Désambiguïsation syntaxique des groupes nominaux en anglais médical : étude des structures adjectivales à partir d'un corpus bilingue.

François Maniez

Centre de Recherche en Terminologie et en 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. En utilisant un corpus bilingue d'articles médicaux anglais traduits vers le français, nous examinons divers moyens de résoudre l'ambiguïté du rattachement de l'adjectif à l'un des deux noms qui le suivent dans les tournures anglaises de forme adjectif-nom-nom.

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. Using a bilingual corpus of medical research articles translated into French, we examine various ways in which disambiguation can be achieved in English adjective-noun-noun structures, where the adjective may modify either of the following nouns.

Mots clés

adjectif, ambiguïté syntaxique, anglais médical, corpus bilingue, découpage, groupe nominal, traduction, traduction automatique.

Keywords

adjective, bilingual corpus, machine translation, medical English, noun clause, syntactic ambiguity, syntactic structure, translation.

1 Introduction

L'ambiguïté syntaxique est un phénomène inhérent à de nombreuses 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 (1985:169) -- 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). 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. A partir de l'étude d'un corpus bilingue aligné, nous nous proposons ici d'envisager les diverses manières dont l'ambiguïté des structures Adjectif - Nom - Nom 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 automatique utilisés par les logiciels d'aide à la traduction.

2 Ressources utilisées

2.1 Constitution du corpus de langue de spécialité

Pour cette étude des structures adjectivales, 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.

Après avoir évalué divers programmes d'étiquetage automatique, nous avons soumis un échantillon représentant 58% de notre corpus (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 du type Adjectif - Nom - Nom ont ensuite été isolées et recopiées dans une base de données afin d'effectuer des mesures statistiques et des regroupements par indexation automatique.

2.2 Exploitation des données

Un total de 272 suites de type Adjectif - Nom - Nom a pu être extrait du corpus, totalisant 552 occurrences. L'étiquetage automatique a ensuite été vérifié manuellement, afin d'éliminer les formes contenant des étiquetages erronés. On a dénombré un total de 34 erreurs. Six erreurs étaient dues au choix de l'adjectif comme premier constituant pour des mots appartenant dans le contexte à une autre catégorie grammaticale (*following*, *separate*, *spare*, *called*, *involved*, *studied*). Quatre erreurs concernaient l'identification d'adjectifs du lexique spécialisé en tant que noms pour le deuxième constituant (*sedentary*, *codominant*, *viridans*[2], *hyperacute*). La majorité des erreurs d'étiquetage provenait néanmoins de l'identification de verbes en tant que noms (*benefit*, *cause(s), contracts*, *correlates*, *exhibit*, *favor*, *increase(s), pertain*, *plays*, *pose*, *results*) pour le troisième constituant. Signalons toutefois que le système d'étiquetage du corpus LOB est prévu pour l'anglais général, ce qui explique les erreurs constatées. Outre l'ajout des termes du vocabulaire médical, une amélioration facile à mettre en œuvre 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[3].

On a donc finalement isolé 234 suites de type Adjectif - Nom - Nom correctement étiquetées, totalisant 490 occurrences. Le Tableau 1 regroupe les formes les plus fréquemment rencontrées dans le corpus.

On a ensuite procédé manuellement à un étiquetage syntaxique binaire des formes isolées, en ajoutant un champ à la base de données, auquel on a attribué la valeur N1 si l'adjectif qualifiait le deuxième constituant (c'est-à-dire le premier nom) de la séquence, et la valeur N2 si l'adjectif en qualifiait le troisième constituant (le deuxième nom). Cet étiquetage a pu être effectué sans avoir recours au contexte pour 204 séquences sur 234 (87% des cas). Les 30 séquences restantes ont été examinées en contexte afin d'être désambiguïsées.

A l'issue de l'étiquetage, on dénombrait 80 formes (34%) dans lesquelles l'adjectif qualifiait le premier nom et 152 formes (65%) dans lesquelles l'adjectif qualifiait le deuxième nom. Dans deux cas de figure (*the nearly normal erythrocyte sedimentation rate* → la quasi-normalisation de la vitesse de sédimentation, *positive thallium stress test* → test d'effort au thallium positif), l'adjectif modifiait un troisième nom qui constituait le nœud d'une lexie composée, et ces formes n'ont donc pas été comptabilisées. On a relevé un seul cas de structure de type Adjectif - Nom - Nom - Nom (*chronic pacing threshold rise* → augmentation du seuil chronique de stimulation) dans laquelle le deuxième nom était modifié. Un nombre non négligeable des suites analysées (29, soit 12%) faisaient partie de syntagmes

---

[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] Les mots latins, quand ils sont reconnus par l'étiqueteur, reçoivent l'étiquette &FW (foreign word).

