Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français

Aurélie Névéol[1,2], Sylwia Ozdowska[3]

[1]Laboratoire PSI – FRE CNRS 2645
INSA de Rouen - BP8 - Avenue de l'Université
76801 Saint Etienne du Rouvray Cedex , France
aneveol@insa-rouen.fr

[2]Equipe CISMeF et L@STICS
Faculté de Médecine de Rouen, 1 rue de Germont
76031 Rouen – France.

[3] Equipe de Recherche en Syntaxe et Sémantique
Université de Toulouse le Mirail, 5 allées Antonio Machado
31058 Toulouse Cedex 1 – France.
ozdowska@univ-tlse2.fr

Résumé. Le Catalogue et Index des Sites Médicaux Francophones (CISMeF) recense les principales ressources institutionnelles de santé en français. La description de ces ressources, puis leur accès par les utilisateurs, se fait grâce à la terminologie CISMeF, fondée sur le thésaurus américain Medical Subject Headings (MeSH). La version française du MeSH comprend tous les descripteurs MeSH, mais de nombreux synonymes américains restent à traduire. Afin d'enrichir la terminologie, nous proposons ici une méthode de traduction automatique de ces synonymes. Pour ce faire, nous avons constitué deux corpus parallèles anglais/français du domaine médical. Après alignement semi-automatique des corpus paragraphe à paragraphe, nous avons procédé automatiquement à l'appariement bilingue des termes. Pour cela, le lexique constitué des descripteurs MeSH américains et de leur traduction en français a fourni les couples amorces qui ont servi de point de départ à la propagation syntaxique des liens d'appariement. 217 synonymes ont pu être traduits, avec une précision de 70%.

1 Introduction

La recherche d'information, l'indexation, et la manipulation de ressources multimédia en général sont des domaines qui s'appuient sur l'utilisation d'une terminologie pour décrire les ressources disponibles et y accéder. Dans le domaine bio-médical, de nombreux travaux ont été réalisés en ce sens et plusieurs terminologies (par exemple, le MeSH[1] pour la gestion de connaissances, ou la SNOMED CT[2] pour les termes cliniques) ou ontologies (par exemple,

[1] Medical Subject headings. cf. http://www.nlm.nih.gov/mesh/meshhome.html
[2] SNOMED Clinical Terminology. cf. http://www.nhsia.nhs.uk/snomed/pages/default.asp

GO[3]) sont disponibles. Bien que ces différentes terminologies soient complémentaires, on observe également des recoupements conceptuels qui s'avèrent toujours intéressants au niveau lexicographique, car un même concept peut être désigné et décrit de manière différente d'une terminologie à l'autre. Le projet UMLS (Unified Medical Language System) a pour objectif d'exploiter ces complémentarités pour les terminologies anglophones du domaine Médical. La plupart de ces terminologies, d'abord développées en anglais, sont ensuite traduites dans d'autres langues par des experts du domaine. Ainsi, la création d'un Vocabulaire Unifié Médical Français (Darmoni *et al*., 2003) est en cours pour compléter les ressources terminologiques médicales disponibles en français, et étendre les réalisations de l'UMLS à cette langue.

Le Catalogue et Index des Sites Médicaux Francophones (CISMeF) bénéficie directement de ces travaux, dans la mesure où la terminologie CISMeF, utilisée pour l'indexation des ressources et pour la recherche d'information au sein du catalogue, est fondée sur le MeSH (Darmoni *et al*., 2000). Ce travail s'inscrit également dans la continuité du développement de ressources médicales pour le système d'indexation automatique de CISMeF (Névéol, 2004).

Dans ce contexte, nous proposons une méthode de traduction automatique des synonymes américains du MeSH, afin d'enrichir la terminologie CISMeF. Nous avons donc isolé les synonymes américains non traduits en français pour les mots clés MeSH utilisés par CISMeF (5166), et constitué deux corpus parallèles du domaine médical afin d'en extraire la traduction en français des synonymes qui y sont présents : 216 synonymes dans le premier corpus, et 247 dans le second.

Nous présentons dans un premier temps les étapes de la constitution et de l'alignement des deux corpus paragraphe à paragraphe. Nous détaillons ensuite la méthode utilisée pour aligner les termes, et extraire la traduction des synonymes. Dans un deuxième temps, nous faisons un bilan des résultats obtenus, et nous discutons de l'apport terminologique réalisé avant de conclure sur les perspectives de poursuite de ce travail.

