L'ambiguïté syntaxique due aux structures coordonnées en anglais médical : analyse de la performance d'un logiciel d'aide à la traduction.

François Maniez

Centre de Recherche en Terminologie et Traduction
Université Lumière Lyon 2
86, rue Pasteur
69007 LYON
maniezf@univ-lyon2.fr

Résumé

L'ambiguïté syntaxique constitue un problème particulièrement délicat à résoudre pour les analyseurs morphosyntaxiques des logiciels d'aide à la traduction, en particulier dans le cas des longs groupes nominaux typiques des langues de spécialité. Le rattachement des adjectifs et celui des groupes prépositionnels sont deux des difficultés principales de l'analyse automatique des structures syntaxiques de ces groupes nominaux. En exploitant un corpus bilingue d'articles médicaux anglais traduits vers le français et en comparant la traduction obtenue par un humain à celle fournie par un logiciel d'aide à la traduction, nous examinons les cas où l'ambiguïté syntaxique est due à des structures coordonnées impliquant plusieurs noms, et nous suggérons quelques pistes pouvant contribuer à l'amélioration de la performance de ce logiciel dans le domaine médical.

Abstract

Syntactic ambiguity is a particularly difficult problem to solve for the morpho-syntactic analysis programs of machine translation software, especially in the case of the long noun phrases that are typical of technical or scientific writing. Adjective and prepositional phrase attachment are two of the major hurdles in the automatic analysis of the syntactic structure of such noun phrases. Using a bilingual corpus of medical research articles translated into French and comparing human translation with the output of an automatic translation program, we examine the cases in which syntactic ambiguity is due to the presence of coordinated structures involving nouns, and we suggest some directions that may help the disambiguation process in medical texts.

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) -- 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). La plupart des ambiguïtés observées en langue de spécialité concernent la portée de la modification adjectivale ou nominale, la multiplicité des solutions envisageables étant parfois augmentée par le phénomène de la coordination. Si le problème du rattachement des groupes prépositionnels et celui des phrases pouvant induire en erreur les programmes d'analyse automatique (parfois appelées *garden path sentences* dans la littérature anglo-saxonne) a donné lieu à de nombreux travaux, comme ceux de Stevenson (1993, 1994) Ersan (1995) ou Trueswell (1996), le problème de l'ambiguïté syntaxique générée par les structures coordonnées en langue de spécialité semble générer moins d'intérêt, sans doute parce que la résolution de ce type d'ambiguïté nécessite un fort apport d'information lexicale dépendant du sous-langage de spécialité et difficilement généralisable à d'autres domaines.

A partir de l'étude d'un corpus bilingue aligné, nous nous proposons ici d'envisager les diverses manières dont l'ambiguïté des structures impliquant la prémodification de deux noms coordonnés 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 morpho-syntaxique utilisés par les logiciels d'aide à la traduction.

2 L'ambiguïté provoquée par la coordination

La structure de type ADJ N_1 AND N_2 est d'usage très fréquent en anglais, et génère une ambiguïté dans la mesure ou l'adjectif peut qualifier les deux noms coordonnés ou uniquement le premier, d'où deux découpages possibles pour l'expression *old friends and acquaintances*, que l'on peut symboliser à l'aide de crochets de la manière suivante : *[old*

friends] and acquaintances (l'adjectif ne qualifie que le premier nom) et *old [friends and acquaintances]* (il y a "distributivité" de la prémodification aux deux noms).

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 :

(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.*

Au décodage, le traducteur de la phrase (1) est amené à se poser plusieurs questions concernant la structure syntaxique de la dernière partie de la phrase, *increased nucleic acid and protein synthesis and glycolysis*. Certaines de ces questions trouveront chez le traducteur humain une réponse immédiate dans l'examen du contexte et des relations sémantiques qui lient les mots entre eux, mais poseront problème à l'analyseur syntaxique automatique

- *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. 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 "l'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 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. Ainsi, si l'analyseur a accès à une telle base, les découpages supposant l'existence de *nucleic protein* ou de *nucleic synthesis* seront invalidés puisque ces séquences n'y figureront pas, et *nucleic acid synthesis* sera inversement validé. Quant à la résolution de l'ambiguïté concernant la portée de la prémodification par *increased*, elle repose fortement sur la connaissance extralinguistique du domaine de spécialité, même si la fréquence des structures coordonnées à la suite de participes passés comme *decreased* et *increased* peut donner lieu à une analyse statistique de la probabilité de distributivité de la prémodification. Nous nous bornerons dans cet article à suggérer la manière dont les données lexicales peuvent aider la machine à résoudre ce type d'ambiguïté.