[3] L'appartenance de certains mots à plusieurs catégories grammaticales donne lieu à des variations typiques de la langue de spécialité. Ainsi, "lipoprotéines résiduelles" se dit indifféremment *lipoprotein remnants* ou *remnant lipoproteins*.

nominaux complexes contenant des syntagmes prépositionnels imbriqués (*subnormal cardiac output response to exercise* → réponse anormale du débit cardiaque à l'effort).

Séquence Adjectif - Nom - Nom	N
coronary artery disease	48
ischemic stroke subtype	8
coronary risk factors	7
overt heart failure	7
normal ejection fraction	7
low ejection fraction	7
diastolic blood pressure	5
systolic blood pressure	5
mitral valve prolapse	5
visceral fat accumulation	4
sexual function questionnaire	4
giant cell arteritis	4
valvular heart disease	4
congestive heart failure	4
transvenous lead systems	4
conjugate eye deviation	3
coronary artery lesions	3
atherogenic lipoprotein particles	3
prosthetic heart valves	3

Tableau 1 : Séquences du type Adjectif - Nom - Nom les plus fréquentes.

3 Recherche d'éléments communs aux mêmes structures syntaxiques

On a vu plus haut qu'il était possible d'arriver à une analyse syntaxique correcte dans 87% des cas au seul vu des trois constituants des séquences automatiquement sélectionnées. Dans les 13% de cas restants, la consultation du contexte élargi et / ou celle de la traduction française ont permis de lever l'ambiguïté. Les mécanismes qui permettent à l'humain de résoudre les ambiguïtés syntaxiques sont complexes et multiples, et dépendent d'une interprétation du contexte qui dépasse souvent le seul groupe nominal et parfois la phrase qui le contient. Ainsi, dans la phrase "*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*.", il est nécessaire de franchir plusieurs étapes successives afin de déterminer si *altered* qualifie *substrate* ou *requirements*. La première est l'examen de la suite *altered substrate*, qui n'est pas immédiatement identifiée comme étant un terme ou une collocation de la langue médicale (même si cette expression a cours en géologie). On peut ensuite chercher à donner un sens à la décomposition issue d'un autre modèle syntaxique, celui qui résulte de la modification du deuxième nom (*altered substrate requirements* = "*changes in substrate requirements* "). La dernière étape, pour le profane qui ne connaît pas les mécanismes décrits, consiste à identifier le lien sémantique entre *altered* et *increased*. Dans la mesure où *alteration* peut être considéré comme un hyperonyme du nom *increase*, la relation de cause à effet paraît plus facilement explicable si *altered* qualifie *requirements*.

Dans le cas des structures relevées dans notre corpus, la plupart des cas sont heureusement beaucoup moins complexes pour l'humain, et peuvent être désambiguïsés grâce à la connaissance préalable du lexique spécialisé. Ainsi, *coronary artery disease* est directement

analysable par qui a déjà rencontré *coronary artery* isolément. Le fait qu'il existe une formulation elliptique (*coronary disease*, traduit alternativement par "maladie coronaire" ou "maladie coronarienne") ne change rien au fait que cette séquence sera décodée comme signifiant *disease of the coronary artery* plutôt que *coronary disease of an artery*. Quant au degré de probabilité de co-occurrence des deux noms, il ne semble pas influer sur l'interprétation de la structure syntaxique : *coronary risk factors* est traduit par "facteurs de risque coronaire / facteurs de risque coronarien" dans notre corpus, mais *hypertensive heart disease* est traduit par "cardiopathie hypertensive". Le fait que *risk factor* et *heart disease* soient tous deux des collocations très fréquemment employées ne signifie pas que l'adjectif modifiera nécessairement l'un ou l'autre des deux noms dans la traduction.

