

# Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés

Sylwia Ozdowska<sup>1</sup>, Aurélie Névéol<sup>2,3</sup>, Benoît Thirion<sup>3</sup>

<sup>1</sup> ERSS – CNRS & Université de Toulouse le Mirail

<sup>2</sup> Laboratoire PSI FRE CNRS 2645 – INSA & Université de Rouen

<sup>3</sup> Equipe CISMef & L@STICS, CHU de Rouen

ozdowska@univ-tlse2.fr, aneveol@insa-rouen.fr, Benoit.Thirion@chu-rouen.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 avons mis en œuvre une méthode de traduction compositionnelle automatique des synonymes. Pour ce faire, nous avons constitué deux corpus alignés anglais/français du domaine médical puis nous avons procédé automatiquement à l'appariement bilingue de mots et à la traduction compositionnelle des bitermes de type Adj N. La technique d'appariement de mots utilisée est basée sur une analyse syntaxique de corpus alignés. 915 synonymes distincts de type Adj N ont pu être traduits, avec une précision globale de 73%, moyennant une sélection des traductions des composants au rang 1.

**Mots-clés :** traduction de terminologies, traduction automatique, traduction compositionnelle, alignement de mots, corpus alignés, anglais, français, domaine médical

---

## 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<sup>1</sup> pour la gestion de connaissances, ou la SNOMED CT<sup>2</sup> pour les termes cliniques) ou ontologies (par exemple GO<sup>3</sup>) sont disponibles. Bien que ces différentes terminologies soient complémentaires, on observe également des recouvrements 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

---

<sup>1</sup> Medical Subject headings. cf. <http://www.nlm.nih.gov/mesh/meshhome.html>

<sup>2</sup> SNOMED Clinical Terminology. cf. <http://www.nhsia.nhs.uk/snomed/pages/default.asp>

<sup>3</sup> Gene Ontology - cf. <http://www.geneontology.org>

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 dans 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 nous intéressons à l'enrichissement de la terminologie CISMeF par la traduction automatique des synonymes américains du MeSH. Le Tableau 1 présente un échantillon des termes MeSH actuellement disponibles en français et en anglais.

Mot clé MeSH américain	Mot clé MeSH français	Synonyme MeSH américain (à traduire)
cardiovascular agents	agents cardiovasculaires	cardiovascular drugs
cell division	division cellulaire	cytokineses
milk, human	lait femme	breast milk
skin diseases, vesiculobullous	dermatoses bulleuses	sneddon wilkinson disease
terpenes	terpenes	isoprenoids

Tableau 1 - Extrait du MeSH 2004

Après une première expérience positive sur la traduction directe des synonymes dans les cas où ceux-ci sont effectivement présents dans les corpus (Névéol et Ozdowska, 2005), nous étendons ce travail en proposant une méthode de traduction compositionnelle. Cette dernière concerne les synonymes qui n'apparaissent pas dans les corpus en tant que tels mais dont les composants y sont néanmoins présents. Nous présentons dans un premier temps le principe de traduction compositionnelle mis en œuvre (section 2). Puis, nous décrivons les corpus de travail (section 3) et détaillons la méthode utilisée pour aligner les mots et extraire la traduction compositionnelle des synonymes à partir de ces corpus (section 4). Enfin, nous faisons un bilan des résultats obtenus et nous discutons de l'apport terminologique réalisé (sections 5 et 6) avant de conclure sur les perspectives de poursuite de ce travail (section 7).

## 2. Traduction compositionnelle : pourquoi ?

Nous avons constaté que sur l'ensemble des termes à traduire, soit 20 048, seule une petite partie est effectivement présente dans les corpus de travail décrits en section 3, soit environ 300 pour le corpus CISMeF<sup>4</sup> et 407<sup>5</sup> pour le corpus des résumés des caractéristiques du produit (Névéol et Ozdowska, 2005). Par ailleurs, bien qu'un grand nombre de termes complexes n'apparaissent pas dans ces corpus, il n'en va pas de même pour leurs constituants. Ainsi, nous avons évalué que les constituants de près de 5000 synonymes se trouvent dans chacun des corpus de travail. Par exemple, bien que le terme *premature birth* ne soit présent ni dans l'un ni dans l'autre des deux corpus, on trouve ses composants, *premature* et *birth*, notamment dans les contextes suivants :

<sup>4</sup> Cet article présente les résultats sur une sous partie du corpus CISMeF/Hansard utilisé dans (Névéol et Ozdowska, 2005).

