MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Traduction de FrameNet par dictionnaires bilingues avec évaluation sur la paire anglais-français

Claire Mouton[1 2], Benoit Richert[1], Gaël de Chalendar[1]

1 : CEA LIST - 18, route du Panorama - BP6, FONTENAY AUX ROSES - F-92265 France.
2 : Exalead - 10 place de la Madeleine - 75008 Paris - France

Contact : {Claire.Mouton, Benoit.Richert, Gael.de-Chalendar}@cea.fr

Résumé
L'analyse sémantique de texte a pour but d'apporter aux machines de l'information leur permettant de traiter intelligemment du texte, au-delà des mots qui ne sont que des symboles. L'annotation sémantique d'un texte en rôles (*Semantic Role Labeling, SRL*) consiste à attribuer des rôles sémantiques aux différent syntagmes du texte. Ces rôles sont prédéfinis par des ressources sémantiques de référence décrivant des situations unitaires standards et les différents rôles qui peuvent y être associés. Les principales ressources de *Semantic Role Labeling* décrivant les situations de référence sont anglophones, et rares sont les ressources multilingues. Ce travail a pour but de transposer une ressource de SRL (FrameNet) dans une autre langue. Ici, la transposition se fait vers le français mais la méthode est applicable à la langue de son choix. Cette transposition a été réalisée à partir de l'extraction de paires de traduction de deux dictionnaires bilingues différents (le dictionnaire collaboratif multilingue Wiktionnaire et un dictionnaire standard français-anglais), puis par filtrage des paires obtenues. L'évaluation, réalisée sur la langue française, a montré l'obtention d'une ressource à forte précision contenant après filtrage autant d'entrées que la ressource d'origine en anglais.

Abstract
Words are only symbols to machines. Semantic analysis of text brings them the information they need to process text cleverly. Semantic Role Labeling (*SRL*) consists in affecting roles to syntactically parsed text. These predefined roles are part of standard unitary situations described in a semantic resource. Most of the Semantic Role Labeling references are in English, and multilingual resources are insufficiently developped. Our goal is to transfer an existing SRL resource (FrameNet) in another language. We have chosen in this work to translate it in French, but our method can be applied to any other language. This transfer has been performed by extracting translation pairs from two different bilingual dictionaries (Wiktionnary, a collaborative multilingual resource, and a standard French-English dictionary), and by filtering the extracted pairs. An evaluation has been carried out for French, showing the good quality of the obtained resource.

Mots-clés : FrameNet, Annotation Sémantique en Rôles, Wiktionnaire

Keywords: FrameNet, Semantic Role Labeling, Wiktionary

1. Introduction

L'annotation sémantique de texte a pour but de rendre un texte interprétable par une machine afin qu'elle puisse effectuer des traitements plus complexes tels que l'analyse d'une question en langage naturel et la recherche de ses réponses potentielles, l'extraction d'information formatée à partir de texte non structuré, le résumé de texte ou bien même du raisonnement. Pour parvenir à ces fins, une des approches consiste à annoter le texte en rôles (*Semantic Role Labeling, SRL*). Plusieurs ressources sémantiques ont ainsi été construites, s'essayant à décrire de façon plus ou moins exhaustive les rôles que peuvent jouer les différents élements d'un texte pour chaque mot considéré comme l'expression d'une action (verbe, nom, adjectif...). On citera entre autres :

– VerbNet [7] qui décrit un ensemble de rôles commun à tous les verbes, et attribue à chacun des verbes de son dictionnaire un sous-ensemble de ces rôles. VerbNet contient actuellement plus de 3 700 verbes ;
– Propbank [4] qui utilise une méthode assez semblable à celle de VerbNet mais où les rôles ne sont pas autant typés sémantiquement.
– FrameNet [1] définit un certain nombre de cadres appelés *Frames* décrivant chacun une situation précise susceptible d'apparaître dans notre perception du monde. Les rôles impliqués dans cette situation sont répertoriés, ainsi que les unités lexicales (*Lexical Unit, LU*) déclencheuses de cette situation, c'est-à-dire les prédicats (verbes, noms, adjectifs, ou mêmes adverbes) régisseurs de ces rôles. La base contient actuellement plus de 10 000 *LUs*.

