TALN 2008, Avignon, 9–13 juin 2008

Construction d'un wordnet libre du français à partir de ressources multilingues

Benoît Sagot[1] Darja Fišer[2]
(1) Alpage, INRIA / Paris 7, 30 rue du Ch. des rentiers, 75013 Paris, France
(2) Fac. des Lettres, Univ. de Ljubljana, Aškerčeva 2, 1000 Ljubljana, Slovénie
benoit.sagot@inria.fr, darja.fiser@guest.arnes.si

Résumé. Cet article décrit la construction d'un Wordnet Libre du Français (WOLF) à partir du Princeton WordNet et de diverses ressources multilingues. Les lexèmes polysémiques ont été traités au moyen d'une approche reposant sur l'alignement en mots d'un corpus parallèle en cinq langues. Le lexique multilingue extrait a été désambiguïsé sémantiquement à l'aide des wordnets des langues concernées. Par ailleurs, une approche bilingue a été suffisante pour construire de nouvelles entrées à partir des lexèmes monosémiques. Nous avons pour cela extrait des lexiques bilingues à partir de Wikipédia et de thésaurus. Le wordnet obtenu a été évalué par rapport au wordnet français issu du projet EuroWordNet. Les résultats sont encourageants, et des applications sont d'ores et déjà envisagées.

Abstract. This paper describes the construction of a freely-available wordnet for French (WOLF) based on Princeton WordNet by using various multilingual resources. Polysemous words were dealt with an approach in which a parallel corpus for five languages was word-aligned and the extracted multilingual lexicon was disambiguated with the existing wordnets for these languages. On the other hand, a bilingual approach sufficed to acquire equivalents for monosemous words. Bilingual lexicons were extracted from Wikipedia and thesauri. The merged wordnet was evaluated against the French WordNet. The results are promising, and applications are already intended.

Mots-clés : Wordnet, corpus alignés, Wikipédia, sémantique lexicale.

Keywords: Wordnet, aligned corpora, Wikipedia, lexical semantics.

1 Introduction

Wordnet (Fellbaum, 1998) est une base de données lexicale à large couverture, dans laquelle les mots sont répartis en catégories et organisés en une hiérarchie de nœuds. Chaque nœud a un identifiant unique, et représente un *concept*, ou *synset* (ensemble de synonymes). Il regroupe un certain nombre de lexèmes synonymes dénotant ce concept. Ainsi, dans le Princeton WordNet (voir ci-dessous), le synset ENG20-02853224-n comprend les lexèmes {*car, auto, automobile, machine, motorcar*}. Les synsets sont précisés par une brève définition et sont liés à d'autres synsets (ainsi, ce synset est lié au synset {*motor vehicle, automotive vehicle*} par un lien d'hypéronymie, et au synset {*cab, hack, taxi, taxicab*} par un lien d'hyponymie). Les lexèmes peuvent être simples ou composés. Les usages métaphoriques et idiomatiques sont pris en compte.

Historiquement, le premier wordnet est le Princeton WordNet (PWN), développé pour l'anglais à la Princeton University. Au fil du temps, il est devenu une des ressources les plus utiles pour de nombreuses applications de compréhension et d'interprétation automatique des langues, telles que la désambiguïsation sémantique, l'extraction d'informations, la traduction automatique, la classification de documents et le résumé de textes. Ceci a provoqué le développement de wordnets pour de nombreuses autres langues (Vossen, P., 1999; Tufiş, 2000).

À l'heure actuelle, la Global WordNet Association[1] répertorie des wordnets pour plus de 50 langues. Bien que la construction manuelle d'un wordnet produise les meilleurs résultats en termes de pertinence et de précision linguistiques, une telle entreprise est très coûteuse en temps et en ressources. C'est pourquoi des approches semi-automatiques ou totalement automatiques ont été proposées, qui tirent parti des ressources existantes. Mis à part le « goulot d'étranglement de l'acquisition de connaissance » (*knowledge acquisition bottleneck*), un problème majeur au sein de la communauté wordnet est la disponibilité des wordnets développés. Actuellement, seuls quelques-uns d'entre eux sont librement disponibles (le PWN, mais également les wordnets de l'arabe, de l'hébreu et de l'irlandais). Bien qu'un wordnet pour le français ait été développé dans le cadre du projet EuroWordNet (EWN), il ne comporte que des lexèmes verbaux et nominaux, mais ni adjectif ni adverbe. De plus, il n'a pas été largement utilisé, principalement en raison de problèmes de licence. Enfin, aucun projet n'a pris le relais pour poursuivre l'extension et l'amélioration de cet EWN français (Jacquin *et al.*, 2007). Ce sont là les motivations du travail présenté dans cet article : exploiter les ressources multilingues librement disponibles pour construire automatiquement un wordnet à large couverture et librement disponible pour le français, nommé WOLF (Wordnet Libre du Français)[2].

