Marqueurs de la relation cause-effet : stabilité et variation dans des corpus de nature différente

Elizabeth Marshman[1], Marie-Claude L'Homme[2] et Victoria Surtees[2]

[1] Observatoire de linguistique Sens-Texte / Université d'Ottawa,
elizabeth.marshman@uottawa.ca
[2] Observatoire de linguistique Sens-Texte, Université de Montréal,
mc.lhomme@umontreal.ca, vsurtees@gmail.com

Résumé : Les marqueurs lexicaux figurent parmi les indicateurs les plus intéressants pour identifier et classer finement des relations terminologiques. Cependant, l'exploitation des marqueurs pour repérer automatiquement ou semi-automatiquement les relations dans des corpus, notamment des corpus associés à des domaines ou à des genres textuels différents, peut soulever des problèmes. Le présent article fait le bilan d'une analyse de 37 marqueurs verbaux de la relation cause-effet dans trois corpus représentant trois domaines (médecine, informatique et psychologie) et deux genres textuels (textes didactiques/vulgarisés et textes spécialisés). Il décrit les variations observées dans l'identification de sens spécifiques indiquant différents sous-types de la relation, et commente la difficulté de prévoir les marqueurs les plus efficaces pour extraire ces relations de différents corpus.

Mots-clés : Relations terminologiques, Marqueurs de relations, Cause-effet, Domaine, Genre textuel.

1 Introduction

Les marqueurs lexicaux sont des unités lexicales ou des combinaisons d'unités lexicales qui expriment de façon récurrente une relation terminologique ; elles apparaissent le plus souvent comme la composante centrale d'un patron de connaissances (*knowledge pattern* dans la terminologie de Meyer (2001)) et relient deux unités partageant un lien sémantique. Par exemple, le patron de connaissances X *stimule* Y, comme dans *l'hormone stimule la croissance*, indique la présence d'une relation causale. L'association récurrente d'unités lexicales et l'expression formelle d'une relation terminologique (réalisée par le marqueur) permet de repérer des relations terminologiques (semi-)automatiquement dans des textes. Bien qu'il ne s'agisse pas des seuls indicateurs de relations terminologiques, les marqueurs lexicaux sont particulièrement intéressants à exploiter dans la mesure où ils permettent d'identifier clairement et de classer finement les relations terminologiques. Cependant, la récurrence de ces marqueurs ne veut pas dire qu'ils constituent des indices d'une portée universelle ; leur utilisation dans des corpus de nature différente (c'est-à-dire dans des domaines ou genres textuels distincts) soulève des difficultés.

Des travaux antérieurs ont montré que les marqueurs linguistiques se caractérisent par une certaine instabilité quant à leur contenu sémantique et ne sont donc pas toujours aisément portables d´un corpus à l´autre. Le travail dont nous rendons compte dans les pages qui suivent vise à étudier l´ampleur de ce phénomène.

Dans ce travail, nous analysons des marqueurs de la relation de cause-effet. Cette dernière est une relation fondamentale dans l'expérience humaine et centrale dans de nombreux domaines de spécialité, notamment la médecine (qui étudie les causes des maladies et l'effet des traitements), la psychologie (qui cherche la source de phénomènes psychiques et les facteurs qui influencent des processus mentaux) et aussi l'informatique (qui décrit des manipulations de données et les outils utilisés pour atteindre les objectifs visés). Cette relation, bien que considéré dans certains travaux terminologiques comme étant secondaire par rapport aux liens hyperonymiques et méronymiques, a suscité ces dernières années de plus en plus d'intérêt (e.g. Nuopponen, 1994, 2005 ; Garcia 1997 ; Barrière 2001 ; Feliu 2004 ; Marshman 2007). Notre travail veut contribuer à la mise au point d'outils de repérage de relation cause-effet dans les textes spécialisés. Le terminologue, qui doit souvent représenter cette relation dans les ressources (ontologies, bases de données terminologiques) qu'il conçoit, doit d'abord repérer les éléments reliés dans les textes. Une meilleure compréhension des facteurs linguistiques intervenant dans l'expression de la relation contribuera au développement d'outils mieux adaptés.

