**	**172**	**43**
	N.	syns	10,4	54,4	1 722	936	478	194	68	15
		trads	5,5	**79,7**	916	730	472	**245**	**94**	**20**
		gloses non pondérées	**95,8**	20,6	**15 828**	3 268	607	116	16	2
		gloses pondérées	**95,8**	22,5	**15 828**	**3 560**	693	127	21	3
		contextes	84,0	20,9	13 882	2 898	**721**	181	34	5
		$s+t$	15,2	**66,9**	2 511	1 681	983	480	166	33
		$10.s+10.t+g$	96,5	33,3	15 948	5 303	1 956	735	219	50
		$10^2.s+10^2.t+g$	96,5	33,2	15 948	5 298	1 952	736	218	52
		$10^2.s+10^2.t+10.g+c$	**98,5**	33,1	**16 274**	5 394	1 908	649	196	38
		$\mathbf{10^3.s+10^3.t+10^2.g+c}$	98,4	36,7	16 273	**5 980**	**2 240**	**825**	**260**	**56**
	V.	syns	10,0	68,0	412	280	172	86	30	5
		trads	19,0	**90,4**	785	710	544	**352**	**146**	**38**
		gloses non pondérées	**95,6**	41,2	**3 947**	1 628	530	149	38	3
		gloses pondérées	**95,6**	44,9	**3 947**	**1 773**	**638**	198	45	8
		contextes	81,8	35,3	3 378	1 192	426	126	28	3
		$s+t$	25,7	**85,6**	1 062	909	669	418	165	48
		$10.s+10.t+g$	96,6	55,9	3 989	2 229	1 161	580	216	**58**
		$10^2.s+10^2.t+g$	96,6	55,8	3 989	2 226	1 160	580	214	**58**
		$10^2.s+10^2.t+10.g+c$	**98,1**	53,2	**4 053**	2 158	1 004	433	146	43
		$\mathbf{10^3.s+10^3.t+10^2.g+c}$	**98,1**	58,4	**4 053**	**2 368**	**1 243**	**604**	**223**	53

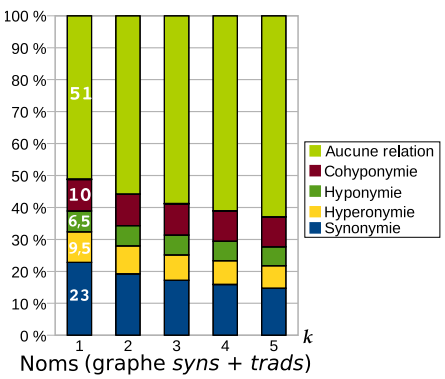
Tableau 7. *Impact des sources de données sur le calcul de similarité*

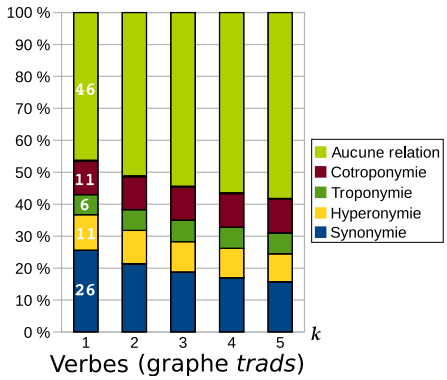
attendu également, l'information véhiculée par ces données étant moins spécifique. Elle permet cependant de proposer des candidats pour la quasi-totalité des lexèmes. La pondération des gloses par les fréquences améliore légèrement les résultats et l'on peut penser qu'une pondération plus fine (*e.g.* en favorisant les termes situés à l'initiale des gloses) produirait une nouvelle amélioration. Plus étonnamment, les cooccurrents