Cet article est organisé comme suit : une brève description des travaux antérieurs est fournie à la section suivante. La section 3 décrit la méthodologie utilisée. La section 4 décrit et évalue la ressource obtenue. La dernière section présente les conclusions et les perspectives envisagées.

2 Travaux antérieurs

On peut répartir les techniques automatiques de développement ou d'aide au développement de wordnets en deux familles : les approches par fusion (*merge approach*) et les approches par extension (*extend approach*). Les approches par fusion, où l'on construit un wordnet à partir de ressources monolingues avant de le mettre en relation avec d'autres wordnets, sont très complexes et coûteuses en ressources. Nous avons donc opté pour l'approche par extension (Vossen, P., 1999). Cette approche prend en entrée un ensemble donné de synsets du Princeton WordNet (PWN) et les traduit dans la langue cible, en préservant la structure du PWN. L'inconvénient de l'approche par extension est que les wordnets obtenus sont biaisés par rapport au PWN, ce qui est moins problématique lorsque les langues source et cible ne sont pas trop éloignées l'une de l'autre, comme c'est le cas pour l'anglais et le français. C'est en raison de sa grande simplicité que l'approche par extension a été utilisée dans nombre de projets, tels que BalkaNet (Tufiş, 2000) et MultiWordNet (Pianta *et al.*, 2002).

Les équipes de recherche développant des wordnets de cette façon tirent parti de toutes les ressources à leur disposition, qu'il s'agisse de dictionnaires électroniques bilingues ou monolingues utilisés pour la désambiguïsation et la structuration du lexique ou bien de taxonomies

[1] `http://www.globalwordnet.org/`

[2] Le WOLF est disponible sous licence Cecill-C (compatible LGPL), à `http://wolf.gforge.inria.fr`

et d'ontologies qui fournissent en général une description plus détaillée et plus formalisée des termes. Le wordnet français développé dans le cadre du projet EuroWordNet a été construit ainsi : un sous-ensemble des synsets du PWN a été automatiquement traduit au moyen d'une base de données sémantique multilingue propriétaire, puis validé manuellement.

Pour la construction du WOLF, nous avons exploité trois types de ressources librement disponibles : le corpus parallèle multilingue JRC-Acquis, Wikipédia et d'autres ressources wiki, ainsi que les descripteurs EUROVOC (utilisé dans le thésaurus européen EUROVOC). Des traductions pour les lexèmes monosémiques, qui ne requièrent pas de désambiguïsation sémantique, ont été extraites des ressources wiki et EUROVOC. Le corpus parallèle a été utilisé pour obtenir des informations sémantiquement pertinentes à partir de la relation multilingue de traduction, afin de traiter également les lexèmes polysémiques. L'idée que des informations sémantiques peuvent être extraites de la relation de traduction a été déjà explorée par (Resnik & Yarowsky, 1997; Ide *et al.*, 2002; Diab, 2004). Elle a également déjà donné des résultats prometteurs pour la construction de synsets pour le wordnet slovène (Fišer, 2007).

3 Méthodologie

3.1 Présentation générale

La difficulté principale lors de la construction d'un wordnet est la polysémie du vocabulaire de base. Dans ce travail, nous avons traité cette question par l'approche par alignement. L'idée repose sur l'hypothèse que les différents sens des mots ambigus dans une langue donnée donnent souvent lieu à des traductions différentes dans une autre langue. À l'inverse, nous supposons que si deux mots ou plus sont traduits par le même mot dans une autre langue, ils partagent souvent un élément de sens. En outre, ces phénomènes sont renforcés par l'utilisation de plus de deux langues, d'où l'intérêt d'une approche par alignement multilingue. La méthode repose sur le lexique multilingue extrait à partir d'un corpus aligné automatiquement en mots, et utilise les wordnets des autres langues pour désambiguïser les entrées lexicales et leur assigner un identifiant de sens. Les relations sémantiques entre synsets sont directement récupérées du PWN (version 2.0, celle utilisé par BalkaNet).

