Constitution d'une ressource sémantique arabe à partir de corpus multilingue aligné

Authoul Abdul Hay[1] Olivier Kraif[2]
(1) Alzaytoonah University of Jordan - 11733 Jordan
(2) Univ. Grenoble Alpes, LIDILEM, F-38040 Grenoble
authoul@voila.fr, olivier.Kraif@u-grenoble3.fr

RÉSUMÉ

Cet article porte sur la mise en œuvre etsur l'étudede techniques d'extraction de relations sémantiques à partir d'un corpus multilingue aligné, en vue de construire une ressource lexicale pour l'arabe. Ces relations sontextraites par transitivité de l'équivalence traductionnelle, deux lexèmes qui possèdent les mêmes équivalents dans une langue cible étant susceptibles de partager un même sens. A partir d'équivalences extraites d'un corpus multilingue aligné, nous tâchons d'extraire des "cliques", ou sous-graphes maximaux complets connexes, dont toutes les unités sont en interrelation, du fait d'une probable intersection sémantique. Ces cliques présentent l'intérêt de renseigner à la fois sur la synonymie et la polysémie des unités, et d'apporter une forme de désambiguïsation sémantique. Ensuite nous tâchons de relier ces cliques avec un lexique sémantique (de type Wordnet) afin d'évaluer la possibilité de récupérer pour les unités arabes des relations sémantiques définies pour des unités en d'autres langues (français, anglais ou espagnol). Les résultats sont encourageants, et montrent qu'avec des corpus adaptés ces relations pourraient permettrede construire automatiquement un réseau utile pour certaines applications de traitement de la langue arabe.

ABSTRACT

The constitution of an Arabic semantic resource from a multilingual aligned corpus

This paper aims at the implementation and evaluation of techniques for extracting semantic relations from a multilingual alignedcorpus, in order to build a lexical resource for Arabic language. We first extract translational equivalents froma multilingual aligned corpus. From these equivalences, we try to extract "*cliques*", which are maximum complete related sub-graphs, where all units are interrelated because of a probable semantic intersection. These cliques have the advantage of giving information on both the synonymy and polysemy of units, providing a kindof semantic disambiguation. Secondly, we attempt to link these cliques with a semantic lexicon (like WordNet) in order to assess the possibility of recovering, for the Arabicunits, a semantic relationships already defined for English, French or Spanish units. These relations would automatically build a semantic resource which would be useful for different applications of NLP, such as Question Answering systems, Machine Translation, alignment systems, Information Retrieval…etc.

MOTS-CLÉS : Corpus multilingues alignés, désambigüisation sémantique, cliques, lexiques multilingues, réseaux sémantiques, traitement de l'arabe.
KEYWORDS :Multilingual aligned corpus, semantic disambiguation, cliques, multilingual lexicons, word net, Arabic Language Processing

1 Introduction

Ce travail vise à étudier une méthode pour la constitution d'une ressource sémantique arabe qui pourrait refléter la richesse de la langue arabe, et son importance en termes de diffusion et de nombre de locuteurs (≈300 millions).

Pour un large éventail d'applications s'appuyant sur l'élaboration d'une ressource sémantique, le réseau sémantique WordNet(Fellbaum, 1998) de l'université de Princetonest devenu un standard *de facto*, malgré certaines limites et imperfections qu'on peut lui reprocher, tels que ses incohérences, la confusion entre sens et concept oul'inadéquation de son organisation des sens à d'autres langues que l'anglais (Mallak, 2011).

Bien avant WordNet, les réseaux sémantiques ont été très utilisés dans le domaine de l'intelligence artificielle, et notamment pour le TAL, depuis les années 1960. Ils montrent une bonne adaptation à la représentation du langage naturel, et la capacité de modéliser toute forme de connaissances que l'on peut représenter dans un système symbolique (Hendrix, 1979). Mais la création d'un réseau sémantique en fonction d'objectifs spécifiques est une opération complexe et coûteuse à mettre en œuvre. C'est pour cela qu'il devient primordial, pour une langue donnée, de bénéficier de réseaux sémantiques génériques déjà développés, afin de rattraper l'écart technologique en termes de contenus, de services et d'usages entre les langues et les cultures du monde sur les réseaux d'information.

Dans la perspective de développer un tel réseau sémantique pour la langue arabe, nous présentons dans cet article une méthode visant à tirer parti de réseaux déjà existants pour d'autres langues, en s'appuyant sur l'extraction préalable de relations d'équivalences lexicales à partir d'un corpus multilingue aligné.

Après une brève description des travaux antérieurs dans le domaine, ci-dessous, la partie 3 de notre articledétaille la méthodologie suivie dans notre travail d'expérimentation. Dans lapartie 4, nous faisons état des résultats obtenus et en donnons une évaluationqualitative et quantitative (sur un petit échantillon). La dernière section enfin présente les conclusions et les perspectives envisagées.

