Adjectifs relationnels et langue de spécialité : vérification d'une hypothèse linguistique en corpus comparable médical

Louise Deléger Bruno Cartoni
LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
{louise.deleger,bruno.cartoni}@limsi.fr

Résumé. Cet article présente une étude en corpus comparable médical pour confirmer la préférence d'utilisation des adjectifs relationnels dans les langues de spécialité et examiner plus finement l'alternance entre syntagmes nominaux avec adjectifs relationnels et syntagmes avec complément prépositionnel.

Abstract. This paper presents a study in medical comparable corpora that aims to confirm the preferred use of relational adjectives in specialised languages and to examine in a more fine-grained manner the alternance between phrases with adjective and noun phrases with prepositional complement.

Mots-clés : corpus comparables monolingues, morphologie constructionnelle, langue de spécialité.

Keywords: monolingual comparable corpora, constructional morphology, specialised language.

1 Introduction : adjectifs relationnels et corpus comparables

Il est souvent énoncé, mais rarement validé empiriquement, que les adjectifs relationnels sont l'apanage des langues de spécialité (*cancer mammaire*), et que la langue courante préférera une construction prépositionnelle (*cancer du sein*). Nous proposons ici de confirmer cette hypothèse en comparant les emplois de ces deux types de construction dans des corpus comparables du domaine médical s'adressant à des publics différents (spécialistes d'une part, grand public d'autre part).

Les adjectifs relationnels sont réputés être un type d'adjectifs à part, qui possède des propriétés bien connues (Mélis-Puchulu, 1991)[1]. Syntaxiquement, ils ne peuvent être utilisés comme attribut et sont considérés comme non-gradables. Morphologiquement, ils sont le plus souvent construits sur une base nominale par suffixation, parfois avec la forme supplétive (*président* → *présidentiel*, ou *coeur* → *cardiaque*), et peuvent être utilisés comme base formelle dans des préfixations nominales, le préfixe portant alors sur la base nominale de l'adjectif (*anticancéreux* = *qui est contre le cancer*). Sémantiquement, ils instaurent une relation entre l'entité dénotée par leur nom base et l'entité modifiée, nom recteur du syntagme[2]. D'un point de vue discursif, les syntagmes impliquant un adjectif relationnel (N+Adj_rel) peuvent donc être glosés par un syntagme avec complément prépositionnel (N+prep+N, la préposition pouvant varier). Enfin, les adjectifs relationnels, de part leur prédominance dans certains discours de langue de spécialité (comme le domaine médical (L'Homme, 2004)) permettent également de caractériser le degré de spécialité d'un discours (Maniez, 2009).

[1] Certaines de ces propriétés, bien que souvent nuancées, sont généralement reconnues par la majorité des linguistes.
[2] Dans *élection présidentielle*, l'adjectif permet de désigner la relation entre *élection* et *président*.

Les adjectifs relationnels sont au coeur de nombreuses applications de TAL, comme l'acquisition de terminologie, particulièrement lorsqu'il s'agit d'acquérir différentes variantes d'un même terme (Daille, 1999; Morin & Daille, 2010). D'autres travaux étudient les adjectifs relationnels en discours, mais sans unité thématique (comme (Maniez, 1995)), ou sans opposer langue de spécialité et langue générale (L'Homme, 2004). L'utilisation des corpus comparables a déjà fait ses preuves dans de nombreux champs d'étude. Traditionnellement, ils sont multilingues et utilisés pour l'extraction de lexiques bilingues. Mais ils peuvent également être monolingues et sont alors généralement employés pour l'acquisition de paraphrases : par exemple, Elhadad & Sutaria (2007) et Deléger & Zweigenbaum (2009) recherchent des paraphrases entre expressions spécialisées et grand public. La distinction spécialisé/grand public est aussi étudiée par Goeuriot *et al.* (2008) qui définissent des critères de distinction des deux types de discours utilisés par des algorithmes d'apprentissage ; ainsi que par Morin *et al.* (2007) dans des corpus multilingues pour y extraire des lexiques bilingues. Ici, nous exploitons des corpus comparables monolingues de discours spécialisé et grand public dans le domaine médical pour valider l'hypothèse linguistique préalablement posée. L'intérêt de ces corpus est qu'ils portent sur une même thématique tout en présentant une variation de type de discours.

