Repérage des collocations adjectivales en anglais médical

François MANIEZ

Abstract. Since many noun-based collocations are to be found in the lexicon of medical English, it is worth considering the criteria that govern their inclusion in specialized dictionaries. Considering the case of adjective-noun collocations, the author points out the relative ease with which such collocations can be located and counted through the automatic research of adjectival suffixes, and explains the choice of the *-al* suffix for the present study. The comparison of two corpora of similar sizes indicates that such collocations are more prevalent in the medical English corpus than they are in a corpus drawn from the non-specialized press. The author then goes on to compare various measures of frequency (relative frequency of the adjectival collocate in combination with a given noun, Tact z-score for the immediate co-occurrence of both collocates) and attempts to draw conclusions regarding the figures obtained through those measures and their relationship with the inclusion of the corresponding forms in three reference books in the field of medicine.

Keywords: Adjective, collocation, co-occurrence, frequency, lexicalisation, lexeme, suffixation, z-score.

Mots-clés : Adjectif, collocation, cooccurrence, fréquence, lexicalisation, lexie, suffixation, z-score.

Introduction

La consultation des dictionnaires bilingues anglais-français spécialisés dans le domaine de la langue médicale nous révèle que les lexiques de ce domaine sont essentiellement composés de formes nominales, chose peu surprenante si l'on considère le caractère descriptif des deux branches de la médecine que sont l'anatomie et la pathologie. Une étude plus approfondie de ces formes nominales nous montre également, à l'intérieur de celles-ci, une forte fréquence de groupements constitués autour de nœuds nominaux,

✉ Université de Lyon 2; Centre de recherche en terminologie et en traduction; F–69365 Lyon Cedex 07 (France).
E-mail : maniezd@diogene.lyon2.fr

en particulier d'expressions de structure grammaticale ⟨NOM – ADJECTIF⟩ en français et ⟨ADJECTIF – NOM⟩ en anglais. Signalons au passage, bien que notre propos soit ici essentiellement l'analyse de l'anglais, la fréquence vraisemblablement supérieure de ces formes en français, due au procédé dit de la transposition, qui fait que l'on traduit souvent par un adjectif un prémodificateur nominal (« concentration sérique » pour *serum concentration*, « apport sodé » pour *sodium intake*).

De nombreux auteurs ont évoqué le caractère arbitraire des critères d'inclusion des collocations dans le lexique. La langue de spécialité n'échappe pas à ce problème. Comme l'ont fait remarquer Thoiron et Béjoint (1989), le critère de fréquence d'emploi dans un corpus ne coïncide pas forcément avec celui d'utilité. Dans le cas d'un lexique médical bilingue anglais-français, l'abondance des racines grecques et latines conduisant à un grand nombre de traductions quasi-littérales de toute une partie du lexique, on pourrait envisager, en considérant que l'utilité pour le traducteur est fonction du degré de différence lexicale entre l'expression de la langue de départ et celle de la langue d'arrivée, de n'inclure dans le lexique que les seules formes qui ne donnent pas lieu à une traduction littérale. Ainsi, selon ce critère, on inclurait dans le dictionnaire les formes *acute abscess* (abcès chaud) et *acute iritis* (iridocyclite aiguë), mais on pourrait se dispenser d'y faire figurer les expressions dans lesquelles l'adjectif et le nom se traduisent par leurs équivalents littéraux (par exemple, *acute alcoholism*, *acute bronchitis*, etc.). Cette solution aurait pour avantage de réduire la taille des ouvrages de référence, mais pose la question du critère d'équivalence des formes. Ainsi, l'expression *acute delirium* (délire aigu) mérite peut-être de figurer dans un lexique établi selon ce critère, puisque le mot *delirium* peut avoir pour équivalent une traduction totalement littérale dans une expression comme *delirium tremens*. Dans de nombreux cas de figure, ce critère de non-inclusion de formes quasi-identiques entrera justement en conflit avec le critère de la fréquence d'emploi. Examinons par exemple les entrées d'un lexique anglais-français, celui du dictionnaire Flammarion de la Langue Médicale, qui commencent par l'adjectif *coronary* :

coronary	coronaire, coronarien
coronary artery fistula	communication coronaroventriculaire
coronary failure	insuffisance coronarienne
coronary sinus	sinus coronaire
coronary sinus rhythm	rythme du sinus coronaire