2 Travaux antérieurs

Peu de travaux, à notre connaissance, ont contribué à l'élaboration d'un wordnet pour l'arabe. Parmi ces travaux, nous citerons la contribution la plus importante, qui est celle de (Alkateb et al., 2006).Il faut noter qu'AWN (pour ArabWord Net), la ressource proposée par ces auteurs, est une des rares ressources pour la langue générale arabe consultable en ligne. Les auteurs ont élaboré un wordnet basé sur l'architecture etle contenu du Princeton WordNet (PWN 2.0) et qui peut être relié directement avec son extension multilingue EuroWordNet (EWN, Vossen, 1998). Dans cette architecture, les wordnets des différentes langues sont indexés par des « ILI » (des relations d'équivalence pointant vers des entrées de PWN) et l'ontologie SUMO (*Suggested Upper Merged Ontology,* Niles et Pease, 2001),une ontologie supérieure formelle quicontient 1 000 termes et 4 000 formules définitionnelles exprimées dans la logique du 1er ordre.Dans la construction d'AWN, les auteurs ont suivi la méthode élaborée pour EWN, avec une approche descendante : ils sont partis des concepts communs de base (partagés par les langues d'EWN et de Balkanet), qu'ils ont étendus vers des concepts plus spécifiques, en suivant les relations d'hyponymie. Les correspondances

avec les synsets[1]anglais sont obtenus grâce à un lexique bilingue, et en suivant différentes heuristiques, de l'arabe vers l'anglais ou réciproquement. Des candidats sont proposés automatiquement, mais la validation des correspondances reste une étape manuelle.La possibilité d'automatiser ce processus de validation permettrait de diminuer de façon notable le cout d'une telle ressource, afin d'en augmenter la couverture.

Dans la perspective d'une plus grande automatisation, Sagot et Fišer (2008)proposent une méthode intéressante faisant intervenir des ressources multilingues (des corpus alignés) afin de construire un réseau sémantique de type wordnet pour le français(le WOLF). Les auteurs appliquent une approche par extension (Vossen, 2008), en partant de PWN pour en traduire les synsets. Pour les mots monosémiques, la traduction est triviale, via des lexiques bilingues (tirés de Wikipedia etdu thésaurusEUROVOC20). Pour les entrées polysémiques (les mots qui appartiennent à plusieurs synsets), ils utilisent le corpus parallèle CCR-Acquis19comportant 5 langues alignées. Ils se basent sur l'idée suivante : *"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. "* (Sagot et Fišer 2008 :3).Ainsi chaque mot simple français se retrouve aligné avec des équivalents en anglais, roumain, tchèque et bulgare. Chacun de ces équivalents est alors rattaché à un ou plusieurs ILI dans EWN et BalkaNet. Les auteurs font alors l'hypothèse que l'intersection de ces ILI indique probablement un ou plusieurs sens rattachables au mot français. Avec cette technique, les auteurs obtiennent respectivement pour les noms et les verbes des précisionsde 77,2% et 65,8%, et desrappels de 68,7% et 54,7% (en prenant le wordnet français d'EWN comme référence). L'approche par extension est cependant assez contestable, car elle présuppose que le wordnet cible soit isomorphe à WPN, comme si l'organisation des sens de la langue cible pouvait correspondre exactement à celle de l'anglais. C'est d'autant plus problématique que les sens dans PWN sont organisés en fonction de 4 parties du discours (nom, verbe, adverbe, adjectif), catégories qui ne correspondent pas au système catégoriel de langues génétiquement éloignées, telles que l'arabe (qui connait 3 catégories principales : nom, verbe et particule).

Citons enfin le modèle géométrique des atlas sémantique de Ploux (2007) qui s'appuie sur l'extraction de cliques de mots synonymes (construites à partir de dictionnaires de synonymes), analogue à des synsets, pour réaliser un découpage plus findes sous-sens des mots. En utilisant un dictionnaire bilingue, et une méthode de projection dans un espace sémantique commun (Ploux et Ji, 2003), l'auteure montre comment les cliques obtenues dans chaque langue peuvent être appariées, ce qui permet d'enrichir à la fois le dictionnaire bilingue, et d'identifier de nouveaux candidats synonymes dans chaque langue. Notre approche, très voisine de ces travaux par sa représentation géométrique du sens, en est complémentaire, dans la mesure où c'est à partir de corpus multilingues parallèles, et non de dictionnaires, que nous allons chercher à extraire ce type de cliques.

[1]Les *synsets* correspondent aux nœuds sémantiques du réseau.

3 Hypothèses et méthodologie

3.1 Présentation générale

Pour l'organisation des sens en arabe, nous nous somme inspirés de l'architecture de réseaux sémantiques de type wordnet préexistants, notamment les réseaux d'EuroWordNet.Notons que les synsets, dans l'architecture de PWN, représentent l'intersection sémantique d'un ensemble d'unités, et constituent l'identification implicite d'une acception (sens), en organisant les unités selon deux propriétés : synonymie et polysémie. La synonymie dans un wordnet monolingue est basée sur l'équivalence (souvent partielle) entre les unités regroupées dans synset.

La FIGURE 1 – Exemple de synsets de WordNet pour le nom anglais*situation* donne les différents synsets de WordNet pour le nom anglais*situation*: *(situation, state of affairs) (situation, position) (situation) (site, situation) (position, post, berth, office, spot, billet, place, situation).* Chacun de ces synsets correspond à une certaine acception (*sense*), explicitée par une glose, mais surtout caractérisée par un ensemble d'unités synonymes susceptibles de partager cette acception.