3.1 Comparaison des probabilités de co-occurrence

Examinons à présent quelques critères pouvant entrer dans la désambiguïsation automatique des structures syntaxiques qui sont l'objet de notre étude. Le premier est la comparaison de la probabilité de co-occurrence des deux groupes de deux mots (ADJ-N1 et N1-N2) considérés isolément. En effet, dans les cas où l'adjectif qualifie le premier nom, on peut s'attendre à ce que la suite ainsi formée (si elle est un constituant terminologique de l'ensemble de la séquence) soit présente plus d'une fois dans le corpus. Si l'adjectif qualifie le deuxième nom (ou l'ensemble N1-N2) dans d'autres occurrences du corpus, on peut supposer qu'il en va de même pour la séquence de deux noms qu'il précède.

Nous avons donc cherché à vérifier cette hypothèse en utilisant les séquences citées dans le Tableau 1. Nous avons pour cela utilisé l'indice de probabilité de co-occurrence nommé z-score du logiciel Tact (les valeurs ont été arrondies à l'entier le plus proche)[4]. La valeur du z-score peut aller jusqu'à 200 pour des mots que l'on ne retrouve qu'en association dans un corpus, et elle est proche de 0 pour les mots dont la co-occurrence est extrêmement rare. Par exemple, dans le Tableau 2, les valeurs très élevées observées pour *mitral valve* (193) et *valve prolapse* (172) sont dues au fait que *mitral* apparaît cinq fois dans le corpus, toujours suivi de *valve*, lui-même employé sept fois (*prolapse* n'apparaît que quatre fois, toujours dans la séquence *mitral valve prolapse* = prolapsus de la valve mitrale).

L'examen du Tableau 2 fait apparaître que dans la quasi-totalité des cas (18/19, soit 90%) cette probabilité de co-occurrence est un indicateur fiable de la portée de l'adjectif. Le seul cas de figure dans lequel le z-score ne confirme pas la portée de la modification adjectivale (*conjugate eye deviation* = déviation conjuguée du regard) est d'ailleurs dû à un manque de finesse de l'étalon que nous avons choisi. En effet, *conjugate* et *deviation* n'étant employés qu'une seule fois dans un autre environnement, les z-scores sont identiques. Il faut également signaler que le codage que nous avons choisi opère une dichotomie dans certains cas où la portée de la modification adjectivale n'est pas clairement définie. Ainsi, dans *coronary risk factors*, il semble bien que l'expression *risk factors* tout entière soit qualifiée par l'adjectif

---

[4] La formule de calcul du z-zcore est la suivante :
Z =(fréquence du collocant-E)/Ecart type
Ecart type = Racine carrée de [Longueur du mini-texte * P * (1-P)]
E = P * longueur du mini-texte, et :
P = Fréquence du collocant sur l'ensemble du texte / longueur du texte.
Le mini-texte représente le nombre total de mots à gauche et à droite du mot-cible (10 par défaut) constituant la "fenêtre" à l'intérieur de laquelle les collocants sont recherchés. La prise en compte d'une fenêtre de plus faible taille augmente donc la valeur du z-score.

coronary, mais *coronary risk* fait également partie d'un découpage syntaxique plausible. La traduction française emploie d'ailleurs systématiquement "facteurs de risque coronaire".

Séquence Adjectif - Nom - Nom	Codage	z-score ADJ-N1	z-score N1-N2
coronary artery disease	N1	115	88
ischemic stroke subtype	N1	67	61
coronary risk factors	N2	11	33
overt heart failure	N2	56	104
normal ejection fraction	N2	21	210
low ejection fraction	N2	32	210
diastolic blood pressure	N2	16	130
systolic blood pressure	N2	17	130
mitral valve prolapse	N1	193	172
visceral fat accumulation	N1	118	63
sexual function questionnaire	N1	40	17
giant cell arteritis	N1	93	25
valvular heart disease	N2	15	41
congestive heart failure	N2	24	105
transvenous lead systems	N1	38	26
conjugate eye deviation	N2	118	118
coronary artery lesions	N1	115	13
atherogenic lipoprotein particles	N2	17	30
prosthetic heart valves	N2	24	26

Tableau 2 : Probabilité de co-occurrence des suites ADJ-N1 et N1-N2 des séquences du type Adjectif - Nom - Nom les plus fréquentes.

L'indice de probabilité de co-occurrence des suites ADJ-N1 et N1-N2 semble donc un indicateur fiable. Il convient de remarquer que sa prise en compte nécessite l'utilisation d'un corpus lemmatisé et étiqueté grammaticalement. Ainsi, dans la séquence *transvenous lead systems*, l'indice de probabilité de co-occurrence du segment ADJ-N1 *transvenous lead* (électrode endocavitaire) diminue fortement si les occurrences du verbe *lead* sont prises en compte et si celles de la séquence contenant le nom au pluriel (*transvenous leads*) ne le sont pas.