On remarquera que les deux collocations adjectivales utilisant *coronary* que l'on imagine *a priori* comme étant les plus fréquentes (*coronary artery* et *coronary disease*) ne figurent pas dans ce lexique. On imagine que la présence

de la forme de l'adjectif seul est due à sa fréquence d'emploi, mais aussi à sa double traduction. Par comparaison, si l'on veut considérer le cas de deux autres adjectifs qui peuvent être associés à *artery* dans le cadre d'une collocation adjectivale, *pulmonary* figure dans le même lexique, mais non pas *pulmonary artery*, que l'on retrouve toutefois dans l'expression *pulmonary artery wedge pressure* (pression capillaire pulmonaire); quant à *mesenteric*, il est absent du lexique comme adjectif isolé, et de même la collocation *mesenteric artery*. On remarquera au passage que l'absence de *coronary artery* et de *pulmonary artery* sous leur forme isolée s'explique peut-être par leur présence à l'intérieur d'une expression de taille supérieure. Il n'en reste pas moins que l'absence du lexique des formes *coronary artery* et *coronary disease* présuppose que l'utilisateur connaît la distribution préférentielle de l'emploi des adjectifs « coronaire » et « coronarien » respectivement pour l'anatomie et la pathologie, donnant ainsi « artère coronaire » et « maladie coronarienne ».

Rôle de l'informatique dans la conception du lexique

Dans le cas de l'utilisation de corpus sous forme encodée, l'informatique s'avère être un instrument efficace de mesure de la fréquence d'emploi des collocations. Elle est également un support commode pour les ouvrages de référence dont le lexique est en grande partie composé de collocations. En effet, les problèmes de coût et de maniabilité qui imposent certaines restrictions à la taille des ouvrages imprimés ne se posant pas de manière aussi aiguë, on pourra prétendre à une exhaustivité totale en incluant les formes jugées (peut-être à tort) comme étant connues parce que très fréquentes, comme on l'a vu ci-dessus, ainsi que celles qui sont omises en raison même de leur rareté. Enfin, dans le cas des collocations adjectivales, l'accès à la double indexation est particulièrement intéressant. La plupart des dictionnaires utilisent en effet l'ordre alphabétique de la collocation vue comme un continuum (c'est à dire pour l'anglais le classement en fonction du premier terme, l'adjectif). L'informatisation du dictionnaire nous permettra un double accès qui est bien sûr impossible dans une version imprimée : outre les collocations adjectivales commençant par *coronary*, on peut également choisir de consulter toutes celles dont le deuxième terme est *artery*. Idéalement, un dictionnaire informatisé des collocations pourrait également inclure un indice de fréquence d'utilisation à l'intérieur d'un corpus représentatif, calculé en fonction de la taille de celui-ci, du nombre de collocations adjectivales qu'il contient, ou bien encore en fonction de la fréquence d'emploi de chacun des termes isolés. Nous

verrons plus loin comment ces données peuvent également être utilisées pour l'élaboration de critères d'inclusion dans les dictionnaires imprimés.

Critères de repérage des adjectifs

Comparé aux langues, telles le français, dont la grammaire nécessite l'accord des adjectifs, l'anglais se prête particulièrement bien à un repérage s'appuyant sur la recherche automatisée de chaînes de caractères correspondant aux divers suffixes adjectivaux. À titre de recherche préalable, nous avons relevé dans un article de longueur moyenne[1] les occurrences de la forme ⟨ADJECTIF – NOM⟩ dans le but de déterminer les fréquences respectives des divers suffixes adjectivaux.

La méthode de repérage automatisé qui isole systématiquement le mot se terminant par une chaîne de caractères pouvant être un suffixe adjectival et le mot suivant à l'intérieur d'un corpus se heurte cependant à plusieurs problèmes.

1) Tout d'abord, le mot ainsi sélectionné peut ne pas être un adjectif. Ainsi, la recherche du suffixe *-al* sous la forme ⟨*a* + *l* + espace⟩ nous donnera une majorité d'adjectifs, que ce soit en langue générale ou en langue médicale, mais également de nombreux autres mots appartenant à d'autres catégories grammaticales (*several*, *hospital*, *meal*, etc.). On peut toutefois optimiser ce type de recherche en éliminant du champ d'investigation les termes les plus fréquents que l'on aura au préalable repérés dans un corpus de taille moyenne.
2) Le mot qui suit l'adjectif peut ne pas être un nom. Dans ce cas, il s'agit le plus souvent d'un syntagme prépositionnel, comme dans la séquence *polyps that are proximal to the splenic flexure*, ou d'une conjonction de coordination, comme dans la séquence : *adenomatous or hyperplastic polyps*. Une recherche automatisée peut facilement éliminer de telles occurrences du champ d'étude, mais il est plus difficile d'inclure systématiquement les collocations disjointes, comme *adenomatous polyps* dans l'exemple précédent. Dans certains cas, les deux éléments de la collocation peuvent être séparés par des marques de ponctuation, comme dans la séquence : *detection of a distal polyp (either adenomatous or hyperplastic)*. Ce dernier

[1] 1992, "Significance of Distal Polyps Detected With Flexible Sigmoidoscopy in Asymptomatic Patients". *Arch. Intern. Med.*, 152, pp. 1776–1780.