FIGURE 1 – Exemple de synsets de WordNet pour le nom anglais*situation*

Ainsi la mise en évidence de l'équivalence de certaines unités, autour d'un sens donné, conduit également à une prise en compte du fait polysémique, par le fait qu'une même unité est susceptible d'intervenir dans différents synsets. Or, comme Sagot et Fišer (2008),nous faisons l'hypothèse que ce type de structuration du sens peut être déduit des relations d'équivalence traductionnelle observées sur des corpus de textes traduits (que nous nommerons désormais corpus parallèles). En effet, nous pensons qu'une approche multilingue basée sur des corpus parallèles permet de donner des renseignements utiles tant sur le plan de la polysémie (un lexème possédant des équivalents différents étant susceptible d'avoir différentes acceptions) que sur celui de la synonymie (deux lexèmes possédant les mêmes équivalents dans une langue cible étant susceptibles de partager un même sens).

La traduction, par le réseau de relations qu'elle constitue, peut peut-être jouer le rôle de révélateur par rapport à la structuration interne des sens d'une langue, vue sous l'angle de cette dialectique entre synonymie et polysémie. En outre, nous pensons que l'utilisation de plus de deux langues permet de renforcer des hypothèses concordantes issues de sources

d'information différentes : une unité très polysémique aura sans doute de nombreux équivalents dans différentes langues cibles. Et si deux unités partagent un même sens, elles partageront sans doute des équivalents dans leurs traductions vers plusieurs langues.

Ainsi une seule langue cible n'est pas forcément suffisante pour porter un éclairage sur la variabilité sémantique d'une unité : il peut arriver qu'une unité polysémique puisse être traduite dans une langue cible par une unité équivalente présentant le même type de polysémie. Mais il est peu vraisemblable que cette même structuration se retrouve à l'identique dans plusieurs langues cibles. Par exemple, le nom *terme* en français présente la même ambiguïté que l'anglais *term* : il peut (entre autres) prendre les sens de /mot/ ou de /fin, échéance/. On trouve la même ambiguïté dans d'autres langues romanes, comme l'espagnol ou l'italien. Mais les équivalents allemands *Begriff* ou*Abschluss*, qui correspondent à ces sens, ne présentent pas cette ambiguïté.C'est pour tirer parti de ces « discordances révélatrices » que nous avons constitué corpus parallèle en 4 langues : français, anglais, espagnol et arabe.

Par ailleurs, dans la perspective d'extraire des unités de sens qui puissent être rapprochées des synsets de WordNet, nous pensons que le lexème n'est pas une entrée consistante pour l'organisation des sens, notamment pour l'enregistrement des relations d'*équivalence traductionnelle* en détachant les unités de leurs contextes d'occurrence. Comme Ploux(2008) nous proposons plutôt de nous appuyer sur *des cliques* de lexèmes, c'est-à-dire des ensembles de lexèmes qui partagent tous, pris deux à deux, un certain contenu sémantique. En effet, nous croyons que ces cliques permettent d'organiser le lexique en fonction des sens, à un niveau de granularité plus fin que celui des lexèmes, qui demeurent très ambigus hors contexte. A la différence de Sagot et Fišer (2008), qui utilisent des relations d'équivalence entre une langue (le français) et les autres, nous pensons que les cliques, en impliquant simultanément tous les couples de langues, imposent un degré de cohésion supérieur.

En extrayant de telles cliques à partir de notre corpus parallèle, nous espérons trouver une organisation des unités suffisamment cohérente pour apporter des informations fiables sur les deux propriétés qui nous intéressent, à savoir la synonymie et polysémie. Nous tenterons notamment de vérifier que les cliques extraites à partir des ensembles d'unités liées par des relations d'équivalences automatiquement extraites, grâce à des techniques d'alignement de corpus, sont apparentées aux synsets des réseaux sémantiques multilingues tels qu'EuroWordNet. De plus, ces cliques, par leur structuration fondamentalement multilingue, sont peut-être moins ancrées dans les particularités du découpage sémantique d'une langue donnée, et pourraient former de meilleurs candidats, pour un ajustement mutuel de différents wordnets, que les synsets de WordNet.Une telle ressource peut donc avoir des applications directes pour la désambigüisation en traduction automatique, dans le contexte spécifique d'une traduction en langue tierce connaissant déjà d'autres traductions outre le texte original (nous pensons aux traductions des textes de l'Union européenne, qui mettent en jeu jusqu'à 23 langues différentes).

Une fois ces cliques multilingues obtenues, nous tenterons de les associer aux synsets existants d'EuroWordNet. Les retombées seraient multiples :

- d'une part, cela permettrait d'établir un lien entre des lexèmes arabes et ces synsets. A partir de cette association, on pourrait envisager de projeter certaines informations du

réseau(non seulement synonymie et polysémie, mais aussi hyper/hyponymie, antonymie, méronymie, etc.) vers l'arabe. Sans présumer qu'un wordnetarabe doit être congruent à PWN, cette possibilité permettrait d'amorcer la construction d'un nouveau réseau, et de récupérer automatiquement un grand nombre d'informations de nature sémantique.

- d'autre part, cela permettrait de mettre au point une méthode pour l'enrichissement automatique d'un réseau de type EuroWordNet, et consolidant des liens interlingues existants (qui seront nommés plus loin ILI, pour Inter Lingual Index), voire en y ajoutant des nouveaux.

