

# L'AMBIGUÏTE SYNTAXIQUE DU GROUPE NOMINAL COMPLEXE EN ANGLAIS MÉDICAL : PRÉMODIFICATION ET COORDINATION\*

François Maniez  
Centre de Recherche en Terminologie et Traduction  
Université Lumière Lyon 2

## 1 INTRODUCTION

L'ambiguïté syntaxique est un phénomène inhérent à toutes les langues naturelles. L'introduction des langages contrôlés dans le domaine des sciences et des techniques, même si elle en diminue la fréquence, ne parvient jamais à totalement l'éliminer. Plusieurs caractéristiques font que l'anglais est tout particulièrement générateur d'ambiguïtés syntaxiques. Il y a tout d'abord le fait qu'un pourcentage non négligeable du lexique de l'anglais - 8% selon J. Tournier (1985, p. 169) - consiste en des mots pouvant appartenir à plusieurs catégories grammaticales, le double statut nominal et verbal étant le cas de figure le plus fréquent. D'autre part, le fait que les adjectifs ne s'accordent pas en genre ni en nombre rend parfois difficile la délimitation exacte de la modification adjectivale. Enfin, le phénomène de la prémodification du nom par un autre nom, partagé par d'autres langues germaniques, obscurcit la relation entre ces deux noms, que d'autres langues expriment parfois plus clairement par l'intermédiaire de syntagmes prépositionnels (*oil well* = puits DE pétrole, *sugar cane* = canne À sucre, *plastic cup* = gobelet EN plastique), et qui nécessite un choix au décodage entre le singulier ou le pluriel pour le nom prémodifiant (*horse manure* = crottin de cheval, *horse race* = course de chevaux).

Si ces ambiguïtés se résolvent facilement en contexte dans le cas de la langue générale, elles sont souvent génératrices d'opacité dans les langues de spécialité, en particulier en raison de la longueur et de la complexité des groupes nominaux de la prose scientifique et technique. Le problème du découpage terminologique est bien connu des traducteurs de textes spécialisés, et a été exposé pour la langue médicale par Van Hoof (1986) et Rouleau (1994), et dans le cadre de l'extraction terminologique par Bourigault (1992). Un court exemple suffira dans un premier stade à illustrer le nombre d'ambiguïtés que génèrent la prémodification nominale et la coordination en anglais :

(1) The ability of PET to detect cancer is based on the altered substrate requirements of malignant cells, which result from increased nucleic acid and protein synthesis and glycolysis.

---

\* Pour faire référence à cet article : Maniez François, « L'ambiguïté syntaxique du groupe nominal complexe en anglais médical : prémodification et coordination », revue électronique *Texte et corpus*, n°3 / août 2008, Actes des Journées de la linguistique de Corpus 2007, p. 207-217 (disponible sur [http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES\\_JLC07\\_maniez.pdf](http://web.univ-ubs.fr/corpus/jlc5/ACTES/ACTES_JLC07_maniez.pdf))

Au décodage, le traducteur de la phrase (1) est amené à se poser plusieurs questions, dont certaines trouveront une réponse immédiate dans l'examen du contexte et des relations sémantiques qui lient les mots entre eux. En voici quelques-unes :

- *altered* qualifie-t-il *substrate* ou *requirements*?
- l'antécédent de la relative est-il *malignant cells*, ou bien l'ensemble du groupe nominal *the altered substrate requirements of malignant cells*?
- *increased* qualifie-t-il *acid*, *synthesis* ou bien l'ensemble *synthesis and glycolysis* ?
- *nucleic* qualifie-t-il *acid*, l'ensemble *acid and protein* ou bien *synthesis* ?
- *protein* est-il un prémodificateur du seul nom *synthesis* ou de l'ensemble *synthesis and glycolysis* ?

On voit que la plupart de ces questions concernent la portée de la modification adjectivale ou nominale, la multiplicité des solutions envisageables étant ici augmentée par le phénomène de la coordination. À partir de l'étude d'un corpus bilingue aligné, nous nous proposons ici d'envisager les diverses manières dont l'ambiguïté de telles structures peut être résolue en anglais médical, et de tenter de déterminer quelles méthodes de désambiguïsation sont envisageables pour améliorer la performance des programmes d'analyse morfo-syntaxique automatique utilisés par les logiciels d'aide à la traduction.