2 Expérience

2.1 Matériel de départ : listes de paires Adj_rel-Nom et corpus médicaux

Pour rechercher des syntagmes N+Adj_rel ou N+prep+N, nous avons élaboré des listes de paires Adj_rel-Nom, à partir des adjectifs relationnels présents dans l'UMLF (lexique spécialisé du français médical (Zweigenbaum *et al.*, 2003)), analysés avec l'analyseur Dérif (Namer, 2009) pour récupérer leur base nominale. Une vérification manuelle a été effectuée. Deux listes ont été créées : une liste de 2091 paires Adj_rel - Nom (*cardiaque* | *cœur*) et une liste de 864 paires Adj_rel préfixé - Nom (*anticancéreux* | *cancer*).

Nous avons utilisé trois corpus comparables médicaux (Deléger & Zweigenbaum, 2009), avec des textes à destination de spécialistes et des textes à destination du grand public, portant sur les thématiques du diabète, du tabac et du cancer. Le corpus cancer provient d'une même source (le site Standard Options Recommandations[3]) qui fournit des documents comparables (recommandations pour les médecins et guides pour le grand public). Les autres corpus ont été construits à travers une recherche guidée sur le web : nous avons interrogé des moteurs spécialisés (CISMeF, HON[4]) à l'aide de mots-clés, en spécifiant le public visé (ces deux portails le permettant), et effectué une recherche sur des sites institutionnels et des sites de santé dédiés au grand public. Les tailles des corpus sont reportées dans le tableau 1.

	Diabète		**Tabac**		**Cancer**	
	S	G	S	G	S	G
Documents	135	600	62	620	22	16
Occurrences	580 712	461 066	595 733	603 257	641 584	228 742

TAB. 1 – Taille des corpus (Nombre de documents et mots ; S=spécialisé, G=grand public)

[3] `http://www.sor-cancer.fr` ; NB : les guides grand public incluent des glossaires qui n'ont pas été conservés ici

[4] `http://www.cismef.org` et `http://www.hon.ch`

2.2 Identification de paires N+Adj_rel / N+prep+N avec même nom recteur

Le but est donc de rechercher en corpus des correspondances entre syntagmes impliquant un adjectif relationnel dans la partie spécialisée et syntagmes impliquant le nom dont cet adjectif est dérivé dans la partie grand public, tous les deux avec le même nom recteur. Le patron lexico-syntaxique suivant, construit autour des paires Adj_rel-Nom décrites précédemment, a été défini[5]:
– N_2+$A_1 \leftrightarrow N_2$+Prep+(Det)+N_1 (ex : *atteinte ganglionnaire* $\leftrightarrow$ *atteinte des ganglions*)
Dans une 1ère phase, la partie gauche du patron est appliquée du côté spécialisé[6], et la partie droite correspondante du côté grand public. Ceci nous fournit une liste de paires de syntagmes équivalents, sur laquelle repose nos mesures. Dans une 2ème phase, chaque syntagme est recherché dans les deux parties des corpus pour pouvoir comparer leurs proportions. Ainsi, si la paire « myocarde-myocardique » permet de trouver « infarctus du myocarde » dans la partie grand public et son équivalent « infarctus myocardique » dans la partie spécialisée (phase I), la phase II recherche « infarctus myocardique » du côté grand public et « infarctus du myocarde » du côté spécialisé, et mesure leur fréquence.
Cette méthode automatique de repérage de paires N+Adj_rel / N+prep+N relève également quelques paires qui ne sont pas équivalentes (sens non strictement identiques (*nombre annuel* / *nombre d'ans*) ou syntagme N+prep+N issu d'un syntagme plus étendu[7]) et la comparaison n'est donc pas pertinente, nous les avons exclus manuellement (respectivement 7,4%, 18,7% et 7,5% pour les corpus diabète, tabac et cancer).