Tableau 1

Collocations adjectivales de l'article *Significance of Distal Polyps Detected With Flexible Sigmoidoscopy in Asymptomatic Patients*

N_C	Collocation
20	*flexible sigmoidoscopy*
20	*colorectal cancer*
13	*flexible sigmoidoscope*
10	*polypoid lesion*
9	*hyperplastic polyp*
8	*proximal neoplasm*
7	*colonic polyp*
7	*asymptomatic patient*
6	*distal polyp*
5	*proximal adenoma*
5	*colonic neoplasm*
4	*current study*
3	*total colonoscopy*
3	*interhaustral fold*
3	*histologic type*
3	*hepatic flexure*
3	*gastrointestinal tract*
3	*distal adenoma*
2	*splenic flexure*
2	*hyperplastic lesion*
2	*frequent cause*
2	*asymptomatic population*
2	*anatomic segment*
2	*adenomatous polyp*

exemple pose également le problème du repérage et de la classification des collocations à adjectifs multiples, très fréquentes en anglais médical.

3) Certaines des formes ainsi isolées, en dépit de leur fréquence d'emploi, n'auront que peu d'intérêt du point de vue de leur inclusion dans un lexique (*actual number*, *additional lesions*). C'est évidemment du côté de la sémantique qu'il faudrait se tourner à la recherche d'éventuels critères d'exclusion. Il s'agira cependant de se livrer à une tâche beaucoup plus exhaustive que celle qui consiste à déterminer si l'adjectif de telles collocations appartient ou non au vocabulaire « spécialisé ». En effet, certaines collocations qui comptent de fort nombreuses occurrences dans notre article (*hyperplastic polyps*, *flexible sigmoidoscopy*) ne se trouvent dans aucun dictionnaire médical. C'est ici que l'étude de la fréquence d'occurrence de chacun des deux termes de la collocation à l'intérieur d'un corpus de grande taille peut s'avérer utile. On posera comme hypothèse que l'intérêt

lexical d'une collocation adjectivale en langue de spécialité est fonction inverse de la diversité des environnements collocatifs dans lesquelles chacun de ses constituants est présent. On verra plus loin si les dictionnaires existants corroborent la prise en compte d'un tel critère.

À l'échelle relativement restreinte de notre article, 213 séquences ⟨ADJECTIF – NOM⟩ ont pu être isolées. Dans les collocations faisant intervenir plusieurs adjectifs, on n'a pris en compte que la séquence débutant par l'adjectif qui précédait directement le nom. Le suffixe *-al* était présent dans 34 collocations adjectivales qui totalisaient 92 occurrences, et le suffixe *-ic* dans 23 collocations totalisant 55 occurrences. Les chiffres obtenus pour les suffixes *-ible* et *-able* (regroupées en un seul groupe symbolisé par la terminaison *-ble*) [3, 34] et *-oid* [1, 10] ne peuvent donner lieu à une interprétation puisque la quasi-totalité des occurrences est due à la présence de deux adjectifs (*flexible* et *polypoid*) sémantiquement liés au sujet de l'article. On trouvera une liste des collocations présentes plus d'une fois dans l'article à la figure 1, dans l'ordre de fréquence d'occurrence. Il est intéressant de noter que certains adjectifs, bien que d'occurrence fréquente dans l'article, se retrouvent rarement en position prénominale. C'est le cas de *distal*, qui est le plus fréquemment employé comme premier adjectif d'une collocation polyadjectivale, et à un moindre degré de *proximal*.