3 Ressources utilisées

3.1 Constitution du corpus de langue de spécialité

Pour cette étude des structures coordonnées, nous avons utilisé un corpus bilingue aligné constitué de 58 articles du *Journal of the American Medical Association* et de leurs traductions. Ces dernières, n'étant pas disponibles sous forme électronique, ont été encodées à l'aide d'un logiciel de reconnaissance optique de caractères. Elles sont extraites de la version française du *Journal of the American Medical Association*. La longueur totale de la partie anglaise du corpus est de 134 000 mots. Pour la vérification des informations statistiques

concernant la co-occurence lexicale des mots du vocabulaire médical anglais , nous avons utilisé un corpus d'articles du *New England Journal of Medicine* totalisant 2 millions de mots.

Après avoir évalué divers programmes d'étiquetage automatique, nous avons soumis un échantillon représentant 58% de notre corpus bilingue (77 600 mots) et regroupant tous les articles ayant trait à la cardiologie à l'étiquetage selon les normes du corpus LOB (Lancaster-Oslo/Bergen)[1]. Toutes les suites contenant deux noms coordonnés par *and* ou *or* et précédés d'un adjectif ont ensuite été isolées et recopiées dans une base de données afin d'effectuer des mesures statistiques et des regroupement par indexation automatique.

3.2 Exploitation des données

Nous avons extrait du corpus bilingue un total de 110 suites du type ADJ N_1 AND N_2, ADJ N_1 OR N_2, N_1 N_2 AND N_3, ou N_1 N_2 OR N_3. L'étiquetage automatique a ensuite été vérifié manuellement, afin d'éliminer les formes contenant des étiquetages erronés. Le système d'étiquetage du corpus LOB étant prévu pour l'anglais général, le nombre d'erreurs était relativement élevé, et il ne subsistait que 68 énoncés ambigus à l'issue de cette vérification[2]. Nous avons ensuite soumis l'ensemble de ces énoncés au logiciel de traduction automatique Systran Classic.

Les logiciels d'aide à la traduction livrent des résultats imparfaits, et il est donc facile de se concentrer sur leurs faiblesses. Dans la mesure où cet article a pour objet de suggérer quelques pistes d'amélioration du système d'analyse de l'un d'entre eux, nous serons nécessairement amenés à nous concentrer sur certains découpages erronés de son analyseur syntaxique. Il est donc important de souligner que lors du test effectué, l'analyseur a le plus souvent effectué le découpage correct : seules 21 structures ont été mal interprétées, dont 3 en raison d'un étiquetage grammatical erroné. Les exemples (2) et (3) montrent un résultat qui représente réellement une aide au traducteur, l'analyseur arrivant au découpage correct (respectivement <ADJ [N_1 AND N_2]> et <[ADJ N_1] AND [N_2]>) dans les deux cas. L'exemple (4) constitue un exemple de découpage correct d'une structure <N_1 [N_2 OR N_3]> (dans les exemples qui suivent, nous ferons figurer l'original en italiques, sa traduction française dans notre corpus bilingue, puis la traduction du logiciel en gras).

(2) *[…] it is difficult to draw firm conclusions about the relative efficacy and safety of these agents.*
[…] il est difficile de tirer les conclusions précises sur l'efficacité et l'innocuité de ces médicaments.
[…] il est difficile de tirer des conclusions fermes au sujet de l'efficacité et de la sûreté relatives de ces agents.

(3) *[…] other factors such as social support and depression strongly influence patient function.*
[…]d'autres facteurs comme le soutien social et la dépression influent fortement sur l'état fonctionnel du patient.
[…] d'autres facteurs tels que l'appui social et dépression influencent fortement la fonction patiente.