2.3 Analyses quantitatives et qualitatives

Nous examinons d'abord les résultats à travers un *indice de préférence* qui représente la proportion de syntagmes du type N+Adj_rel par rapport à l'ensemble des cas possibles (N+Adj_rel et N+prep+N). Pour chaque paire de syntagmes identifiée, l'indice se calcule ainsi : $I = \frac{F_{NA}}{F_{NA}+F_{NN}}$ avec F_{NA} la fréquence du syntagme N+Adj_rel et F_{NN} la fréquence du syntagme N+prep+N. Nous calculons l'ensemble des indices pour chaque paire dans chaque partie des corpus, et leur moyenne. Suivant notre hypothèse de départ (le discours spécialisé emploie plus volontiers les constructions avec adjectif que le grand public), l'indice devrait être fort (proche de 1) dans la partie spécialisée et faible (proche de 0) dans la partie grand public. Nous établissons ensuite une comparaison entre les indices de préférence des deux parties des corpus, en calculant pour chaque paire de syntagmes l'écart entre l'indice de la partie spécialisée (I_s) et celui de la partie grand public (I_g): $E = I_s - I_g$. Plus l'écart est grand (et positif), plus la différence entre les deux types de discours est marquée, ce qui confirme l'hypothèse.
D'un point de vue qualitatif, nous examinons d'une part les cas où l'écart entre indices de préférence va à l'encontre de l'hypothèse (écart nul ou négatif), et d'autre part les prépositions utilisées dans les syntagmes nominaux (particulièrement lorsqu'il s'agit de gloser un adjectif relationnel préfixé). L'analyse de ces prépositions permettrait l'amélioration des patrons pour la recherche de paraphrases ou de gloses de mots morphologiquement construits.

[5]Les indices indiquent les mots en correspondance (Adj_rel-Nom=indice 1) ; aucune restriction n'est mise sur la préposition.
[6]Chaque corpus est préalablement étiqueté morphosyntaxiquement et lemmatisé.
[7]Dans la paire *carie dentaire* / *carie de dents*, le deuxième syntagme provient en fait de *carie de dents de lait*.

3 Résultats

Les moyennes des indices de préférence pour l'ensemble des corpus sont reportées dans le tableau 2. On constate que, dans tous les corpus, l'indice moyen est fort pour la partie spécialisée et faible pour la partie grand public[8], ce qui vient confirmer l'hypothèse de départ. Des exemples de paires de syntagmes avec leurs indices sont donnés dans le tableau 3. La ligne (4) montre un cas où, contrairement à ce qu'on attendrait, l'indice de préférence est faible dans la partie spécialisé. La ligne (5) constitue également un cas inverse, avec un fort indice dans la partie grand public.

	Diabète (150 paires)	**Tabac (74 paires)**	**Cancer (86 paires)**
Spécialisé	0,78	0,69	0,82
Grand public	0,32	0,25	0,14

TAB. 2 – Indice de préférence moyen dans chaque partie (spécialisé, grand public) des corpus

		Indice de préférence		**Ecart**
		Spécialisé	Grand public	
(1)	risque infectieux / risque d'infection	1	0	1
(2)	vaccination antigrippale / vaccination contre la grippe	0,62	0	0,62
(3)	échographie abdominale / échographie de l'abdomen	1	0,47	0,53
(4)	traitement antidiabétique / traitement du diabète	0,16	0,04	0,12
(5)	rythme cardiaque / rythme du coeur	1	0,94	0,06
(6)	campagne publicitaire / campagne de publicité	0,33	0,5	-0,17
(7)	greffe rénale / greffe du rein	0,71	0,8	-0,09

TAB. 3 – Exemples d'indices calculés pour quelques paires

Le calcul des écarts entre les indices de chaque paire nous permet d'observer de manière plus fine la préférence. La figure 1 présente le nombre de paires de syntagmes par tranche d'écart pour l'ensemble des corpus, avec en abscisse la valeur de l'écart entre les indices de préférences (regroupés par tranches de 0,1) et en ordonnée le nombre de paires pour chacune de ces tranches de valeur. Des exemples d'écarts sont reportés dans la dernière colonne du tableau 3. Dans l'ensemble, une majorité de paires présente un écart positif, ce qui va dans le sens de l'hypothèse et montre une différence entre les deux types de discours (lignes (1) à (3) du tableau 3). Il est intéressant de noter le fort nombre de paires dont l'écart est de 1 (ou proche), donc où la préférence est très nettement marquée. Il existe également un nombre assez important de cas où l'écart est (quasi-)nul, donc où l'on n'observe aucune différence entre les deux types de discours. Parmi les écarts nuls, on distingue deux cas : soit les deux types de discours préfèrent une construction N+Adj_rel (ex. (4) du tableau 3) ; soit ils préfèrent une construction N+prep+N (ex. (5)). Dans un cas, les exemples semblent témoigner de syntagmes fréquents couramment utilisés dans la langue générale. Le deuxième cas montre que l'usage d'un syntagme N+Adj_rel n'est pas automatique dans la langue spécialisée ; à première vue, il est difficile d'identifier une raison évidente au choix du syntagme N+prep+N. En outre, pour quelques rares cas, l'écart est négatif, montrant ainsi une préférence inverse (plus forte dans le

[8]La différence des moyennes des indices, testée avec un T-test unilatéral apparié, est statistiquement significative ($p<0.01$).