Le *Semantic Role Labeling* consiste à attribuer les rôles sémantiques décrits dans ces ressources aux différents syntagmes distinguables dans les phrases du texte. Le tableau 1 présente les caractéristiques des trois ressources principales anglaises, ainsi qu'un exemple associé. Parmi ces ressources, les rôles sémantiques peuvent être peu nombreux et communs à toutes les situations (5 arguments principaux dans PropBank) ou nombreux et spécifiques à chaque situation (environ 250 rôles dans FrameNet).

Plusieurs méthodes ont été employées pour traduire des ressources SRL dans différentes langues.

Ressource	Rôles principaux	Exemples de rôles	Exemple de situation :	Exemple de phrase
VerbNet	21	agent, source, instrument...	keep-15.2 : agent, theme, location	He *[agent]* left *[keep-15.2]* the car *[theme]* in the park *[location]*
PropBank	5	Arg0, Arg1, Arg2...	leave.02 : Arg0, Arg1, Arg2, Argm-TMP	He *[Arg0]* left *[leave.02]* the car *[Arg1]* in the park *[Arg2]*
FrameNet	250	source, theme, buyer...	Departing : source, theme, place, circumstances, etc...	He *[theme]* left *[Departing]* the car *[source]* in the park *[place]*

TAB. 1 – Comparaison des différents rôles sémantiques

La méthode la plus employée est la projection de rôles sémantiques de l'anglais vers la langue cible. Elle nécessite la traduction d'un corpus anglais déjà annoté, la reconnaissance des rôles de la langue cible. Elle permet donc l'obtention d'une traduction de la ressource et d'un corpus, pouvant ensuite servir à un apprentissage supervisé du moteur d'annotation. Sur la base de la méthode de S. Padó et M. Lapata [5], des traductions ont ainsi été obtenues pour l'allemand, l'italien [8] et le français [6]. Dans la langue française, les ressources utilisables dans le cadre d'une méthode *SRL* sont plutôt rares. On peut citer Volem [2], une base de 1 500 verbes constituée manuellement, ainsi que ce que nous nommerons par la suite FrameNet.Fr, une traduction automatique de FrameNet réalisée par [6]. Volem a son propre paradigme, se rapprochant de celui de VerbNet. FrameNet.Fr est une première approche de la transposition de FrameNet, exploitant l'alignement de corpus bilingues parallèles pour générer les unités lexicales françaises à partir de leur alignement avec les unités lexicales anglaises.

Nous nous proposons ici d'adopter une seconde approche, fondée sur l'obtention de paires de traduction dans des dictionnaires existants, et leur filtrage à partir des informations de polysémie. Dans [6], Padó et G. Pitel comparent leur ressource obtenue à l'aide de corpora parallèles avec une traduction obtenue par dictionnaire bilingue. Cette dernière s'avère de moindre qualité. Nous reprenons cette idée mais en affectant un score de confiance aux paires de traduction, ce qui permet de ne conserver que les plus pertinentes. Notre but est de pouvoir traduire la ressource FrameNet

..., sip.n, devour.v, feed.v, sip.v, swig.n, gobble.v, **quaff.v**, breakfast.v, slurp.n, ingest.v, **drink.v**, put away.v, down.v, slurp.v, eat.v, ...

FIG. 1 – Échantillon de la liste des unités lexicales du cadre *Ingestion* de FrameNet

dans n'importe quelle langue, autant dans l'étape de l'extraction des paires de traduction (à partir notamment de la ressource libre Wiktionnaire, comportant plus de 700 langues) que dans l'étape de filtrage. Nous avons généré des traductions de FrameNet dans différentes langues, mais seule la traduction vers le français a été évaluée.

La section 2 décrit notre procédé d'extraction de paires de traduction. Nous présentons dans la section 3 les heuristiques de filtrage que nous utilisons ainsi que l'évaluation de leurs résultats dans la section 4. Enfin la section 5 dresse le bilan de ces travaux et propose les perspectives à venir.

2. Extraction de paires de traduction à partir de ressources lexicographiques bilingues

Le paradigme de FrameNet est fondé sur la théorie des cadres sémantiques de CJ Fillmore [3]. Un cadre possède un sens assez général, il comprend des mots déclencheurs et un ensemble de rôles sémantiques ; il est utilisé lorsque le système de *SRL* repère dans une phrase l'une des unités lexicales déclencheuses de celui-ci. Les unités lexicales sont l'association d'un mot et de sa catégorie grammaticale (par exemple *boire.v* pour le verbe *boire*, *boisson.n* pour le nom *boisson*). Le tableau 2 donne les rôles du cadre *Ingestion*, tandis que la figure 1 présente un extrait des unités lexicales qui déclenchent ce cadre. Ce sont ces associations mot-catégorie grammaticale que nous voulons transposer dans une autre langue, car les cadres et les rôles sont sémantiques et donc considérés en première approximation comme indépendants de la langue utilisée.

