Traduction automatique et ambiguïté syntaxique : le cas de la coordination dans les groupes nominaux complexes en anglais médical

François Maniez

Centre de Recherche en Terminologie et Traduction, Université Lumière Lyon 2

Abstract

Specialized languages are characterized by the very frequent use of terms, and consequently of noun phrases. In English, using nouns or adjectives to modify noun phrases favors economy, so that such structures are ideally suited for the coinage of scientific terms. However, using coordination between noun phrases which have such modifiers generates syntactically ambiguous structures that may induce errors by human translators as well as automatic translation systems.

Focusing on the case of medical language, we describe various syntactic structures that are potentially ambiguous, and we assess the performance of an automatic translation program, using sentences with noun phrases whose syntactic structure may be interpreted in four different ways. We also consider the possibility of using the numbers returned by the Web's search engines in order to solve ambiguous syntactic dependencies, by comparing the numbers of results obtained for queries using expressions whose components are contiguous to those that are created by validating long-distance syntactic dependencies.

Résumé

Les langues de spécialité se caractérisent par une haute densité terminologique qui favorise l'usage du groupe nominal. En anglais, le phénomène de la prémodification nominale permet une économie d'expression qui en fait un outil commode dans la dénomination des termes scientifiques. Toutefois, la coordination de groupes nominaux faisant intervenir la prémodification nominale est génératrice d'ambiguïtés syntaxiques susceptibles de provoquer des erreurs à la traduction chez l'humain comme en traduction automatique.

En prenant l'exemple de la langue médicale, nous décrivons brièvement les diverses structures syntaxiques pouvant générer de telles ambiguïtés, et nous examinons la performance d'un logiciel de traduction automatique auquel ont été soumises des phrases contenant des groupes nominaux dont le patron syntaxique peut donner lieu à quatre découpages distincts. La prise en compte des données recueillies sur les moteurs de recherche du Web est envisagée comme piste de résolution des ambiguïtés de rattachement, par l'intermédiaire de la comparaison des chiffres obtenus pour les requêtes contenant des expressions dont les composants sont contigus et pour celles qui sont obtenues par la validation des dépendances syntaxiques à distance.

Mots-clés : ambiguïté syntaxique, anglais scientifique, coordination, groupe nominal, prémodification nominale, traduction automatique.

1. Introduction

Les langues de spécialité abondent en groupes nominaux complexes, et le domaine médical ne fait pas exception à cette règle (Maniez, 2000). La conjugaison de la prémodification et de la coordination à l'intérieur de ces groupes nominaux complexes demeure encore à l'heure actuelle un obstacle majeur pour les systèmes de traduction automatique confrontés à la traduction de textes scientifiques (Maniez, 2001). Quant à l'humain apprenant de langue étrangère de spécialité, on sait qu'il évite volontiers le recours à l'outil de décodage classique qu'est le dictionnaire bilingue lorsqu'il a le sentiment de maîtriser les constituants des

groupes complexes (Cormier, 1990 ; Thoiron, 2000). Pour l'humain comme pour la machine, il importe donc de développer la maîtrise des lexies complexes et des collocations en langue de spécialité, dont le décodage est souvent susceptible de devenir problématique.

2. 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).

3. 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., 1999). 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 préconstruite.

4. 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.

4.1. Analyse de la traduction de la structure ADJ N1 N2 par la version 6 du logiciel Systran

Dans un test effectué sur la version 6 du logiciel Systran à partir de 107 structures de ce type tirées du corpus Europarl (dont les résultats sont consignés dans le Tableau 1), un pourcentage important de ces ambiguïtés (45%) semblent de fait avoir été résolues par la présence de la suite ADJ N1 ou N1 N2 (voire de la totalité de la séquence ADJ N1 N2) dans la mémoire de traduction utilisée. Dans 40% des cas, la traduction indique qu'un découpage ADJ (N1 N2) a été correctement effectué sans qu'un équivalent de traduction prémémorisé ait été utilisé. Les

erreurs de découpage résultant d'un découpage (ADJ N1) N2 incorrect ne représentent que 11% des cas. Deux traductions incorrectes sont dues à l'interprétation de l'adjectif en tant que nom, deux autres cas correspondant à l'absence de traduction de l'un des éléments du groupe nominal. Le logiciel livre une traduction résultant du découpage correct dans 85% des cas.