Cependant, cette approche est limitée au vocabulaire du corpus, et il n'est pas réaliste d'espérer la disponibilité de corpus alignés pour un large éventail de domaines dans le but d'accroître la couverture. Heureusement, tous les lexèmes du PWN ne sont pas polysémiques : parmi les 145 627 lexèmes du PWN, 82% n'apparaissent que dans un seul synset. Ces lexèmes ne nécessitent pas une technique aussi élaborée que celle décrite précédemment : ils peuvent être traduits directement à l'aide de ressources bilingues, construisant ainsi par correspondance directe de nouveaux synsets sans que des erreurs de sens ne soient à craindre.

3.2 Approche par alignement

Pour l'approche par alignement, nous avons utilisé le corpus SEE-ERA.NET[3], sous-corpus du corpus JRC-Acquis (Ralf *et al.*, 2006) aligné en phrases. Parmi les 8 langues du corpus, et mis

[3]Corpus créé au sein du projet SEE-ERA.NET ICT 10503 RP (2007-2008), *Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages*.

à part le français, nous avons utilisé les langues pour lesquelles nous disposions de toutes les ressources nécessaires : l'anglais, le roumain, le tchèque et le bulgare. Ce corpus fait environ 1 million et demi de mots, le chiffre exact variant d'une langue à l'autre. Lors des travaux décrits dans cet article, ce corpus n'avait pas encore été annoté morphosyntaxiquement dans le cadre du projet SEE-ERA.NET. Nous avons donc utilisé TreeTagger[4] pour étiqueter et lemmatiser le corpus. Les fichiers de paramètres par défaut ont été utilisés pour l'anglais et le français, alors que les fichiers pour le roumain et le bulgare ont été entraînés sur le corpus MULTEXT-East[5]. Pour le tchèque, nous avons utilisé l'analyseur morphologique FMorph pour annoter le corpus de façon ambiguë, puis l'étiqueteur Morče pour désambiguïser cette annotation (Hajič, 2008).

Le corpus a été aligné en mots par Uplug (Tiedemann, 2003), qui repose sur l'outil standard GIZA++. Pour augmenter la qualité de l'alignement, l'anglais n'a pas été utilisé comme langue pivot : nous avons aligné des paires de langues appartenant à la même famille (français-roumain et tchèque-bulgare), puis nous avons utilisé l'anglais comme pont (français-anglais et tchèque-anglais). Le résultat de l'alignement est un ensemble de liens entre mots qui, dans une phrase donnée, sont en relation de traduction. Chacun de ces liens est complété d'identifiants uniques (numéro de phrase, numéro de mots) et d'un indicateur du degré de confiance qui lui est attribué.

Les identifiants des mots ont permis de remplacer, dans ces liens, les formes fléchies par les lemmes tels que fournis par les lemmatiseurs. Afin de réduire le bruit dans les lexiques obtenus, seuls les liens entre lexèmes de même partie du discours ont été pris en compte, et toutes les entrées contenant des caractères autres que des lettres ont été éliminées. Les lexiques bilingues obtenus contiennent toutes les variantes de traduction d'un lexème source dans le corpus, assorties d'informations de fréquence, de partie du discours, ainsi que des identifiants des phrases et mots concernés. La taille de ces lexiques va de 43 024 entrées pour le lexique tchèque-anglais à 50 289 entrées pour le lexique tchèque-bulgare.

Les lexiques bilingues extraits ont été alors combinées pour créer cinq lexiques multilingues comportant de trois à cinq langues. Les lexèmes anglais et leurs identifiants de mots ont été utilisés comme pivots. La taille des lexiques multilingues va de 49 356 entrées pour le lexique regroupant toutes les langues (français-roumain-tchèque-bulgare-anglais) à 59 019 entrées pour le lexique français-tchèque-bulgare-anglais[6].

Une fois les lexiques obtenus, il reste à associer un synset à chacune de leurs entrées. Pour cela, nous recherchons pour chaque composant d'une entrée donnée, à l'exception du lexème en français, l'ensemble des identifiants des synsets auxquels il appartient. Nous avons utilisé pour cela le PWN 2.0 et les wordnets développés dans le cadre du projet BalkaNet (Tufiş, 2000), qui utilisent les mêmes identifiants de synsets que le PWN 2.0. On calcule alors, pour chaque entrée de lexique multilingue, l'intersection des ensembles d'identifiants de synsets associés aux différents composants de l'entrée. Si l'intersection est non vide, les synsets qu'elle contient sont attribués au lexème français de l'entrée. L'utilisation de plusieurs langues permet de désambiguïser les lexèmes polysémiques et élimine la plupart des erreurs d'alignement. Il est en effet peu probable qu'une même polysémie se retrouve dans de nombreuses langues différentes, ou qu'une erreur d'alignement induise une intersection non vide.