Type de rôle	rôles
Core	Ingestor, Ingestibles
Non-Core	Degree, Duration, Instrument, Manner, Means, Place, Purpose, Source, Time

TAB. 2 – Les rôles du cadre *Ingestion*

Nous avons utilisé deux dictionnaires bilingues différents afin de traduire les unités lexicales de FrameNet. Le premier est le Wiktionnaire, dictionnaire créé bénévolement à la manière de l'encyclopédie Wikipedia et avec la même plate-forme. Le second est *SCI-FRAN-EURADIC*, un dictionnaire réalisé et validé par des linguistes dans le cadre du projet EuRADic.

2.1. Extraction de paires de traductions à partir du Wiktionnaire

Le Wiktionnaire[1] est un ensemble de dictionnaires multilingues collaboratifs. Il existe un dictionnaire multilingue par langue d'édition. dont les standards laissent les contributeurs très libres sur la façon d'y écrire les données. De plus, ces standards changent selon la langue originale de chaque dictionnaire. Les entrées de chaque dictionnaire peuvent être multiples. Un mot de la langue originale du dictionnaire peut constituer une entrée et être traduit dans sa section traduction. Mais un mot d'une autre langue peut aussi se trouver en entrée, et être alors traduit dans la langue originale. Ainsi, en utilisant les données du dictionnaire anglais et du dictionnaire de la langue dans laquelle l'on veut traduire FrameNet, on réalise l'extraction à partir de quatre informations possibles, comme illustré figure 3. Nous appellerons désormais ces quatres sources d'information *source Wiktionnaire*. D'autre part, les entrée du dictionnaire répertorient souvent les traductions selon les différents sens du mot source. Nous nous sommes concentrés sur l'extrac-

[1] `http://fr.wiktionary.org`

Langue du Wiktionnaire	Entrée Wiktionnaire	Langue du mot source	Langue du mot cible	Exemple de syntaxe Wiktionnaire
français	boire	français	anglais	`* {{T|en}} : to {{trad+|en|drink}}`
français	drink	anglais	français	`# [[boire|Boire]]`
anglais	boire	français	anglais	`# to [[drink]]`
anglais	drink	anglais	français	`* French: {{t+|fr|boire}}`

TAB. 3 – Quatre possibilités d'extraire des paires de traduction dans les Wiktionnaires

tion des traductions et des catégories grammaticales. Cette extraction inclut une petite partie de bruit dû à l'aspect collaboratif de la ressource, comme des erreurs de traduction ou l'emploi par un contributeur d'une syntaxe wiktionnaire non standard. Ce bruit semble ne représenter qu'une très faible quantité d'unités lexicales mais nous ne l'avons pas évalué. D'autre part, l'analyseur n'a pas été capable de reconnaître toutes les paires de traduction, notamment parmi les entrées des langues étrangères au dictionnaire extrait où la syntaxe wiktionnaire n'est pas standardisé et moins souvent employée.
Pour la traduction de l'anglais au français, en n'utilisant que des unités lexicales anglaises présentes dans FrameNet, nous obtenons 19 912 *LUs* françaises issues de 27 109 paires de traduction.

Il est déjà possible avec l'extraction du Wiktionnaire anglais d'obtenir des paires de traduction anglais-langue cible avec de nombreuses langues. Cependant dans la pratique, le plus grand nombre de paires a été trouvé dans le Wiktionnaire de la langue cible. Par exemple, 67% des paires anglais - français ont été extraites à partir du Wiktionnaire français. Adapter l'analyseur à la syntaxe du Wiktionnaire de la langue cible est donc une étape importante pour la quantité de résultats, mais qui n'est pas indispensable dans un premier temps. À partir du seul Wiktionnaire anglais, nous avons par exemple extrait 8 220 paires anglais - italien, soit 7 749 unités lexicales dans la ressource FrameNet italienne constituée.