syntaxiques donnent une précision moindre, ce qui ne va pas dans le sens des résultats de Van der Plas et Bouma (2005). Filtrer les contextes de basse fréquence permettrait probablement d'améliorer ce score : en effet, un lexème apparaissant avec un unique contexte syntaxique tend à avoir avec lui une information mutuelle élevée sans que le rapprochement avec un autre lexème apparaissant avec ce même contexte ne soit réellement significatif. Pour l'anglais, agglomérer les graphes de synonymie et de traduction améliore le rappel sans détriment notable pour la précision. Pour le français, cela conduit à une baisse de précision par rapport à l'utilisation du seul graphe de traduction. Dès lors que l'on introduit les gloses, des candidats sont proposés pour quasiment tous les lexèmes. Les meilleurs résultats sont obtenus en combinant les graphes de synonymie, de traduction et de gloses pour l'anglais (graphe $10.s + 10.t + g$) et celui de synonymie, traduction et des contextes syntaxiques pour le français (graphe $10^3.s+10^3.t+10^2.g+c$). L'utilisation de ces graphes nous permet de proposer 5 candidats pour presque tous les lexèmes et 35 à 60 % des listes proposées comportent au moins 1 candidat validé par un étalon. Rappelons que ces listes sont proposées sur la base d'une ressource en cours de construction encore très incomplète. On peut penser que la prédiction de liens manquants s'améliorera avec la densification de la ressource. D'autre part, l'objectif n'est pas de suggérer des candidats pour un nombre restreint de lexèmes avec un score de confiance élevé (nous atteignons dans ce cas 80 à 90 % de listes pertinentes en ne recourant qu'aux graphes de traduction pour le français, mais pour seulement 6 à 20 % des lexèmes), mais d'effectuer les suggestions les plus pertinentes pour une nombre maximal de lexèmes. Notons que l'utilisation des gloses et des contextes syntaxiques dans les graphes combinés, tout en permettant de traiter l'ensemble des lexèmes, fait chuter la précision seulement au niveau global : au niveau local, les lexèmes pour lesquels les liens de synonymie et de traduction conduisaient à proposer des candidats pertinents ne sont pas nécessairement affectés par cette baisse (la proportion P_5 baisse mais pas le cardinal N_k^{+i}).

Les résultats sont meilleurs pour le français que pour l'anglais. Cela pourrait s'expliquer par la densité inégale des réseaux initiaux, mais la différence est surtout dûe aux étalons utilisés : DicoSyn, étant une compilation de sept dictionnaires, est plus dense que WN. Tirer des conclusions définitives quant à la qualité globale de la méthode est difficile : une analyse qualitative montre que certains candidats proposés et rejetés par les étalons ne semblent pas déraisonnables (*e.g.* <*force, poigne*>, <*salade, bobard*>, <*drogue, psychotrope*>, etc.). À l'inverse, des candidats validés comme synonymes par les étalons pourraient être discutés (*e.g.* <*rongeur, mulot*>, <*sens, toucher*> semblent être de bons candidats à l'hyperonymie). Les étalons restent néanmoins utiles pour la sélection des sources de données à utiliser.

La figure 3 montre le type de relations calculées par notre méthode. Pour tous les noms et verbes communs à Wiktionary et WN, nous avons considéré des listes de 1 à 5 candidats, puis supprimé ceux absents de WN. Nous pouvons voir qu'une partie des candidats non pertinents pour la synonymie recouvre une autre réalité sémantique. Sans que les résultats soient directement comparables (enrichissement endogène *vs.* extraction en corpus, WN *vs.* WordNet néerlandais), nous pouvons noter que la similarité de Heylen *et al.* (2008) produit autant de cohyponymes que de synonymes alors que nos marches aléatoires produisent deux fois plus de synonymes que de cohyponymes.





Catégories	Lexèmes	Candidats	Relations
Noms	*hound*	*greyhound*	hyponymie
	law	*rule*	hyponymie
	fool	*idiot, dummy*	cohyponymie (synset hyperonyme : *{simpleton, simple}*)
Verbes	*to represent*	*to depict*	troponymie
	to negociate	*to bargain*	troponymie
	to blame	*to incriminiate*	cotroponymie (synset hyperonyme : *{to charge, to accuse}*)

Figure 3. *Exemples et proportions de relations capturées par marches aléatoires*

6. Implémentation et application à Wiktionary : le système WISIGOTH

Pour mettre en œuvre l'enrichissement et son système de validation décrits plus haut, nous avons implémenté le système WISIGOTH (WIktionarieS Improvement by Graph-Oriented meTHods), constitué d'une chaîne de traitement qui, à partir des *dumps* de Wiktionary, extrait son réseau de synonymie et calcule les nouvelles relations candidates, et d'une extension Firefox. Lorsqu'un internaute installe cette extension et visite une page française ou anglaise de Wiktionary, notre serveur est interrogé et renvoie une liste ordonnée de candidats pour le(s) lexème(s) en cours de consultation. Ces candidats sont alors présentés à l'internaute (*cf.* figure 4). S'il les valide (bouton '+'), l'extension prend en charge l'ajout du synonyme dans Wiktionary et l'édition du code wiki correspondant. Un champ texte libre permet également de proposer un synonyme non suggéré. Indépendamment de notre méthode d'enrichissement, WISIGOTH permet ainsi aux internautes non familiers avec le code wiki de contribuer. Pour rester proche du principe wiki, nous n'avons pas mis en place de système de validation croisée et les ajouts sont publiés immédiatement. Néanmoins, facilitant l'ajout de synonymes, nous devons également en faciliter leur suppression : un bouton '-' est ajouté à chaque synonyme présent à cette fin. Enfin, un système de liste noire permet à un contributeur de signaler une suggestion jugée non pertinente (bouton '×'). Cette suggestion ne lui sera plus faite et permettra aux autres candidats de remonter dans la liste. Au-delà d'un seuil (réglé pour l'instant arbitrairement à 3) de mises sur listes noires individuelles, la proposition est insérée dans une liste noire globale de manière à ne plus être proposée à aucun utilisateur.