L'examen des traductions obtenues par un programme d'aide à la traduction (Systran Classic 3.0) confirme qu'une identification préalable des termes spécialisés ou des lexies composées formées à partir de ceux-ci éviterait la plupart des erreurs de découpage observées. Nous avons proposé à ce programme les 19 séquences du tableau 2. Le programme a effectué un découpage correct dans 13 cas, employant même la terminologie correcte à plusieurs reprises (nous avons considéré comme correctes les traductions ne comprenant pas d'erreur manifeste, comme "arrêt du cœur congestif" pour *congestive heart failure*, où l'on peut comprendre que "congestif" qualifie "arrêt" ou "cœur"). Les erreurs de découpage sont résumées dans le Tableau 3.

Dans les séquences du Tableau 3, l'absence de reconnaissance des collocations ADJ-N1 (*ischemic stroke*, *visceral fat*, *sexual function*, *giant cell*, *transvenous lead*) ou N1-N2 (*ejection fraction*) conduit donc à un découpage incorrect, dont l'adoption est favorisée par d'autres facteurs : la polysémie de *stroke* (le problème est facilement réglé par l'utilisation d'un dictionnaire personnalisé, puisque ce mot signifie presque toujours "accident vasculaire cérébral" en anglais médical), l'interprétation de *fat* comme adjectif ou encore l'absence de termes du vocabulaire spécialisé (*arteritis* et *transvenous*) du dictionnaire du logiciel.

Original	**Traduction Systran**	**Traduction correcte**
ischemic stroke subtype	sous-type ischémique de course	sous-type d'AVC ischémique
normal ejection fraction	fraction normale d'éjection	fraction d'éjection normale
visceral fat accumulation	grosse accumulation viscérale	accumulation de graisse viscérale
sexual function questionnaire	questionnaire sexuel de fonction	questionnaire concernant l'activité sexuelle
giant cell arteritis	arteritis géant de cellules	artérite à cellules géantes
transvenous lead systems	systèmes transvenous de fil	systèmes utilisant des électrodes endocavitaires

Tableau 3 : Découpage incorrect des suites ADJ-N1-N2 par le logiciel Systran Classic.

3.2 Comparaison des catégories syntaxiques et sémantiques

La comparaison des groupes distingués par les deux types de découpages fait apparaître un certain nombre de distinctions d'ordre morphologique et sémantique entre les adjectifs utilisés. Ainsi, lorsque l'adjectif qualifie le deuxième nom, les adjectifs utilisés rentrent presque toujours dans l'une des catégories suivantes :

- adjectifs courts d'utilisation fréquente (*deep*, *high*, *low*, *new*, *rare*) : lorsqu'un tel adjectif est utilisé dans une séquence de type ADJ-N1-N2, il ne qualifie le premier nom que dans 6% des cas.
- participes passés employés comme adjectifs (*diminished*, *improved*, *increased*, *shortened*). Un grand nombre d'entre eux expriment un changement d'état, et leur traduction fait très souvent apparaître une transposition vers le nom (*diminished blood flow* → diminution du débit sanguin, *shortened platelet survival* → raccourcissement de la durée de vie plaquettaire). Dans cette catégorie, l'adjectif ne qualifie le premier nom que dans 8% des cas.
- comparatif et superlatif des adjectifs : dans notre corpus, l'adjectif qualifiait le deuxième nom pour la totalité de ces formes (13 occurrences).
- adjectifs à sens temporel, exprimant la fréquence, la vitesse, l'ordre chronologique, etc. (*current*, *daily*, *frequent*, *original*, *previous*, *progressive*, *prompt*, *recent*, *rare*, *subsequent*).
- adjectifs exprimant le degré (*absolute*, *complete*, *considerable*, *extensive*, *important*, *significant*).
- adjectifs exprimant la quantité (*additional*, *cumulative*, *numerous*, *various*).
- adjectifs exprimant un jugement de valeur (*adequate*, *appropriate*, *defective*, *effective*, *negative*, *positive*).
- adjectifs exprimant la modalité assertive (*eventual*, *potential*).