<sup>5</sup> Données MeSH 2004.

*They may care for immunocompromised patients (including **premature** infants)...*  
*Ils peuvent s'occuper de patients immunodéprimés (y compris de bébés **prématurés**)...*

*The infant can be vaccinated at **birth**...*  
*L'enfant pourra être vacciné après sa **naissance**...*

C'est pourquoi, afin d'augmenter la couverture des termes complexes traduits, nous avons envisagé de déduire la traduction des synonymes à partir des traductions respectives des mots qui les composent : c'est la traduction compositionnelle (Figure 1).

*They may care for immunocompromised patients (including **premature** infants)...*  
*Ils peuvent s'occuper de patients immunodéprimés (y compris de bébés **prématurés**)...*

*premature* ⇔ *prématuré*

*The infant can be vaccinated at **birth**...*  
*L'enfant pourra être vacciné après sa **naissance**...*

*birth* ⇔ *naissance*

*premature birth* ⇔ *naissance prématurée*<sup>6</sup>

Figure 1 – Traduction compositionnelle

Nous reprenons ainsi le principe de compositionnalité (Frege, 1982 ; Janssen, 1997) selon lequel le sens d'un terme complexe résulte directement du cumul des sens des unités qui le composent et nous l'appliquons à la traduction des termes complexes. En effet, nous tenons pour acquis que ce principe s'observe dans les deux langues mises en correspondance. Autrement dit, nous faisons l'hypothèse que la traduction des composants des termes complexes anglais pris individuellement permet d'inférer des termes complexes équivalents corrects en français.

### 3. Corpus

Afin de constituer des corpus de travail adaptés à notre problématique, nous avons porté une attention particulière aux critères suivants : la qualité de la traduction et l'adéquation du contenu avec le domaine médical (plus spécifiquement, avec les concepts concernés par les synonymes à traduire).

#### 3.1. CISMef

Le catalogue CISMef<sup>7</sup> indexe uniquement des ressources<sup>8</sup> 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 URL 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 contiennent un résumé anglais d'un article développé en français, ou présentent 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<sup>9</sup>, ce qui est une garantie de la qualité de la traduction disponible. De plus, ces sites observent un classement régulier et

<sup>6</sup> Le changement de flexion a fait l'objet d'un ajustement manuel le cas échéant.

<sup>7</sup> <http://www.cismef.org>

<sup>8</sup> Nous entendons par "ressource" un document en ligne (au format html, pdf, etc.) ou un site web entier.

<sup>9</sup> 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>).

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 URL, nous avons téléchargé les pages correspondantes (150), puis nous les avons converties au format texte. Nous avons ensuite découpé le corpus parallèle ainsi obtenu en phrases et l'avons aligné à ce niveau de segmentation<sup>10</sup>.

Ce premier corpus aligné anglais/français du domaine médical (corpus CISMef) comporte 7141 couples de phrases, soit environ 260 000 mots.

### **3.2. RCP**

Le second corpus parallèle a été constitué dans le cadre du projet PERTOMed<sup>11</sup> 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) qui contiennent des informations sur les indications, les effets indésirables ou encore les interactions des médicaments. Dans ce domaine, l'EMA (European Medicines Agency)<sup>12</sup> 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'EMA dans chacune des langues de l'Union 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. Ce corpus (corpus RCP) répond donc aux mêmes critères de qualité que ceux retenus pour la construction du corpus CISMef.

Il est constitué de 94 résumés dans chaque langue, le français et l'anglais, et a également fait l'objet d'un découpage puis d'un alignement au niveau des phrases. Il compte 17 806 couples de phrases pour un total d'environ 600 000 mots.

## **4. Traduction des synonymes MeSH**

### **4.1. Appariement de mots**

#### *4.1.1 Principe de base*

Pour la recherche des traductions en français des constituants des synonymes MeSH américains, nous avons mis en œuvre une méthode d'appariement dite « appariement par propagation syntaxique » (Ozdowska, 2004). 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

---

<sup>10</sup> Nous avons utilisé l'aligneur Japa développé au Laboratoire de Recherche Appliquée en Linguistique Informatique (<http://rali.iro.umontreal.ca/Japa>)

<sup>11</sup> Sous la responsabilité scientifique de Marie-Christine Jaulent, INSERM ERM 202 (<http://www.spim.jussieu.fr>, rubrique "Projets de Recherche")

<sup>12</sup> <http://www.emea.eu.int>

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 appairer (*clothing, vêtements*) (Figure 2).