## 2 LES STRUCTURES DE LA PRÉMODIFICATION NOMINALE

### 2.1 La prémodification nominale simple

On sait que la majorité des termes employés en anglais de spécialité sont de longueur 2 (Frantzi *et al.*, 1999), et sont généralement formés selon un modèle comprenant un élément modificateur (adjectif ou nom) qui précède le nom constituant le « nœud » du terme. Le passage d'une langue utilisant la prémodification (comme les langues germaniques) à une langue utilisant la post-modification (langues romanes) par le biais d'un groupe prépositionnel nécessite en effet l'explicitation de cette relation. Le problème est toutefois compliqué par l'utilisation des adjectifs dits « relationnels », qui peuvent également traduire cette prémodification. Les relations entre les deux noms concernés par la prémodification peuvent être de natures multiples. Le prémodificateur peut signifier la localisation anatomique (*back pain* = douleur dorsale ou dorsalgie, *brain stem* = tronc cérébral), la fonction (*taste buds* = papilles gustatives, *sweat glands* = glandes sudoripares), la cause (*heat rash* = erythème calorique) ou la forme (*sickle cell* = cellule falciforme).

### 2.2 La double prémodification nominale (N1 N2 N3)

On sait que les termes de longueur 2 sont fréquemment imbriqués à l'intérieur de termes ou de collocations de longueur supérieure, certains programmes d'extraction terminologique se fondant d'ailleurs sur cette caractéristique (Frantzi *et al.*, 2001). L'analyse du schéma de prémodification par l'apprenant présuppose un choix entre deux découpages possibles de cette relation de prémodification : N1 prémodifie N2-N3 (*placebo control group*) ou N1-N2 prémodifie N3 (*bone marrow transplant*). Ce choix dépend de l'identification du lien privilégié entre les deux noms qui forment la séquence pré-construite (qu'il s'agisse d'un terme ou d'une collocation).

### 2.3 La prémodification adjectivale (ADJ N1 N2)

Le patron syntaxique ADJ N1 N2 est d'usage extrêmement fréquent en anglais médical. Il pose intrinsèquement des problèmes de décodage similaires à ceux des groupes nominaux formés par la concaténation de trois noms. Le découpage syntaxique, de nature binaire, s'effectue en fonction de la reconnaissance éventuelle d'un lien de prémodification entre l'adjectif et le premier nom de la chaîne. Ainsi, on reconnaît la suite *coronary artery* dans le groupe nominal [*coronary artery*] *disease*, alors que l'absence de reconnaissance d'un lien de prémodification dans la suite *coronary heart* (qui peut se conjuguer à la connaissance préalable de l'unité terminologique *heart disease*) impose un découpage inverse pour *coronary* [*heart disease*]. L'utilisation des fréquences comparées des suites ADJ N1 et ADJ N2 en corpus (Maniez, 2001) semble constituer un indice fiable pour le découpage à adopter.

## 3 LES STRUCTURES FAISANT INTERVENIR LA COORDINATION ET LA PRÉMODIFICATION NOMINALE

L'apprenant est souvent confronté à une multiplicité de découpages potentiels par le biais des phénomènes de coordination et de prémodification. Afin de tenter de décrire les mécanismes de compréhension qui favorisent l'élimination des découpages incorrects et le choix du découpage attendu, nous avons utilisé pour la détection de ces deux patrons syntaxiques un corpus constitué à partir du CD-ROM *Annals of Internal Medicine*. Ce corpus totalise 4,5 millions d'occurrences (*tokens*) et contient l'intégralité des articles publiés dans les revues suivantes pendant l'année 1993 : *New England Journal of Medicine*, *Journal of the American Medical Association*, *Annals of Internal Medicine*, *Lancet*, *British Medical Journal*. Ce corpus a subi une catégorisation en partie du discours (*part-of-speech tagging*) à l'aide du logiciel Winbrill.