Rattachement correct de ADJ à N2	43
Identification de la suite ADJ N1	19
Identification de la suite N1 N2	17
Identification de la suite ADJ N1 N2	12
Rattachement incorrect de ADJ à N1	6
Rattachement incorrect de ADJ à N2	6
Etiquetage incorrect de ADJ comme N	2
Parties de la suite non traduites	2
	107

Tableau 1 : Découpage de 107 structures ADJ N1 N2 par Systran

4.1.1. Rattachement correct de ADJ à N2

Le rattachement a été considéré comme correct sans considération des choix lexicaux. Ainsi dans l'exemple (1), le mot « entrée » n'est pas l'équivalent correct du mot « lobby » au sens où il est employé, mais la traduction de Systran 6 montre que le rattachement de *huge* (ADJ) à *lobby* (N2) a été correctement effectué[1] :

(1) EN : There is a **huge disability lobby** in the European Union

FR : Le groupe de pression de l 'Union européenne est très puissant.

Systran 6 : Il y a une entrée énorme d'incapacité dans l'Union européen.

4.1.2. Identification de la suite ADJ N1

Elle concerne des termes ou collocations dont la haute fréquence d'emploi provoque la dictionnairisation (**early retirement** = *retraite anticipée*, **solid fuel** = *combustible solide*, **single market** = *marché unique*, **good quality** = *bonne qualité*, **central bank** = *banque centrale*, **cold war** = *guerre froide*). Le prétraitement des suites ADJ N1 motive parfois le changement de catégorie grammaticale pour l'adjectif (**human rights** = *droits de l'homme*), ou l'adoption d'une traduction non littérale du nom (**public procurement** = *marchés publics* dans **public procurement directives** = *directives de marchés publics*, **foreign exchange** = *devises étrangères* dans **foreign exchange reserves** = *réserves de devises étrangères*).

4.1.3. Identification de la suite N1 N2

Elle est elle également repérable par le fait qu'une traduction non littérale est utilisée (**tourist industry** = *l'industrie du tourisme*, **warning signs** = *signaux d'alarme*, **safety hazards** = *risques en matière de sécurité*), qu'il y a changement de catégorie grammaticale de N1 (**budget lines** = *lignes budgétaires*, **welfare benefits**. = *avantages sociaux*, **family policy** =

[1] Dans les exemples qui suivent, EN désigne la version anglaise et FR la version française des énoncés étudiés. Nous avons sélectionné ces énoncés sans considération de la langue originale de l'intervention des députés du parlement européen. Les formes ADJ N1 N2 de l'anglais et les sous-chaînes qui en sont extraites apparaissent en gras dans les exemples et dans les commentaires qui les suivent. Les équivalents de traduction français apparaissent en italiques.

politique familiale) ou que l'ensemble N1 N2 est traduit par un lexème unique (**police officers** = *policiers*, **market place** = *marché*, **bottom line** = *résultat*).

4.1.4. Identification de la suite ADJ N1 N2

Dans ces cas de figure, l'ensemble d'un terme de haute fréquence semble être tiré des mémoires de traductions utilisées par le logiciel (**nuclear power stations** = *centrales nucléaires*, **former Soviet Union** = *ex-Union soviétique*, **collective bargaining agreements** = *conventions collectives de travail*, **international arrest warrant** = *mandat d'arrêt international*, **mad cow disease** = *maladie de la vache folle*). Ces mémoires de traductions sont en effet alimentées par de grands corpus bilingues alignés au niveau de la phrase, et la segmentation des phrases alignées permet le stockage en mémoire des équivalents de traduction des groupes nominaux récurrents. Le corpus Europarl faisant partie des ressources utilisées en traduction automatique (Boitet, 2007), il n'est pas surprenant de constater une traduction correcte par Systran de tels groupes nominaux récurrents.

4.1.5. Rattachement incorrect de ADJ à N1

La phrase (2) fournit un exemple de ce type de rattachement, qui s'est produit dans 6 cas (5,5%) de notre échantillon. Le mot composé *question mark* (point d'interrogation) semble être absent du dictionnaire utilisé par le logiciel.

(2) EN : There has to be a **big question mark** as to whether we are in for a similar disappointment as far as EMU is concerned.

FR : Il y a de quoi se demander sérieusement s'il faut s'attendre aux mêmes déboires avec l'UEM, étant donné que le marché unique et l'UEM sont les produits de la même façon de concevoir les choses.