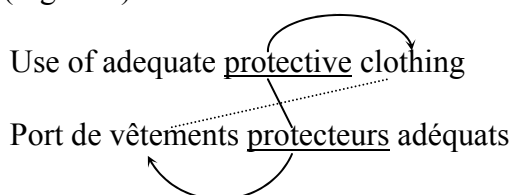


Figure 2 - Principe d'appariement par propagation syntaxique

Le repérage des relations syntaxiques est effectué par les analyseurs Syntex (Bourigault et Fabre, 2000) qui prennent en entrée un corpus étiqueté<sup>13</sup> et identifient, pour 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.

Différentes techniques peuvent être utilisées pour obtenir des couples amorces qui servent de point de départ à la propagation.

#### 4.1.2 Repérage des couples amorces

Pour identifier les couples amorces, il est possible d'utiliser des ressources lexicales bilingues existantes, 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 observé dans notre étude précédente (Névéol et Ozdowska, 2005) que l'utilisation d'un lexique bilingue, qui comprend aussi bien un vocabulaire propre au domaine et/ou spécifique au corpus qu'un vocabulaire plus général, combinée à la recherche de cognats permettait une couverture plus grande (plus de synonymes traduits) pour une précision équivalente. Nous avons donc utilisé un lexique composé de :

- descripteurs MeSH américains et leur traduction en français. Parmi l'ensemble des descripteurs possibles, nous n'avons retenu que les mots simples<sup>14</sup>. Il s'agit donc d'un lexique limité aux noms qui relèvent du vocabulaire spécialisé de la médecine.
- couples de mots en relation de traduction dans nos corpus. Afin de les isoler, nous avons utilisé un calcul de cooccurrence des mots dans les couples de phrases alignées (Ahrenberg *et al.* 2000 ; Gale et Church, 1991). La mesure d'association choisie est le Jaccard ; les seuils et les techniques de filtrage de la liste des associations obtenues sont identiques à ceux décrits dans (Ozdowska, 2004b).

Le nombre de couples amorces obtenu est de 33 351 et 112 322 respectivement pour le corpus CISMef et le corpus RCP.

#### 4.1.3 Propagation syntaxique

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

<sup>13</sup> L'étiqueteur utilisé pour les deux langues est Treetagger (<http://www.ims.uni-stuttgart.de>)

<sup>14</sup> Les règles de propagation utilisées actuellement sont fondées uniquement sur les mots simples.

syntaxique qui sert de base à la propagation ainsi que du sens dans lequel cette dernière s'effectue. Dans la mesure où nous nous sommes limités pour le moment à la traduction compositionnelle des synonymes de type Adj N (section 4.2), seuls les patrons qui visent à aligner des noms et des adjectifs ont été pris en compte.

#### 4.2. Application du principe de traduction compositionnelle

Le principe de traduction compositionnelle introduit en section 2 présente bien évidemment des limites comme le soulignent de nombreux travaux en traduction automatique, notamment (Rosetta, 1994). En effet, même si l'on connaît la traduction française de chacun des constituants du terme *breast milk* (*breast* se traduit par *sein* et *milk* par *lait*), il est difficile d'en déduire la traduction du terme qui est *lait maternel*. A la non-compositionnalité du sens s'ajoute alors la non-correspondance structurelle entre les termes (ici, la structure N N du terme en anglais diffère de la structure N Adj de sa traduction en français). Néanmoins, comme le montre Daille (1994), il existe des régularités dans la manière de rendre les structures syntaxiques des termes anglais en français. Par exemple, les termes anglais de structure Adj N sont régulièrement traduits en français par des termes de structure N Adj. D'autres correspondances, telles que N1 Prep (Det) N2 ou N1 de N2, sont beaucoup plus rares. Gaussier (2001) estime la probabilité de correspondance anglais/français Adj N/N Adj à 0,84. Nous faisons donc l'hypothèse que la correspondance structurelle va de pair, dans la plupart des cas, avec le principe de compositionnalité.