2 Constitution et alignement des corpus

2.1 Corpus CISMeF/Hansard

Afin de constituer un premier corpus de travail adapté à notre problématique, nous avons porté une attention particulière aux critères suivants: la qualité de la traduction, l'adéquation du contenu avec le domaine médical (plus spécifiquement, avec les concepts concernés par les synonymes à traduire) et la qualité de l'alignement, au niveau des textes, dans un premier temps, puis au niveau des paragraphes. Ainsi, une partie du corpus provient d'un corpus juridique préalablement aligné (le Hansard[4]), et l'autre partie d'un corpus médical spécialisé (CISMeF[5]).

Le Hansard est un concordancier bilingue français/anglais rassemblant les débats à la chambre des communes du parlement canadien, ainsi que leur traduction. Nous avons effectué des recherches sur des termes MeSH à l'aide de l'outil TransSearch afin de sélectionner des textes ayant trait au droit de la santé.

[3] Gene Ontology - cf. http://www.geneontology.org
[4] http://www.tsrali.com
[5] http://www.cismef.org

Le catalogue CISMeF indexe uniquement des ressources francophones spécialisées dans le domaine de la santé, et précise si ces ressources sont également disponibles dans d'autres langues. Nous avons extrait les ressources bilingues anglais/français sous forme d'une liste de 1510 URLs correspondant à la version française des ressources. Certaines ressources, comme les sites des hôpitaux, ne présentent pas d'intérêt pour l'acquisition de traduction de synonymes et ont donc été écartées. D'autres ressources contenaient un résumé anglais d'un article développé en français, ou présentaient les textes sans séparation nette entre les deux langues. Elles ont été également écartées. Parmi les ressources restantes, plusieurs émanent de sites éditeurs bilingues affiliés au ministère de la santé canadien[6], ce qui est une garantie de la qualité de la traduction disponible. De plus, ces sites observent un classement régulier et organisé des documents dans les différentes langues. Nous sommes donc en mesure de déduire l'URL de la version anglaise de la ressource à partir de l'URL de la version française, ou bien dans certains cas, à partir de la ressource elle-même, lorsque celle-ci contient un lien vers la version anglaise.

Après avoir procédé à un alignement des ressources par l'intermédiaire de leurs URLs, nous avons téléchargé les pages correspondantes (150), puis nous les avons converties au format texte depuis HTML ou PDF[7]. Nous avons ensuite utilisé une méthode d'alignement au niveau des paragraphes fondée sur le parallélisme entre la structure d'une ressource et celle de sa traduction. En effet, pour la majorité des ressources, le premier paragraphe de la version française constitue la traduction du premier paragraphe de la version anglaise, et ainsi de suite. Nous avons donc procédé à l'alignement au niveau des paragraphes de manière automatique, modulo quelques ajustements réalisés manuellement pour rétablir le parallélisme de structure dans certaines ressources.

A l'issue de ces opérations, nous avons obtenu un corpus parallèle anglais/français du domaine médical d'environ 370.000 mots (soit ~2,9 Mo), aligné au niveau des paragraphes.

2.2 Corpus RCP

Le second corpus parallèle, RCP, a été constitué dans le cadre du projet PERTOMed[8] dont l'objectif est de produire et d'évaluer des ressources terminologiques et ontologiques dans plusieurs secteurs de la médecine tels que la réanimation chirurgicale, la périnatalité ou encore la pharmacovigilance, d'une part, et de développer des méthodes innovantes d'appariement de ces ressources, d'autre part.

La principale ressource développée l'a été dans le secteur de la pharmacovigilance, à partir de résumés des caractéristiques du produit (RCP). Dans ce domaine, l'EMEA (European Medicines Agency)[9] est une agence européenne qui assure une évaluation des données scientifiques sur les médicaments à l'échelle européenne. Le RCP de chaque médicament qui a fait l'objet d'une procédure d'autorisation de mise sur le marché au niveau européen est mis à disposition sur le site de l'EMEA dans chacune des langues de l'Union

[6] La société canadienne de pédiatrie (http://www.cps.ca), Santé Canada (http://www.hc-sc.gc.ca) et le ministère de la santé et des soins de longue durée de l'Ontario (http://www.gov.on.ca/health/indexf.html).