(4) *Catheter drainage was occasionally complicated by catheter occlusion or infection.*
Dans quelques cas, le drainage s'est compliqué d'une occlusion ou d'une infection du cathéter.
Le drainage de cathéter a été de temps en temps compliqué par l'occlusion ou l'infection de cathéter.

---

[1] Nous tenons à remercier les concepteurs du projet AMALGAM (décrit par Atwell et al., 2000), qui met à la disposition des internautes plusieurs programmes d'étiquetage grammatical de l'anglais. Il est accessible à l'adresse suivante : http://agora.leeds.ac.uk/amalgam/

[2] Outre l'ajout des termes du vocabulaire médical, une amélioration contextuelle de l'étiquetage automatique consisterait à résoudre statistiquement un certain nombre d'ambiguïtés concernant les items lexicaux pouvant appartenir à plusieurs catégories grammaticales : ainsi, dans la prose médicale de langue anglaise, *novel* est toujours employé comme adjectif et *exhibit* presque exclusivement en tant que nom.

Toutefois, dans les phrases contenant plusieurs structures coordonnées, un découpage incorrect peut provoquer une traduction très éloignée de l'original et empêcher le logiciel de remplir sa fonction d'aide à la traduction. Dans l'exemple (5), la distributivité de la prémodification n'est pas appliquée par l'analyseur pour les deux premières occurrences de coordination (*English-language studies and reviews*, *LDL composition and size*) alors qu'elle est nécessaire. Inversement, elle est appliquée pour la troisième occurrence (*a MEDLINE search and reference citations*) alors qu'elle est impossible syntaxiquement.

(5) *English-language studies and reviews pertaining to LDL composition and size were identified through a MEDLINE search and reference citations.*
Les études et revues de langue anglaise, se rapportant à la composition et à la taille des LDL ont été identifiées par l'intermédiaire de Medline et des références citées.
Des études de langue anglaise et les revues concernant la composition en LDL et la taille ont été identifiées par des citations d'une recherche et de référence de MEDLINE

La relative fréquence des structures coordonnées en langue scientifique justifie que l'on se penche sur ce problème. Dans la suite de cet article, nous examinerons divers types de structures syntaxiques et tenterons d'explorer quelques pistes susceptibles de lever l'ambiguïté qu'elles génèrent grâce à l'apport de données lexicales dont l'accès peut être automatisé.

4 Structures syntaxiques ambiguës incorrectement interprétées

4.1 Structure ADJ N_1 AND N_2

Dans l'exemple (6) la traduction de Systran Classic applique la distributivité de la coordination alors que l'adjectif *weight-reducing* qualifie uniquement *diet*. L'amélioration de la performance du programme dépend ici de l'apport de connaissances lexicales (*weight-reducing diet* est un terme composé qui peut être intégré à un dictionnaire spécialisé) et grammaticales (la chlortalidone est un médicament, et le nom n'est donc pas compatible avec l'emploi d'une détermination indéfinie). La difficulté de traitement des adjectifs composés s'ajoute ici à celle de l'interprétation de la structure coordonnée.

(6) *[…] (24% compared with 5% in women who received a weight-reducing diet and chlorthalidone).*
[…] (24 %, contre 5 % des patientes soumises à un régime amaigrissant et traitées par la chlortalidone).
[…] (24% a rivalisé avec 5% chez les femmes qui ont reçu un régime et un chlorthalidone poids-weight-reducing).

La plupart des erreurs de découpage de ces structures par l'analyseur sont dues au mécanisme inverse, c'est-à-dire la non-application de la prémodification au deuxième terme coordonné. Dans les exemples (7) et (8), l'humain interprète les noms des combinaisons telles que *time and expense* ou *sensitivity and specificity* comme étant associés sémantiquement et applique la prémodification aux deux noms coordonnés.

(7) *[…], causing patients the extra time and expense of physicians' office visits.*
[…], ce qui impose aux patients de consulter un médecin.
[…], causant à des patients le temps supplémentaire et dépenses des visites du bureau des médecins.