Afin de tester le principe de traduction compositionnelle, nous nous sommes limités dans un premier temps aux termes de structure Adj N. Le corpus CISMef contient 489 synonymes de ce type pour lesquels chaque constituant a pu être traduit grâce à la procédure d'appariement de mots décrite en 4.1 (resp. 635 synonymes pour le corpus RCP<sup>15</sup>). Nous avons calculé la traduction de ces synonymes selon la méthode suivante :

**T(Adj N) = T(N) + T(Adj) où T(N) et T(Adj) correspondent à la traduction de N et Adj**

Si on reprend l'exemple de la Figure 1 :

$T(\text{premature birth}) = T(\text{birth}) + T(\text{premature})$

$T(\text{premature birth}) = \text{naissance prématurée}$

Dans le cas où plusieurs équivalents français sont proposés pour l'un et/ou les deux composants N et Adj du synonyme à traduire, nous avons fait le choix de ne retenir que la première traduction proposée, à savoir la plus fréquente, nommée par la suite « candidat de rang 1 ». Nous discuterons dans la section 6 des avantages et inconvénients qui découlent de ce choix.

## 5. Résultats et évaluation de la traduction compositionnelle

### 5.1. Méthode d'évaluation

L'évaluation de notre travail s'effectue en deux étapes. Dans un premier temps, nous avons évalué la traduction des termes simples (N et Adj) puis la traduction compositionnelle des synonymes MeSH du type Adj N qui en est déduite. L'évaluation des traductions des composants puis des synonymes a été réalisée manuellement sur un échantillon représentatif de taille 100. Chaque traduction extraite est jugée correcte ou erronée par un traducteur (AN), et une vérification en dictionnaire ou en corpus est effectuée. Dans le cas de composants polysémiques, toutes les traductions possibles ont été considérées comme correctes dans la

---

<sup>15</sup> Soit au total 915 synonymes distincts, si on tient compte des termes présents dans les deux corpus.

mesure où, hors contexte, il n'est pas possible de préférer l'une à l'autre. Ainsi, les traductions *médicament* et *drogue* du composant *drug* sont considérées comme correctes. Par contre, la traduction *drogue cardiovasculaire* du terme *cardiovascular drug* a été considérée comme erronée car, dans ce contexte, seul le terme *médicament cardiovasculaire* est une traduction correcte. Pour la traduction compositionnelle, nous avons classé les erreurs de traduction rencontrées en deux catégories : erreurs dues à la traduction erronée d'au moins un composant et erreurs dues à la non-compositionnalité.

Dans un second temps, nous avons procédé à une validation manuelle des synonymes obtenus par traduction avec un expert en terminologie médicale (BT), afin de les inclure dans la terminologie CISMéF. L'expert ne prend en compte que le terme et le synonyme proposé et juge si ce dernier est valable ou non.

Les deux étapes de la validation sont manuelles et il faut remarquer qu'il existe un biais inévitable constitué par le jugement (dans certains cas subjectif) de l'expert, traducteur ou terminologue.

## **5.2. Résultats**

Le Tableau 2 présente la précision de la méthode employée pour la traduction des composants simples (noms et adjectifs) utilisés par la suite pour la traduction compositionnelle. La colonne 1 indique la précision obtenue si tous les candidats extraits sont pris en compte. En moyenne, 1,5 candidats ont été proposés pour chaque adjectif du corpus CISMéF (resp. 2,1 pour le corpus RCP), et 2,7 candidats ont été proposés pour chaque nom du corpus CISMéF (resp. 3,4 pour le corpus RCP). La colonne 2 indique la précision obtenue en considérant seulement les candidats de rang 1 (les plus fréquents).

	<b>Tous les candidats</b>	<b>Candidats rang 1</b>
	<b>Corpus CISMéF</b>	
Noms	39%	80%
Adjectifs	64%	85%
	<b>Corpus RCP</b>	
Noms	31%	89%
Adjectifs	57%	91%

*Tableau 2 – Précision de la traduction des composants par corpus.*

Le Tableau 3 présente une évaluation quantitative de cent traductions compositionnelles pour chaque corpus.