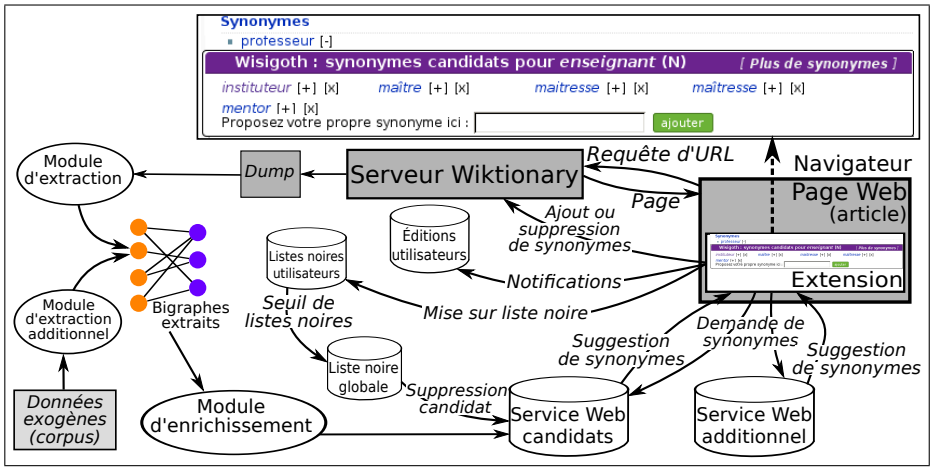


Figure 4. *Architecture du système WISIGOTH*

7. Conclusion et perspectives

Nous avons pointé dans ce travail les obstacles qui s'opposent à la création de ressources lexicales sémantiques et notamment les problèmes liés à leur coût, leur disponibilité et leur évaluation. À la lumière de ces observations, nous avons proposé un système d'enrichissement pour les ressources déjà amorcées, fondé sur une validation manuelle. Nous fournissons une implémentation pour appliquer cette méthode à Wiktionary, dont le contenu est déjà significatif et misons sur une validation *par les foules*. Ce pari pourrait permettre de régler les problèmes de coût, de disponibilité, de pérennisation et de mise à jour d'une telle ressource. Par ailleurs, fondée sur les graphes et l'utilisation de données endogènes, notre méthode est indépendante de la langue considérée et peut s'appliquer à d'autres types de réseaux lexicaux. Elle pourrait aider, par exemple, au développement des WordNets actuellement en cours de construction. Nous ne nous interdisons cependant pas l'utilisation de ressources exogènes lorsqu'elles sont disponibles. Bien qu'ayant souligné les limites des ressources étalons, nous avons néanmoins procédé à une évaluation des candidats proposés par notre système fondée sur des étalons. Ceux-ci ont permis de montrer d'une part la viabilité de notre méthode pour l'application envisagée et également d'observer l'impact des sources de données utilisées. L'évaluation réellement significative aura lieu après un ou deux ans d'utilisation de cette application. Nous pourrons alors juger si notre système a permis d'accélérer la croissance du réseau de synonymie de Wiktionary.
Ces premiers résultats laissent envisager plusieurs extensions. Un apprentissage automatique permettrait d'obtenir une meilleure pondération des arcs des graphes utilisés. À plus court terme, une modification de l'interface permettra, pour un candidat jugé non synonyme, de l'ajouter le cas échéant comme cible d'autres relations. Nous comptons adapter le calcul de similarité pour produire à la demande ces relations (*e.g.*

hyperonymie, traductions). Il serait finalement intéresant d'opérer une désambiguïsation pour produire des relations entre lexèmes (et non entre formes). Les travaux dans ce sens de Meyer et Gurevych (2010) tirent parti d'une meilleure structuration du Wiktionnaire allemand (présence systématique du numéro du lexème source des relations), mais rencontrent pour l'instant un succès mitigé. Nous sommes enfin ouverts à toute collaboration pour porter WISIGOTH à d'autres langues et d'autres réseaux lexicaux. De plus, nous pouvons intégrer des candidats produits par d'autres équipes : en les hébergeant ou en faisant interroger un service Web supplémentaire par notre interface.