### 3.1 Le patron syntaxique <ADJ1 ADJ2 N1 (AND/OR) N2 N3>

L'étude détaillée des 338 formes correspondant au patron syntaxique <ADJ1 ADJ2 N1 (AND/OR) N2 N3> a permis d'identifier cinq découpages possibles :

[ADJ1 ADJ2 N1] and [N2 N3]  
[**central nervous system**] and [**bone marrow**].

ADJ1 ( [ADJ2 N1] and [N2 N3]  
**good** ( [**hygienic practices**] and [**hospital policies**] )

( [ADJ1 ADJ2 N1] and [N2] ) N3  
( [**serious adverse event**] or [**withdrawal**] ) **form**

ADJ1 ( ADJ2 [N1 and N2] ) N3  
**persistent** ( **central** [ **excitability and pain** ] ) **behaviours**

ADJ1 ( ADJ2 [ N1 and [N2 N3] ] )  
**rheumatoid** ( **synovial** [ **fluids and** [ **tissue cultures** ] ] )

Ce patron syntaxique n'a pas été étudié ici et fera l'objet d'une publication ultérieure.

## 3.2 Le patron syntaxique <ADJ N1 (AND/OR) N2 OF N3>

### 3.2.1 LES QUATRE DÉCOUPAGES POSSIBLES

L'étude détaillée des 589 formes correspondant au patron syntaxique <ADJ N1 (AND/OR) N2 OF N3> a permis d'identifier quatre découpages possibles :

a) [ADJ [N1 AND N2] ] OF N3

L'adjectif modifie N1 et N2 et la complémentation s'applique aux deux noms coordonnés. Il s'agit du découpage le plus fréquent (55% des occurrences).

**effective [identification and care] of patients**  
**optimal [dosage and duration] of therapy**

b) [ADJ N1] AND [N2 OF N3]

Aucune dépendance syntaxique ne « traverse » la conjonction. L'adjectif ne modifie que le premier nom et la complémentation par OF N3 s'applique uniquement à N2. Ce découpage concerne 26% des occurrences.

**[maternal age] and [length of gestation]**  
**[ethnic group] and [place of birth]**

c) ADJ [N1 AND [N2 OF N3] ]

L'adjectif modifie N1 et N2, mais la complémentation s'applique uniquement à N2. Ce découpage concerne 11% des occurrences.

**substantial [overlap and [loss of discrimination] ]**  
**invaluable [ context and [source of information] ]**

d) [ [ADJ N1] AND [N2] ] OF N3

L'adjectif ne modifie que le premier nom et la complémentation s'applique aux deux noms coordonnés. Ce découpage concerne 8% des occurrences.

**[ [absolute number] and percentage] of lymphocytes**  
**[ [natural history] and epizootiology] of Lyme disease**

### 3.2.2 ANALYSE DE LA PERFORMANCE DU LOGICIEL D'AIDE À LA TRADUCTION SYSTRAN (VERSION 6)

Les deux derniers types de découpage (les plus rares) sont fréquemment mal reconnus par Systran, qui opte généralement pour l'un des deux premiers. Le logiciel effectue cependant le découpage correct dans 52% des cas. Lorsque l'ambiguïté syntaxique est correctement résolue, la traduction proposée par le logiciel est de bonne qualité. L'exemple (2) nécessitera ainsi une intervention minimale de la part du réviseur :

(2) This technique is very attractive conceptually, although many important questions about the physiologic function and duration of effect of the expressed gene will need to be addressed before its importance in the modification of vascular diseases can be determined.	Cette technique est très attrayante conceptuellement, bien que beaucoup de questions importantes au sujet de la fonction physiologique et de la durée de l'effet du gène exprimé doivent être abordées avant que son importance dans la modification des maladies vasculaires puisse être déterminée.
---	---

Dans notre décompte, sont considérées comme correctes les phrases dans lesquelles la traduction française suggère le découpage exact de la structure concernée, sans tenir compte d'autres erreurs éventuelles. Ainsi, l'exemple (3) est considéré comme une instance de découpage correct, en dépit du rattachement erroné de « seul » (l'adjectif *alone* ne fait pas partie de la structure étudiée). Notons au passage que la séquence *efficacy and safety* est l'une des collocations les plus fréquentes dans le vocabulaire des essais médicaux (plus de deux millions d'occurrences sur *Google*<sup>1</sup>), et que le découpage correct de telles structures repose probablement chez l'humain sur la connaissance de ces collocations.