Ainsi, nous espérons que l'analogie entre nos cliques et les synsets de wordnets déjà créés, nous permettra de dégager une méthode pour amorcer automatiquement la construction d'une ressource sémantique arabe. A terme, une telle ressource serait utile pour de nombreuses applications du traitement de la langue arabe, comme la recherche d'information, la traduction automatique, les moteurs de question-réponse, la veille informationnelle, l'analyse d'opinion, etc.

3.2 Étapes suivies

3.2.1 Constitution de corpus multilingue parallèle

Notre corpus, qui provient des archives des Nations Unies (NU)[3], est constitué de 185 textes traitant de sujets différents (ex. commerce international, droit de la femme, santé…etc.)dans chacune des quatre langues suivantes : français, anglais, espagnol et arabe classique.

Notre corpus a subi des étapes de reformatage, d'étiquetage et de lemmatisation (sauf pour l'arabe, qui est segmenté mais pas lemmatisé). Nous avons utilisé l'étiqueteur treetagger(Schmid, 1995) pour les trois langues indo-européenneset Amira1.0pour l'arabe (Diab et al., 2007).

Ensuite une étape d'alignement phrastique avec Alinea(Kraif, 2001)a été appliquée sur notre corpus. Certains textes comportant des tableaux, des index, etc.brisant le parallélisme, le taux d'alignements erronés était d'environ 28% : nous avons procédé à une étape de filtrage des alignements problématiques avant de lancer l'alignement lexical avec Giza++ (Ochet Ney, 2003). Le nombre de paires de mots (ou groupes de mots) alignés pour chaque couple de languesvarie entre 73 823 (couple fr-ar) et 98 303 (couple en-es).

3.2.2 Extraction de cliques multilingues

Notre méthode d'extraction de cliques, ou des sous-graphes maximaux complets connexes, s'appuie sur l'extraction automatique des équivalents traductionnels. Les correspondances extraites à partir de tous les alignements deux à deux des textes du corpus forment un immense graphe reliant des unités des quatre langues considérées. Pour ne retenir que les arcs les plus pertinents de ce graphe, nous avons d'abord procédé à un filtrage des correspondances lexicales (en fonction de leur fréquence). Dans un deuxième temps, nous avons procédé à l'extraction de toutes les cliques autour d'une unité donnée. Enfin, pour éviter l'éparpillement des cliques voisines mais disjointes du fait de l'absence d'une ou deux relations dans notre corpus, une phase de clusterisationascendante hiérarchique a été mise

[3] Téléchargé depuis le site http://unbisnet.un.org

en œuvre (la proximité de deux cliques étant calculées sur avec une formule de Dice).

Pour interpréter les cliques obtenues, nous émettons l'hypothèse de *centralité des cliques*, l'interrelation entreles éléments de la clique pris *2 à 2*étant probablement une conséquence de l'existence d'une intersection sémantique *commune*non vide.

Par exemple dans la clique : (*fr-N-économie, en-N-saving, it-N-risparmio, de-N-Einsparung*[4]), il est probable que les sens partagés par (*fr-N-économie, en-N-saving*) et (*fr-N-économie, it-N-risparmio*) aient une intersection commune. En effet si tel n'était pas le cas, cela signifierait que les équivalences (*fr-N-économie, en-N-saving*) et (*fr-N-économie, it-N-risparmio*) correspondent à deux acceptions distinctes de *fr-N-économie*. Et du coup il serait peu probable que *en-N-saving* et *it-N-risparmio*soient eux-même des équivalents potentiels. L'ajout d'un équivalent commun à ces trois unités, avec l'allemand *de-N-Einsparung,* renforce encore cette hypothèse de convergence des intersections. L'appartenance à une clique, qui implique une relation avec tous les éléments de la clique, révèle, lorsqu'un grand nombre de langues est mis en jeu, une propriété centripète de la clique, le "centre" qui en assure la cohésion pouvant être interprété comme l'intersection commune à tous ses membres.

Mais que peut-on dire pour deux éléments de la même langue au sein d'une clique ? Par définition, ils ne sont pas en relation d'équivalence traductionnelle. Doivent-ils nécessairement être synonymes, c'est-à-dire avoir une intersection sémantique (un sens dénotationnel commun) correspondant au centre de la clique ? On peut imaginer certains cas où une langue opère une distinction non marquée dans les autres, utilisant par exemple deux lexèmes concurrents là où les autres n'en n'utilisent qu'un seul. Dans ce cas, les deux lexèmes peuvent être considérés comme cohyponymes, et ce n'est pas leur intersection mais plutôt leur union qui doit correspondre au centre de la clique.

3.2.3 Rattachement de sens et de relations à des unités arabes

Ces cliques maximales où toutes les unités sont en interrelation, du fait d'une probable intersection sémantique (des sens voisins ou connexes), ressemblent auxsynsets d'un réseau sémantique tel que PWN. En effet, dans PWN, les sens sont caractérisés de manière similaire, par l'intersection sémantique d'un ensemble de nœuds fortement liés et activés simultanément : chaque synset dénote un "concept" différent situé au croisement d'un ensemble d'unités lexicales susceptible de porter ce sens, décrit par une courte définition appelée *gloss*. De la même manière qu'avec nos cliques, l'appartenance d'une unité lexicale à plusieurs synsets constitue une manifestation explicite de sa polysémie.