[4] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

[5] `http://nl.ijs.si/ME/`

[6] Nous utilisons une option du logiciel d'alignement Uplug qui ne produit que les liens les plus vraisemblables. C'est pour cette raison que le nombre d'entrées varie peu lorsque le nombre de langues augmente : l'augmentation à laquelle on pourrait s'attendre est compensée par la diminution du nombre d'entrées qui incluent un lexème pour chacune des langues.

frq	pos	Fr	Cs	Bg	En
18	n	droit	právo	законодателство	law
56	n	droit	právo	право	law
4	n	loi	právo	закон	law
4	n	loi	právo	законодателство	law
6	n	loi	právo	право	law
33	n	loi	zákon	закон	law
8	n	loi	zákon	закона *	law
19	n	législation	právo	законодателство	law
7	n	législation	právo	право	law
4	n	législation	předpis	законодателство	law

Fr : droit	Cs : právo	Bg : право	En : law
droit	06129345-n	04893549-n	00577416-n
	05559593-n	04888072-n	05529208-n
	05791721-n	07928837-n	05531141-n
	04617988-n	00577416-n	**05791721-n**
	07928837-n	**05791721-n**	06129345-n
		01000872-n	07712371-n
		04881053-n	07928837-n
		04617988-n	

TAB. 1 – Exemple d'entrées lexicales et de désambiguïsation pour le nom *droit* (le lexème bulgare identifié par une étoile résulte d'une erreur de lemmatisation ; la partie commune ENG20 a été omise dans les identifiants de synsets).

Illustrons ce processus sur un exemple. Soit le nom français *droit*, qui est polysémique (il peut ainsi être traduit en anglais, entre autres, par *right*, *law*, *droit*, *royalty*, *entitlement*, *claim*). Comme le montre la table 1(a), 56 de ses occurrences ont été alignées avec *právo* en tchèque, право en bulgare et *law* en anglais. L'intersection des ensembles d'identifiants de synsets contenant chacun de ces mots dans le wordnet correspondant ne contient qu'un seul identifiant de synset, ENG20-05791721-n. Il est donc attribué aux occurrences correspondantes du mot *droit* (cf. table 1(b)). Ce synset, défini dans le PWN comme définissant *the branch of philosophy concerned with the law and the principles that lead courts to make the decisions they do*, correspond bien à un des sens de *droit*.

Appliquée aux lexiques multilingues décrits précédemment, cette technique nous a permis de construire cinq ensembles différents de synsets comportant au moins un lexème français. Ils contiennent entre 1 338 (français-roumain-tchèque-bulgare-anglais) et 5 073 synsets (français-roumain-anglais). En raison des erreurs induites par les phases de pré-traitement (étiquetage, lemmatisation) et d'alignement, on s'attend à ce que ces synsets comportent certaines erreurs. Cependant, cette approche permet de traiter les lexèmes polysémiques, fréquents dans le vocabulaire de base, ce que l'approche par traduction, décrite ci-dessous, ne permet pas.

3.3 Approche par traduction

Wikipédia (`http://www.wikipedia.org`) est une encyclopédie collaborative et multilingue. Elle contient actuellement plus de deux millions d'articles en anglais, et plus de 600 000 en français. Les articles écrits dans différentes langues sont reliés entre eux à l'aide de liens inter-langues, qui permettent l'extraction d'informations multilingues. L'idée de faire correspondre des articles de Wikipédia aux synsets d'un wordnet n'est pas nouvelle, et a été utilisée à diverses fins (Declerck *et al.*, 2006; Ruiz-Casado *et al.*, 2005). Nous l'avons appliquée au PWN pour la création de nouveaux synsets dans le WOLF.

Comme dans toute encyclopédie, les titres des articles de Wikipédia sont principalement des noms (communs et propres), simples ou composés. Leur casse a été normalisée automatiquement à partir du corps des articles. Nous avons alors extrait un lexique bilingue à partir des titres anglais et de leurs équivalents français en suivant les liens inter-langues. Le lexique obtenu comporte 314 713 entrées. Nous l'avons enrichi par des synonymes et des définitions à l'aide de la première phrase de chaque article[7].