Systran 6 : Il doit y a une marque d'importante question de savoir si nous sommes dedans pour une déception semblable en ce qui concerne l'EMU.

4.1.6. Rattachement incorrect de ADJ à N2

La phrase (3) fournit un exemple de ce type de rattachement, qui s'est produit dans 6 cas (5,5%) de notre échantillon.

(3) EN : **Permanent status negotiations** resumed in September 1999.

FR : Les négociations concernant le statut permanent ont repris en septembre 1999.

Systran 6 : Négociations permanentes de statut reprises en septembre 1999.

5. La coordination des groupes nominaux complexes

Les difficultés de traduction imputables au phénomène de prémodification ont été abondamment traitées dans la littérature (cf. Rouleau, 2003). Un bref exemple suffira à illustrer le nombre d'ambiguïtés que génère la combinaison de la prémodification nominale et de la coordination en anglais de spécialité :

(4) *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.*

Au décodage, le traducteur de la phrase (4) est amené à se poser plusieurs questions :

- *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* ?

- *increased* qualifie-t-il *acid*, *synthesis* ou bien l'ensemble *synthesis and glycolysis* ?

Les sources possibles d'erreurs se conjuguant, les chances d'arriver au découpage correct sans l'apport de connaissances lexicales sont réduites (il existe en effet une douzaine de découpages possibles). Si l'on symbolise la portée des prémodifications à l'aide de crochets, le découpage correct est le suivant : *increased [[[[nucleic acid] and [protein]] synthesis] and glycolysis]*, et ce segment peut donc se traduire par « augmentation de la glycolyse et de la synthèse des protéines et de l'acide nucléique ». Les mécanismes de désambiguïsation du traducteur humain dépendent partiellement de sa connaissance de la réalité extralinguistique (la médecine) mais aussi d'une connaissance lexicale transmissible à la machine sous forme d'une base de données contenant les termes et les collocations de la langue de spécialité.

6. Deux exemples de GN complexes : les patrons syntaxiques <ADJ1 ADJ2 N1 (AND/OR) N2 N3> et <ADJ N1 (AND/OR) N2 OF N3>

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.

6.1. 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 :

a) ADJ1 et ADJ2 ne modifient que N1 :

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

b) ADJ 1 modifie N3 :

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

c) ADJ 1 modifie le groupe N2 N3 :

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

d) le groupe ADJ1 ADJ2 N1 modifie N3 :

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

e) ADJ 1 modifie le groupe disjoint ADJ2 N2 N3 :

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

Les effectifs observés pour certains de ces découpages étant insuffisants, cette structure devra toutefois être traitée dans le cadre d'une étude utilisant un corpus de taille supérieure.

6.2. Le patron syntaxique <ADJ N1 (AND/OR) N2 OF N3>

6.2.1. 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

6.2.2. 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 (5) nécessitera ainsi une intervention minimale de la part du réviseur :

(5) 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 (6) 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 e*fficacy and safety* est l'une des collocations les plus fréquentes dans le vocabulaire des essais médicaux (plus de deux millions d'occurrences sur Google),[2] et que le découpage correct de telles structures repose probablement chez l'humain sur la connaissance de ces collocations.

(6) Consequently, we compared the **antiemetic efficacy and safety of ondansetron alone** with […]	En conséquence, nous avons comparé **l'efficacité et la sûreté antiémétiques seul de l'ondansetron** à […]

De la même manière, l'exemple (7) est considéré comme une instance de découpage correct :

(7) Management algorithms would be more useful than figures showing anecdotal responses to **cholesterol-lowering medications or photographs of patients** doing push-ups.[3]	Les algorithmes de gestion seraient plus utiles que des figures montrant des réponses anecdotiques aux **traitements hypolipidémiants ou aux photographies des patients** faisant des poussées.

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 (8), 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.

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

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 (9), 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* et *management of patients with gastrointestinal bleeding* :

[2] 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.

[3] 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.

(9) The acquisition of information on the **overall diagnosis and management of patients** with gastrointestinal bleeding has become easier with the availability of this book.	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.

L'exemple (10) 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.

(10) 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.	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/recommandation pratiques.

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 (11), 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*)[4]. Ce type de structure n'a donc pas été pris en compte.

(11) **Repeated endoscopic examinations and analyses of stools** did not reveal pathologic findings.	**Les examens endoscopiques et les analyses répétés des tabourets** n'ont pas indiqué des résultats pathologiques.