Par comparaison, les adjectifs qualifiant le premier nom font presque toujours partie du vocabulaire spécialisé (*atherogenic*, *atrial*, *cardiac*, *coronary*, *ischemic*, *mitral*, *systolic*, *temporal*, *visceral*, etc.). La longueur moyenne de ces adjectifs est légèrement supérieure à celle des adjectifs qualifiant le deuxième nom, et 17% d'entre eux sont des adjectifs composés contre 12% pour les adjectifs qualifiant le premier nom. On a relevé un seul contre-exemple concernant les catégories décrites plus haut (*large artery atherosclerosis* → athérosclérose des gros troncs).

Il semble donc que cette méthode, plus chronophage que la précédente dans la mesure où elle nécessite l'étiquetage préalable des adjectifs en fonction de la probabilité de leur rattachement au premier ou au deuxième nom des séquences de type ADJ-N1-N2, soit aussi performante que celle qui consiste à repérer les collocations ADJ-N1 et N1-N2 séparément et à comparer leurs indices de probabilité de co-occurrence.

3.3 Comparaison des équivalents de traduction à partir du corpus bilingue

Les corpus bilingues se prêtent à deux utilisations distinctes. En cas de doute sur le rattachement de l'adjectif, l'examen des diverses traductions d'une même séquence (ou des traductions d'une séquence employant le même adjectif) peut fournir des indications statistiques sur une fréquence de rattachement supérieure pour l'un des deux noms de la séquence. En dépit de certains indices propres à la traduction française de la séquence (place ou accord de l'adjectif), cette traduction peut toutefois elle-même consister en une structure syntaxique ambiguë, qui nécessitera donc une analyse propre.

On peut également tenter d'extraire des corpus bilingues des traductions déjà employées pour les séquences étudiées dans le but de court-circuiter le processus de désambiguïsation. Cette approche fonctionne bien dans tous les cas où la traduction de la séquence ADJ-N1-N2 subit peu de variations. Ainsi, la séquence la plus fréquente de notre corpus, *coronary artery disease*, est traduite dans la majorité des cas par "maladie coronaire" (traduction qui tend récemment à supplanter "maladie coronarienne"). Cette réalité statistique fait que la question de savoir lequel des deux noms est qualifié par l'adjectif *coronary* a peu d'importance (il est toutefois intéressant de noter un certain nombre de cas dans lesquels cette traduction n'apparaît pas en raison soit d'une ellipse, soit de l'utilisation du nom "coronaropathie" ou de "coronarien" (employé comme nom ou comme adjectif): *vasospastic coronary artery disease* → vasospasme coronaire, *greater risk for coronary artery disease* → risque coronarien accru, *patients with coronary artery disease* → les coronariens, *atherosclerotic coronary artery disease* → coronaropathie artériosclérotique).

(1) Thus, **ischemic stroke subtype** is often never established with certainty, and acute therapeutic decisions must often be made with the knowledge that the **ischemic stroke subtype** diagnosis may be inaccurate.	Le **sous-type de l'AVC ischémique** reste donc souvent incertain, et les décisions thérapeutiques au stade aigu se font souvent en sachant que le **diagnostic du sous-type** est peut-être erroné.
(2) However, only slight diagnostic precision resulted when **ischemic stroke subtypes** were collapsed into a single category.	Cependant, la précision diagnostique ne s'est améliorée que légèrement lorsque **tous les accidents ischémiques** ont été regroupés en une seule catégorie.
(3) However, […] differences between the initial and final […] **ischemic stroke subtype diagnosis** frequently occur .	Cependant, […] **les diagnostics de sous-type** formulés initialement diffèrent souvent des diagnostics finals.
(4) Accurate diagnosis of **ischemic stroke subtype** requires a more extensive laboratory investigation.	**Un diagnostic exact du sous-type d'AVC ischémique** exige des examens complémentaires plus approfondis.
(5) The distinction between **ischemic stroke subtypes** (ie, large artery thrombosis vs cardiogenic embolism) is also an important guide […].	La distinction entre **sous-types d'AVC ischémiques** (athérosclérose des gros troncs ou embolie cardiogénique) est également un élément important […].
(6) Precision in **the classification of ischemic stroke subtype** […] is higher […]	**La classification des accidents ischémiques en sous-types** […]a une meilleure précision […]
(7) […] differences between initial and final **ischemic stroke subtype diagnoses** frequently occur.	[…] le **diagnostic** final **du sous-type d'un AVC ischémique** est souvent différent de ce qu'il était initialement.

Tableau 4 : Traductions de la séquence *ischemic stroke subtype*.