[7]Par exemple, l'article intitulé *Langue construite* débute par la phrase suivante : *Une **langue construite** ou*

Une autre ressource que nous avons utilisée pour extraire des équivalents de traduction pour les lexèmes monosémiques du PWN est Wiktionary (`http://www.wiktionary.org`). Complément lexical de Wikipédia, Wiktionary existe aussi pour différentes langues. Chaque Wiktionary contient pour chaque mot répertorié une définition ainsi que différentes informations complémentaires, telles que son étymologie, sa prononciation, des exemples, des synonymes et antonymes, et, plus important pour nous, des traductions dans d'autres langues. Les Wiktionaries anglais et français (« le Wiktionnaire », `http://www.wiktionary.org`) nous ont permis d'extraire deux lexiques bilingues de 24 464 et 24 873 entrées respectivement. Ces lexiques couvrent toutes les parties du discours, contrairement à Wikipédia.

Nous avons également exploité Wikispecies, taxonomie des espèces vivantes qui inclut à la fois les noms des espèces en latin, que l'on retrouve dans le PWN, et (pour les plus communes) en anglais ou en français. Ceci nous a permis d'identifier 129 509 termes latins indépendants de la langue ainsi qu'un équivalent français pour 2 648 d'entre eux.

Enfin, les descripteurs EUROVOC (`http://europa.eu/eurovoc`) font partie d'un thésaurus multilingue qui couvrent les champs d'activité de l'Union Européenne. Ils sont disponibles dans les 21 langues officielles de l'Union. La version 4.2 du thésaurus contient une liste de 6 802 descripteurs et de leurs équivalents dans les autres langues. En revanche, ils recouvrent une large variété de thèmes, et incluent de nombreuses expressions multi-mots.

Puisque nous avons utilisé ces ressources pour traduire les lexèmes monosémiques du PWN, l'identification du sens (c'est-à-dire de l'identifiant du synset) de leur équivalent français est triviale. Wikipédia a permis la construction de 18 721 synsets, Wikispecies de 6 848 synsets, le Wiktionnaire français de 6 215 synsets, le Wiktionary anglais de 4 363 synsets et les descripteurs EUROVOC de 1 319 synsets.

3.4 Fusion des résultats

L'ensemble des synsets obtenus par les deux approches ont été fusionnés. Si un même synset a été créé *via* plusieurs approches, ou au moyen de plusieurs sources (ensemble de langues pour l'approche par alignement, ressource lexicale pour l'approche par traduction), les ensembles de lexèmes sont unifiés. L'information sur les sources d'où provient chacun des lexèmes est conservée, pour permettre un filtrage en fonction de leur nombre de la fiabilité de chacune d'entre elles. En effet, les différentes sources de lexèmes ne donnent pas des résultats homogènes en termes de précision (cf. section 4). Si un lexème est associé à un identifiant de synset par tous les lexiques multilingues, et que la fréquence des entrées correspondantes dans ces lexiques est élevée, le résultat est fiable. Si à l'inverse un lexème n'est affecté à un synset que par un seul des lexiques multilingues (probablement par un lexique trilingue) et que l'entrée correspondante a une fréquence faible, alors le résultat est douteux. Nous avons donc mis en place un filtre pour éliminer les lexèmes les plus douteux, en nous fondant sur le nombre de sources, la nature des sources, et la fréquence des entrées correspondantes. Enfin, d'autres informations indépendantes de la langue (domaine, relations sémantiques) ont été héritées des synsets du PWN.

---

langue artificielle *(étymologiquement « faite par l'art ») est une langue créée par une ou plusieurs personnes dans un temps relativement bref, contrairement aux langues naturelles dont l'élaboration est largement inconsciente.* Ce texte, y compris les informations typographiques, nous on permis de considérer *langue construite*, mais également *langue artificielle*, comme des lexèmes associés au concept décrit par l'article, et d'extraire la définition correspondante, destinée à compléter le futur synset : *langue créée (…) largement inconsciente*.

La construction automatique de synsets conduit inévitablement à des trous dans la hiérarchie, un certain nombre de synsets n'ayant été attribués à aucun lexème français par manque de couverture des ressources de départ. En raison de l'importance des principes de densité conceptuelle et de préservation de la hiérarchie pour les applications des wordnets (Tufiş, 2000), nous avons récupéré les synsets manquants du PWN, en en éliminant les lexèmes anglais, mais en conservant toutes les relations structurelles et les autres informations indépendantes de la langue. Ces synsets vides seront remplis par des travaux futurs, mais l'intérêt est que l'ontologie obtenue est dense, en ce sens qu'elle ne contient aucun trou structurel. Si une application utilisant WOLF était amenée à rencontrer un de ces synsets vides, elle pourrait malgré tout utiliser les informations relationnelles pour accéder à un concept plus général ou plus spécifique.

4 Résultats et évaluation

4.1 Résultats

Nous avons étudié les caractéristiques de la ressource obtenue, le WOLF, par rapport au PWN et à l'EWN français. L'évaluation a été effectuée par rapport à l'EWN français[8].