[7] A l'aide de gratuiciels disponibles sur Internet.

[8] Sous la responsabilité scientifique de Marie-Christine Jaulent, INSERM ERM 202 (http://www.spim.jussieu.fr, rubrique "Projets de Recherche")

[9] http://www.emea.eu.int

Européenne. La procédure d'autorisation doit respecter des impératifs scientifiques, d'une part, car les médicaments doivent être validés pour une indication donnée, et linguistiques, d'autre part, car l'information disponible dans chaque pays doit être la même quelle que soit la langue. Le corpus RCP répond donc aux mêmes critères de qualité que ceux retenus pour la construction du corpus CISMef/Hansard.

Il est constitué de 94 résumés dans chaque langue, le français et l'anglais. Il compte au total environ 600 000 mots (soit ~4,5 Mo). Chaque RCP étant organisé suivant une même structure hiérarchique de dix sections[10], nous avons pu mettre en place une procédure d'alignement automatique au niveau des paragraphes similaire à celle utilisée pour le corpus CISMef/Hansard.

3 Traduction des synonymes MeSH

3.1 Principe de base de la procédure d'appariement

Pour la recherche des traductions en français des synonymes MeSH américains, nous avons mis en oeuvre une méthode d'appariement de mots et de syntagmes dite « appariement par propagation syntaxique » (Ozdowska, 2004a ; Ozdowska 2004b). Il s'agit d'une approche linguistique d'appariement de segments sous-phrastiques basée sur l'analyse syntaxique bilingue de corpus parallèles anglais/français. Son principe est le suivant: à partir de deux mots qui sont en relation de traduction dans des phrases alignées, appelés couple amorce, le lien d'équivalence est propagé vers d'autres mots en suivant les relations syntaxiques préalablement mises en évidence. Plus précisément, en partant du couple amorce (*protective*, *protecteurs*), dont chaque élément est en relation syntaxique avec un nom, on peut apparier (*clothing, vêtements*) (FIG. 1)[11].

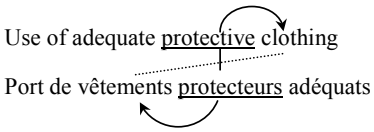

FIG. 1- Principe d'appariement par propagation syntaxique

La technique d'appariement par propagation syntaxique requiert par conséquent que l'on dispose d'un corpus aligné au niveau des phrases, d'outils d'analyse pour les deux langues de travail, le français et l'anglais, ainsi que de couples amorces servant de point de départ à la propagation.

3.2 Traitement des corpus

Le repérage des relations syntaxiques est pris en charge par les analyseurs Syntex (Bourigault et Fabre, 2000) qui prennent en entrée un corpus étiqueté[12] et identifient, pour

[10] Un RCP peut éventuellement contenir les caractéristiques de plusieurs dosages d'un même médicament, au quel cas seule la description d'un dosage a été prise en compte.
[11] Le sens des flèches correspond à celui des relations de dépendance syntaxique.
[12] L'étiqueteur utilisé pour les deux langues est Treetagger (http://www.ims.uni-stuttgart.de)

chaque phrase du corpus, des relations syntaxiques telles que sujet, objet direct et indirect, modifieur, etc. L'appariement s'effectue par conséquent entre des mots lemmes et non des mots formes.

Comme la plupart des méthodes travaillant au niveau sous-phrastique, la méthode d'appariement par propagation syntaxique nécessite un corpus préalablement aligné au niveau des phrases. Comme décrit dans la section 2, les corpus de travail dont nous disposons sont alignés de manière fiable uniquement au niveau des paragraphes. Le découpage en phrases étant pris en charge de manière indépendante dans chacune des deux langues par les étiqueteurs, l'alignement à ce niveau de segmentation est susceptible de présenter des erreurs. Nous avons pris le parti de ne pas corriger les éventuels décalages et avons ignoré, lors du processus de recherche des couples amorces ainsi que de celui de propagation, les phrases non alignées.