(8) *The aggregate sensitivity and specificity of ultrasonography and IPG in the medical literature are shown in Table 3.*
Les sensibilité et spécificité cumulées de l'échographie Doppler et de la PGI dans la littérature médicale sont détaillées dans le Tableau 3.
la sensibilité d'agrégat et la spécificité de l'ultrasonography et de l'IPG dans la littérature médicale sont montrées dans le tableau 3.

Cette relation d'association sémantique est toutefois difficile à formaliser dans la mesure où les mots coordonnés ne sont pas des synonymes mais font plutôt référence à des notions complémentaires ou fréquemment mises en opposition. Le recensement de telles paires de mots par la statistique lexicale semble donc une solution plus viable que celle d'une base de données lexicale comme Wordnet, par exemple.

4.2 Structure N_1 N_2 AND N_3

En dehors des trois occurrences décrites dans l'exemple (5), les structures de ce type (par exemple, *catheter occlusion or infection*) ont été correctement interprétées. L'analyseur n'applique pas systématiquement le même découpage, comme l'exemple (5) le démontre, mais le nombre d'occurrences de cette structure à l'intérieur de notre corpus est trop faible pour que l'on puisse déduire l'algorithme qui sous-tend les choix de l'analyseur. L'apport d'information lexicale concernant les collocants de *laboratory* permettrait sans doute d'invalider *laboratory treatment* comme étant peu probable, puisque *treatment* ne figure pas parmi les collocants postérieurs immédiats de *laboratory* dans notre corpus unilingue.[3]

(9) *[…] that will guide laboratory assessments and treatment.*
[…] afin de guider les examens complémentaires et le traitement.
[…] qui guideront des évaluations et le traitement de laboratoire.

4.3 Structure N_1'S N_2 AND N_3

Nous n'avons relevé que deux structures ambiguës impliquant une prémodification par un génitif. Dans ces deux cas, l'analyseur de Systran n'a pas choisi la distributivité de la prémodification qui s'imposait dans les expressions *the patients' symptoms and signs* et *the patient's condition and tumor type*.

(10) *[…] the likelihood that the patients' symptoms and signs could be assigned to a specific vascular territory.*
[…] la probabilité avec laquelle les symptômes pouvaient être attribués à l'atteinte d'un territoire vasculaire donné.
[…] la probabilité qui les symptômes des patients et signe pourraient être assignées à un territoire vasculaire spécifique.

(11) *[…] the patient's condition and tumor type, the success rates and risks of the various modalities, and local availability and expertise.*
[…] de l'état du patient, du type de la tumeur, du taux de succès et de complications de chaque méthode, de leur disponibilité dans chaque centre et de leur évaluation.
[…] l'état du patient et le type de tumeur, les taux de succès et les risques des diverses modalités, et disponibilité locale et expertise.

Une analyse de contrôle de 9 structures équivalentes a révélé que l'analyseur n'applique pas systématiquement l'absence de distributivité. Ainsi, elle a été correctement appliquée pour *patients' finances and insurance*, *the patient's parents and six siblings*, *the patient's insurability and employment*, mais elle ne l'a pas été pour *the patient's plasma and purified protein S*, *this patient's lymphoma and bone marrow disorder*, *the patient's blood and bone marrow* et *the patient's blood and frozen tumor specimen*. (ce dernier cas n'étant pas

[3] L'utilisation de ces données devrait idéalement prendre en compte l'appartenance à une classe de synonymes. Ainsi, dans notre corpus unilingue, *assessment* n'est utilisé que 3 fois après *laboratory* sur 679 occurrences, ce qui ne justifierait pas nécessairement l'automatisation de la distributivité dans une structure du type *laboratory N1 and assessments*. Toutefois, on relève dans la même position 97 emplois de ses proches synonymes (*evaluation(s)*, *examination(s)*, *measures*, *measurements*, *testing*, *values*) et 107 emplois de mots sémantiquement reliés (*data*, *diagnosis*, *evidence*, *findings*, *records*, *results*).

réellement ambigu syntaxiquement puisque le mot *specimen*, indénombrable, est nécessairement prémodifié par le génitif *the patient's* en l'absence d'article). Au vu de ces quatre derniers exemples, il semble que l'analyseur n'applique pas la distributivité de la prémodification dans les cas où le deuxième groupe nominal coordonné est long ou complexe.