Le WOLF contient actuellement 32 351 synsets non vides regroupant 38 001 lexèmes distincts. Ce nombre est bien plus élevé que le nombre de synsets de l'EWN français (22 121[9]), en raison des résultats de l'approche par traduction : rappelons que le PWN contient 115 424 synsets pour 145 627 lexèmes, dont 82% sont monosémiques, soit 119 528 lexèmes que nous avons pu essayer de traduire directement. Toutefois, ces synsets ne sont pas toujours les plus intéressants.

Pour évaluer la couverture du WOLF sur les concepts de base, nous nous sommes appuyés sur les BCS (*Basic Concept Sets*, Ensembles de Concepts de Base) définis par le projet BalkaNet (Tufiş, 2000). Les synsets de base sont répartis en trois niveaux de BCS : BCS1 rassemble les 1 218 concepts les plus importants, BCS2 regroupe 3 471 autres concepts importants, et BCS3 inventorie 3 827 autres concepts relativement importants. Tous les autres concepts sont hors BCS. Outre la couverture du WOLF, les résultats présentés à la table 2, ces chiffres valident *a posteriori* la définition de ces 3 BCS, les synsets BCS1 étant les mieux couverts (71,4%). Le détail par approche montre sans surprise que l'approche par alignement a permis la construction des synsets les plus fondamentaux, alors que l'approche par traduction, restreinte aux lexèmes monosémiques, a fait mieux pour BCS3 et hors BCS que pour BCS1 et BCS2.

Contrairement à WOLF, l'EWN français ne comporte ni adjectifs ni adverbes. Les résultats par partie du discours sont donnés à la table 2. La polysémie moyenne est de 1,23% synsets par lexème (10,5% des lexèmes sont polysémiques, dont 1,2% des lexèmes composés[10]). Ces

[8]L'EWN français n'est utilisé *que* pour l'évaluation. *Aucune* information issue de cette ressource n'a été utilisée dans le WOLF, ni pour y être intégrée, ni pour éliminer des lexèmes erronés du WOLF. Du reste, tous les scripts ayant servi à construire le WOLF sont librement disponibles sur son site. Les ressources d'entrée étant également librement disponibles, tout un chacun peut reconstruire le WOLF tel qu'il est aujourd'hui. Nous envisageons bien sûr un travail manuel de validation et de complétion dans l'avenir, mais la mise à disposition de fichiers décrivant chaque modification manuelle permettra de perpétuer la reproductibilité de la construction du WOLF.

[9]L'EWN français comporte en réalité 22 857 synsets, mais ils ne sont pas en correspondance biunivoque avec les synsets du PWN 2.0 (contrairement au WOLF ou aux wordnets du projet BalkaNet). Une fois mis en correspondance *via* un mapping entre PWN 1.5 et PWN 2.0, il arrive que plusieurs synsets d'EWN correspondent à un seul synset du PWN 2.0. Les chiffres que nous donnons sont pour l'EWN français *après* ce mapping.

[10]Les lexèmes composés sont tous obtenus par traduction de lexèmes anglais monosémiques. Mais parfois, deux

wordnet	PWN	WOLF		EWN français	
BCS1	1 218	870	71,4%	1 211	99,4%
BCS2	3 471	1 668	48,0%	3 022	87,1%
BCS3	3 827	1 801	47,1%	2 304	60,2%
hors BCS	106 908	28 012	26,2%	15 584	14,6%
synset nominal	79 689	25 559	35,8%	17 381	21,8%
synset verbal	13 508	1 544	11,5%	4 740	35,1%
synset adjectival	18 563	1 562	8,4%	0	0,0%
synset adverbial	3 664	676	18,4%	0	0,0%
total	*115,424*	***32,351***	***28.0%***	*22,121*	*19.2%*

TAB. 2 – Comparaison du nombre de synsets du WOLF et de l'EWN français par rapport au PWN, par BCS et par partie du discours.

	WOLF/align		WOLF/transl		WOLF	
	Précision	Rappel	Précision	Rappel	Précision	Rappel
noms	77,2%	68,7%	82,6%	74,9%	**80,4%**	**74,5%**
verbes	65,8%	54,7%	54,8%	35,8%	**63,2%**	**52,5%**
total	*74,6%*	*65,4%*	*78,8%*	*69,6%*	***77,1%***	***70,3%***

TAB. 3 – Évaluation par rapport à l'EWN français.

chiffres sont à comparer aux 1,74 synsets par lexème de l'EWN français (1,39 dans le PWN).