<b>Corpus</b>	<b>CISMéF</b>	<b>RCP</b>
Traduction correcte	<b>72</b>	<b>75</b>
<i>dont synonymes validés</i>	46	47
Traduction erronée	<b>28</b>	<b>25</b>
<i>dont traduction erronée d'un composant</i>	24	12
<i>dont erreur due à la compositionnalité</i>	4	13

*Tableau 3 – Évaluation de 100 traductions compositionnelles par corpus.*

## 6. Discussion

### 6.1. Performances de la méthode de traduction des synonymes MeSH

#### 6.1.1 Analyse globale des résultats

Nous constatons que la précision de la traduction des composants est nettement meilleure si on ne considère que les candidats de rang 1, plutôt que l'ensemble des candidats (pour les adjectifs, 85% au rang 1 vs. 64% en considérant tous les candidats pour le corpus CISMéF, et 91% au rang 1 vs. 57% en considérant tous les candidats pour le corpus RCP). On peut cependant remarquer une différence de précision notable sur l'ensemble des candidats entre les adjectifs et les noms. Cette différence vient en partie du grand nombre de candidats proposés pour certains noms – par exemple, pour le corpus RCP, 15 candidats pour *vaccine* (dont trois traductions correctes) ou, pour le corpus CISMéF, 26 candidats pour *maladie* (dont trois traductions correctes).

En terme de précision, la sélection des candidats de rang 1 constitue un avantage certain pour la traduction compositionnelle et semble par conséquent pleinement justifiée. Par contre, elle peut représenter un inconvénient en terme de rappel. En effet, elle ne permet de proposer qu'une traduction et une seule pour chaque synonyme américain alors que, dans certains cas, plusieurs traductions sont disponibles et pourraient être retenues par le terminologue. Ainsi, pour le terme *stable population*, la traduction compositionnelle proposée suite à la sélection des candidats de rang 1 est *population stable*. Or, si l'on prend en compte les candidats de rang supérieur à 1, on obtient une autre traduction correcte, à savoir *population constante*. Il en est de même pour le synonyme *vaginal injuries*, correctement traduit par *lésions vaginales*, pour lequel une autre traduction pourrait être obtenue en tenant compte des autres candidats pour la traduction de *injuries* : *atteintes vaginales*. L'impact de la sélection au rang 1 sur le rappel, mais aussi sur la précision (section 6.1.2), mérite donc d'être étudié de manière plus approfondie, notamment à la lumière des travaux tels que ceux menés en recherche d'information par Hull et Grefenstette (1996), par exemple, qui tiennent compte de tous les candidats et opèrent un filtrage en corpus pour ne conserver que les termes pertinents pour la requête.

Statistiquement, on pouvait attendre une précision de  $0,84 * 0,85 * 0,80^{16}$ , soit environ 57% pour le corpus CISMéF (resp. 68% pour RCP). Les résultats obtenus (72% et 75% respectivement) sont de cet ordre, et même supérieurs. La correspondance entre Adj N et N Adj anticipée est effectivement observée : au moins 72 cas sur 100 pour le corpus CISMéF (resp. 75 pour RCP). Pour les cas où la traduction compositionnelle proposée était incorrecte, nous avons étudié plus précisément les sources d'erreur. Celles-ci sont liées soit à la traduction des composants, soit aux limites de la compositionnalité.

#### 6.1.2 Erreur de traduction des composants

Nous avons identifié trois principaux types d'erreur pour ce qui est de la traduction des composants :

---

<sup>16</sup> La probabilité qu'un terme de la forme Adj N soit correctement traduit par compositionnalité selon notre méthode dépend de la probabilité qu'il y ait une correspondance entre ce terme et sa traduction en N Adj (0,84 d'après (Gaussier, 2001)), de la probabilité que le premier terme soit correctement traduit (estimée à 0,85 pour le corpus CISMéF d'après les résultats du tableau 1) et de la probabilité que le second terme soit correctement traduit (estimée à 0,80).

## Traduction compositionnelle automatique de bitermes

- erreur de traduction de l'un/des deux composants. Par exemple, dans *agricultural crops*, la traduction proposée pour *crops* est *poussées*. Il s'agit en effet de l'une des traductions possibles pour ce mot dans les corpus ; on la trouve notamment dans le contexte *appear in successive crops/apparaissent en poussées successives*. Par contre, ce n'est pas la traduction attendue dans *agricultural crops*. L'équivalent français résultant de la traduction compositionnelle, *poussées agricoles*, est donc erroné.
- erreur de traduction de l'un/des deux composants ayant pour origine la sélection au rang 1. La traduction proposée pour *alcoholic cirrhosis* est *cirrhose alcoolisée* alors que la traduction correcte est *cirrhose alcoolique*. Le bon équivalent pour *alcoholic, alcoolique*, apparaît en seconde position dans la liste des traductions possibles. La sélection se faisant au rang 1, seule la traduction la plus fréquente, *alcoolisée*, est prise en compte.
- erreur de traduction de l'un/des deux composants ayant pour origine sa/leur polysémie. Le terme *drug*, par exemple, se traduit en français soit par *drogue* soit par *médicament*. Dans les corpus de travail, c'est son emploi au sens de *drogue* qui est le plus fréquent, le sens *médicament* étant moins représenté. Ainsi, le terme *cardiovascular drug* a été traduit par *drogue cardiovasculaire, médicament cardiovasculaire* étant la traduction correcte attendue.