4.4 Structure ADJ_1 ADJ_2 N_1 AND N_2

Dans l'exemple (12) la traduction de Systran Classic n'applique la distributivité de la coordination alors que les adjectifs *coronary* et *angiographic* qualifient à la fois *equipment* et *techniques*. Ici, la démarche qui consisterait à s'assurer de la fréquence statistique de la suite obtenue à partir de la distribution au deuxième terme de l'expression coordonnée (*coronary angiographic techniques*) n'est pas la mieux adaptée. En effet, les mots *equipment* et *techniques* sont d'un usage trop fréquent pour que leurs emplois à la suite d'adjectifs ou de noms désignant des interventions en chirurgie ou en imagerie soient systématiquement répertoriés et lexicalisés. De plus, l'emploi de l'adjectif *coronarographic* constitue ici une forme particulière d'hypallage[4], puisque l'expression anglaise signifie *equipment and techniques that are used in coronary angiography*, ce dernier terme pouvant être traduit par "angiographie coronaire" ou "coronarographie".

(12) *[…] when coronary angiographic equipment and techniques were not as sophisticated as they are today.*
[…] lorsque la technique et les appareils de coronarographie n'avaient pas atteint le niveau de perfectionnement actuel.
[…] où l'équipement angiographique coronaire et les techniques n'étaient pas aussi sophistiqués qu'ils sont aujourd'hui.

La raison pour laquelle le traducteur humain n'hésite guère devant ce genre de structure réside plutôt dans le fait que les mots *equipment* et *techniques* appartiennent au même champ lexical. On ne peut dire pour autant qu'ils sont fréquemment associés, puisque nous n'avons trouvé que deux exemples de co-occurrence dans notre corpus unilingue.[5] Plutôt que le recours à des données statistiques brutes concernant la co-occurrence des deux termes coordonnés en corpus, l'appartenance au même champ d'une base de données organisée selon le modèle d'un thésaurus semble plus appropriée. Une autre approche consisterait à répertorier toutes les structures coordonnées pour lesquelles s'applique la distributivité de la prémodification.

4.5 Structure N_1 N_2 AND N_3 N_4

Dans l'exemple (13) , le fait que le nom *autoantibody* soit vraisemblablement interprété par l'analyseur comme un adjectif n'est pas essentiel, cette erreur pouvant être éliminée par son ajout au dictionnaire spécialisé. Il s'agit plutôt de déterminer comment empêcher l'analyseur de l'interpréter comme modifiant les deux termes de l'expression coordonnée (*studies* et *complement levels*).

---

[4] Certains exemples de ce type de glissement existent déjà dans la terminologie médicale. Ainsi, *smooth muscle cells* se traduit par "cellules musculaires lisses", alors qu'il s'agit des cellules des muscles lisses (par opposition aux muscles striés) et "infarctus ventriculaire droit" s'emploie pour désigner l'infarctus du ventricule droit.

[5] A titre de comparaison, les deux noms coordonnés dans l'exemple (2), *efficacy* et *safety*, entrent en co-occurrence 76 fois dans le même corpus.

(13) *Apart from a positive rheumatoid factor, results of autoantibody studies and complement levels were unremarkable.*
A l'exception de la positivité de la recherche du facteur rhumatoïde, les résultats des recherches d'auto-anticorps et des dosages du complément n'ont pas révélé d'anomalies.
Indépendamment d'un facteur rheumatoid positif, les résultats des études et des niveaux autoantibody de complément étaient unremarkable.

Même si le découpage <N [[N] AND [NN]]> est beaucoup moins fréquent que le découpage <ADJ [[N] AND [NN]]>, il est théoriquement possible. La statistique lexicale peut ici aider à la désambiguïsation en invalidant la suite *autoantibody complement* puisque *complement* ne figure pas parmi les collocants postérieurs du nom *autoantibody* (*affinity*, *determinations*, *formation*, *levels*, *production* et *response*) dans notre corpus.