Plus spécifiquement, notre analyse a porté sur 37 marqueurs verbaux de la relation de cause-effet dans trois corpus associés à trois domaines (médecine, informatique et psychologie) et à deux genres textuels (textes didactiques/vulgarisés et spécialisés). À partir d'une typologie des sens causaux et d'une liste de marqueurs qui les expriment de manière récurrente, nous étudions les variations qui peuvent s´observer dans les différents corpus, notamment dans les fréquences. Ces observations serviront à évaluer : 1. la productivité des différents marqueurs ; 2. l'intérêt de les utiliser dans les applications qui cherchent à repérer des relations terminologiques automatiquement ou semi-automatiquement ; 3. leur stabilité d'un corpus à l'autre.

La section 2 présente un bref survol de quelques études dans le domaine. La méthodologie est expliquée dans la section 3 et les résultats sont résumés dans la section 4. Enfin, la section 5 présente quelques remarques en guise de conclusion, ainsi que des suggestions pour de futurs travaux.

2 Concepts clés et études antérieures

Depuis les travaux de Hearst (1992), de nombreux chercheurs ont étudié le potentiel des marqueurs linguistiques pour extraire – de manière (semi-)automatique – des relations sémantiques de corpus spécialisés (par exemple, Ahmad & Fulford, 1992 ; Meyer *et al*., 1999 ; Condamines & Rebeyrolle, 2001 ; Meyer, 2001 ; Marshman *et al*., 2002 ; Malaisé *et al*., 2005), et ce, dans plusieurs langues.

Malgré l'attrait indéniable que présentent les marqueurs pour repérer des relations terminologiques (et pour les étiqueter de manière précise), leur exploitation soulève des difficultés importantes. Parmi celles-ci, nous pouvons citer : 1) des variations

dans les sous-types de relations terminologiques qui peuvent être véhiculées par ces marqueurs (notées par exemple dans le cas de relations de cause-effet par Barrière (2002) et Marshman (2007)), comme le marqueur INHIBER, qui peut indiquer la prévention ou la réduction; 2) la polysémie de marqueurs qui compromet la précision avec laquelle des occurrences peuvent être identifiées (décrite entre autres dans Condamines, 2000 ; Marshman *et al.*, 2002 et Marshman, 2007), comme les marqueurs CONDUIRE et AUGMENTER, qui peuvent indiquer des sens causaux ou non causaux ; et 3) la présence de marqueurs distincts pour exprimer une même relation, nécessitant la prise en compte d'une gamme relativement large de marqueurs afin de permettre l'identification d'une proportion acceptable des occurrences. (Des exemples sont présentés dans la section 4.)

Évidemment, l'investissement requis pour dresser des listes de marqueurs devient plus rentable lorsque ces marqueurs permettent d'extraire de l'information utile dans divers corpus. Toutefois, on a observé (par exemple, dans Séguéla, 1999 ; Condamines, 2000, 2002, 2008 et Jacques & Aussenac-Gilles, 2006), que la productivité des marqueurs peut varier de façon significative dans des corpus associés à divers domaines ou composés de textes différents (*genres textuels*, cf. Biber, 1988, ou liés à différentes situations communicatives, cf. Pearson, 1998).

Des études antérieures au travail présenté dans ces pages (Marshman *et al.*, 2008, 2008a) ont analysé des marqueurs de relations causales dans des corpus associés à trois domaines et à deux genres textuels. Nous avons évalué : 1) la fréquence des occurrences des marqueurs dans les corpus ; 2) la proportion des occurrences qui indiquaient de véritables relations causales ; 3) la polysémie des marqueurs à un niveau plus fin, notant plusieurs sens véhiculés par les différents marqueurs (ainsi que des variations quant au nombre d'occurrences associées à ces sens spécifiques). Dans les tous les cas, nous avons observé des différences importantes dans le cas de certains des marqueurs, ainsi que des variations entre domaines et genres textuels. Il était malheureusement difficile de confirmer et de lier définitivement ces variations à l'un des facteurs en raison de la variabilité individuelle des marqueurs.