(3) <b>Consequently, we compared the antiemetic efficacy and safety of ondansetron alone with [...]</b>	<b>En conséquence, nous avons comparé l'efficacité et la sûreté antiémétiques seul de l'ondansetron à [...]</b>
---	---

De la même manière, l'exemple (4) est considéré comme une instance de découpage correct :

(4) <b>Management algorithms would be more useful than figures showing anecdotal responses to cholesterol-lowering medications or photographs of patients doing push-ups.<sup>2</sup></b>	<b>Les algorithmes de gestion seraient plus utiles que des figures montrant des réponses anecdotiques aux traitements hypolipémiants ou aux photographies des patients faisant des poussées.</b>
---	--

La traduction suggère en effet un rattachement de *photographs* à *responses*, alors que *photographs* est coordonné à *figures*, mais l'erreur de rattachement ne concerne pas la structure qui fait l'objet de notre étude.

Certains cas ont été classés comme indéterminés, car il est difficile de savoir si la traduction donnée est le résultat d'un découpage syntaxique adéquat. Ainsi, dans l'exemple (5), la traduction est incorrecte en raison de l'étiquetage du mot *increase* en tant que nom, et le découpage de la structure concernée semble correct dans la mesure où *daily* ne modifie pas *duration*, mais il est impossible de savoir si la complémentation par *of treatment* s'applique dans la traduction aux deux groupes nominaux coordonnés, ce qui est le cas dans l'original.

(5) <b>In general, the probability of physical dependence increases as the daily dose and duration of treatment increase.</b>	<b>Généralement la probabilité de la dépendance physique augmente à mesure que la dose quotidienne et la durée de l'augmentation de traitement.</b>
---	---

<sup>1</sup> Une brève clarification s'impose à propos des chiffres donnés par *Google*, dont l'exactitude a été souvent mise en doute (voir notamment l'explication de Jean Véronis sur <http://aixtal.blogspot.com/2005/02/web-le-mystre-des-pages-manquantes-de.html>). Nous les utiliserons ici uniquement afin d'évaluer la fréquence d'emploi de certaines formes coordonnées et de comparer entre eux plusieurs découpages possibles.

<sup>2</sup> Le push-up, mouvement vulgairement connu sous le nom de « pompe », a pour dénomination scientifique le terme « redressement brachial ». Le terme est toutefois peu usité.

Les erreurs de découpage se produisent parfois parce que les structures étudiées sont elles-mêmes imbriquées dans des groupes nominaux de taille supérieure, qui génèrent à leur tour d'autres ambiguïtés. Dans l'exemple (6), le découpage correct serait [ADJ [N1 AND N2] ] OF N3, mais le lien ADJ-N2 n'est pas validé car le découpage adopté prend en compte une coordination entre deux groupes nominaux complexes qui seraient tous deux compléments du nom acquisition, *information on the overall diagnosis and management of patients with gastrointestinal bleeding* :

<p><b>(6) The acquisition of information on the overall diagnosis and management of patients with gastrointestinal bleeding has become easier with the availability of this book.</b></p>	<p><b>L'acquisition d'information sur le diagnostic global et de gestion des patients présentant le saignement gastro-intestinal est devenue plus facile avec la disponibilité de ce livre.</b></p>
---	---

L'exemple (7) concerne un cas de figure proche du précédent, puisque la présence de la chaîne *Investigation of* devant la structure étudiée rend possible un découpage qui scinde la structure étudiée, et qui s'avère être correct, les noms *investigation* et *analysis* appartenant au même niveau de coordination. Les exemples correspondant à ces deux types de structures n'ont donc pas été comptabilisés.