4.6 Structure ADJ N_1 AND N_2 N_3

La suite *ventricular size and wall thickness* dans l'exemple (14) est susceptible de subir quatre découpages distincts : [*ventricular size*] *and* [*wall thickness*], *ventricular* [[*size*] *and* [*wall thickness*], *ventricular* [[*size and wall*] *thickness*] et [*ventricular* [*size and wall*]] *thickness*. Le deuxième découpage est correct, ce qui n'apparaît pas dans la version française, le traducteur ayant choisi une formule elliptique plus proche du premier découpage (l'épaisseur pariétale = l'épaisseur de la paroi du ventricule).

(14) *By contrast, the echocardiogram allows measurement of ventricular size and wall thickness and evaluation of valvular function and pericardial disease.*
L'échocardiogramme, au contraire, permet d'évaluer les dimensions ventriculaires, l'épaisseur pariétale, le fonctionnement des valves et d'éventuelles anomalies péricardiques.
En revanche, l'échocardiogramme permet la mesure de l'épaisseur ventriculaire de taille et de paroi et l'évaluation de la fonction valvular et de la maladie péricardique.

L'analyseur de Systran a opté pour le troisième découpage, qui pourrait être éliminé par des méthodes lexicales invalidant *size thickness*. L'élimination du premier découpage est beaucoup plus subtile et requiert l'apport d'information de type sémantique (une paroi peut être ventriculaire). On remarquera que le long groupe nominal qui suit *allows* et termine la phrase consiste en une double structure coordonnée associant deux groupes nominaux de structure quasi-identique. Cette structure complexe a été correctement interprétée par l'analyseur, l'absence de virgules imposant un découpage intermédiaire parfaitement symétrique: [*measurement of [ventricular size AND wall thickness]] AND [evaluation of [valvular function AND pericardial disease]*. Soulignons au passage que la tâche de l'humain cherchant à analyser ces structures complexes est grandement facilitée par la synonymie des mots *measurement* et *evaluation*, ce type d'information n'étant vraisemblablement pas accessible à l'analyseur. Dans l'exemple (15), c'est le quatrième type de découpage mentionné plus haut qui est correct, c'est-à-dire [*left* [*arm and leg*]] *weakness*.

(15) *[…] a 68-year-old white man admitted with left arm and leg weakness.*
[…] un homme âgé de 68 ans, de race blanche, hospitalisé pour une diminution de la force musculaire du bras et de la jambe gauches.
[…].un homme 68-year-old blanc admis avec la faiblesse gauche de bras et de jambe […].

De nouveau, l'analyseur opte pour le troisième type de découpage. Notre corpus ne contenant pas d'autres exemples de structures de ce type, nous n'avons pu déterminer si l'analyseur choisit systématiquement ce découpage, qui semble le plus fréquent pour cette structure relativement rare.

5 Intégration de ressources lexicales

On a vu plus haut que l'adoption du découpage adéquat par le traducteur humain pour les structures de type ADJ [N_1 AND N_2] ou [N_1 AND N_2] N_3 dépend de l'identification des deux noms coordonnés comme formant une unité indissociable, cette identification provenant de la perception de similitudes sémantiques entre les deux noms. Le stockage de ces structures coordonnées et de leurs équivalents dans la langue cible par le programme de traduction peut donc être bénéfique, s'il s'avère que les énoncés les contenant donnent systématiquement lieu au même découpage syntaxique. Nous avons donc relevé les structures de type N_1 AND N_2 employées au moins trois fois dans un sous-ensemble de notre corpus unilingue comptant 500°000 mots, soit un quart du corpus. Un total de 101 structures coordonnées a ainsi été isolé. Le Tableau 1 regroupe les structures de fréquence supérieure à 7.

N_1 AND N_2	FRQ
head and neck	15
control and prevention	14
hematoxylin and eosin	13
women and men	12
trioleate and trierucate	12
patients and controls	11
lung and blood	11
lumpectomy and radiation	11
fat and cholesterol	11
data and safety	11
safety and efficacy	10
food and drug	9
education and income	8
blood and bone	8

Tableau 1 : Séquences du type N_1 AND N_2 les plus fréquentes dans le corpus unilingue.