Les collocations adjectivales à suffixe *-al* représentent donc 40 % du total des séquences isolées par notre méthode. Il nous a semblé intéressant de vérifier si le chiffre obtenu était semblable dans un corpus de plus grande échelle, en comparant le nombre d'occurrences de ces collocations à la taille du document. Afin de vérifier si l'homogénéité lexico-sémantique de la langue de spécialité exerce une influence sur la fréquence du processus collocatif, nous avons étudié les collocations adjectivales à suffixe *-al* dans deux corpus distincts (langue générale, langue médicale) de taille voisine (entre 850 000 et 900 000 caractères). Le corpus de langue générale était composé d'articles tirés de la presse contemporaine anglo-saxonne (*Time*, *Newsweek*, *The Economist*), et le corpus de langue médicale d'une cinquantaine d'articles tirés de revues spécialisées de langue anglaise (*The Lancet*, *Journal of the American Medical Association*). Dans ces derniers, seul le texte était encodé, à l'exclusion des tableaux et figures, qui contiennent un grand nombre d'expressions nominales sous formes de liste dont la prise en compte aurait pu fausser la comparaison entre les deux corpus. Le nombre de collocations ainsi isolées s'élevait à 1 339 pour le corpus de langue générale et à 2 061 pour le corpus de langue médicale, qui comportait 122 361 mots.

Tableau 2

Fréquence d'occurrence collocationnelle des 40 adjectifs en *-al* les plus utilisés dans les deux corpus

Corpus médical		Corpus général	
170	*clinical*	153	*political*
98	*medical*	108	*national*
72	*colorectal*	100	*international*
42	*fecal*	47	*global*
36	*total*	46	*real*
34	*prenatal*	45	*social*
34	*general*	40	*general*
31	*potential*	38	*central*
31	*normal*	34	*federal*
30	*myocardial*	31	*industrial*
29	*physical*	29	*local*
28	*social*	28	*royal*
28	*renal*	28	*medical*
28	*national*	23	*special*
27	*cerebral*	21	*annual*
26	*menstrual*	17	*chemical*
25	*statistical*	16	*presidential*
25	*periconceptional*	16	*personal*
24	*maternal*	16	*internal*
23	*local*	16	*congressional*
22	*viral*	15	*total*
22	*oral*	15	*official*
20	*endometrial*	15	*natural*
19	*proximal*	15	*liberal*
19	*additional*	15	*financial*
18	*biochemical*	14	*final*
17	*verbal*	14	*critical*
17	*original*	13	*viral*
17	*intestinal*	13	*individual*
17	*initial*	13	*commercial*
16	*surgical*	13	*clinical*
16	*mental*	12	*traditional*
16	*individual*	12	*moral*
16	*experimental*	11	*regional*
15	*peripheral*	11	*naval*
15	*pathological*	11	*environmental*
15	*fetal*	11	*conventional*
14	*serological*	11	*constitutional*
14	*final*	10	*sexual*
14	*environmental*	10	*mutual*

Le tableau 2 nous indique la fréquence d'utilisation des 40 adjectifs en *-al* les plus utilisés en collocation à l'intérieur des deux corpus. D'un point de vue strictement quantitatif, on remarque un parallélisme frappant entre le nombre d'occurrences des dix adjectifs les plus fréquemment utilisés. Pour le reste du tableau, on observe, à rang égal, un chiffre constamment plus élevé pour les adjectifs du corpus médical, chose naturelle au vu de la disparité entre les chiffres déjà cités, mais qui dénote néanmoins une plus grande fréquence collocationnelle en langue médicale. Douze adjectifs sont présents dans les deux listes (par ordre de fréquence dans le corpus médical : *clinical*, *medical*, *total*, *general*, *national*, *social*, *local*, *viral*, *central*, *individual*, *final*, *environmental*). Ce fait est intéressant dans la mesure où il indique, pour le moins dans le cas de l'adjectif, qu'une part importante d'un corpus de langue de spécialité est constitué d'éléments présents dans la langue dite générale. D'autre part, il nous permettra d'évaluer plus loin l'intérêt des collocations adjectivales en anglais médical en fonction de la fréquence d'emploi de l'adjectif dans le corpus de langue générale (on s'attendra à un moindre intérêt pour les collocations utilisant des adjectifs présents dans les deux listes).

Les tableaux 3 et 4 indiquent les 40 collocations les plus fréquemment utilisées dans chaque corpus, précédées du nombre d'occurrences de la collocation (N_C), du nombre de collocations distinctes contenant cet adjectif dans le corpus, et du pourcentage d'emplois de l'adjectif représenté par l'ensemble des occurrences de la collocation ($T_C = N_C/N_{W_1}$). Là encore, la comparaison entre les deux tableaux montre qu'à rang égal, le nombre de collocations est constamment plus élevé dans le corpus médical (trois fois plus pour les sept premiers items, puis environ deux fois plus sur le reste du tableau). En ce qui concerne la fréquence d'emploi de l'adjectif dans la collocation par rapport à la totalité de ses usages collocationnels (T_C), on remarque qu'elle est également beaucoup plus élevée dans le corpus médical que dans le corpus général, dont quatre collocations seulement (par ordre de fréquence relative d'emploi de l'adjectif : *menstrual cycle*, *oval office*, *clinical trial* et *liberal democrat*) dépassent 0,50. La moyenne de cette valeur pour ces 30 collocations les plus fréquentes est de 0,50 pour le corpus médical et de 0,21 pour le corpus général.