Cette première analyse a permis d'évaluer l'efficacité des marqueurs pour identifier et classer les relations. Toutefois, nous avions envisagé le problème du point de vue du marqueur linguistique et sous l'angle des sens linguistiques associés à la relation causale. Il est alors souhaitable de raffiner ce portrait au moyen d'une étude ayant comme point de départ les sens exprimés par les différents marqueurs. Cela permettra d'étudier les préférences dans le choix de marqueurs en fonction du domaine ou du genre textuel. La comparaison des différents corpus (en fonction, par exemple, de la ressemblance plus étroite entre la médecine et la psychologie) pourra aussi permettre d'étudier des facteurs qui expliquent les variations.

3 Méthodologie

Les données de l'étude sont tirées de trois corpus portant sur des domaines différents : un corpus de médecine (600 000 occurrences) composé d'articles spécialisés, un corpus d'informatique (1 000 000 occ.), d'articles didactiques, et un

corpus de psychologie, de deux sous-corpus, le premier d'articles spécialisés (165 000 occ.) et le deuxième d'articles didactiques et vulgarisés (420 000 occ.).

À partir de ces corpus nous avons extrait des occurrences de 37 marqueurs verbaux de relations de cause-effet identifiés dans une étude antérieure (Marshman, 2007). Cette étude avait comme objectif, entre autres, de découvrir les marqueurs de relations de cause-effet dans un corpus médical[1]. Des marqueurs verbaux, qui sont parmi les plus prototypiques et fréquents pour cette relation (cf. Garcia 1997 ; Barrière 2001 ; Marshman 2007) ont été retenus pour cette étude plus approfondie. À l'aide du concordancier WordSmith Tools, nous avons sélectionné un échantillon aléatoire d'environ 50 occurrences de chaque marqueur dans chaque corpus, et avons éliminé manuellement des occurrences qui correspondaient à du bruit[2]. Ce tri préliminaire a produit un nombre variable d'occurrences (entre une seule occurrence et une soixantaine, selon le corpus et le marqueur) à analyser pour chaque marqueur. (Pour le nombre d'occurrences analysées pour chaque marqueur dans chaque corpus, voir Marshman *et al*., 2008.)

Les occurrences retenues ont été analysées par trois terminologues (chacune prenant en charge les données d'un des corpus) et ont d'abord été classées en deux catégories principales : celles qui exprimaient des sens causaux (c'est-à-dire, dont une paraphrase du sens contenait un élément tel que *cause* ou *à cause de*) et celles qui exprimaient des sens non causaux. Ensuite, les deux catégories ont été subdivisées en utilisant un système de paraphrases : celles-ci faisaient appel à des variables (X, Y, Z) pour représenter les arguments, et à une décomposition du sens au moyen d'un vocabulaire simplifié (entre autres, *causer* et *à cause de* pour indiquer la relation de cause-effet, *être*, *avoir lieu*, *plus*, *moins*, et *différent* pour d'autres éléments du sens).

Dans l'analyse, nous avons identifié plusieurs marqueurs qui partagent une même paraphrase causale. Les paraphrases qui apparaissent plusieurs fois dans au moins deux des corpus et les marqueurs qui les véhiculent sont présentées dans le Tableau 1.

Tableau 1. Paraphrases des sens et marqueurs qui les expriment

Paraphrase	**Marqueurs**
X cause que Y ait lieu/soit	ABOUTIR, ASSURER, CAUSER, CONDUIRE, DECLENCHER, ENTRAINER, EXPLIQUER, EXPRIMER, INDUIRE, PRODUIRE, PROVOQUER, REALISER, RESULTER, STIMULER
X cause que Y soit différent (grâce à Z)	AGIR, ALTERER, INFLUENCER, MODIFIER, MODULER

[1] Pour une discussion détaillée de la méthodologie de l'identification initiale des marqueurs et de plusieurs typologies disponibles de la relation cause-effet, ainsi que la typologie retenue pour l'identification initiale des marqueurs et les motifs de ce choix, voir Marshman (2007).