4.2 Évaluation par rapport à l'EWN français

Nous avons effectué deux types d'évaluation. Tout d'abord, nous avons comparé automatiquement, pour chaque lexème présent dans WOLF *et* dans l'EWN français, les synsets qui lui sont attribués. Cette évaluation n'a pu se faire que sur les noms et les verbes, seules parties du discours présentes dans l'EWN français. De plus, les synsets absents de l'EWN français mais présents dans WOLF sont inévitablement laissés de côté. Puis nous avons étudié manuellement quelques cas, pour identifier les sources de non-correspondance entre WOLF et l'EWN français.

Le tableau 3 présente les résultats de l'évaluation automatique. La précision est le pourcentage moyen de synsets attribués par WOLF à un lexème et que l'EWN français lui attribue également. Le rappel est le pourcentage moyen des synsets attribués à un lexème par l'EWN français et que WOLF lui attribue également. On constate que les verbes ont conduit à des résultats moins bons que les noms, en raison de leur forte polysémie[11].

Cependant, une précision inférieure à 100% pour un lexème donné ne signifie pas nécessairement une erreur : WOLF peut attribuer à juste titre un synset à un lexème, là où l'EWN français ne le fait pas. En revanche, un rappel inférieur a de bonnes chances d'être dû à une incomplétude du WOLF. C'est ce que montre l'évaluation manuelle que nous avons effectuée sur 100 lexèmes choisis aléatoirement pour lesquels le WOLF et l'EWN français sont en désaccord. Ils

lexèmes anglais monosémiques se traduisent en un même lexème, qui est donc polysémique. Exemple : *fatty liver* et *foie gras*, tous deux monosémiques en anglais, se traduisent en le terme français polysémique *foie gras*.

[11] Cette évaluation laisse de côté les synsets non vides du WOLF absents de l'EWN français. Pourtant, la majorité de ces synsets sont remplis grâce à Wikipédia, et sont donc de bonne qualité, comme le montrent la précision de 94,6% obtenue par cette seule source par rapport à l'EWN français et son rappel de 87,8%.

Catégorie	nom	verbe	adjectif	adverbe	all
dans l'EWN français	76 (88%)	33 (46%)	0 (0%)	0 (0%)	109 (60%)
hors de l'EWN français					
correct	16	18	*4*	*0*	38
sem. proche	10	6	*0*	0	17
sem. relié	2	6	*0*	0	7
morph. relié	2	*0*	0	*0*	2
non relié	5	5	*0*	*0*	10
total	111	68			183
total correct (précision réelle du WOLF)	**92 (83%)**	**51 (75%)**	***4***	***0***	**147 (80%)**

TAB. 4 – Évaluation manuelle du WOLF (les nombres en italique sont à prendre avec précaution compte tenu du nombre insuffisant de couples lexème-synset concernés).

correspondent à 183 couples lexème-synset. Nous avons vérifié, pour chacun de ceux qui ne sont pas présents dans les deux ressources, s'ils sont bel et bien corrects ou non. Les erreurs du WOLF ont été classées en quatre catégories :

– le lexème est sémantiquement proche du synset (hyperonyme, hyponyme, quasi-synonyme ; p.ex. *absence* dans le synset dont l'identifiant est celui de {*lack*, *deficiency*, *want*} du PWN),
– il est relié sémantiquement (autre relation sémantique ; p.ex. *abri* dans le synset dont l'identifiant est celui du synset {*penthouse*} du PWN),
– il est relié morphologiquement (il fait partie d'un composé qui aurait été correct dans le synset, ou il s'agit d'une variante morphologique d'un mot qui aurait été correct ; p.ex. *affaire* dans le synset correspondant à {*things*}, alors que le pluriel *affaires* serait correct ; *aisance* dans le synset correspondant à {*toilet*, *lavatory*, *lav*, *can*, *john*, *privy*, *bathroom*} alors que le composé *cabinet d'aisances* serait correct),
– il n'est pas relié du tout (ceci peut venir d'une erreur d'alignement ou de désambiguïsation, p.ex. *abattre* dans le synset correspondant à {*excavate*, *dig up*, *turn up*}).

Les résultats pour les différentes catégories sont donnés à la table 4. Environ la moitié des désaccords sont des couples synset-lexème qui manquent dans l'EWN français, et non des erreurs de WOLF. Sans surprise, les synsets les moins problématiques sont les concepts les plus spécifiques, et les plus difficiles sont ceux qui contiennent des lexèmes très polysémiques décrivant des concepts vagues. Pour une évaluation plus détaillée, y compris une évaluation source par source, on pourra se reporter à (Fišer et Sagot, 2008, soumis).