Même si la taille des deux corpus que nous avons comparés est extrêmement faible par rapport aux corpus utilisés en linguistique quantitative, la disparité entre les deux chiffres cités laisse à penser que non seulement l'homogénéité lexico-sémantique de la langue de spécialité médicale favorise la présence de collocations adjectivales (leur fréquence d'occurrence est de 65 % plus élevée qu'en langue générale sur notre échantillon), mais que la fréquence d'utilisation des adjectifs en combinaison avec le même nom est

Tableau 3

Collocations en *-al* du corpus médical, classées par fréquence d'occurrence

N_C	N_{W_1}	T_C	Collocation
32	72	0,44	*colorectal cancer*
30	42	0,71	*fecal blood*
24	34	0,71	*prenatal care*
24	30	0,80	*myocardial infarction*
23	25	0,92	*periconceptional period*
18	170	0,11	*clinical trial*
17	72	0,24	*colorectal neoplasia*
16	17	0,94	*verbal ability*
14	25	0,56	*statistical analysis*
14	15	0,93	*fetal growth*
13	72	0,18	*colorectal evaluation*
12	12	1,00	*portacaval shunt*
12	13	0,92	*nasal spray*
12	26	0,46	*menstrual period*
11	29	0,38	*physical examination*
10	22	0,45	*viral hepatitis*
10	36	0,28	*total mortality*
10	11	0,91	*neural tube*
10	17	0,59	*intestinal polypeptide*
9	16	0,56	*mental health*
9	98	0,09	*medical research*
9	10	0,90	*beneficial effects*
8	12	0,67	*visual loss*
8	9	0,89	*visceral leishmaniasis*
8	28	0,29	*renal lesion*
8	98	0,08	*medical anthropology*
8	14	0,57	*environmental factor*
8	20	0,40	*endometrial resection*
8	170	0,05	*clinical course*
7	19	0,37	*proximal colon*
7	10	0,70	*observational study*
7	8	0,88	*natural history*
7	28	0,25	*national institute*
7	98	0,07	*medical school*
7	17	0,41	*intestinal peptide*
7	34	0,21	*general population*
7	170	0,04	*clinical variable*
7	170	0,04	*clinical practice*
7	170	0,04	*clinical feature*
6	36	0,17	*total number*

N_C est le nombre d'occurrences de la collocation dans le corpus;
N_{W_1}, le nombre total d'occurrences de l'adjectif dans le corpus;
$T_C = \frac{N_C}{N_{W_1}}$.

Tableau 4

Collocations en *-al* du corpus général, classées par fréquence d'occurrence

N_C	N_{W_1}	T_C	Collocation
12	108	0,11	*national institute*
9	100	0,09	*international community*
8	28	0,29	*royal family*
8	15	0,53	*liberal democrat*
7	45	0,16	*social democrat*
7	108	0,06	*national security*
7	34	0,21	*federal reserve*
7	13	0,54	*clinical trial*
7	38	0,18	*central government*
6	108	0,06	*national interest*
6	38	0,16	*central bank*
5	13	0,38	*viral sample*
5	108	0,05	*national product*
5	28	0,18	*medical researcher*
5	100	0,05	*international economist*
5	47	0,11	*global economy*
5	34	0,15	*federal government*
5	17	0,29	*chemical industry*
4	153	0,03	*political scientist*
4	153	0,03	*political party*
4	153	0,03	*political campaign*
4	7	0,57	*oval office*
4	108	0,04	*national*
4	6	0,67	*menstrual cycle*
4	28	0,14	*medical news*
4	100	0,04	*international police*
4	100	0,04	*international conference*
4	31	0,13	*industrial countries*
4	40	0,10	*general agreement*
4	34	0,12	*federal spending*
4	11	0,36	*conventional forces*
4	4	1,00	*bacterial infection*
4	21	0,19	*annual rate*
4	10	0,40	*agricultural policy*
3	23	0,13	*special role*
3	46	0,07	*real problem*
3	153	0,02	*political system*
3	153	0,02	*political malaise*
3	153	0,02	*political leadership*
3	153	0,02	*political establishment*

N_C est le nombre d'occurrences de la collocation dans le corpus;
N_{W_1}, le nombre total d'occurrences de l'adjectif dans le corpus;
$T_C = \frac{N_C}{N_{W_1}}$.