C'est pourquoi nous allons tenter de relier nos cliques avec le lexique sémantique d'EuroWordNet, afin d'évaluer la possibilité de récupérer pour les unités arabes des relations sémantiques déjà déclarées pour des unités en anglais, français et espagnol, dans leurs réseaux respectifs. Voici lesprincipes suivis pour le rattachement des sous-sens et des relations d'EWN aux unités arabes:

1 **Principe de clôture transitive intra-clique** : rattachement des unités arabes à un ILI.

Si toutes les unités d'une même clique partagent un et un seul sens d'EuroWordNet(via les ILI) alors la clique est désambigüisée et on rattache l(es)

[4]Pour éviter les ambiguïtés, nous préfixons chaque lexème par sa langue et sa catégorie.

unité(s) arabe(s) à ce sens commun.

Par exemple, dans la clique (*en-N-science fr-N-science es-N-cienciaar-N-*علم)leslexèmes anglais, français et espagnol sont tous les trois rattachés à un seul ILI glosé par /*a particularbranch of scientificknowledge*/. On peut donc également lui rattacher l'unité arabe, car il n'y a pas d'ambiguïté.

2 **Principe de clôture transitive inter-clique** : ajout d'une relation entre deux unités arabes.

Si deux cliques ontchacune été rattachéesà un seulILI, respectivement,et si pour une langue donnée il existe une relation sémantique entre deux unités appartenant à ces deux cliques, pour uneacception liée auILI retenu, alors la relation peut être étendue pour les unités arabes contenues dans ces cliques, sauf si une relation contradictoire peut être inférée à partir d'une autre paire de lexèmes.

Par exemple si on considère les deux cliques suivantes :: (*ar-N-*قسم*fr-N-fragment en-N-snippet es-N-recorte*) et(*ar-N-*حصة*fr-N-morceau es-N-pedazo en-N-piece*), sachant qu'on a une relation *'has_hyperonym'*entre en*-N-snippet et en-N-piece*, et qu'il n'existe pas de relation différente pour les unités des autres langues (il se trouve qu'on a la même relation pour le français et l'espagnol, même si ce n'est pas une condition nécessaire ici), on peut étendre la relation aux unités arabes*ar-N-*قسمet*ar-N-*حصة.

4 Evaluation des résultats

Nous n'avons évalué à ce jour que les résultats de l'application du principe de clotûre transitive intra-clique : la projection des relations sémantiques fera l'objet de recherches ultérieures.

4.1 Evaluation quantitative

Dans ce travail, nous avons testé seulement les cliques contenant des unités de deux catégories (noms et verbes) puisque leFREWN (French EuroWordNet) ne comporte ni adjectifs ni adverbes. Nous n'avons pas traité non plus les unités pour lesquelles les catégories Nom et Verbe n'étaient pas complètement désambiguïsées ex. (N/Adj), (V/N/Adj)…etc.

Nous avons appliqué notre approche sur les 100 noms et les 100 verbes français les plus fréquents dans notre corpus. Parmi les clusters obtenus, nous en avons prélevé, par tirage au sort, 100 pour les verbes et 100 pour les noms(voir tableau 1).

	Nom	Verbe
Nb clusters traités	100	100
Nb clusters valides (désambigüisés et non-désambigüisés)	56	29
Nb lemmes arabes dans les clusters désambigüisés	**74**	**37**
Nb lemmes validés complètement (VC)	59	21
Nb lemmes validés partiellement (VP)	8	6
Nb lemmes non validés	7	10
Nb Total d'unités arabes validés (VC+VP)	**94 / 111** ≈ 84,7%	

TABLE1 – Tableau récapitulatif des résultats pour l'arabe

Les clusters valides désambiguïsés ou non désambiguïsés sont respectivement les clusters qui sont reliés à unou plusieurs ILI (certains clusters n'étant reliés à aucun, du fait de la couverture d'EWN). Nous en avons obtenu 56 pour les noms et 29 pour les verbes. Le nombre de lemmes arabes (et non de formes fléchies) dans les clusters précédents est de 74 pour les noms et 37pour les verbes. Parmi ces lemmes, nous avons vérifié manuellement que le sens de l'ILI correspondait bien à une acception du lexème (nous nous sommes référés au dictionnaire Alwaseet). Nous avons ainsi trouvé 59 lemmes arabes validés complètement pour les noms, et 21 pour les verbes.Quand le sens de l'ILIest voisin, mais correspond à une catégorie plus générale ou plus spécifique, nous avons considéré les lemmes comme partiellement validés. Nous en avons trouvé 8 pour les noms et 6 pour les verbes. Le nombre de lemmes arabes non valides, c.-à.-d. pour lesquels l'ILI est trop éloignédes différents sens attestés par le dictionnaire, est de 7 pour les noms et 6 pour les verbes.

Au final nous avons calculé le pourcentage d'unités arabes (noms et verbes) validées (ou partiellement validées) au sein des cliques(Nb.de lemmes validés complètement+partiellement/Nb de lemmes arabes dans les clusters valides)et nous avons obtenu un pourcentage d'environ 84,7% de rattachements sémantiques valides.

4.2 Evaluation qualitative

4.2.1 Validitésémantique des clusters

Dans cette partie de l'évaluation, nous cherchons à examiner la validité au plan sémantique des clusters obtenus.

Plusieurs cas de figure ont été rencontrés, que nous listons ci-dessous.