[2] En raison de variations dans le nombre d'occurrences des marqueurs et dans la taille des corpus et aussi de l'usage de la fonction de sélection aléatoire dans le corpus offert par WordSmith Tools, le nombre d'occurrences initialement extraites n'est pas toujours égal à 50. Le bruit éliminé correspondait surtout à des occurrences qui n'étaient pas des formes verbales (par exemple, des formes adjectivales et nominales) ou qui contenaient la construction causale *faire* + verbe.

X cause que Y ait moins lieu/soit moins	CONTROLER, DIMINUER, INHIBER, LIMITER, REDUIRE
X cause que Y puisse avoir (plus) lieu (plus facilement)	FACILITER, FAVORISER, PERMETTRE
X cause que Y soit plus	ACCROITRE, AUGMENTER, STIMULER
X (est l'un des agents qui) cause(nt) que Y puisse avoir lieu/être	AIDER, INTERVENIR
X cause que Y fonctionne d'une certaine manière	CONTROLER, ENTRAINER
X cause que Y n'ait pas lieu	BLOQUER, EMPECHER
X cause que Y ne fonctionne pas/plus	BLOQUER, INHIBER
X cause que Z fasse Y	CONDUIRE, STIMULER

Enfin, nous avons comparé le nombre d'occurrences des différents marqueurs exprimant ces sens dans les trois domaines et deux genres textuels (regroupant ensemble les articles didactiques et vulgarisés pour les fins de la comparaison).

4 Résultats

Cette section présente les proportions des occurrences des sens causaux analysés qui étaient associés aux marqueurs. Les tableaux présentent : a) les marqueurs exprimant un sens spécifique ; et b) le nombre absolu d'occurrences ainsi que le pourcentage des occurrences du sens associé à chaque marqueur dans chaque corpus.

Un sous-type « de base » de relations de cause-effet a été identifié, correspondant au sens « X cause que Y ait lieu/soit ». Ce sens est exprimé par un nombre élevé de marqueurs (Tableau 2, Figure 1).

Tableau 2. Marqueurs exprimant le sens « X cause que Y ait lieu/soit »

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**[3]	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
ABOUTIR	39	6	32	63	14	9	2	8	9	3
ASSURER	33	51	31	59	56	7	14	8	8	11
CAUSER	33	15	45	54	39	7	4	11	8	8
CONDUIRE	41	35	32	65	43	9	9	8	9	8
DECLENCHER	26	46	37	45	64	6	12	9	6	13
ENTRAINER	37	52	43	59	73	8	14	10	8	14
EXPLIQUER	29	20	18	44	23	7	5	4	6	5
EXPRIMER	23	0	0	23	0	5	0	0	3	0
INDUIRE	39	4	22	56	9	9	1	5	8	2

[3] Dans les en-têtes des tableaux, *M*, indique le corpus de médecine, *I* le corpus d'informatique, *P* le corpus de psychologie, *S* les textes spécialisés, et *DV* les textes didactiques ou vulgarisés.

PRODUIRE	27	47	48	67	55	6	13	12	9	11
PROVOQUER	36	48	45	52	77	8	13	11	7	15
REALISER	44	49	48	85	56	10	13	12	12	11
RESULTER	2	0	2	4	0	0	0	0	1	0
STIMULER	36	0	8	42	2	8	0	2	6	0
Total[4]	445	373	411	718	511	99	100	100	100	101