3.3 Expérimentation : identification des couples amorces et propagation syntaxique des liens d'appariement

Les couples amorces permettant d'initialiser le processus de propagation peuvent être fournis au système de différentes manières. Il est possible d'utiliser des ressources lexicales bilingues préexistantes, de construire de telles ressources à partir du corpus ou encore de repérer des cognats, c'est-à-dire des chaînes de caractères identiques ou très proches dans les deux langues. Nous avons, dans un premier temps, choisi de combiner la projection d'une ressource lexicale existante et la recherche de cognats (autres que ceux présents dans la ressource) au niveau des phrases alignées. En effet, nous disposions d'une liste constituée des descripteurs MeSH américains et de leur traduction en français (liste 1), dont nous avons extrait les mots simples[13]. Nous avons ainsi obtenu, à partir d'une liste de 6127 mots, 28139 couples amorces sur un ensemble de 10299 phrases alignées (TAB 1). Il convient de noter que seuls 556 couples de la liste de départ sont effectivement présents dans le corpus et ont donc pu être utilisés pour la recherche des amorces.

Dans un second temps, cette ressource nous est apparue comme insuffisante et ce principalement pour deux raisons. Premièrement, elle ne contient que des noms, ce qui implique que l'alignement ne peut concerner que des mots qui sont en relation syntaxique avec un nom. Par conséquent, si l'on considère l'exemple ci-dessous (FIG. 2), il apparaît clairement que l'équivalent français de *nightmare*, qui est l'un des synonymes dont on cherche la traduction, ne pourra être trouvé que si l'on dispose du couple amorce constitué des verbes *continue*/*durer*, à moins que l'on ne trouve ailleurs dans le corpus une ou plusieurs autres occurrences de *nightmare* et *cauchemar*, toutes deux en relation syntaxique avec des noms amorces.

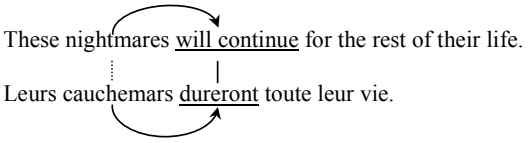

FIG. 2 Propagation syntaxique à partir d'un couple amorce de verbes

[13] Les règles de propagation utilisées actuellement sont fondées uniquement sur les mots simples.

Deuxièmement, les noms que cette liste contient relèvent pour la plupart d'un vocabulaire spécialisé relatif au domaine de la santé (FIG. 3), ceux relevant de la langue générale et susceptibles d'être présents dans le corpus étant peu représentés.

bromine	brome
bromizovalum	bromizoval
bromouracil	bromouracile
bronchography	brochographie

FIG. 3 – Descripteurs MeSH et leur traduction – extrait de la liste 1

Nous avons donc fait le choix de compléter la liste existante avec des données extraites du corpus (FIG. 4) afin d'étudier l'influence du nombre et de la diversité des couples amorces sur les alignements obtenus, en termes de catégories grammaticales (restriction aux noms pour la liste 1 *versus* toutes catégories confondues pour la liste 2) et de type de vocabulaire (spécialisé pour la liste 1 *versus* spécialisé et général pour la liste 2). Pour ce faire, nous avons utilisé une méthode largement répandue dans les travaux sur l'alignement basée sur l'hypothèse que les mots qui apparaissent fréquemment dans des segments de texte alignés ont de fortes chances d'être en relation de traduction (Gale et Church, 1991 ; Ahrenberg *et al*., 2000). Afin d'isoler des couples de mots en relation de traduction dans nos corpus, nous avons utilisé une mesure d'association. Comme pour les expériences précédentes, nous avons utilisé le Jaccard avec des seuils et des techniques de filtrage de la liste des associations obtenues identiques à ceux décrits dans (Ozdowska 2004a ; Ozdowska, 2004b) :

- calcul du Jaccard pour les mots dont la fréquence sur l'ensemble du corpus est égale ou supérieure à 5 ;
- sélection des associations pour lesquelles la valeur du Jaccard est égale ou supérieure à 0,2 ;
- filtrage de la liste de associations par reconnaissance des cognats et par vérification de la réciprocité de l'appariement.

die	mourir
monitor	surveiller
next	prochain
often	souvent

FIG. 4 –Traductions extraites du corpus par calcul des cooccurrences – extrait de la liste 2

Enfin, nous avons fusionné les deux listes (liste 2) pour obtenir au total un ensemble de 8866 couples de mots qui ont à leur tour été projetés au niveau des phrases alignées, ce qui a permis d'identifier 39903 couples amorces (TAB. 1). 2093 des couples de la liste 2 ont pu être pris en compte lors de la phase de repérage des couples amorces.