2.2. Extraction de paires de traductions à partir de la ressource EuRADic

Nous utilisons aussi un dictionnaire bilingue plus standard, issu du projet EuRADic[2], afin de comparer les apports respectifs des dictionnaires manuels et collaboratifs. Celui-ci contient 243 539 paires français-anglais avec leurs catégories grammaticales. On peut remarquer une grosse proportion de locutions par rapport aux termes simples.
L'extraction des données de ce dictionnaire en vue de leur traitement était triviale, nous ne la développerons pas. Nous avons obtenu 80 666 paires anglais-français, aboutissant à 57 787 unités lexicales françaises.

3. Filtrage

Les *LUs* obtenues dans la langue cible ne sont pas parfaites pour diverses raisons : polysémie du mot anglais qui génèrera une mauvaise affectation, erreurs dans la ressource... Pour tenter de ne conserver que les meilleures, nous leur attribuons des scores. On ne conservera par la suite que les *LUs* dont les scores sont supérieurs à un certain seuil.

3.1. Les filtres

3.1.1. S1 : score initial

Le score *S1* d'une *LU* cible correspond au nombre de sens des *LUs* anglaises de ce cadre qui ont la *LU* cible pour traduction. On compte donc le nombre de paires *LU*_source_#sens *LU*_cible que l'on a extraites des ressources (ou l'inverse, selon la source Wiktionnaire utilisé) : e.g. la paire *boire.v - drink.v* est dotée du score 2, car *boire* est une traduction de *drink* pour les sens *consume liquid through the mouth* et *consume alcoholic beverages*. Au final, *boire.v* se trouve dans le cadre *Ingestion*

[2] `http://www.technolangue.net/article.php3?id_article=203`

...
remettre.v {put back.v : 1} 1.0
ingérer.v {ingest.v : 1} 1.0
barboter.v {lap.v : 1} 1.0
absorber.v {sip.v : 1} 1.0
s'alimenter.v {eat.v : 1} 1.0
clapoter.v {lap.v : 1} 1.0
boire.v {quaff.v : 1 , drink.v : 2} 3.0
alimenter.v {feed.v : 1} 1.0
banqueter.v {feast.v : 1} 1.0
mâchonner.v {munch.v : 1} 1.0
déjeuner.v {lunch.v : 1 , dine.v : 1 , feed.v : 1 , eat.v : 1} 4.0
...

FIG. 2 – Échantillon du cadre *Ingestion* extrait du Wiktionnaire

avec le score 3, composé du score de 2 de la paire *boire.v - drink.v* et de 1 de la paire *boire.v - quaff.v*, *drink.v* et *quaff.v* étant deux unités lexicales du cadre *Ingestion*.

Quand une paire de traduction apparaît dans plusieurs sources Wiktionnaire différentes, nous prenons le score maximum produit par cette paire dans une des sources, considérant que cela correspond le mieux au nombre de sens possibles. Par exemple, *boire.v* est une traduction de *drink.v* avec un score de 2 dans le Wiktionnaire anglais, car cette paire illustre les deux sens *consume liquid through the mouth* et *consume alcoholic beverages*. Mais elle a un score de 1 dans les trois autres façons de l'extraire de la ressource, car les deux sens n'ont pas été ainsi séparés. Son score total sera donc le maximum, soit 2.

3.1.2. S2 : La division par le nombre d'unités lexicales

Nous considérons que moins un cadre comporte d'unités lexicales (*LU*), plus une unité lexicale individuelle est importante pour ce cadre. Ce filtre vise à augmenter l'exigence requise (comparé au filtre sur le score *S1*) pour les unités lexicales présentes dans les cadres les plus grands. En conséquence, après utilisation de ce filtre la distribution du nombre d'unités lexicales dans les cadres aura un écart-type plus faible.

$$S2 = \frac{S1}{\mathrm{Nombre\_de\_LUs\_dans\_le\_cadre\_cible}^{\alpha}}$$

Le score de chaque unité lexicale traduite est divisé par le nombre d'unités lexicales traduites dans le cadre considéré. Ce nombre est élevé à une puissance α, qui permet de moduler l'impact du filtre sur le score. Notre exemple applique ce filtre avec $\alpha = 1$ sur l'unité lexicale *boire.v* de *Ingestion*, cadre possédant 47 unités lexicales traduites. Le score initial de *boire.v* (le nombre de paires de cette LU dans ce cadre) est $S1 = 3$ (voir figure 2 pour plus de détails). Son score S2 est donc $S2 = \frac{3}{47} = 0,064$.