Ressources : les ressources présentées ou mentionnées dans cet article (extension WISIGOTH, *dumps* convertis au format XML, corpus Wikipédia) sont disponibles en ligne sur le site REDAC : `http://redac.univ-tlse2.fr/`

Remerciements

Nous remercions Yannick Chudy, pour son travail de développement de WISIGOTH et sa contribution à l'ensemble de nos travaux. Nous avons également bénéficié depuis le début de ce projet de la collaboration précieuse de Laurent Prévot.

8. Bibliographie

Albert R., Barabasi A.-L., « Statistical Mechanics of Complex Networks », *Reviews of Modern Physics*, vol. 74, p. 74-47, 2002.

Bourigault D., Un analyseur syntaxique opérationnel : SYNTEX, Mémoire d'habilitation à diriger des recherches, Université de Toulouse, 2007.

Brunello M., « The Creation of Free Linguistic Corpora from the Web », *Proceedings of WAC5 : 5th Workshop on Web As Corpus*, San Sebastian, p. 37-44, 09, 2009.

Cruse D., *Lexical Semantics*, Cambridge University Press, 1986.

Curran J. R., Moens M., « Improvements in Automatic Thesaurus Extraction », *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, p. 59-66, 2002.

Edmonds P., Hirst G., « Near-Synonymy and Lexical Choice », *Computational Linguistics*, vol. 28, n° 2, p. 105-144, 2002.

Fellbaum C. (ed.), *WordNet : An Electronic Lexical Database*, MIT Press, 1998.

Gaume B., « Balades aléatoires dans les petits mondes lexicaux », *I3 : Information Interaction Intelligence*, 2004.

Gaume B., Duvignau K., Prévot L., Desalle Y., « Toward a Cognitive Organization for Electronic Dictionaries, the Case for Semantic Proxemy », *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008)*, Manchester, p. 86-93, 2008.

Gaume B., Mathieu F., « PageRank Induced Topology for Real-World Networks », *Complex Systems*, 2008.

Gaume B., Mathieu F., Navarro E., « Building Real-World Complex Networks by Wandering on Random Graphs », *Information - Interaction - Intelligence*, 2010.

Gaume B., Venant F., Victorri B., « Hierarchy in Lexical Organization of Natural Language », *in* D. Pumain (ed.), *Hierarchy in Natural and Social Sciences*, Methodos series, Kluwer Academic Publishers, p. 121-143, 2005.

Giles J., « Internet Encyclopaedias go Head to Head », *Nature*, vol. 438, p. 900-901, 2005.

Habert B., Naulleau E., Nazarenko A., « Symbolic Word Clustering for Medium-Size Corpora », *Proceedings of the 16th conference on Computational linguistics*, Association for Computational Linguistics, Morristown, NJ, USA, p. 490-495, 1996.

Hearst M. A., « Automatic Acquisition of Hyponyms from Large Text Corpora », *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Nantes, p. 539-545, July, 1992.

Heylen K., Peirsman Y., Geeraerts D., Speelman D., « Modelling Word Similarity : an Evaluation of Automatic Synonymy Extraction Algorithms. », *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may, 2008.

Hughes T., Ramage D., « Lexical Semantic Relatedness with Random Graph Walks », *Proceedings of EMNLP-CoNLL*, p. 581-589, 2007.

Jacquin C., Desmontils E., Monceaux L., « French EuroWordNet Lexical Database Improvements », *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing (CICLING)*, p. 12-22, 2007.

Kilgarriff A., « I don't Believe in Word Senses », *Computers and the Humanities*, vol. 31, n° 2, p. 91-113, 1997.

Kilgarriff A., « Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs », *Computer Speech & Language*, vol. 12, n° 4, p. 453-472, 1998.

Lafourcade M., « Making People Play for Lexical Acquisition with the JeuxDeMots prototype », *SNLP'07 : 7th International Symposium on Natural Language Processing*, Pattaya, Thailand, 12, 2007.

Meyer C. M., Gurevych I., « Worth its Weight in Gold or Yet Another Resource – A Comparative Study of Wiktionary, OpenThesaurus and GermaNet », *Proceedings of the 11^{th} International Conference on Intelligent Text Processing and Computational Linguistics*, vol. 6008 of *LNCS*, Springer, p. 38-49, Mar, 2010.