Une fois les couples amorces repérés, la propagation syntaxique des liens d'appariement repose sur différents patrons de propagation dont on a pour le moment limité la définition aux cas de correspondance directe, c'est-à-dire ceux où la configuration syntaxique est identique dans les deux langues. Comme décrit dans (Ozdowska 2004a), chaque patron rend compte de la catégorie grammaticale des mots sources et des mots visés par la propagation,

de la relation syntaxique qui sert de base à la propagation ainsi que du sens dans lequel cette dernière s'effectue.

Corpus	CISMeF / Hansard	RCP
nombre de phrases alignées	10299	18034
nombre de couples amorces (liste 1)	28139	51600
nombre de couples amorces (liste 2)	39903	72136

TAB. 2– Repérage des couples amorces : influence du lexique bilingue utilisé.

4 Résultats

A l'aide des deux corpus, une traduction a été proposée pour 217 synonymes au total, soit 4,2% de l'ensemble des synonymes à traduire. Après validation par un expert en terminologie médicale, 133 nouveaux termes ont été inclus dans la terminologie (plus des flexions et/ou dérivations de ces termes le cas échéant – soit 190 termes au total). La précision est de 71% sur le corpus Hansard/CISMeF, et de 70% sur le corpus RCP. Le rappel est respectivement de 54% et 60%. Ainsi, la méthode utilisée offre une précision globale de 70% pour un rappel global de 57%.

La figure FIG. 5 présente les performances obtenues en fonction du nombre d'occurrences des synonymes dans le corpus CISMeF/Hansard. Les points (*/°) représentent les valeurs de précision et de rappel pour un nombre d'occurrences donné. Afin de simplifier la représentation, les courbes en trait plein montrent la précision moyenne pour les termes de basse fréquence (une seule occurrence dans le corpus), de faible fréquence (entre deux et cinq occurrences), de moyenne fréquence (entre six et dix occurrences) et de haute fréquence (plus de 10 occurrences).

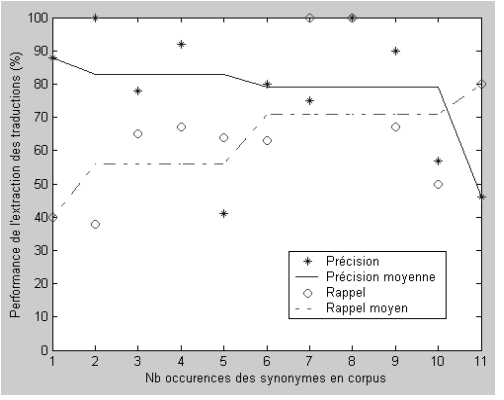

FIG. 5–Performance de la méthode en fonction du nombre d'occurrences des synonymes dans le corpus CISMeF/Hansard.

Le tableau TAB. 3 présente le nombre de traductions extraites pour chaque corpus, en fonction du lexique utilisé pour repérer les couples amorces. Les traductions proposées se recoupent parfois: la traduction d'un synonyme peut être extraite des deux corpus ou à l'aide des deux lexiques pour un corpus donné Nous indiquons ci-dessous le nombre total de synonymes distincts pour lesquels au moins une traduction a été extraite pour chaque corpus, puis pour l'ensemble des deux corpus.

	Lexique MeSH seul (liste 1)	Lexique MeSH + cooccurrences (liste 2)
	Corpus CISMeF/Hansard	
Nb traductions extraites	102	115
Nb total traductions	116 distincts	
Précision et Rappel	P = 71% - R = 54%	
	Corpus RCP	
Nb traductions extraites	139	146
Nb total traductions	148 distincts	
Précision et Rappel	P = 70% - R = 60%	
	Corpus CISMeF/Hansard et RCP	
Nb total traductions	217 distincts	
Précision et Rappel	P = 70% - R = 57%	

TAB. 3– Nombre de traductions extraites de chaque corpus en fonction du lexique .

5 Discussion

5.1 Performances de la méthode de traduction des synonymes MeSH

5.1.1 Analyse globale des résultats

Des résultats complémentaires sont proposés si l'on utilise la liste 2[14] pour le repérage des couples amorces (102 synonymes traduits vs. 115 avec la liste 1 seule pour le corpus CISMeF/Hansard). Il est donc souhaitable de travailler avec des couples amorces appartenant à des registres de langue et à des catégories grammaticales diverses afin d'optimiser les résultats.