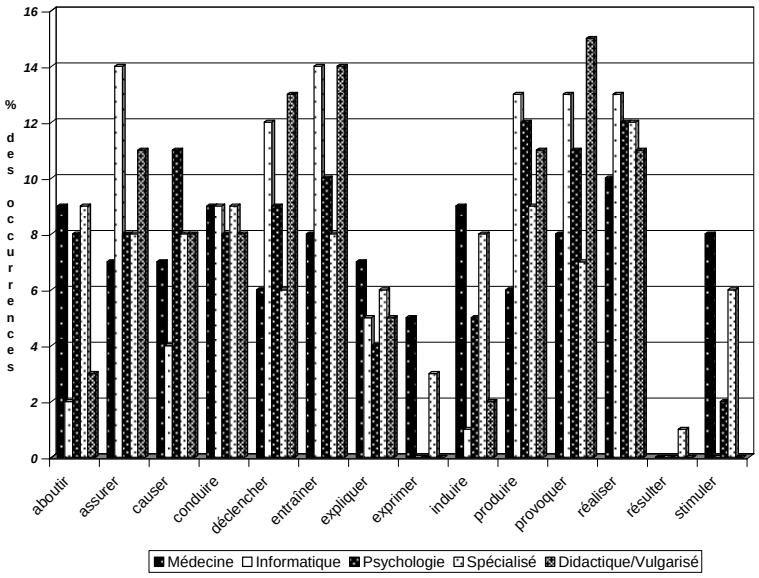

Fig. 1 — Marqueurs exprimant le sens « X cause que Y ait lieu/soit »

L'analyse révèle quelques variations généralisées (ex. CONDUIRE, PROVOQUER et ENTRAINER exprimant plus souvent le sens en question qu'EXPLIQUER). Mais il y a des écarts plus importants pour d'autres marqueurs, ASSURER, DECLENCHER et ENTRAINER étant proportionnellement plus fréquents dans les données tirées du corpus d'informatique, CAUSER dans le corpus de psychologie, et STIMULER et INDUIRE dans le corpus médical. Une corrélation intéressante peut être faite pour certains marqueurs : les corpus de médecine et de psychologie (dont le contenu est plus étroitement lié que la médecine ou la psychologie et l'informatique) présentent des proportions d'occurrences similaires pour certains marqueurs (ex. ABOUTIR, ASSURER, ENTRAINER), tandis que ces mêmes marqueurs sont plus rares ou absents, ou au contraire plus fréquents, dans le corpus d'informatique. Ce n'est pas toujours le cas cependant (par exemple, dans les cas de ENTRAINER, PRODUIRE et PROVOQUER, apparaissent plus souvent en psychologie et en informatique qu'en médecine).

En ce qui concerne le genre textuel, on note beaucoup plus de différences que de similitudes entre les marqueurs exprimant ce sens : CAUSER, CONDUIRE, PRODUIRE et

[4] Puisque ces pourcentages ont été arrondis, le total n'est pas nécessairement exactement égal à 100 p. cent.

REALISER sont les seuls marqueurs vraiment stables dans les deux genres. Dans les autres cas, il existe des différences d'importance variable : ex. ABOUTIR, DECLENCHER, ENTRAINER, PROVOQUER et STIMULER présentent tous des variations assez éloquentes.

Le sens « X cause que Y soit différent (grâce à Z) » est exprimé par 5 marqueurs (Tableau 3). Le marqueur MODIFIER est le plus souvent observé dans les trois corpus, mais il est particulièrement fréquent dans le corpus d'informatique. Les autres marqueurs montrent une certaine variation aussi d'un corpus à l'autre, la plupart du temps avec une correspondance plus étroite entre le corpus médical et celui de psychologie qu'entre l'informatique et les deux autres corpus (l'exception étant ALTERER). En fonction de ces observations, la dominance du marqueur MODIFIER est très prononcée dans le genre didactique/vulgarisé ; la distribution des occurrences entre les autres marqueurs dans les textes spécialisés est plus équilibrée.

Tableau 3. Marqueurs exprimant le sens « X cause que Y soit différent (grâce à Z) »

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
AGIR	15	2	7	20	4	8	3	7	8	4
ALTERER	33	10	7	37	13	19	14	7	15	13
INFLUENCER	39	9	26	59	15	22	13	26	24	14
MODIFIER	51	49	45	79	66	29	68	45	32	63
MODULER	40	2	14	50	6	22	3	14	20	6
Total	178	72	99	245	104	100	101	99	99	100

L'analyse des occurrences du sens « X cause que Y soit moins/ait moins lieu » (Tableau 4), exprimé par cinq marqueurs distincts, montre également des variations. En général, REDUIRE est le marqueur le plus souvent observé, mais DIMINUER apparaît souvent (en fait, encore plus souvent) dans le corpus d'informatique. Par contraste, les marqueurs CONTROLER et INHIBER ne servent pas à exprimer ce sens dans le corpus d'informatique, mais étaient présents en médical et en psychologie. Cette différence suggère qu'il existe des similitudes entre les corpus de médecine et de psychologie, mais il existe aussi des contre-exemples de ressemblances relativement étroites entre l'informatique et la psychologie pour les marqueurs LIMITER et REDUIRE.