Même si la précision est globalement meilleure lorsque l'on sélectionne la traduction des composants au rang 1, il apparaît que, tout comme pour le rappel (section 6.1.1), ce critère est dans certains cas trop contraignant.

### 6.1.3 Limites de la compositionnalité

Les erreurs de traduction rencontrées confirment deux limites du principe de compositionnalité :

- échec de la correspondance structurelle Adj N/N Adj. Par exemple, le synonyme *bacterial count* a été traduit par *nombre bactérien*, la traduction attendue étant *nombre de bactéries*. Nous avons donc affaire à une correspondance structurelle de type Adj N/N de N, dont la probabilité est estimée à seulement 0,13 dans (Gaussier, 2001). On voit ici que la correspondance des structures limite le nombre de traductions qui peuvent être obtenues avec notre méthode. Il peut arriver que la "formulation préférée" ne soit pas trouvée, au profit d'une forme moins usitée. Par exemple, le synonyme *medical school* a été traduit par *école médicale* qui est un terme peu employé par rapport à la traduction la plus usitée, *école de médecine*, qui a une structure différente de N Adj.
- échec de l'équivalence des sens. Le synonyme *cold sore* a été traduit par *lésion froide vs. feu sauvage* qui est l'une des traductions correctes. Il s'agit ici d'une collocation du même type que *breast milk* : la traduction du tout ne peut être déduite de la traduction des composants. L'autre traduction correcte, *bouton de fièvre*, présente à la fois le problème de l'équivalence structurelle et sémantique.

## 6.2. Enrichissement de la terminologie

Lors de la phase de validation des synonymes traduits, 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. Comme dans notre expérience précédente, ces cas relevaient soit de la différence de complexité lexicale entre les deux langues, soit d'une différence d'usage.

### 6.2.1 Différence de complexité lexicale

Certains concepts peuvent être désignés par deux termes synonymes en anglais pour lesquels il n'existe qu'une seule et même désignation en français. Par exemple, le synonyme <microbiological phenomena> du mot clé américain <microbiologic phenomena> (en français, <phénomène microbiologique>) a été traduit par *phénomène microbiologique* grâce à notre méthode. Cette traduction est correcte. Cependant en français, il n'existe qu'un seul terme pour désigner le concept «phénomène microbiologique» contrairement à l'anglais où il en existe deux : le terme formé sur l'adjectif *microbiological* et le terme formé sur l'adjectif *microbiologic*. Ainsi, en français, le terme et le synonyme américains ne peuvent être différenciés.

### 6.2.2 Différence d'usage

Bien que certains termes soient des équivalents traductionnels en français et en anglais, l'usage des termes dans chaque langue met au jour un glissement de sens, et la traduction, bien que correcte, peut finalement se révéler ambiguë en français. Par exemple, le synonyme <intravenous infusion> du mot clé américain <intravenous perfusion> (en français, <perfusion intraveineuse>) a été traduit par *injection intraveineuse* grâce à notre méthode. Cette traduction est acceptable dans la mesure où, en anglais, les termes *perfusion/injection/infusion* semblent être plus facilement utilisés les uns pour les autres, en dépit de la différence conceptuelle entre *perfusion/infusion* (administration continue) et *injection* (administration instantanée) qui, en français, est beaucoup plus marquée. La traduction *injection intraveineuse* reflète l'imprécision dont font preuve certains locuteurs anglophones dans l'usage de ces termes. Par ailleurs, il existe bien un mot clé MeSH différent pour désigner le concept d'administration instantanée : <intravenous injection> en anglais, et <injection intraveineuse> en français. Il n'est donc pas possible d'utiliser <injection intraveineuse> comme synonyme de <perfusion intraveineuse> sans introduire une ambiguïté dans la terminologie.