3.1.3. S3 : La division par la polysémie

Plus une unité lexicale est présente dans beaucoup de cadres, moins elle a du sens dans un cadre donné. Ainsi, ce filtre augmente l'exigence requise pour les unités lexicales possédant aussi d'autres sens n'appartenant pas au cadre considéré. En conséquence, dans la ressource finale, la présence d'unités lexicales initialement dans plusieurs cadres est diminuée.

$$S3 = \frac{S1}{\mathrm{Nb\_de\_cadres\_contenant\_la\_LU}^{\alpha}}$$

Le score de chaque unité lexicale traduite est divisé par sa polysémie. Celle-ci se traduit par le score de cette unité lexicale tous cadres confondus. De la même façon que le score précédent, un coefficient α vient régler la force du filtre. Nous prenons ici comme exemple, l'unité lexicale *rue.n*, qui a un score $S1 = 1$ dans le cadre *Roadways* (traduite de *street*) et un score $S1 = 1$ dans le cadre *Measure_linear_extent* (traduite de *block*). On a donc dans le cadre *Roadways* $S3 = \frac{1}{2} = 0,5$.

4. Évaluation

4.1. Méthode d'évaluation

Pour l'évaluer les ressources obtenues, nous avons sélectionnés 10 cadres de façon à ce que leur nombre d'unités lexicales oit représentatif de l'ensemble (quantiles). Ces 10 cadres ont été corrigés pour chaque ressource initiale, en supprimant manuellement les unités lexicales jugées incorrectes.

Les indicateurs qui nous intéressent sont la $Couverture$, la $Précision$, la $Fc-mesure$ et la $Fc_{0,5}-mesure$. La $Couverture$ représente l'exhaustivité de la ressource, la $Précision$ sa qualité, les $Fc-mesures$ sont les indicateurs à maximiser, moyennes harmoniques pondérées de la $Précision$ et de la $Couverture$. Le poids de R et de P sont identiques pour Fc, la précision est favorisée pour $Fc_{0,5}$. Les formules de ces indicateurs peuvent être retrouvées dans le tableau 4. Chaque

$$C = \frac{Nombre\_de\_LUs\_correctes\_présentes}{Nombre\_de\_LUs\_correctes\_avant\_filtrage} \qquad P = \frac{Nombre\_de\_LUs\_correctes\_présentes}{Nombre\_de\_LUs\_présentes}$$

$$Fc_\beta = (1+\beta^2)\frac{PC}{\beta^2 P + C} \qquad (\, Fc = Fc_1 \,)$$

TAB. 4 – Indicateurs de la qualité du filtrage et des ressources

mesure rapportée ici correspond à la moyenne de la mesure correspondante calculée sur chacun des cadres d'évaluation. Cela permet d'accorder une importance identique à chaque cadre indépendamment de sa taille.

Une évaluation de la ressource est possible en précision mais pas en rappel, puisque l'on ne peut pas déterminer exhaustivement toutes les unités lexicales possibles dans un cadre (exemple significatif, un nombre infini de mots doit être présent dans le cadre *Ordinal_Numbers*). Nous proposons donc de calculer une mesure de couverture en considérant la ressource corrigée non filtrée comme vérité terrain. La couverture des ressources non filtrées vaudra donc toujours 1, car elles sont évaluées par rapport à leur propre correction, sans ajout de nouvelles unités lexicales. Les couvertures (ainsi que les mesures *Fc*) des différentes ressources ne sont donc pas comparables, il s'agit uniquement de quantifier la perte engendrée par le filtrage de chacune des ressources. Quant à la précision, elle évalue la qualité de la ressource obtenue. Son augmentation par rapport à la précision de la ressource non filtrée évalue la qualité du filtre.

4.2. Les résultats des filtres sur les ressources

Les filtres peuvent être combinés afin de maximiser l'effet de chacun. Nous avons choisi de calculer les combinaisons linéaires des scores S1, S2 et S3. Pour cela, les scores filtrés sont normalisés (écarts-type à 1, moyennes égales et de telle manière que le score le plus petit soit nul). Ainsi, les scores deviennent comparables. Nous faisons ensuite varier les paramètres α, les coefficients de pondération des filtres ainsi qu'un seuil au dessous duquel les unités lexicales sont éliminées car considérées non fiables.