Cas n°1 : identification correcte de plusieurs acceptions

Nous avonsobtenu des clusters qui peuvent permettre d'identifier différents sens pour une même unité arabe. Prenons l'exemple suivant qui illustre ce point :

Nous avons obtenu deux clusters pour le lemme arabe علم:

Cluster 1: (ar-N-العلوم ar-N-العلم en-N-science fr-N-science es-N-ciencia).

Il se trouve que les noms arabes العلم، العلوم sont des formes fléchies pour le lemme علم.

Par ailleurs, les trois unités *en-N-science, es-N-ciencia*et*fr-N-science*, en se référant aux wordnetsde chaque langue pris indépendamment, sont polysémiques. Mais toutes les unités (fr-en-es) de ce cluster partagent le seul lien ILI suivant: */a particularbranch of scientificknowledge/*. Un des sens de l'unité arabe *ar-N-*علم mentionné dans le dictionnaire Alwaseetest : */un groupe de connaissances scientifiques dans un domaine particulier/* (nous traduisons) ce quicorrespond bien au ILImentionné et valide donc le rattachement.

> Cluster 2 pour le même lemme arabeعلم: (ar-N-علمar-N-تعلمfr-N-apprentissage en-N-learning es-N-aprendizaje).

Toutes les unités arabes du cluster précédent sont des formes fléchies du même lemme علم : (ar-N-علمar-N-تعلم).L'ILIcommun des troisunités (en-es-fr) estglosé par/*the cognitive process of acquiring skill or knowledge; "the child's acquisition of language"& 03 09 2ndOrderEntity Agentive Cause Dynamic Experience Mental Property SituationType Static*/.Cet ILI est donc assez proche de l'un des sens de l'unité arabe علمdans le dictionnaire Alwaseet, également lié à la notion d'apprentissage: */l'acquisition et la connaissance de la vérité des choses*/.Le sens identifié par ce cluster semble donc pertinent pour rattacher le lemmearabeà un deuxième ILI.

Mais dans certains cas les sens représentés par les clusters sont incomplets, puisque certaines acceptions très communes ne sont pas représentées, du fait des limitations de nos ressources, wordnets et corpus. Nous avons relevé les cas suivants :

Cas n°2 : Insuffisance de couverture des WNs

1. Absence d'un lexème dans les WNs

Certains verbes français, pourtant assez communs, qui se trouvent dans nos cliques, n'apparaissent pas dans FREWN : *adjoindre, s'approprier, figurer, spécialiser*...

2. Absence d'un sens dans les WNs

Le cluster obtenu pour le mot arabe فلسفة est : (ar-N-فلسفة es-N-filosofíafr-N-philosophie en-N-philosophy). Notonsque les deux unités en-N-philosophyet es-N-filosofía, se référant auWN de chaque langue, sont polysémiques alors quel'unité fr-N-philosophie serait monosémique (d'après FREWN).Mais cela est simplement dû à une lacune de FREWN puisque l'on trouve dans la langue française d'autres sens pour le mot *philosophie* (p.ex. */sagesse/*).

3. Spécificité du découpage des sens dans les WNs

Dans certains cas, assez marginaux, il se peut que l'hypothèse de centralité des cliques ne soit pas vérifiée. Par exemple chaque couple d'unités dans la cliquesuivante (en-N-fundfr-N-fonds es-N-fondo) partage un lien ILI, mais aucun ILI n'est partagé par les trois unités considérées ensemble :

- en-N-fundETfr-N-fonds: */a reserve of money set aside for some purpose& 03 1stOrderEntity 21 Artifact Function MoneyRepresentation Origin Possession/.*
- en-N-fundET es-N-fondo: */a supply of something available for future use& 03 1stOrderEntity 21 Function Possession/.*
- fr-N-fondsETes-N-fondo: */assets in the form of money& 03 1stOrderEntity 21 Function Possession/*

On constate que les sens 1 et 3 sont cependant assez voisins, et qu'avec un découpage un peu moins spécifique des sens on pourrait cependant les identifier (tout découpage comportant une certaine part d'arbitraire).

4. Rattachement à un sens trop générique (*top-ontology*)

Dans certains cas, l'ILI commun est beaucoup trop générique pour donner une indication utile sur l'acception commune des unités. Par exemple, la clique : (es-N-disposición en-N-provision fr-N-disposition) est liée au ILI-RECORD suivant de la *top-ontology :/& 03 10 2ndOrderEntity 3rdOrderEntity AgentiveBoundedEvent Cause Communication Dynamic Mental Purpose Relation Situation Type Social Static/.*

Cas n°3 : Insuffisance de couverture du corpus

Prenons les deux clusters obtenus pour le mot arabeمادة:

Cluster1: (ar-N-مادةfr-N-article en-N-article es-N-artículo).

Les trois unités fr-N-article en-N-article es-N-artículo partagent l'ILIsuivant: */one of a class of artifacts; "an article of clothing"/.*

Cluster2: (ar-N-مادةfr-N-matériaux en-N-material es-N-material).