Dans une optique de validation semi-automatique des résultats obtenus, il semble alors pertinent de filtrer les cas où le synonyme traduit sera automatiquement rejeté. Pour cela, il suffit de vérifier d'une part que le synonyme traduit est différent du mot clé auquel il se rapporte et d'autre part que le synonyme traduit ne correspond à aucun autre mot clé de la terminologie.

## 7. Conclusion

Afin d'enrichir une terminologie médicale francophone, nous avons proposé et mis en œuvre une méthode de traduction automatique compositionnelle de termes MeSH américains à l'aide de deux corpus parallèles du domaine. Nous avons pu ajouter 91 synonymes à la terminologie CISMef issus de l'échantillon de 200 traductions validées. La méthode de traduction compositionnelle offre une précision globale de 73% sur cet échantillon, et ce moyennant une sélection des traductions des composants au rang 1. L'application du principe de compositionnalité et de correspondance structurelle entre les termes s'avère par conséquent pertinente pour la traduction en français des synonymes MeSH américains de structure Adj N.

Une analyse plus fine des résultats montre que si, de manière globale, elle améliore considérablement la précision, l'application du critère de sélection au rang 1 s'avère parfois trop contraignante. En effet, dans certains cas elle ne permet pas de proposer la bonne traduction alors que celle-ci peut être trouvée en tenant compte des équivalents de rang inférieur ; dans d'autres, elle ne permet de retenir qu'une seule traduction là où plusieurs sont possibles et seraient susceptibles d'être retenues dans la terminologie.

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 d'affiner les critères de sélection des traductions des composants mais aussi d'étendre la traduction compositionnelle aux termes de structure autre que Adj N, qu'il s'agisse de termes constitués de deux mots pleins ou plus. Enfin, nous souhaitons mettre nos techniques d'extraction automatique des traductions à l'épreuve d'un nouveau corpus : le corpus parallèle CESART.

## **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).

## **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 Corpora*, Dordrecht: Kluwer Academic Publishers, pp. 97-138.
- BOURIGAUT D., FABRE C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, *Cahiers de Grammaire*, 25, pp. 131-151, Université Toulouse le Mirail.
- DAILLE B. (1994), *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*, Thèse de doctorat en Informatique fondamentale, Université Paris 7.
- DARMONI S. J., LEROY J. P., 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, S. J., JAROUSSE E., ZWEIGENBAUM P., LE BEUX P., NAMER F., BAUD R., JOUBERT M., VALLÉE H., COTE R. A., BUEMI A., BOURIGAUT D., RECOURCÉ G., JENNEAU S., and RODRIGUES J. M. (2003), VUMeF: Extending the French Involvement in the UMLS Metathesaurus, *Proceedings of AMIA Symp. 2003*.
- FREGE G. (1982), On Sense and Reference. In P. Geach and M. Black eds. 1970. *Translations from the Philosophical Writings of Gottlob Frege*, Oxford: Blackwell.
- GALE W. A., CHURCH K. W. (1991), Identifying Word Correspondences in Parallel Text, *Proceedings of the DARPA Workshop on Speech and Natural Language*.
- GAUSSIÉ E. (2001), General considerations on bilingual terminology extraction, D. Bourigault, Ch. Jacquemin, M.-C. L'Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 167 – 183.
- HULL D. A., GREFFENSTETTE G. (1996), Querying Across Languages: A Dictionary-based Approach to Multilingual Information Retrieval, *Proceedings of the 19<sup>th</sup> ACM SIGIR*, pp. 49-57.
- JANSSEN T. M. V. (1997), Compositionality, *Handbook of Logic and Language*, Elsevier, Amsterdam and MIT Press, Cambridge, J. Van Benthem & A. ter Meulen (Eds), p417-473.
- NEVEOL A. (2004), Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé, *Actes de RECITAL*, pp 105-14.
- NEVEOL A., OZDOWSKA S. (2005), Extraction bilingue de termes médicaux dans un corpus parallèle anglais/français, *Actes des 5<sup>èmes</sup> Journées d'Extraction et Gestion des Connaissances*, Paris.
- 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 corpora, *Proceedings of the Workshop on Multilingual Linguistic Resources*, COLING'04.

ROSETTA M. T. (1994), *Compositional Translation*. Kluwer Academic Publishers, Dordrecht, The Netherlands, chap. 9. – d'après: DORR, B.J. (1995), Compositional Translation by Rosetta: a book review, *Computational Linguistics* Vol 21:4.