également bien supérieure, le chiffre correspondant étant 2,5 fois plus élevé. Ces résultats concordent avec ceux d'autres études sur la fréquence comparée des suites de deux mots en anglais scientifique et en anglais littéraire (Huizhong, 1986).

Outre la faible taille de notre échantillon, on peut néanmoins relever deux biais méthodologiques inhérents aux limites de notre étude : d'une part, le suffixe *-al* est assurément le plus fréquemment utilisé en langue médicale anglaise (voir Van Hoof pour la diversité de ses traductions en français), mais il reste à prouver que c'est effectivement le cas dans la langue générale. La vérification de cette hypothèse nécessite un corpus de taille conséquente quand on sait qu'aucun adjectif ne figure parmi les 100 mots les plus employés de la langue anglaise (Cf. Sinclair, 1989). Les adjectifs les plus employés (*new, first, last*) ne rentrent d'ailleurs pas dans le cadre de l'étude des suffixes adjectivaux. D'autre part la mesure de la fréquence relative d'emploi d'un adjectif sur l'ensemble des collocations n'est pas forcément la mesure la plus exacte de l'intérêt lexicographique d'une collocation adjectivale. On peut imaginer bien d'autres critères, comme par exemple la fréquence d'emploi par rapport à l'ensemble des occurrences de l'adjectif dans chaque corpus. La fréquence d'emploi du nom est probablement un facteur important, comme nous allons essayer de le montrer.

En nous concentrant à présent sur les collocations tirées de notre corpus médical, nous en avons extrait les 50 les plus fréquentes, et nous avons vérifié si elles figuraient dans trois ouvrages de référence : le dictionnaire anglais–français des Sciences Médicales et Paramédicales de Gladstone, le *Webster's Medical Dictionary* et le lexique anglais–français du Dictionnaire de la Médecine de Flammarion. Le tableau 5, outre les données précédemment citées, fait apparaître la présence des collocations dans ces deux ouvrages, ainsi que la valeur du z-score du nom par rapport à l'adjectif tel qu'elle a été calculée par le logiciel *Tact*[2], les collocations étant classées par z-score décroissant.

Les résultats ainsi obtenus montrent de manière frappante que la fréquence d'emploi d'une collocation n'est certainement pas le premier critère d'inclusion dans un ouvrage de référence. En effet, seules 15 collocations sur 43 ayant obtenu un z-score de *Tact* supérieur à 30 constituent une entrée de l'un des trois dictionnaires, et 6 d'entre elles seulement figurent dans

[2] *Tact* : logiciel développé en 1989 par John Bradley de l'Université de Toronto. John Bradley, University of Toronto, Computing Service, Rm 201, 4 Bancrofe Ave, Toronto, Ontario, Canada M5S 1A1. Fax. 416.978.6620, Email : bradley@gpu.ulcs.utoronto.ca.

Tableau 5

Présence des collocations adjectivales dans trois dictionnaires

N_C	N_{W_1}	N_{W_2}	T_C	z-score	Collocation	DMF	WMD	GLA
10	11	14	0,91	281,859	*neural tube*	×	×	×
12	13	14	0,92	280,043	*nasal spray*			
8	9	10	0,89	279,824	*visceral leishmaniasis*	×	×	×
14	15	17	0,93	275,700	*fetal growth*			
16	17	26	0,94	266,182	*verbal ability*			
12	12	22	1,00	219,994	*portacaval shunt*		×	×
24	30	54	0,80	182,054	*myocardial infarction*	×	×	×
23	25	81	0,92	168,858	*periconceptional period*			
8	12	27	0,67	155,433	*visual loss*			
24	34	89	0,71	146,152	*prenatal care*			×
6	15	14	0,40	140,198	*pathological staging*			
17	72	22	0,24	138,609	*colorectal neoplasia*			
30	42	189	0,71	119,052	*fecal blood*			
5	10	24	0,50	118,974	*variceal bleeding*			
7	28	35	0,25	118,223	*national institutes*			
8	20	27	0,40	114,763	*endometrial resection*			
6	34	40	0,18	110,583	*general practitioner*	×	×	×
6	34	40	0,18	104,888	*general anaesthesia*	×	×	×
14	25	85	0,56	98,111	*statistical analysis*			
13	72	25	0,18	97,480	*colorectal evaluation*			
7	19	14	0,37	96,127	*proximal colon*			
7	8	39	0,88	93,316	*natural history*		×	
12	26	81	0,46	88,015	*menstrual period*			×
10	22	81	0,45	87,423	*viral hepatitis*	×	×	×
32	72	210	0,44	86,889	*colorectal cancer*			
11	29	41	0,38	84,228	*physical examination*		×	×
8	28	37	0,29	81,241	*renal lesion*			
8	14	68	0,57	79,294	*environmental factor*			×
6	34	22	0,18	63,878	*prenatal visit*			
7	10	151	0,70	63,074	*observational study*			
9	16	202	0,56	60,406	*mental health*		×	
10	17	12	0,59	60,265	*intestinal polypeptide*			
9	10	91	0,90	54,013	*beneficial effects*			
8	98	13	0,08	46,573	*medical anthropology*			
7	98	36	0,07	44,708	*medical school*			
7	34	56	0,21	44,220	*general population*			
18	170	117	0,11	44,067	*clinical trial*			
5	36	27	0,14	41,005	*total resection*			
10	36	93	0,28	39,672	*total mortality*			
6	42	18	0,14	38,554	*fecal hemoglobin*			
9	98	69	0,09	36,233	*medical research*			
7	170	36	0,04	33,516	*clinical practice*			×
6	98	36	0,06	33,486	*medical practice*			×