<p><b>(7) Investigation of physiological effects and analysis of outcome for infants in various categories will point to optimal management of cord clamping and enable the establishment of practical guidelines.</b></p>	<p><b>La recherche sur des effets et l'analyse physiologiques des résultats pour des enfants en bas âge dans diverses/varié catégories indiquera la gestion optimale de la corde/cordon maintenant et permettra l'établissement des directives/recommandations pratiques.</b></p>
--	---

La présence d'un autre adjectif précédant ADJ peut révéler une dépendance syntaxique erronée qui serait passée inaperçue. C'est le cas du participe passé employé comme adjectif *repeated* dans l'exemple (8), dont la traduction à la suite du nom *analyses* implique la prémodification par les deux adjectifs de l'ensemble du groupe nominal, et donc la validation erronée du lien N1 OF N3, donnant *endoscopic examinations of stools*<sup>3</sup>. Ce type de structure n'a donc pas été pris en compte.

<p><b>(8) Repeated endoscopic examinations and analyses of stools did not reveal pathologic findings.</b></p>	<p><b>Les examens endoscopiques et les analyses répétés des tabourets n'ont pas indiqué des résultats pathologiques.</b></p>
---	--

Les ambiguïtés syntaxiques ne peuvent cependant pas toujours être résolues sans recours à la réflexion et/ou à des connaissances de type encyclopédique et non lexical. Ainsi, dans l'exemple (9), *enunciation* et *rehearsal* sont interprétés par l'humain comme appartenant au même champ sémantique, ce qui favorise le découpage (N AND N), mais la validation de la dépendance à distance *internal rehearsal* se pose néanmoins. Le recours à l'expert ou à un corpus est ici nécessaire au traducteur humain pour la validation du lien

<sup>3</sup> La traduction correcte de *stools* dans ce contexte est "selles".

ADJ-N2, qui semble plausible. À défaut d’approbation par un expert du domaine ou d’occurrences attestées présentes en corpus, les moteurs de recherche du Web permettent pour le moins de vérifier l’existence de la collocation concernée (comme c’est le cas dans le contexte suivant : « *Articulation rate is assumed to correlate with the speed of internal rehearsal* »).

<p><b>(9) Speech sounds impede cognition by preempting auditory neural pathways used in the internal enunciation and rehearsal of words related to the performance of mental tasks.</b></p>	<p><b>Les phonèmes empêchent la connaissance/cognition en acquérant des voies neurales auditives utilisées dans l’énonciation et la répétition internes des mots liés à l’exécution des tâches mentales.</b></p>
---	--

Le recours aux résultats des requêtes faites sur les moteurs de recherche du Web peut constituer un mode de validation de dépendances syntaxiques à distance telles que ADJ-N2, l’expression *surgical decompression* ayant environ 200 000 occurrences<sup>4</sup>. La prise en compte de ce type de résultat permettrait d’éviter le découpage incorrect de l’exemple (10) :

<p><b>(10) The options for treatment include anticysticercal drugs, corticosteroids, cerebrospinal fluid shunting procedures, and surgical removal or decompression of cysts.</b></p>	<p><b>Les options pour le traitement incluent les drogues &lt;nfw&gt;anticysticercal&lt;/nfw&gt;, les corticostéroïdes/corticoïde, les procédures de manoeuvre de liquide céphalo-rachidien, et l’ablation chirurgicale ou la décompression des kystes.</b></p>
---	---

Inversement, il arrive que l’interprétation correcte nécessite l’élimination logique du lien ADJ-N2, validé dans la traduction incorrecte de Systran dans l’exemple (11), qui traduit la suite « *therapeutic [...] discussions of infections* » par « examens thérapeutiques des infections ». L’utilisation d’une comparaison entre les ordres de grandeur des résultats obtenus sur les moteurs de recherche du Web pour le lien ADJ-N1 (71 600 pour « *therapeutic recommendations* ») et le lien ADJ-N2 (645 pour « *therapeutic discussions* ») pourrait constituer un indice menant au découpage correct par une pondération restant à déterminer (nous supposons que l’étiquetage incorrect de *detailed* en tant que forme verbale n’a pas d’incidence sur le découpage du groupe nominal qui suit).

<p><b>(11) [...] this book has more detailed therapeutic recommendations and discussions of infections in patients with AIDS.</b></p>	<p><b>[...] ce livre a plus détaillé des recommandations et des examens thérapeutiques des infections dans les patients présentant des SIDAS.</b></p>
---	---

Toutefois, les statistiques d’emploi de la suite N1 AND N2 devraient sans doute être également prises en compte dans la pondération que nous avons évoquée , le découpage

<sup>4</sup> Ce chiffre et les suivants indiquent les résultats de requêtes effectuées sur <http://www.google.fr/> le 18 octobre 2007.