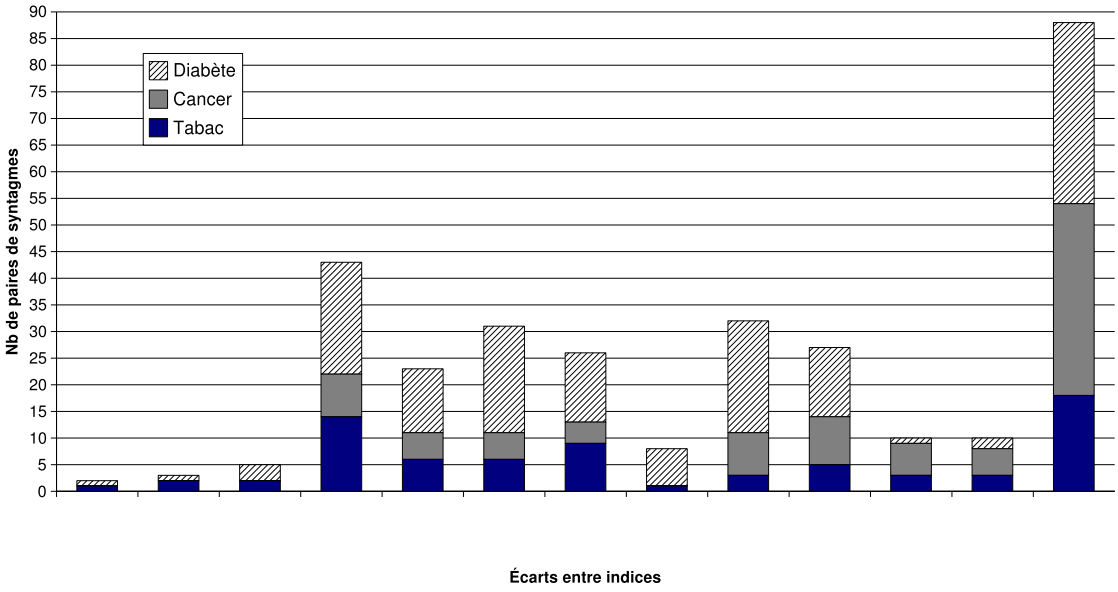

FIG. 1 – Nombre de paires de syntagmes par tranches d'écarts entre indices de préférence

discours grand public, comme dans les lignes (6) à (7) du tableau 3). Il s'agit essentiellement d'adjectifs ne faisant pas partie du vocabulaire de spécialité, comme *publicitaire* (ex. (6)), ce qui semble indiquer que les adjectifs relationnels sont l'apanage de la langue de spécialité seulement quand ils relèvent de la terminologie de la langue en question. Enfin, il existe une poignée d'exemples d'écart négatif impliquant des adjectifs relationnels appartenant bien au vocabulaire spécialisé (ex. (7)), difficilement explicables.

Concernant les prépositions, sur les 296 syntagmes N+prep+N repérés sur la base des paires Adj_rel-Nom, on constate évidemment la grande fréquence de la préposition *de* (78%). Ensuite on trouve les prépositions *par* (5,1%), *dans* (4,1%), et *à* (3,7%). Comme on peut s'y attendre, l'utilisation d'une préposition locative (ex. *dans*) renvoie à un équivalent N+Adj_rel où l'adjectif est préfixé (*injection intraveineuse* ↔ *injection dans la veine*), mais pas uniquement (*métastase cérébral* ↔ *métastase dans le cerveau*). Certains syntagmes composés d'adjectifs relationnels préfixés n'ont pas uniquement des équivalents impliquant une préposition sémantiquement proche de l'instruction de la règle de préfixation. Ainsi, nous avons découvert des cas où l'emploi d'une préposition semble beaucoup plus flexible, comme dans *traitement anticancéreux*, pour lequel on trouve *traitement du cancer* ou *traitement contre le cancer*. Pour d'autres cas, nous n'avons trouvé aucun équivalent avec une préposition « liée » au préfixe de l'adjectif (*tumeur intraprostatique* mais *tumeur de la prostate*). La disparité d'utilisation des préfixes ou des prépositions semblent souligner le besoin de précision de la langue de spécialité, là où la langue générale paraît plus souple.