La formule utilisée pour le calcul du z-score dans *Tact* est la suivante:

$$Z = \frac{(\text{fréq. observée du colloquant} - E)}{\text{Écart type}}$$

où

$$\text{Écart type} = \sqrt[2]{\text{longueur mini-texte} \times P(1-P)},$$

$$E = P \times \text{longueur mini-texte},$$

et

$$P = \frac{\text{fréq. colloquant dans l'ensemble du texte}}{\text{longueur du texte}}.$$

les trois ouvrages. Parmi les expressions ne figurant pas dans les dictionnaires comme entrées, seule *colorectal cancer* (32 occurrences) est citée à titre d'exemple dans la définition des adjectifs correspondants dans le Webster's. Sur le faible nombre de formes présentes dans au moins l'un des trois ouvrages, on remarque cependant une spécificité/sensibilité relativement élevée de l'indice T_C, puisqu'il est supérieur à 0,3 dans tous les cas sauf 5 (*clinical trial*, *clinical practice*, *medical practice*, *general practitioner* et *general anaesthesia*) et que seules 5 formes dépassent 0,7 sans figurer dans les ouvrages de référence (*periconceptional period*, *fetal growth*, *nasal spray*, *fecal blood* et *verbal ability*). La comparaison de l'indice T_C avec les z-scores de *Tact* obtenus pour les noms en tant que colloquants immédiatement consécutifs à l'adjectif nous montre une relative proportionnalité entre ces deux chiffres, le z-score de *Tact* étant en règle générale de 200 à 300 fois plus élevé. Dans certains cas, une collocation peut cependant avoir un indice T_C élevé tout en ayant un z-score de *Tact* relativement faible (c'est le cas de *beneficial effects*, et, à un moindre degré, de *natural history* ou *environmental factor*). En effet, la fréquence d'occurrence du colloquant est prise en compte dans le calcul de ce dernier, alors que T_C ne tient compte que de la fréquence d'utilisation d'un adjectif avec un nom donné par rapport au nombre total d'occurrences de cet adjectif en collocation. Le z-score de *Tact* diffère également de T_C dans la mesure où il prend en compte l'ensemble des occurrences de l'adjectif, et donc certains emplois attributifs de celui-ci. Dans le cas de *natural history*, la divergence s'explique sans doute par la conjugaison de ces deux différences : d'une part, le terme *history* a de nombreux autres emplois dans le vocabulaire médical, et d'autre part, l'adjectif natural est lui aussi fréquemment utilisé comme attribut. Dans d'autres cas de divergence frappante (*mental health*, *observational study*), c'est la forte fréquence d'occurrence du nom qui explique à elle seule les variations observées. La répartition des croix dans le tableau 5 indique que le z-score de *Tact* semble un moyen relativement sensible de repérer les collocations adjectivales pouvant être envisagées comme lexies potentielles. Quant à l'indice T_C, il n'est pas fiable lorsqu'il est utilisé isolément, puisque dans le cas d'un adjectif d'usage très fréquent combiné à un nom d'usage rare (comme dans les lexies *general practitioner*, *general anesthesia*), il peut être jusqu'à 500 fois inférieur au z-score de *Tact* correspondant.