Le but est d'obtenir deux types de ressources : une ressource équilibrée (de grande taille et conservant une précision raisonnable) et une ressource robuste (à très forte précision). Nous nous sommes fondés sur la maximisation de la $Fc_{0,5}-mesure$ pour obtenir la ressource la plus équilibrée, car nous avons privilégié la précision, au vu de la quantité d'unités lexicales présentes sans filtrage, assez importante par rapport au FrameNet original. Pour l'obtention de la ressource robuste, nous avons fait varier les différents paramètres pour obtenir une précision souhaitée de 0,95 à 0,015 près, tout en maximisant la couverture.

Pour les deux types de ressources, c'est une combinaison des scores S2 et S3, avec différents coefficients α et coefficients de pondération, qui nous a permis de produire les meilleurs résultats. Chacun de ces deux filtres contribue donc de façon significative au filtrage des ressources.

Nous avons testé les filtres sur la ressource FrameNet.Fr, qui comportait ses propres scores. Mais la ressource non filtrée donnait toujours de meilleurs résultats, les filtres n'étant pas adaptés à cette distribution de scores.

Ressource	Nombre de LUs (% / FrameNet)	C	P	F	$Fc_{0,5}$
Wikt	19 912 (178%)	1	0,634	0,763	0,678
Wikt $Fc_{0,5} - max$	13 815 (124%)	0,784	0,775	0,759	0,763
Wikt $P = 0,95$	3 476 (31%)	0,350	0,940	0,475	0,649
EuRADic	57 787 (517%)	1	0,509	0,668	0,562
EuRADic $Fc_{0,5} - max$	22 703 (203%)	0,616	0,678	0,632	0,655
EuRADic $P = 0,95$	2 340 (21%)	0,112	0,945	0,178	0,292
Wikt∪EuRADic	65 488 (613%)	1	0,478	0,640	0,531
Wikt∪EuRADic $Fc_{0,5} - max$	26 585 (238%)	0,637	0,666	0,635	0,649
Wikt∪EuRADic $P = 0,95$	4 645 (42%)	0,182	0,937	0,293	0,476
FrameNet.Fr	6 659 (60%)	1	0,611	0,721	0,647

TAB. 5 – Évaluation des filtres

4.3. Analyse des résultats

Les résultats rapportés dans le tableau 5 sont calculés sur les 10 cadres à partir desquels nous optimisons les paramètres, ils ne sont pas validés sur un corpus de test. On peut remarquer que la ressource EuRADic génère un FrameNet français trois fois plus important que le Wiktionnaire. Cela est dû non seulement au fait que ce soit une ressource professionnelle, plus exhaustive, mais aussi à son grand nombre d'expressions idiomatiques (23% contre 4% pour le Wiktionnaire).

Les résultats de l'évaluation des ressources basées sur le Wiktionnaire sont meilleurs en précision que ceux des ressources issues d'EuRADic : on extrait un cœur robuste de 3 476 LUs du Wiktionnaire, contre 2 340 d'EuRADic, alors que la ressource non filtrée d'EuRADic est trois fois plus grande. Pourtant, le nombre de paires de traduction par unité lexicale française (S1) est assez proche : 1,40 pour EuRADic (soit 40% de paires venant enrichir les mot déjà présents) contre 1,36 pour le Wiktionnaire (36% d'enrichissement). En fait, ce score est composé différemment dans les deux types de ressources. Avec EuRADic, ce score par *LU* est composé de 1,34 *LUs* anglaises par *LU* française, chaque *LU* anglaise ayant un sous-score moyen de 1,04. Autrement dit, une paire de traduction est pratiquement unique. À l'inverse avec le Wiktionnaire, on obtient seulement 1,22 *LUs* anglaises par *LU* française, mais avec un sous-score moyen par *LU* anglaise de 1,11. Autrement dit, il n'est pas rare qu'une même paire de traduction apparaisse plusieurs fois. Cela est dû à la différence de constitution des deux dictionnaires : dans le wiktionnaire, chaque mot a plusieurs sens. Chaque sens peut avoir une traduction, parfois similaire. On se retrouve donc avec un nombre de paires de traduction équivalant idéalement au nombre de sens communs entre la *LU* française et la *LU* anglaise. C'est le cas de la paire *boire.v - drink.v*, dont le sous-score est 2, car elle apparaît pour les sens *consume liquid through the mouth* et *consume alcoholic beverages*, alors qu'il est de 1 dans EuRADic. En pratique, on se rend compte que souvent seuls les sens les plus courants ont été traduits. Cela donne deux raisons au Wiktionnaire de donner une meilleure ressource :

– Le Wiktionnaire répertorie les sens des mots. Il permet ainsi une meilleure prise en compte de la polysémie et donc un meilleur filtrage.
– Les sens principaux y sont plus souvent traduits que les sens secondaires. Or les mots dans FrameNet sont souvent utilisés avec leur sens principal.