5 Conclusion

Nous avons présenté la méthodologie employée pour construire la première version d'un nouveau wordnet pour le français, le WOLF (Wordnet Libre du Français). Deux approches ont été utilisées de façon complémentaire : une approche par alignement de corpus multilingues, pour extraire des informations lexicales sémantiques pertinentes sur tous les types de lexèmes, y compris polysémiques ; une approche par traduction, utilisant des ressources telles que Wikipédia ou le Wiktionnaire, a permis la traduction directe d'un grand nombre de lexèmes polysémiques. Nous avons alors évalué le WOLF ainsi construit par rapport à l'EWN français.

Outre la validation et la correction manuelle (déjà en cours sur les BCS1), nous envisageons plusieurs pistes pour améliorer la méthodologie et la ressource construite. Naturellement, l'ap-

proche par alignement pourrait tirer parti de l'utilisation de corpus plus importants (JRC-Acquis dans son ensemble), bien que l'impact réel de la taille du corpus utilisé, lorsqu'il est aussi spécifique, n'est pas clair. Par ailleurs, nous avons déjà commencé des expériences préliminaires pour généraliser aux lexèmes polysémiques l'approche par traduction, par assignation automatique de synsets du PWN à des entrées de Wikipédia (Ruiz-Casado *et al.*, 2005). De plus, l'utilisation de telles ressources a un avantage au fil du temps : puisqu'elles sont en permanente évolution, utiliser notre méthodologie sur une future version de Wikipédia ou des Wiktionnaires devrait également permettre de construire de nouveaux synsets.

Enfin, nous souhaitons valoriser la ressource, notamment en désambiguïsation syntaxique et sémantique et en extraction et recherche d'informations. Outre une validation de l'intérêt de la ressource, nous attendons de ces travaux des retours permettant d'en améliorer la qualité.

Références

DECLERCK T., PÉREZ A. G., VELA O., GANTNER Z. & MANZANO-MACHO D. (2006). Multilingual lexical semantic resources for ontology translation. In *Proc. of LREC'06*, Gênes, Italie.

DIAB M. (2004). The feasibility of bootstrapping an arabic wordnet leveraging parallel corpora and an english wordnet. In *Proc. of the Arabic Lang. Tech. and Res.*, Le Caire, Égypte.

FELLBAUM C. (1998). *WordNet : An Electronic Lexical Database*. MIT Press.

FIŠER D. (2007). Leveraging parallel corpora and existing wordnets for automatic construction of the slovene wordnet. In *Proc. of L&TC'07*, Poznań, Pologne.

HAJIČ J. (2008). *Disambiguation of Rich Inflection – Computational Morphology of Czech*. Charles University Press – Karolinum. À paraître.

IDE N., ERJAVEC T. & TUFIŞ D. (2002). Sense discrimination with parallel corpora. In *Proc. of ACL'02 Workshop on Word Sense Disambiguation*, Philadelphia, PA, USA.

JACQUIN C., DESMONTILS E., & MONCEAUX L. (2007). French eurowordnet lexical database improvements. In *Proc. of CICLing'07 (LNCS 4394)*, Mexico City, Mexico.

PIANTA E., BENTIVOGLI L. & GIRARDI C. (2002). Multiwordnet : developing an aligned multilingual database. In *Proc. of the First Global WordNet Conference*, Mysore, Inde.

RALF S., POULIQUEN B., WIDIGER A., IGNAT C., ERJAVEC T., TUFIŞ D. & VARGA D. (2006). The JRC Acquis : A multilingual aligned parallel corpus with 20+ languages. In *Proc. of LREC'06*, Gênes, Italie.

RESNIK P. & YAROWSKY D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *ACL SIGLEX Workshop Tagging Text with Lexical Semantics : Why, What, and How ?*, Washington, D.C., États-Unis.

RUIZ-CASADO M., ALFONSECA E. & CASTELLS P. (2005). Automatic assignment of wikipedia encyclopedic entries to wordnet synsets. In *Proc. of Advances in Web Intelligence*.

TIEDEMANN J. (2003). Combining clues for word alignment. In *Proc. of EACL'03*, Budapest, Hongrie.

TUFIŞ D. (2000). Balkanet design and development of a multilingual balkan wordnet. *Romanian Journal of Information Science and Technology*, **7**(1–2).

VOSSEN, P. (1999). *EuroWordNet : a multilingual database with lexical semantic networks for European Languages*. Kluwer, Dordrecht.