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 lexicales. Ainsi, dans l'exemple (12), *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 ADJ-N2, qui semble plausible. A 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 »).

(12) 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.	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.

Le recours aux résultats des requêtes faites sur les moteurs de recherche du Web peut constituer un mode de validation de dépendances syntaxique à distance telles que ADJ-N2,

[4] La traduction correcte de *stools* dans ce contexte est "selles".

l'expression « surgical decompression » ayant environ 200 000 occurrences.[5] La prise en compte de ce type de résultat permettrait d'éviter le découpage incorrect de l'exemple (13) :

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

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 (14), 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).

(14) […] this book has **more detailed therapeutic recommendations and discussions of infections** in patients with AIDS.	[…] ce livre a plus détaillé **des recommandations et des examens thérapeutiques des infections** dans les patients présentant des SIDAS.

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 [ADJ [N1 AND N2]] OF N3 étant majoritaire. Ainsi, dans l'exemple (15), 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.[6]

(15) Second, hemochromatosis can be detected before **any clinical signs or symptoms of disease** develop and even before hepatic iron loading occurs.	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.

6.2.3. 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 2).

[5] Ce chiffre et les suivants indiquent les résultats de requêtes effectuées sur http://www.google.fr/ le 18 octobre 2007.

[6] Une autre solution serait d'intégrer ces collocations aux dictionnaires spécialisés utilisés par les logiciels d'aide à la traduction.

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

Tableau 2: 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 3 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.

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 3 : 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). 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 (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 4) 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
therapeutic recommendations and discussions of infections	69 900	643
adverse events and yields of efficacy	1 890 000	27
subjective awareness or type of symptoms	55 600	17 100
additional personnel and continuity of care	1 910 000	10 600

Tableau 4 : 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 5) 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
different methods and criteria of efficacy	68	11 500
financial incentives or methods of remuneration	1	13 600
substantial overlap and loss of discrimination	3	989
reversible hyperglycemia or development of diabetes	643	375 000

Tableau 5 : Résultats obtenus sur Google pour les collocations N1 of N3 et N2 of N3 dans quelques cas de validation erronée du lien ADJ N2

7. 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.

Références

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, 25-38.

Cormier M. (1990). Traduction de textes de vulgarisation et de textes didactiques : approche pédagogique. *Meta 35/ 4*, 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, Thomas editors, *Approches contrastives en lexicologie bilingue*. Éditions Champion, 213-222.

Frantzi K. T., Tsujii J. and Ananiadou S. (1999). Clustering Terms Using the C-value Method for Automatic Term Recognition. In Sandrini P. editor, *TKE '99, Terminology and Knowledge Engineering*, 356-366.

Kocourek R. (1991). *La langue française de la technique et de la science. Vers une linguistique de la langue savante*. 2e édition. Brandstetter Verlag.

Maniez F. (2000). La prémodification nominale en anglais médical : quelques problèmes de traduction . In Banks D. editor, *Le groupe nominal dans le texte spécialisé*. 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. *Actes du colloque TALN de Tours*, 2-5 juillet 2001, Tome 2, 277-286.

Maniez F. (2002). The Use of Electronic Corpora and Lexical Frequency Data in Solving Translation Problems. In Altenberg, B. and Granger S. editors. *Lexis in Contrast*, *Studies in Corpus Linguistics*, *7*. John Benjamins.

Roberts R. P. & Cormier C. (2000). L'analyse des corpus pour l'élaboration du Dictionnaire canadien bilingue. In Szende, T. editor, *Approches contrastives en lexicologie bilingue*. Éditions Champion.

Rouleau M. (1994). *La traduction médicale, une approche méthodique*. Linguatech.

Rouleau M. (2001). La facture des principaux dictionnaires médicaux français : point de vue d'un traducteur. *Meta, 46/1*, 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>.

Selva T., Verlinde S. and Binon J. (2003). Vers une deuxième génération de dictionnaires électroniques. In Zock, M. and Carroll J. editors. *TAL 44 - 2/2003 Les dictionnaires électroniques*, 177-198.

Thoiron P. (2000). La traduction des termes scientifiques : jeu entre concepts et termes. In *Le Langage scientifique, Congrès National des sociétés historiques et scientifiques, 119e, Amiens, 120e, Aix-en-Provence*, 329-339.

Tournier J. (1985). *Introduction descriptive à la lexicogénétique de l'anglais contemporain*. Champion-Slatkine.