Les trois unités *fr-N-matériaux en-N-material es-N-material*partagent l'ILI suivant: */Information (data or ideas or observations) thatcanbereworkedinto a finishedform;"the archives providedrichmaterial for a definitivebiography"/.*

Mais il manque d'autres sens pour l'unité *ar-N*مادة, qui ne sont pas représentés du fait des limitations du corpus, par exemple : /*chose physique, corporelle, par opposition à l'esprit*/. La taille trop réduite du corpus n'a pas permis à notre méthode d'extraire un second cluster contenant d'autres équivalents susceptibles de se référer à cette autre acception.

*Cas n°4 : Ambiguïtés dues à des polysémies parallèles.*Voici un exemple des cas des cliques ambiguës:

Dans la clique (en-N-topicfr-N-sujet ar-N-موضوعen-N-subject es-N-tema), l'unité *en-N-subject* est polysémique et elle partage avec *fr-N-sujet* et *es-N-tema* plusieurs ILI, */some situation or event that is thought about; "he kept drifting off the topic"; "itis a matter for the police"/* et */something (a person or object or scene) selected by an artist or photographer for graphicrepresentation/.* Onconstate que l'ambigüité est partagée par les trois langues. Ce cas, assez courant, n'est pas lié aux ressources mais aux langues impliquées. On peut supposer que plus le nombre de langues est grand, plus ces cas devraient être rares, car la probabilité d'obtenir des polysémies parallèles diminue avec la variété des langues mises en jeu.

Cas n°5 : Bruit lié à la non reconnaissance d'unités polylexicales dans les équivalents traductionnels

Considérons le cluster suivant : (ar-N-**لغة** fr-N-langue en-N-language es-N-idioma fr-N-linguistique). L'unité française *fr-N-linguistique* qui est monosémique (dans FREWN) et qui appartient à un synset totalement différent de celui de *fr-Noun-langue* a comme ILI : */the scientificstudy of language/.*

On peut penser que cette erreur est à des alignements n-n non reconnus

(*languagestudy→ linguistique*) ou à l'ambigüité morphologique (p.ex. *languageresearch→ recherche linguistique*), l'adjectif *linguistique* étant par erreur étiqueté comme un nom).

Cas n°6 : Ambigüités liées à une sur-clusterisation

On observe parfoisle regroupement de deux cliques qui devraient rester séparées, comme dans le cluster suivant pour le nom français *droit*: {(fr-N-droit en-N-right es-N-derechoar-N-حق)(fr-N-droit en-N-Law es-N-derechoar-N-قانون)}

Deux sous-sens existent dans ce cluster :

1. (fr-N-droit en-N-right es-N-derechoar-N-حق)→ */an abstract idea of thatwhichis due to a person or governmental body by law or tradition or nature /.*
2. (fr-N-droit en-N-law es-N-derechoar-N-قانون)→ */the collection of rulesimposed by authority; "civilizationpresupposes respect for the law/.*

Voicien % la répartition observée des causesde non-rattachement pour les noms (pour un total de 44%) :

- Cas 2.Insuffisance de couverture des WNs	18%
- Cas 3, cas 5, cas 6. Pas d'ILI commun à toutes les unités	9%
- Cas 4. Ambiguïtés dues à des polysémies parallèles	17%

Quant aux verbes, nous avons eu des résultats très faibles. La répartition est la suivante (pour un total de 71%) :

- Cas 2.Insuffisance de couverture des WNs	24%
- Cas 3, cas 5, cas 6. Pas d'ILI commun à toutes les unités	30%
- Cas 4. Ambiguïtés dues à des polysémies parallèles	17%

Ainsi, pour les verbes nous avons obtenu des clusters correctement désambiguïsées et rattachées à EWN dans seulement 29% des cas.

4.2.2 Rattachement invalide

L'examen minutieux des cas invalides nous révèlent plusieurs types de cas :

1. Le sens arabe est complètement différent de celui des autres unités, car l'unité a été regroupée par erreur lors de la phase de clusterisation.
2. Le mot arabe est un lexème composé, mais un seul lexème se trouve appartenir à la clique, du faitd'une mauvaise tokenisation.
3. Le dictionnaire arabe qui nous sert de référence est également lacunaire.

D'un point de vue général, dans nos résultats expérimentaux, on voit que 94 / 111 unités arabes (verbes et noms) (voir tableau 1) sont validées partiellement ou complètement par le dictionnaire *Alwaseet*, ce que nous semble être un résultat tout à fait encourageant pour la méthode. Si celle-ci semble mieux fonctionner pour les noms que pour les verbes, c'est peut-être dû au fait que beaucoup de verbes présentent un sens très général, et sont plus polysémiques que les noms. En outre, de nombreux verbes font partie de locutions verbales ou jouent le rôle de collocatifs dans des collocations verbe-nom, et ne prennent leur sens précis qu'au sein d'une expression plus large.

5 Conclusion et perspectives

Nous avons présenté dans cetarticleune méthode visant à la construction d'une ressource sémantique pour la langue arabe.A travers nos expérimentations, nous avons constaté que les cliques créées automatiquement à partir de notre corpus multilingue constituent un guide intéressant pour l'organisation des sens. En effet, ces cliques présentent l'intérêt de renseigner à la fois sur la synonymie et la polysémie des unités, et d'apporter une forme de désambiguïsation sémantique - les équivalents traductionnels apportant souvent des indices intéressants pour la désambiguïsation, puisqu'une même polysémie a peu de chances de se retrouver dans de nombreuses langues différentes.