[ADJ [N1 AND N2] ] OF N3 étant majoritaire. Ainsi, dans l'exemple (12), la prise en compte des résultats obtenus pour les coordinations fréquentes (plus de trois millions d'occurrences sur le Web pour *signs and/or symptoms*) permettrait d'orienter le découpage syntaxique vers la solution correcte<sup>5</sup>.

(12) <b>Second, hemochromatosis can be detected before any clinical signs or symptoms of disease develop and even before hepatic iron loading occurs.</b>	<b>En second lieu, la hémochromatose primitive peut être détectée avant que tous les signes cliniques ou symptômes de la maladie se développent et même avant le chargement hépatique de fer se produit.</b>
---	--

### 3.2.3 VERS UN ALGORITHME DE DÉCOUPAGE SYNTAXIQUE INTÉGRANT LES STATISTIQUES D'EMPLOI DES COLLOCATIONS SUR LE WEB

L'examen des découpages incorrects de la structure ADJ N1 (AND/OR) N2 OF N3 fait apparaître que la majorité des erreurs (ce qui représente un peu plus du quart de l'ensemble des énoncés traités) concerne la non-validation de la dépendance ADJ N2 (Tableau 1).

<b>non-validation de ADJ N2</b>	<b>56,5%</b>
<b>non-validation de N1 of N3</b>	<b>4,5%</b>
<b>validation erronée de ADJ N2</b>	<b>13%</b>
<b>validation erronée de N1 of N3</b>	<b>16%</b>

Tableau 1 : Erreurs de découpage de l'analyseur de Systran pour la traduction des groupes nominaux de type <ADJ N1 (AND/OR) N2 OF N3>

Nous postulons l'existence d'une formule permettant de valider les dépendances syntaxiques à distance (ADJ N2 et N1 of N3) à partir d'une comparaison des fréquences d'emploi des collocations résultant de l'application de ces dépendances à distance avec celles des dépendances de proximité (ADJ N1 et N2 of N3). Le Tableau 2 ci-dessous fournit les chiffres obtenus sur la Toile pour les collocations correspondant aux dépendances de proximité et aux dépendances à distance pour quelques-unes des expressions dans lesquelles les deux liens sont validés et qui n'ont pas été découpées correctement dans la traduction proposée par Systran.

<sup>5</sup> Une autre solution serait d'intégrer ces collocations aux dictionnaires spécialisés utilisés par les logiciels d'aide à la traduction.

Expression	ADJ N1	ADJ N2	N1 of N3	N2 of N3
cardiovascular health and disease of women	1 780 000	32 400 000	529 000	18 400
clinical signs and symptoms of disease	2 100 000	1 790 000	499 000	335 000
possible modes and efficiency of transmission	382 000	147 000	331 000	54 700
clinical evaluation and identification of patients	1 910 000	41 000	857 000	430 000
particular type or mode of pacemaker	2 000 000	666 000	16 600	27
surgical removal or decompression of cysts	1 530 000	196 000	13 200	23
high prevalence and severity of hypertension	1 930 000	254 000	310 000	47 000
effective isolation and treatment of patients	71 600	2 510 000	40 800	2 160 000
overall diagnosis and management of patients	29 600	1 910 000	194 000	1 930 000
safe use and disposal of sharps	1 990 000	756 000	13 900	66 300
experimental methods and standards of proof	1 740 000	14 400	115 000	188 000

Tableau 2 : Résultats obtenus sur Google pour les collocations ADJ N1, ADJ N2, N1 of N3 et N2 of N3 en cas de non-validation du lien ADJN1 et du lien N2 of N3

Les chiffres obtenus peuvent être considérés comme suffisants pour valider les dépendances à distance, dans les cas où les valeurs observées pour ADJ N2 sont supérieures à celles de ADJ N1, et où les valeurs observées pour N1 of N3 sont supérieures à celles de N2 of N3. Par ailleurs, le fait que certaines expressions complètes soient employées à une relativement haute fréquence (plus de 700 occurrences pour *clinical signs and symptoms of disease*) constitue également un indice de figement, et donc vraisemblablement de validation des dépendances à distance. D'une manière générale, le fait que les valeurs observées pour N1 of N3 et N2 of N3 soient proches favorise probablement la validation du lien N1 of N3.