5. Bilan et Perspectives

Nous avons proposé une nouvelle approche de transfert de la ressource FrameNet à une autre langue que l'anglais et validé cette approche pour la langue française.
Ainsi pour chaque dictionnaire nous obtenons deux types de ressources. L'une est robuste (environ 95% de précision) mais de taille inférieure à la ressource anglaise de départ (31% du nombre de *LUs* de FrameNet pour le Wiktionnaire, 21% pour EuRADic). L'autre est équilibrée ($Fc_{0,5}$ = 76% pour le Wiktionnaire, $Fc_{0,5}$ = 66% pour EuRADic), à la fois plus grande et plus précise que la traduction déjà existante [6], mais aussi plus grande que FrameNet (124% du nombre de *LUs* de FrameNet pour le Wiktionnaire, 203% pour EuRADic).
Ces résultats montrent par ailleurs que la ressource anglaise d'origine n'est pas exhaustive. En effet certaines unités lexicales devraient y figurer mais n'appartiennent à aucun cadre (e.g. *taxonomy.n* devrait figurer parmi les unités lexicales de *Categorization*) tandis que d'autres y figurent mais n'appartiennent pas à tous les cadres qu'elles devraient déclencher (e.g. *boom.n* n'apparait qu'en tant que *Sounds* alors qu'on aimerait le voir aussi dans *Progress*). Leurs traductions ne figurent donc généralement pas dans les cadres correspondants non plus.
Nos futurs travaux s'orientent vers l'enrichissement de la ressource française par une classification de nouvelles unités lexicales. Cette classification pourra exploiter un plus grand nombre d'unités lexicales d'entraînement qu'à partir de la ressource anglaise et permettra en outre de valider certaines affectations obtenues par traduction.
Les résultats obtenus lors de l'évaluation montrent que nos scores traduisent bien la confiance que l'on peut avoir dans une affectation *LU française - Cadre*. Nous pouvons maintenant nous attaquer à la tâche d'annotation pour laquelle nous effectuerons de façon semblable une traduction des têtes des syntagmes annotés en Rôles, ces traductions disposant alors d'un score de confiance que l'on exploitera dans notre algorithme d'annotation.

Bibliographie

1. Collin F. Baker, Charles J. Fillmore, et John B. Lowe. The berkeley framenet project.
2. Ana Fernandez, Gloria Vazquez, Patrick Saint-Dizier, Farah Benamara, et Mouna Kamel. The VOLEM project : a framework for the construction of advanced multilingual lexicons. In *Language Engineering Conference, 2002. Proceedings*, pages 89–98, 2002.
3. Charles J. Fillmore. Frame semantics. *Cognitive Linguistics : Basic Readings*, pages 185–238, 2006.
4. Paul Kingsbury et Martha Palmer. From treebank to propbank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 1989–1993, 2002.
5. Sebastian Padó et Mirella Lapata. Cross-lingual bootstrapping for semantic lexicons : The case of framenet. In *Proceedings of AAAI-05*, pages 1087–1092, Pittsburgh, PA, 2005.
6. Sebastian Padó et Guillaume Pitel. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Actes de la 14e conférence sur le Traitement Automatique des Langues Naturelles (communications orales)*, page 271, 2007.
7. Karin Kipper Schuler. VerbNet : A broad-coverage, comprehensive verb lexicon. *Univ. of Pennsylvania-Electronic Dissertations*, 2005.
8. S. Tonelli et E. Pianta. Frame information transfer from English to Italian. *Proceedings of LREC-2008*.

Nous tenons ici à remercier Sebastian Padó et Guillaume Pitel pour nous avoir fourni leur transposition de FrameNet. Ces travaux ont été financés en partie par le projet PASSAGE.