4 Discussion et conclusion

Nous avons montré les possibilités qu'offrent les corpus comparables monolingues pour la validation d'hypothèses linguistiques et décrit différentes mesures permettant d'obtenir une analyse de l'alternance entre deux réalisations linguistiques. Globalement, malgré quelques exemples récalcitrants, les données extraites par notre étude semblent confirmer l'hypothèse d'une préférence pour les constructions avec adjectifs relationnels dans la langue de spécialité : pour une même idée, un syntagme nominal avec adjectif

relationnel est préféré dans un corpus spécialisé, et son « équivalent » avec préposition et nom est davantage présent dans un corpus grand public. En outre, cette tendance se retrouve dans les trois corpus étudiés. Cependant nous avons également montré que cette préférence n'était pas toujours aussi marquée : il existe des cas où l'on n'observe pas de différence entre les deux types de discours que ce soit parce que les deux préfèrent une construction avec adjectif ou parce que le syntagme avec préposition et nom est utilisé majoritairement dans les deux. Les rares cas où la préférence est inversée montrent en général des syntagmes nominaux beaucoup moins spécialisés (*campagne publicitaire*). Ces résultats sont à nuancer au vu de la taille relativement petite des corpus étudiés et de la restriction à trois thèmes particuliers par rapport à l'ensemble du domaine médical. Nous envisageons donc de constituer des corpus plus larges, pour donner plus de généralité à la vérification de l'hypothèse. Une autre perspective intéressante serait d'effectuer la même étude dans des corpus comparables d'autres domaines de spécialité.
En outre, cette expérience n'est qu'une première étape pour la compréhension du rôle et de l'usage de l'adjectif relationnel. Son appartenance à la langue de spécialité n'est qu'un des aspects qui le caractérisent, et les méthodes présentées ici permettrait sans doute d'explorer d'autres de ses caractéristiques, comme l'alternance dans les constructions préfixées : *accord interuniversité* vs. *accord interuniversitaire*.

Références

DAILLE B. (1999). Identification des adjectifs relationnels en corpus. In *TALN 1999*, p. 105–114.

DELÉGER L. & ZWEIGENBAUM P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-Parallel Corpora*, p. 2–10.

ELHADAD N. & SUTARIA K. (2007). Mining a lexicon of technical terms and lay equivalents. In *ACL BioNLP Workshop*, p. 49–56, Prague, Czech Republic.

GOEURIOT L., GRABAR N. & DAILLE B. (2008). Characterization of scientific and popular science discourse in French, Japanese and Russian. In *Proceedings of LREC*, p. 2933–2937.

L'HOMME M. (2004). Adjectifs dérivés sémantiques (ADS) dans la structuration des terminologies. In *Terminologie, ontologie et représentation des connaissances*, Université Jean-Moulin Lyon-3.

MANIEZ F. (1995). Repérage des collocations adjectivales en anglais médical. *Revue Informatique et Statistique dans les Sciences Humaines*, **1 à 4**, 113–127.

MANIEZ F. (2009). L'adjectif relationnel en langue de spécialité : utilisations terminologiques et phraséologiques. *Revue Française de Linguistique Appliquée*, **XIV-2**, 117–130.

MORIN E. & DAILLE B. (2010). Compositionality and lexical alignment of multi-word terms. *Language Resources and Evaluation (LRE)*, **44**(1-2), 79–95.

MORIN E., DAILLE B., TAKEUCHI K. & KAGEURA K. (2007). Bilingual terminology mining - using brain, not brawn comparable corpora. In *Proceedings of ACL*, Prague, Czech Republic.

MÉLIS-PUCHULU A. (1991). Les adjectifs dénominaux: des adjectifs de "relation". *Lexique*, **10**, 33–60.

NAMER F. (2009). *Morphologie, lexique et traitement automatique des langues : l'analyseur DériF*. Paris: Lavoisier.

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., JARROUSSE E., GRABAR N., RUCH P., LE DUFF F., THIRION B. & DARMONI S. (2003). UMLF : construction d'un lexique médical francophone unifié. In *Proceedings of the 10th French Language Medical Informatics Conference*, Tunis.