En observant les performances de notre méthode en fonction du nombre d'occurrences des synonymes traduits dans le corpus, on constate que les résultats sont inégaux. La meilleure précision (88%) est obtenue pour les termes qui n'apparaissent qu'une seule fois dans le corpus. Ce résultat met en évidence un avantage de la méthode utilisée par rapport à des méthodes statistiques, comme par exemple les modèles IBM (Och et Ney, 2003), qui nécessitent un grand nombre d'occurrences des termes pour proposer un alignement fiable. Par contre, pour un grand nombre de ces termes, aucune traduction n'est proposée. La différence de rappel entre les deux corpus peut donc s'expliquer par le fait que le corpus Hansard/CISMeF comporte une proportion plus élevée de synonymes de fréquence 1 (45%) que le corpus RCP (35%). En revanche, pour les termes très fréquents, le rappel est optimal

[14] voir section 3.3

(80%). La précision est cependant moins intéressante du fait du nombre important de candidats erronés proposés.

Nous avons étudié plus précisément les causes de silence (mesure équivalente à 1 – rappel) et de bruit (mesure équivalente à 1 – précision).

5.1.2 Analyse du silence

L'impossibilité de trouver une traduction s'explique par l'une des raisons suivantes :

- Erreur d'alignement au niveau des phrases : dans près de 30% des cas, les termes non traduits se trouvent dans des phrases qui n'ont pu être correctement alignées.
- Différence structurelle entre les deux langues : les phrases présentent une différence de formulation qui ne permet pas la propagation.

Occluded dialysis access ***grafts***[15]
Occlusion des ***courts-circuits*** *artério-veineux (dialyse)*

- Appariement de type *m-n* : le nombre d'unités à mettre en correspondance est différent dans les deux langues :

[...] particularly if diarrhea is accompanied by weight loss, ***hematochezia****, ...*
*[...] en particulier si la diarrhée s'accompagne d'une perte de poids, de l'****émission de selles sanglantes*** *...*

Or, pour le moment, la méthode d'appariement employée ne permet de trouver que des correspondances entre des mots simples ou des syntagmes constitués de deux mots pleins. Tout autre type d'appariement, notamment les appariements *m-n*, n'est pas pris en compte.

- Absence de couple amorce : aucun couple amorce ne permet d'atteindre par propagation les termes que l'on cherche à traduire ou bien le couple amorce trouvé est incorrect :

[...] because of the low ***wages*** *paid to the maker.*
[...] en raison du maigre ***salaire*** *versé au fabricant.*

Ni le couple paid/versé ni low/maigre n'ont pu être extraits par le calcul du Jaccard à cause d'une fréquence de cooccurrence insuffisante.

- Longueur des termes : les termes que l'on cherche à apparier comportent plus de deux mots pleins. Ils ne sont donc pas pris en charge par la méthode d'appariement, comme nous venons de l'évoquer ci-dessus.
- Erreur d'analyse syntaxique : il y a une erreur d'analyse syntaxique dans l'une ou l'autre des deux langues et on ne dispose pas de la relation syntaxique nécessaire à la propagation. C'est le cas pour la préposition *par* en français qui n'a pu être rattachée a son recteur, *meurtre*, dans l'exemple suivant :

The issue of euthanasia, better known to many of us as ***mercy*** *killing...*
L'euthanasie, mieux connue de beaucoup sous le vocable de meurtre par ***compassion****...*

[15] Les couples amorces sont soulignés, les termes que l'on cherche à apparier sont en gras.

- Terme non traduit: le synonyme dont on cherche la traduction n'est pas traduit dans la langue cible mais repris tel que en langue source dans la partie cible du corpus.
- Absence de relation syntaxique : les occurrences de l'un ou des deux termes apparaissent dans des configurations syntaxiques isolées, par exemple entre virgules. Elle ne sont liées aux autres mots de la phrase par aucune relation syntaxique :

*[...] an increase in spontaneous abortion, **stillbirth**, or prematurity.*
*[...]un risque accru d'avortement spontané, de **mortinaissance** ou de prématurité.*