Tableau 4. Marqueurs exprimant le sens « X cause que Y soit moins/ait moins lieu » et nombre et proportions des occurrences correspondant à chaque marqueur

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
CONTROLER	7	0	14	11	10	5	0	15	6	8
DIMINUER	26	32	20	36	42	20	44	21	20	36
INHIBER	33	0	8	36	5	25	0	8	20	4
LIMITER	25	11	13	36	13	19	15	14	20	11
REDUIRE	39	30	41	62	48	30	41	43	34	41

Total	130	73	96	181	118	99	100	101	100	100

Quant aux genres textuels, des variations évidentes sont présentes : les marqueurs DIMINUER et REDUIRE (ainsi que CONTROLER à un moindre degré) expriment plus souvent ce sens dans les données didactiques/vulgarisés, tandis que INHIBER et LIMITER étaient plus souvent identifiés dans les données tirées des textes spécialisées.

L'expression du sens « X (est l'un des agents qui) cause(nt) que Y puisse avoir lieu/être » (Tableau 5) présente aussi des variations importantes : les occurrences tirées du corpus médical sont divisées presqu'également entre les deux marqueurs, tandis que dans le corpus d'informatique le marqueur INTERVENIR est beaucoup plus utilisé (alors que c'est l'inverse dans le corpus de psychologie. Il est difficile avec deux marqueurs seulement de tirer des conclusions sur des corrélations éventuelles entre les résultats observés entre le corpus médical et celui de psychologie. Les variations entre les corpus individuels expliquent sans doute celles des genres textuels, qui révèlent une répartition plus équilibrée dans les textes spécialisés et une fréquence plus élevée dans les occurrences tirées des textes vulgarisés et didactiques.

Tableau 5. Marqueurs exprimant le sens « X (est l'un des agents qui) cause(nt) que Y puisse avoir lieu/être »

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
AIDER	19	1	10	23	7	51	5	83	53	25
INTERVENIR	18	21	2	20	21	49	95	17	47	75
Total	37	22	12	43	28	100	100	100	100	100

Nous observons une variation assez importante dans l'expression du sens « X cause que Y ne fonctionne pas/plus » (Tableau 6). Les occurrences analysées dans le corpus médical sont pour la plupart associées au marqueur INHIBER, avec quelques occurrences du marqueur BLOQUER, tandis que seul le marqueur BLOQUER a été relevé dans le corpus d'informatique. (En fait, le marqueur INHIBER est absent du corpus d'informatique.) Encore une fois, il est difficile de tirer des conclusions sur des correspondances entre les différents corpus avec deux marqueurs seulement[5].

Tableau 6. Marqueurs exprimant le sens « X cause que Y ne fonctionne pas/plus »

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
BLOQUER	4	16	0	4	16	27	100	0	25	100
INHIBER	11	0	1	12	0	73	0	100	75	0
Total	15	16	1	16	16	100	100	100	100	100

[5] Le petit nombre d'occurrences provenant du corpus de psychologie dans ce cas fait qu'une analyse des occurrences en fonction du genre textuel n'apporterait pas de données supplémentaires.

La variation est moins importante mais aussi présente pour certains autres sens observés dans les données analysées. On observe des variations dans les proportions des occurrences du sens « X cause que Y n'ait pas lieu » (Tableau 7), surtout dans le corpus de psychologie, avec 75 p. cent des occurrences analysées indiquées par le marqueur EMPECHER. Ceci est vrai au niveau des corpus et aussi des genres textuels. Quant à la ressemblance entre le corpus de médecine et celui de psychologie, elle est possible mais pas très prononcée.