L'utilisation des données du Tableau 3 peut ainsi servir de base au découpage correct, certaines dépendances à distance ayant des valeurs supérieures aux dépendances de proximité (les cellules correspondantes sont grisées dans le Tableau 2). On remarque que dans sept cas sur dix, le lien N1 of N3 serait validé par la prise en compte de ces chiffres. Toutefois, la validation du lien ADJ N2 ne serait obtenue que dans quatre cas sur dix par cette méthode. Ceci peut s'expliquer partiellement par le fait que l'ordre des deux noms coordonnés est souvent quasi-figé (on dénombre par exemple sur le Web huit fois plus d'occurrences de la suite *health and disease* que de *disease and health*). Ce phénomène est particulièrement frappant dans les cas où les deux noms coordonnés sont des déverbaux désignant des procès généralement consécutifs (*diagnosis and management, evaluation and identification, isolation and treatment*).

Le fait que la suite N1 and N2 soit elle-même d'emploi fréquent est clairement un autre facteur qui influence la probabilité de validation des dépendances à distance, notamment quand les valeurs observées sont largement supérieures à celles de N2 of N3, comme c'est le cas pour les deux derniers exemples du Tableau 2 (830 000 pour *use and disposal* et 262 000 pour *methods and standards*). Ainsi, cette haute fréquence observée, associée à une valeur élevée du ratio N1 AND N2 / N2 OF N3 (supérieure à 10 dans notre cas) pourrait également être prise en compte.

La consultation de données équivalentes dans les cas de validation erronée du lien ADJ N2 (Tableau 3) nous montre toutefois que le seuil à partir duquel la validité du lien ADJ

N2 peut être envisagée est vraisemblablement beaucoup plus bas, puisque dans trois de ces cas, les chiffres obtenus pour ADJ N2 représentent moins de 1% de ceux obtenus pour ADJ N1. Ainsi, on peut estimer au vu de notre échantillon qu'un système qui présupposerait la validation du lien ADJ N2 et ne l'invaliderait que dans les cas où le rapport ADJ N1 / ADJ N2 est supérieur à 20 obtiendrait un taux d'erreur d'environ 5%.

Expression	ADJ N1	ADJ N2
<b>therapeutic recommendations and discussions of infections</b>	<b>69 900</b>	<b>643</b>
<b>adverse events and yields of efficacy</b>	<b>1 890 000</b>	<b>27</b>
<b>subjective awareness or type of symptoms</b>	<b>55 600</b>	<b>17 100</b>
<b>additional personnel and continuity of care</b>	<b>1 910 000</b>	<b>10 600</b>

Tableau 3 : Résultats obtenus sur Google pour les collocations ADJ N1 et ADJ N2 dans quelques cas de validation erronée du lien ADJ N2

La consultation des données correspondant aux cas de validation erronée du lien N1 of N3 (Tableau 4) est encore plus révélatrice de ce point de vue, puisque les chiffres obtenus pour N1 of N3 représentent dans tous les cas moins de 1% de ceux obtenus pour N2 of N3. Ainsi, on peut estimer au vu de notre échantillon qu'un système qui présupposerait la validation du lien N1 of N3 et ne l'invaliderait que dans les cas où le rapport N2 of N3 / N1 of N3 est supérieur à 20 obtiendrait un taux d'erreur proche de 0%.