Un premier avantage lié à la constitution des cliques multilingues est qu'elles contiennent des unités connexes dans plusieurs langues : cette cohésion interne liée a leur propriété de complétude et de connexité permet de filtrer les résultats de l'alignement lexical et d'éliminer la plupart des erreurs d'alignement(de nombreuses correspondances obtenues avec Giza++ étant soit erronées, soit incomplètes, soit difficilement isolables de leur contexte traductionnel particulier). Il est en effet peu probable qu'une telle erreur d'alignement induise une intersection non vide entre plusieurs langues à la fois.

En retour, l'ensemble des cliques multilingues obtenu s'est révélé utile pour mettre à jour et enrichir les relations sémantiques qui sont manquantes dans les réseaux d'EWN. En effet, les cliques offrent la possibilité de maximiserla compatibilité entre les wordnets de différentes langues, en consolidant les relations d'équivalence existantes et en les complétant par de nouvelles relations pour certaines langues.

Au vu de l'évaluation que nous avons effectuée, concernant la possibilité d'étendre les connaissances sémantiques dans d'autres langues vers la ressource arabe que nous voudrions construire *in fine*, les résultats obtenus apparaissent plutôt encourageants.Il nous reste à évaluer laméthode de projection des relations sémantiques, qui offrirait la possibilité destructurer automatiquement les sens des unités arabes par l'intermédiaire de PWN : c'est une perspective de recherche selon nous très prometteuse, même si nous pensons qu'un wordnet arabe doit comporter une structure originale qui ne peut calquer celle de WordNet.

Cetteméthode automatique, et donc peu coûteuse, s'avère suffisamment générale pour être appliquée à d'autre langue que l'arabe, qui est sans doute un cas de figure parmi les plus difficiles étant donné les difficultés posées par cette langue en terme d'ambigüité graphique et de segmentation des mots. Cette technique présente un intérêt pour les langues dites "peu dotées", qui ont besoin de rattraper l'écart technologique concernant les traitements informatiques et les usages sur les réseaux de l'information. La constitution d'une ressource sémantique complète peut en effet être très utile, dans un second temps, pour alimenter et améliorer diverses applications de TAL, comme les moteurs de question-réponse, la traduction automatique, les systèmes d'alignement, la recherche d'information, etc.

Nos expérimentations ont aussi permis de dégager certaines limites inhérentes aux données investies dans notre méthode. A cause de certaines insuffisances au niveau de la couverture de notre corpus, une phase de clusterisationa été rendue nécessaire pour regrouper des cliques artificiellement éparses. Cette phase a engendré des erreurs dans nos résultats, certaines cliques étant indûment regroupées. Afin d'éviter ce type de bruit, probablement, le recours à un corpus plus vaste et plus varié permettrait d'obtenir une couverture suffisante

de la langue générale, de façon à capter de manière plus complète toutes les virtualités sémantiques des unités et toute l'étendue de leurs possibilités de traduction. Par ailleurs, pour diminuer l'effet de ce que nous avons dénommé des polysémies parallèles, le recours à un plus grand nombre de langues différentes ne pourrait qu'améliorer la finesse des résultats.

6 Références

DIAB, M., HACIOGLU K. ETJURAFSKY D. (2007). *Arabic Computational Morphology: Knowledge based and Empirical Methods*, chapter 9, A.Soudi, A. van den Bosch et G. Neumann (Eds.), Springer, pp. 159–179.

ELKATEB S., W. BLACK, H. RODRIGUEZ, M. ALKHALIFA, P. VOSSEN, A. PEASE, ET C. FELLBAUM (2006).Building a WordNet for Arabic.*Proceedings ofThe fifth international conference on Language Resourcesand Evaluation* (LREC 2006).

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

HENDRIX, GARY G. (1979). *Encoding knowledge in partitioned networks*.Findler, pp. 51-92.

KRAIF O. (2001). *Exploitation des cognats dans les systèmes d'alignement bi-textuel: architecture et évaluation*.TAL 42: 3, ATALA, Paris, pp.833-867.

OCH F. J. ET NEY H. (2003) A Systematic Comparison of Various Statistical Alignment Models, *Computational Linguistics*, volume 29, number 1, pp. 19-51.

MALLAK I.(2011). *De nouveaux facteurs pour l'exploitation de la sémantique d'un texte en Recherche d'Information*.Thèse de doctorat àl'Université Toulouse III - Paul Sabatier.

NILES I. ET PEASE A. (2001). *Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology.* 1800 Embarcadero Rd. Palo Alto CA 94303.

PLOUX S. (2007) Enrichir automatiquement des dictionnaires électroniques de synonymes et de traduction : une application du modèle d'appariement multilingue des Atlas sémantiques*Actes des 2èmes journées d'animation scientifique régionales « Élaborer des dictionnaires en contexte multilingue »*, Tunis.

PLOUX, S. ET JI. H.(2003). «A model for matching semantic maps between languages(French/English, English/French) ».*Computational Linguistics*, vol. 29, no. 2, p. 155–178.

SAGOT B. AND FIŠER D. (2008).Building a Free FrenchWordNetfromMultilingual Resources.*Proceeding of Ontolex*, Marrakech, Maroc.

VOSSEN P. (1998). EuroWordNet: A Multilingual Database with Lexical SemanticNetworks.*Computational Linguistics*, Volume 25, Number 4.