Toutefois, dans certains cas, l'extrême diversité des traductions employées fait que le repérage automatique des équivalents de traduction est difficile, le choix de l'équivalent de traduction le plus pertinent l'étant encore plus. Si l'on examine les traductions de la deuxième séquence de notre étude par ordre de fréquence (*ischemic stroke subtype*), regroupées dans le Tableau 4,

on s'aperçoit de la diversité des formes rencontrées. L' exemple (1) fait apparaître le problème du choix du type de détermination ("sous-type **d'un** AVC ischémique" peut être employé) et de l'ellipse (ou condensation, cf. Lethuiller, 1989) qui fait que le mot *stroke* n'est pas traduit. Dans les exemples (2) et (3), ce sont respectivement *subtypes* et *ischemic stroke* qui subissent cette condensation. Les exemples (4) et (5) font apparaître une variation de la détermination au singulier et au pluriel; l'article Ø étant ici employé. Enfin, les exemples (6) et (7) montrent la diversité des traductions lorsque la séquence étudiée se trouve imbriquée dans un groupe nominal de taille supérieure : on observe alors une variation de la préposition employée ("classification des accidents ischémiques en sous-types") et du type de détermination ("diagnostic final du sous-type **d'un** AVC ischémique"). Au total, sur huit occurrences de la séquence, seules deux ont donné lieu à des traductions semblables ("sous-type d'AVC ischémique"), avec une variation singulier/pluriel. Cette extrême diversité atteste de la difficulté d'extraire automatiquement une terminologie en langue de spécialité (cf. Bourigault, 1992).[5]

4 Conclusion

Les deux approches évoquées plus haut (utilisation des données statistiques concernant la co-occurrence et l'appartenance à certaines catégories syntaxiques ou sémantiques des adjectifs) semblent à même de résoudre la plupart des ambiguïtés évoquées. L'approche lexicographique consistant en un recensement aussi exhaustif que possible des lexies à mots multiples et des collocations du domaine de spécialité semble être une première étape nécessaire, mais un certain nombre de questions demeurent quant à 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, 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 nécessaire 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.

[5] Les changements de catégorie grammaticale à la traduction constituent une autre difficulté du repérage automatique des équivalents de traduction. Les adjectifs argumentaux (comme dans l'anglais *criminal lawyer*), ne possèdent pas toujours un équivalent dans la langue cible et donnent souvent lieu à une traduction française utilisant un syntagme prépositionnel (*We also examined **procedural** mortality rates*. → Nous avons aussi pris en compte les taux de mortalité **des différentes méthodes**). Inversement, le nom adjectival anglais est parfois traduit par un adjectif (*deep **vein** thrombosis* → thrombose **veineuse** profonde). Dans quelques rares cas, l'usage terminologique impose un rattachement de l'adjectif différent d'une langue à l'autre (*sinus venous thrombosis* → thrombose des sinus veineux).

Langlois L., Plamondon P. (1998), « Le répérage automatique de collocations équivalentes à partir de bitextes» In Fontenelle T. et al. (eds), *Euralex'98: Proceedings of the Eighth Euralex International Congress* Liège, Université de Liège, pp 175-186

Lerat P. (2000), « Quelles propriétés syntaxiques des adjectifs coder dans un dictionnaire bilingue? » in SZENDE, Thomas (éd.), *Approches contrastives en lexicologie bilingue*, Paris, Editions Champion, pp. 147-154

Lethuiller J. (1989), « La synonymie en langue de spécialité». *Meta, 34-3*.

Maniez F. (1995), « Repérage des collocations adjectivales en anglais médical». In: *Revue Informatique et Statistique dans les Sciences Humaines* , n° 1 à 4, 1995, 113-127.

Maniez F. (2001), « Extraction d'une phraséologie bilingue en langue de spécialité : corpus parallèles et corpus comparables» *Meta, 46-1*

Peters C., Picchi E., (1998), « Bilingual Reference Corpora for Translators and Translation Studies», L. Bowker, M. Cronin, D. Kenny and J. Pearson (Eds), *Unity in Diversity? Current Trends in Translation Studies,* Manchester, St. Jerome Publishing.

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

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

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

Thoiron P., Bejoint H. (1998), « Dénominations, définitions et génériques», In: *Revue française de linguistique appliquée*, 1998, III-2 (57-70).

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

Van Doorslaer L. (1995), « Quantitative and Qualitative Aspects of Corpus Selection in Translation Studies», *Target*, 7 (2), pp. 245-260.

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

LOGICIELS UTILISES