Expression	N1 of N3	N2 of N3
<b>different methods and criteria of efficacy</b>	<b>68</b>	<b>11 500</b>
<b>financial incentives or methods of remuneration</b>	<b>1</b>	<b>13 600</b>
<b>substantial overlap and loss of discrimination</b>	<b>3</b>	<b>989</b>
<b>reversible hyperglycemia or development of diabetes</b>	<b>643</b>	<b>375 000</b>

Tableau 4 : Résultats obtenus sur Google pour les collocations N1 of N3 et N2 of N3 dans quelques cas de validation erronée du lien ADJ N2

## 4 CONCLUSION

Nous préconisons la mise en oeuvre de méthodes de recherche systématique de groupes nominaux coordonnés dans les corpus spécialisés et de stockage des fréquences observées sur le web pour les collocations résultant des divers découpages syntaxiquement possibles de ces groupes nominaux. Leur traitement, dans lequel les critères statistiques habituels (fréquence et z-score notamment) pourraient dans un deuxième stade être associés aux analyses linguistiques (notamment concernant le degré de synonymie des substantifs coordonnés), conduira à l'établissement d'une liste des groupes potentiellement problématiques pour le traducteur humain comme pour les programmes d'aide à la traduction.

## 5 RÉFÉRENCES

- Scott M. & Tribble C. (2006). *Textual Patterns. Key words and corpus analysis in language education*, Amsterdam : John Benjamins.
- Blanche-Benveniste C. (2005). « L'étude grammaticale des corpus de langue parlée en français », in : Williams G. (dir.), *La linguistique de corpus*, Rennes : PUR.
- Boitet C. (2007). « Corpus pour la TA : Types, tailles et problèmes associés, selon leur usage et le type de système », *Revue française de linguistique appliquée*, vol. XII, 2007/1. p. 25-38.

- Bourigault D. (1992). « Lexter: un logiciel d'extraction de terminologie. », in : *TAMA '92, Actes du 2<sup>o</sup> Symposium TermNet : Applications terminologiques et microordinateurs*, Vienne, Autriche.
- Cormier M. (1990). « Traduction de textes de vulgarisation et de textes didactiques : Approche pédagogique. », *Meta*, vol. 35/ 4. p. 676-688.
- Cormier M. and Roberts R.P. (2000). « Lexicographie comparée du français et de l'anglais au Canada : Le dictionnaire canadien bilingue. », in : Szende T. (ed), *Approches contrastives en lexicologie bilingue*, Éditions Champion, p. 213-222.
- Frantzi K. T., Tsujii J. & Ananiadou S. (1999). « Clustering Terms Using the C-value Method for Automatic Term Recognition », in : Sandrini, P. (ed), *TKE '99, Terminology and Knowledge Engineering*. p. 356-366.
- Maniez F. (2000). « La prémodification nominale en anglais médical : Quelques problèmes de traduction », in : Banks, D. (ed.), *Le groupe nominal dans le texte spécialisé*, Paris : L'Harmattan.
- Maniez F. (2001). « L'ambiguïté syntaxique due aux structures coordonnées en anglais médical : analyse de la performance d'un logiciel d'aide à la traduction. », in : *Actes du colloque TALN de Tours, 2-5 juillet 2001, Tome 2*, p. 277-286.
- Maniez F. (2002). « The Use of Electronic Corpora and Lexical Frequency Data in Solving Translation Problems. », in : Altenberg B. & Granger S. (ed.). *Lexis in Contrast, Studies in Corpus Linguistics*, n°7, Amsterdam : John Benjamins.
- Roberts R. P. & Cormier C. (2000). « L'analyse des corpus pour l'élaboration du Dictionnaire canadien bilingue. », in : Szende T. (ed.), *Approches contrastives en lexicologie bilingue*. Éditions Champion.
- Rouleau M. (1994). *La traduction médicale, une approche méthodique*. Linguattech.
- Rouleau M. (2001). « La facture des principaux dictionnaires médicaux français : Point de vue d'un traducteur. », *Meta*, vol. 46/1, p. 34-55.
- Rouleau M. (2003). « La terminologie médicale et ses problèmes. », *Panacea*, vol. IV, n°12 [[http://www.medtrad.org/panacea/PanaceaPDFs/Panacea12\\_junio2003.pdf](http://www.medtrad.org/panacea/PanaceaPDFs/Panacea12_junio2003.pdf)].
- Selva T, Verlinde S. & Binon J. (2003). « Vers une deuxième génération de dictionnaires électroniques. », in : Zock M. & Carroll J. (eds).. *TAL 44 - 2/2003 Les dictionnaires électroniques*, p. 177-198.
- Tournier J. (1985). *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Champion-Slatkine.
- Van Hoof H. (1986). *Précis pratique de traduction médicale*. Maloine.