Murray G. C., Green R., « Lexical Knowledge and Human Disagreement on a WSD Task », *Computer Speech & Language*, vol. 18, n° 3, p. 209-222, 2004.

Navarro E., Sajous F., Gaume B., Prévot L., Hsieh S., Kuo I., Magistry P., Huang C.-R., « Wiktionary and NLP : Improving Synonymy Networks », *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, Singapore, p. 19-27, August, 2009.

Newman M. E. J., « The Structure and Function of Complex Networks », *SIAM Review*, vol. 45, p. 167-256, 2003.

Ollivier Y., Senellart P., « Finding Related Pages Using Green Measures : an Illustration with Wikipedia », *AAAI'07 : Proceedings of the 22^{nd} national conference on Artificial intelligence*, AAAI Press, p. 1427-1433, 2007.

Palmer M., Dang H. T., Fellbaum C., « Making Fine-Grained and Coarse-Grained Sense Distinctions, Both Manually and Automatically », *Natural Language Engineering*, vol. 13, n° 2, p. 137-163, 2007.

Pantel P., Pennacchiotti M., « Espresso : Leveraging Generic Patterns for Automatically Harvesting Semantic Relations », *Proceedings of the International Conference on Computational Linguistics*, ACL Press, Sydney, p. 113-120, 17th–21st July, 2006.

Ploux S., Victorri B., « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *Traitement automatique des langues*, vol. 39, n° 1, p. 161-182, 1998.

Sagot B., Fišer D., « Building a Free French Wordnet from Multilingual Resources », *Proceedings of OntoLex 2008*, Marrakech, 2008.

Sajous F., Navarro E., Gaume B., Prévot L., Chudy Y., « Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources : Piggybacking onto Wiktionary », *in* H. Loftsson, E. Rögnvaldsson, S. Helgadóttir (eds), *Advances in Natural Language Processing*, vol. 6233 of *Lecture Notes in Computer Science*, Springer Berlin/Heidelberg, p. 332-344, 2010.

Sekine S., « We Desperately Need Linguistic Resources ! –Based on the Users' Point of View », FLaReNet Forum. Barcelona, Feb, 2010.

Snow R., O'Connor B., Jurafsky D., Ng A. Y., « Cheap and Fast—but is it good ? : Evaluating Non-Expert Annotations for Natural Language Tasks », *EMNLP '08 : Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Morristown, NJ, USA, p. 254-263, 2008.

Soria C., Monachini M., Bertagna F., Calzolari N., Huang C.-R., Hsieh S.-K., Marchetti A., Tesconi M., « Exploring Interoperability of Language Resources : the Case of Cross-Lingual Semi-Automatic Enrichment of Wordnets », *Language Resources and Evaluation*, vol. 40, n° 1, p. 87-96, 2009.

Tufis D., « Balkanet Design and Development of a Multilingual Balkan Wordnet », *Romanian Journal of Information Science and Technology*, 2000.

Van der Plas L., Bouma G., « Syntactic Contexts for Finding Semantically Related Words », *in* T. van der Wouden, M. Poß, H. Reckman, C. Cremers (eds), *Computational Linguistics in the Netherlands 2004 : Selected papers from the fifteenth CLIN meeting*, vol. 4 of *LOT Occasional Series*, Utrecht University, 2005.

Victorri B., Fuchs C., *La Polysémie, construction dynamique du sens*, Hermès, 1996.

Voormann H., Gut U., « Agile Corpus Creation », *Corpus Linguistics and Lingustic Theory*, vol. 4, n° 2, p. 235-251, 2008.

Vossen P. (ed.), *EuroWordNet : a Multilingual Database with Lexical Semantic Networks*, Kluwer Academic Publishers, Norwell, MA, USA, 1998.

Watts D. J., Strogatz S. H., « Collective Dynamics of Small-World Networks », *Nature*, vol. 393, p. 440-442, 1998.

Weale T., Brew C., Fosler-Lussier E., « Using the Wiktionary Graph Structure for Synonym Detection », *Proceedings of the ACL-IJCNLP Workshop on The People's Web Meets NLP : Collaboratively Constructed Semantic Resources*, Singapore, p. 28-31, August, 2009.

Zesch T., « What's the Difference ? –Comparing Expert-Built and Collaboratively-Built Lexical Semantic Resources », FLaReNet Forum. Barcelona, Feb, 2010.

Zesch T., Gurevych I., « Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words », *Journal of Natural Language Engineering.*, vol. 16, n° 01, p. 25-59, Jan, 2010.