Tableau 7. Marqueurs exprimant le sens « X cause que Y n'ait pas lieu »

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
BLOQUER	20	15	5	21	19	45	65	25	40	54
EMPECHER	24	8	15	31	16	55	35	75	60	46
Total	44	23	20	52	35	100	100	100	100	100

Le sens « X cause que Y soit plus » a été observé en conjonction avec trois marqueurs (Tableau 8). Dans tous les corpus, le marqueur STIMULER est moins utilisé que les deux autres, qui varient légèrement quant à leur fréquence relative dans les différents corpus et genres textuels. Si on analyse les correspondances entre le corpus médical et celui de psychologie par rapport à l'informatique, l'hypothèse selon laquelle de plus grandes similitudes pourraient être observées étant donné la relation entre les domaines ne serait pas appuyée par ces données : outre STIMULER, les proportions observées dans les données tirées du corpus médical et du corpus d'informatique sont les plus similaires.

Tableau 8. Marqueurs exprimant le sens « X cause que Y soit plus »

	Nombre d'occurrences					**Pourcentage des occurrences**				
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
ACCROITRE	13	21	27	37	24	39	31	55	54	30
AUGMENTER	19	46	20	29	56	58	68	41	42	69
STIMULER	1	1	2	3	1	3	1	4	4	1
Total	33	68	49	69	81	100	100	100	100	100

Dans certains cas très peu de variation est observée (Tableau 9). L'expression du sens « X cause que Y puisse avoir plus lieu » / « X cause que Y puisse avoir lieu plus facilement » est représenté dans des proportions à peu près égales dans trois corpus et les deux genres textuels par les marqueurs FACILITER, FAVORISER et PERMETTRE. Pour CONDUIRE et STIMULER exprimant le sens « X cause que Z fasse Y », les fréquences demeurent aussi relativement stables. Ces deux verbes ne présentent que des variations mineures dans les données, l'usage de STIMULER au lieu de CONDUIRE pour exprimer ce sens étant rare ou absent dans les trois corpus. Il en va de même pour CONTROLER et ENTRAINER lorsqu'ils expriment le sens « X cause que Y fonctionne d'une certaine manière », puisque la vaste majorité des occurrences

analysées en informatique et en psychologie sont exprimés par CONTROLER (le sens n'ayant pas été identifié dans les occurrences analysées en médecine).

Tableau 9. Marqueurs exprimant les sens « X cause que Y puisse avoir (plus) lieu (plus facilement) », « X cause que Z fasse Y » et « X cause que Y fonctionne d'une certaine manière »

	Nombre d'occurrences					**Pourcentage des occurrences**				
« X cause que Y puisse avoir (plus) lieu (plus facilement) »										
Marqueur	**M**	**I**	**P**	**S**	**DV**	**M**	**I**	**P**	**S**	**DV**
FACILITER	44	51	46	76	65	34	38	38	34	39
FAVORISER	44	44	33	68	53	34	32	27	31	32
PERMETTRE	43	41	42	77	49	33	30	35	35	29
Total	131	136	121	221	167	101	100	100	100	100
« X cause que Z fasse Y »										
CONDUIRE	2	8	19	20	9	100	100	90	95	90
STIMULER	0	0	2	1	1	0	0	10	5	10
Total	2	8	21	21	10	100	100	100	100	100
« X cause que Y fonctionne d'une certaine manière »										
CONTROLER	0	10	3	1	12	0	91	100	100	92
ENTRAINER	0	1	0	0	1	0	9	0	0	8
Total	0	11	3	1	13	0	100	100	100	100

Ces données montrent que la plupart des sens observés sont exprimés par les mêmes ensembles de marqueurs dans les corpus et genres textuels, mais que les proportions des occurrences indiquées par les différents marqueurs varient souvent (mais de manière différente) entre corpus et genres textuels. Cela soulève la possibilité que ces derniers facteurs influencent l'expression des relations ; malheureusement, la variation observée au niveau des différents sens rend très difficile la caractérisation précise de l'effet que ceux-ci pourraient avoir.