5.1.3 Analyse du bruit

Parmi les traductions erronées proposées, on retrouve plusieurs type d'erreurs:
- traduction incomplète: par exemple, "banque" proposé comme traduction du synonyme "databanks" pour lequel la traduction attendue était "banque de données". Ceci s'explique par le fait que la méthode ne peut traiter les appariements de type *m-n*, comme expliqué en 5.1.2.
- traduction par une forme dérivée du terme : par exemple, "varicelleuse" proposé comme traduction du synonyme "varicella" pour lequel la traduction attendue est "varicelle". Comme proposé dans (Debili, 1997), deux types d'appariements sont à distinguer. Les premiers mettent en relation des unités qui peuvent être considérées comme traduction l'une de l'autre aussi bien dans le contexte linguistique où elles apparaissent qu'en dehors de ce dernier. On parlera alors d'appariements non contextuels. Les seconds mettent en correspondance des unités qui peuvent être considérées comme traduction l'une de l'autre seulement dans le contexte linguistique où elles apparaissent. Il s'agit d'appariements contextuels. L'appariement varicella/varicelleuse relève du second type. Considérons l'un des couples de phrases dont cette correspondance a été extraite :

*Recurrences of varicella-like rash have been reported by 4% to 13% of individuals who had previous **varicella** infection.*
*Des cas récurrents d'éruption varicelliforme ont été observés chez 4% à 13% des personnes ayant déjà eu une infection **varicelleuse** .*

Il apparaît clairement que l'appariement varicella/varicelleuse, obtenu à partir du couple amorce infection/infection, ne résulte pas d'une erreur de l'algorithme de propagation mais qu'il s'agit d'un appariement contextuel « seulement recevable en contexte et non reconnu comme étant une traduction acceptée de manière générale, et dès lors répertoriée en principe dans un dictionnaire bilingue » (Debili, 1997).
- traduction par un terme relevant du même champ lexical ou par un hyperonyme: par exemple, "anticoquelucheux" ou "maladie" proposés comme traductions du synonyme "pertussis" pour lequel la traduction attendue était "coqueluche". Il s'agit, comme ci-dessus, d'appariements contextuels :

*[...] varicella rates of 3% to 4% per year are expected to occur after **varicella** vaccination.*
*[...] des taux annuels de varicelle de 3% à 4% après la vaccination **antivaricelleuse**.*

Nous avons, dans les deux cas, des différences de formulation dans les deux langues qui peuvent apparaître de manière récurrente dans le corpus, reflétant un usage établi.

5.2 Enrichissement de la terminologie

La dernière étape de notre travail a consisté à valider les synonymes obtenus par extraction des traductions avec un expert en terminologie médicale, afin de les inclure dans la terminologie CISMeF. Lors de cette phase de validation, nous avons rencontré plusieurs cas où il n'a pas été possible d'inclure la traduction proposée dans la terminologie bien que cette dernière ait été correcte.

Tout d'abord, certaines traductions se sont révélées ambiguës en français. Par exemple, le synonyme <*cirrhosis*> du mot clé américain <*fibrosis*> (en français, <*fibrose*>) a été traduit par "cirrhose" grâce à notre méthode. Le terme "cirrhosis" a donc été correctement traduit par "cirrhose" mais, en français, "cirrhose" a une connotation restreinte qui se limite à la cirrhose du foie – d'ailleurs, le mot clé MeSH américain <*liver cirrhosis*> est traduit en français par <*cirrhose*>. Il n'est donc pas possible d'utiliser "cirrhose" comme synonyme de "fibrose".

Ensuite, certaines traductions se sont révélées redondantes à cause d'une différence de champ lexical entre les deux langues. Par exemple, le synonyme <*scar*> du mot clé américain <*cicatrix*> (en français, <*cicatrice*>) a été traduit par "cicatrice" grâce à notre méthode. Cette traduction est correcte. Cependant en français, il n'existe qu'un seul terme pour désigner le concept « cicatrice » contrairement à l'anglais où il en existe deux : le terme scientifique « cicatrix » et le terme courant « scar ». Ainsi, en français, le terme et le synonyme américains ne peuvent être différenciés.

Inversement, pour certains synonymes, plusieurs des traductions proposées ont pu être incluses dans la terminologie. Par exemple, le synonyme <*cannabis smoking*> du mot clé américain <*marijuana smoking*> (en français, <*consommation de marijuana*>) a été traduit grâce à notre méthode par <*consommation de cannabis*> et <*inhalation de cannabis*>, qui ont tous les deux été inclus dans la terminologie.