Ces derniers exemples montrent à quel point la mesure du seul taux de fréquence d'utilisation de l'adjectif est incomplète si l'on veut évaluer l'utilité de la présence d'une collocation dans un ouvrage de référence. Une mesure comparable en utilisant les emplois collocationnels de *period* nous montre en effet que les collocations adjectivales de notre corpus contenant

ce terme sont nombreuses (*periconceptional period* n'est utilisé que 23 fois sur 81, ce qui représente un taux de fréquence de 0,28 sur l'ensemble des usages de *period*[3]), et que cette base nominale est un terme trop fréquent pour mériter l'inclusion d'une de ses collocations dans un dictionnaire (à l'exception de *menstrual period*, qui relève d'une extension différente du nom). Cela est moins vrai de *growth*, *spray* et *ability*, dont les taux de fréquence sont respectivement de 0,87, 0,86 et 1,00 dans notre corpus, mais on peut inversement se demander si les termes *fetal growth*, *nasal spray* et *verbal ability* ne devraient pas figurer dans les dictionnaires médicaux. En tout état de cause on ne peut envisager la mesure du seul taux de fréquence relatif d'utilisation de l'adjectif comme critère d'inclusion dans les dictionnaires spécialisés, car cela reviendrait à considérer que plus les collocations adjectivales contenant un adjectif donné sont nombreuses, moins celles-ci méritent d'y figurer. Ainsi, dans un corpus de très grande taille, le terme *visceral leishmaniasis* (qui représente 8 des 9 emplois de *visceral* dans notre corpus et figure dans les deux dictionnaires cités) entrerait « en concurrence » avec les autres expressions des dictionnaires qui emploient cet adjectif en association avec *arch*, *cleft*, *muscle*, *pericardium*, *peritoneum*, et bien d'autres, dans la mesure où ces termes devraient « partager » cet indice entre eux. Que dire alors d'adjectifs comme *renal*, dont les emplois totalisent 20 collocations dans le *Webster's Medical Desk Dictionary* ? De toute évidence, les noms ne sont pas tous dotés du même « pouvoir collocatif », et certains, tels *artery*, *bone*, *muscle* pourront être précédés d'un grand nombre d'adjectifs divers pour former des expressions qui doivent nécessairement figurer dans le lexique médical. Notre démarche, si elle peut contribuer au repérage des expressions les plus fréquemment employées, et ainsi enrichir le lexique, passe sans doute par l'élaboration de critères plus fins d'évaluation du rôle de la fréquence d'emploi des collocations adjectivales et des termes qui la composent.

[3] Le calcul du z-score à partir des données indiquées dans le tableau 5 pourra, dans certains cas donner des valeurs différentes de celles qui figurent dans ce même tableau. En effet, pour le calcul du z-score, *Tact* prend en compte l'ensemble des occurrences de l'adjectif, alors que la base de données dont sont extraits les résultats de la colonne N_{W_1} ne prenait en compte que les emplois collocationnels des adjectis dans le corpus (ceci explique que le chiffre des occurrences du nom *period* soit de 54 dans cette base, et de 81 dans l'ensemble du corpus).

Bibliographie

BENSON (Morton) : 1985, "A Combinatory Dictionary of English", *Dictionaries* 7, pp. 189–200.

GARSIDE (R.), LEECH (G.), and SAMPSON (G.) : 1987 (eds.), *The Computational Analysis of English* (London : Longman).

GLADSTONE (W.J.) : 1990, *Dictionnaire Anglais–Français des sciences médicales et paramédicales* (Paris : Maloine).

JOHANSSON, (S.) : 1982, *Computer Corpora in English Language Research* (Bergen : Norwegian Computing Centre for the Humanities).

KERNBAUM (S.) : 1989, *Dictionnaire de Médecine* (Paris : Flammarion).

SINCLAIR (John) : 1991, *Corpus, Concordance, Collocation* (Oxford : Oxford University Press).

THOIRON (Philippe) et BEJOINT (Henri) : 1989, « Pour un index cumulatif et évolutif de cooccurrents », *Meta*, XXXIV, 4.

VAN HOOF (H.) : 1986, *Précis pratique de traduction médicale* (Paris : Maloine).

WEBSTER'S : 1986, *Webster's Medical Desk Dictionary* (Merriam-Webster Inc).

YANG (H.Z.) : 1986, "A New Technique for Identifying Scientific/Technical Terms and Describing Science Texts" *Literary and Linguistic Computing*, Vol. 1, N° 2 : 93–103. (Oxford University Press).