L'apport de ces données brutes peut contribuer à aider l'analyseur à effectuer un découpage correct, mais on peut envisager deux types d'affinement de l'apport d'information lexicale. D'une part, certaines de ces séquences font partie de groupes nominaux plus longs qui peuvent être ajoutés au dictionnaire spécialisé (dans notre corpus, *data and safety* n'est présent qu'à l'intérieur des séquences *data and safety monitoring committee* et *data and safety monitoring board*). Certaines d'entre elles (*National Heart, Lung and Blood Institute*, *Food and Drug Administration*) sont en fait des noms propres désignant des institutions médicales connues et fréquemment citées. D'autre part, certaines séquences (par exemple, *head and neck* ou *blood and bone*) ne sont présentes que dans un seul type de structure ([N_1 AND N_2] N_3), et d'autres ne subissent aucune prémodification dans notre corpus (*safety and efficacy*). Par ailleurs, la fréquence d'utilisation de certaines séquences à l'intérieur d'un groupe nominal de taille supérieure impose la lexicalisation de ce groupe dans son intégralité (*case patients and controls*), alors que d'autres (*education and income*) subissent des modifications diverses qui justifient le stockage de l'expression coordonnée uniquement[6]. Par ailleurs, la distributivité de la prémodification ne peut pas être déduite systématiquement à partir de données statistiques concernant la fréquence de la coordination. Ainsi, dans la séquence fréquemment observée *saturated fat and cholesterol*, l'adjectif *saturated* ne qualifie que le nom *fat*.

[6] A titre d'exemple, l'analyseur de Systran effectue le découpage correct pour *lower [education and income] levels* mais pas pour *a particular [education or income] category.*

6 Conclusion

L'approche évoquée ici, c'est-à-dire l'utilisation des données statistiques concernant la co-occurrence des noms coordonnés et des séquence de forme ADJ N semble à même de résoudre la plupart des ambiguïtés évoquées. L'approche lexicographique consistant en une compilation des lexies à mots multiples et des collocations du domaine de spécialité constituerait sans doute un préalable nécessaire, mais un certain nombre de questions demeurent concernant l'automatisation de l'analyse syntaxique du groupe nominal en langue de spécialité. Pour que les systèmes automatisés d'aide à la traduction deviennent plus fiables en langue de spécialité, un recensement exhaustif de tous les types de structures ambiguës est nécessaire. Une étude détaillée de la traduction des groupes nominaux faisant intervenir la coordination semble constituer un préalable souhaitable en ce domaine.

Références

Atwell Eric et al. (2000), "A comparative evaluation of modern English corpus grammatical annotation schemes" in *ICAME Journal N° 24*, pp 7-23.

Bourigault D. (1992), "Lexter: un logiciel d'extraction de terminologie." In *TAMA '92, Actes du 2° Symposium TermNet : Applications terminologiques et microordinateurs*, Vienne, Autriche.

Ersan M, Charniak E. (1995) "A Statistical Syntactic Disambiguation Program and What It Learns", Technical Report, Department of Computer Science, Brown University

Maniez F. (2001), "Extraction d'une phraséologie bilingue en langue de spécialité : corpus parallèles et corpus comparables" *Meta, 46-2*

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

Sinclair J. (1991), *Corpus, Concordance, Collocation* (Oxford : Oxford University Press)

Stevenson S. (1993). ``A Competition-Based Explanation of Syntactic Attachment Preferences and Garden Path Phenomena." *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, 266--273.

Stevenson S. (1994) "A Competitive Attachment Model for Resolving Syntactic Ambiguities in Natural Language Parsing", Doctoral dissertation, University of Maryland."

Teubert W. (1996), "Comparable or Parallel Corpora? " In: *International Journal of Lexicography* Vol 9, N° 3, pp 238-264.

Trueswell, J.C. (1996). "The role of lexical frequency in syntactic ambiguity resolution." *Journal of Memory and Language*, 35, 566-585.

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

Van Hoof H. (1986), *Précis pratique de traduction médicale*, Maloine.