6 Conclusion

Afin d'enrichir une terminologie médicale francophone, nous avons proposé et mis en œuvre une méthode de traduction automatique de termes MeSH américains à l'aide de deux corpus parallèles du domaine. Nous avons pu ajouter 133 synonymes à la terminologie CISMeF. La méthode d'extraction des traductions par propagation des liens d'équivalence à l'aide des relations syntaxiques offre une précision globale de 70% pour un rappel de 57%.

Une analyse plus fine des résultats montre que ces performances sont optimales pour la traduction de termes de fréquence moyenne (autour de 5 occurrences). Pour les termes de basse fréquence, le rappel est faible, et inversement, pour les termes très fréquents, la précision chute. Ces éléments seront à prendre en compte pour la constitution de futurs corpus pour la poursuite de ce travail.

Par ailleurs, il apparaît également que l'enrichissement de la terminologie ne relève pas d'une simple traduction des synonymes. Il est impératif de tenir compte de la complexité lexicale et de l'usage des termes dans les deux langues afin de valider l'ajout de synonymes correctement traduits.

Pour la poursuite de ces travaux, nous envisageons également la traduction automatique compositionnelle des synonymes ne se trouvant pas dans le corpus, mais dont les constituants se trouvent dans le corpus – ainsi, nous pourrions déduire la traduction d'un synonyme à partir des traductions respectives des mots qui le composent.

Références

Ahrenberg L., Andersson M., Merkel M. (2000), A knowledge-lite approach to word alignment, Véronis J. (Ed.), Parallel Text Processing : Alignment and Use of Translation Corpus, Dordrecht: Kluwer Academic Publishers, pp. 97-138.

Bourigault D., Fabre C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de Grammaire, 25, pp. 131-151, Université Toulouse le Mirail.

Darmoni SJ, Leroy JP, Thirion B, Baudic F, Douyère M, Piot J. (2000), CISMeF: a structured Health resource guide, Methods of Information in Medicine, 39(1), pp 30-5.

Darmoni, SJ., Jarousse, E., Zweigenbaum, P., Le Beux P., Namer, F., Baud, R., Joubert M., Vallée H., Cote RA., Buemi A., Bourigault D., Recourcé G., Jenneau S., and Rodrigues JM. (2003), VUMeF: Extending the French Involvement in the UMLS Metathesaurus. Proceedings of AMIA Symp. 2003;:824.

Debili F. (1997), L'appariement : quels problèmes?, Actes des 1ères JST FRANCIL de l'AUPELF-UREF, pp. 199-206.

Gale WA., Church KW. (1991), Identifying Word Correspondences in Parallel Text, Proceedings of the DARPA Workshop on Speech and Natural Language.

Névéol A. (2004), Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé, Actes de RECITAL, pp 105-14.

Och FJ., Ney H. (2003), A systematic comparison of various statistical alignment models, Computational Linguistics, 29(1), pp. 19-51.

Ozdowska S. (2004a), Appariement bilingue de mots par propagation syntaxique à partir de corpus français/anglais alignés, Actes de RECITAL pp 125-34.

Ozdowska S. (2004b), Identifying correspondences between words: an approach based on a bilingual syntactic analysis of French/English parallel corpus, Proceedings of the Workshop on Multilingual Linguistic Resources, COLING'04.

Remerciements

Ce travail a été réalisé dans le cadre du projet VUMeF, qui bénéficie d'un financement du Réseau National Technologies pour la Santé (RNTS).

Summary

The CISMeF catalogue (French acronym for "catalogue and index of online medical resources in French") displays and indexes the major resources related to health in French. The resources are described and accessed with the CISMeF terminology, which is founded on the Medical Subject Headings (MeSH), the reference thesaurus for the bio-medical domain. The French version of the MeSH includes all headings and subheadings, but numerous synonyms remain to be translated. In this paper, we propose a automatic method for translating these synonyms. First, we have collected a French/English parallel corpus of medical resources. Then, the corpus was aligned semi-automatically at the paragraph level. Eventually, a lexicon composed of MeSH headings and their translation was used as a set of anchor words for the syntactic propagation of alignment links. As a result, 217 synonyms could be translated and the method achieved 70% precision.