Les données permettent tout de même d'observer un certain nombre de corrélations possibles entre corpus (surtout les corpus associés à des domaines étroitement liés, comme la médecine et la psychologie) quant à l'utilisation de certains marqueurs pour exprimer des sens spécifiques, mais il n'a pas été possible de conclure avec certitude que ce genre de ressemblance est prévisible. Un nombre plus important de données seraient nécessaires pour tirer ces conclusions.

5 Conclusions et perspectives

L'analyse réalisée a permis de constater une variation parfois importante dans le nombre d'occurrences de 37 marqueurs verbaux associés à des relations de cause-effet dans trois corpus et deux genres textuels. Cependant, bien que certains sens causaux analysés présentent des variations évidentes, d'autres varient peu. Nous

reconnaissons ainsi la difficulté de dégager des tendances claires quant à la capacité des différents marqueurs à exprimer des sens précis dans des corpus spécifiques.

L'observation qui précède tend à confirmer les conclusions d'études précédentes et laisse supposer que la productivité de marqueurs individuels aura tendance à varier d'un corpus à l'autre selon les caractéristiques des textes qui les composent. Á notre avis, cette variation milite en faveur de l'inclusion d'une gamme aussi vaste que possible de différents marqueurs dans des applications qui y ont recours pour repérer et classer des relations terminologiques.

Nous devons néanmoins reconnaître que cet échantillon de données est limité, et que, s'il permet de dégager certaines tendances, un nombre plus conséquent de données s'impose. Une analyse qui inclut d'autres genres textuels et d'autres domaines permettrait d'étayer nos observations. D'ailleurs, nos résultats reflètent la contribution de plusieurs facteurs aux différentes étapes de notre analyse (par exemple, la fréquence variable des marqueurs en général dans les divers corpus, ainsi que la polysémie des marqueurs, non seulement dans l'attestation de sens causaux ou non causaux, mais aussi dans l'expression de différents sens causaux). Il s'agit donc d'une combinaison complexe de facteurs qui interviennent dans les résultats décrits.

Enfin les différences observées soulèvent une question plus pratique : est-il possible d'exploiter efficacement les marqueurs lexicaux dans des applications automatiques ? Nous persistons à croire que oui. Cela dit, il sera nécessaire de continuer à parfaire les techniques de repérage et de mettre au point des stratégies afin de réduire le bruit et améliorer les résultats de manière générale. Le niveau d'automatisation prévu pour des applications spécifiques aidera sans doute à déterminer ce qui est acceptable comme quantité de bruit ; une application qui vise à présenter à l'utilisateur une liste triée de contextes potentiellement riches en connaissances pourra tolérer davantage de bruit qu'une application entièrement automatique. Des stratégies pour trier des contextes potentiellement utiles (en analysant, par exemple, les structures actancielles et les classes d'actants observés dans les textes, comme décrit dans Marshman & L'Homme (2006)) pourraient aussi contribuer à améliorer les résultats. Il est donc essentiel de continuer des recherches de ce genre pour maximiser la productivité des marqueurs.

Remerciements

Nous remercions le Conseil de recherches en sciences humaines du Canada et le Fonds québécois de recherches sur la société et la culture du Québec pour leur soutien, et Stéphanie Caron pour son travail sur les corpus.

Références

AHMAD K. & FULFORD H. (1992). *Knowledge Processing: 4. Semantic Relations and their Use in Elaborating Terminology*. (Computing Sciences Report CS-92-07). Guildford.

BARRIÈRE C. (2002). Hierarchical refinement and representation of the causal relation. *Terminology*. 8(1), p. 91-111.

BIBER D. (1988). *Variation across Speech and Writing*. Cambridge.

CONDAMINES A. (2000). Chez dans un corpus de sciences naturelles: un marqueur de relation de relation méronymique? *Cahiers de lexicologie* 77, p. 165-187.

CONDAMINES A. (2002). Corpus analysis and conceptual relation patterns. *Terminology*. 8(1), p. 141-162.

CONDAMINES A. (2008). Taking *genre* into account when analysing conceptual relation patterns. *Corpora* 3(2), p. 115-140.

CONDAMINES A. & REBEYROLLE J. (2001). Searching for and identifying conceptual relationships via a corpus-based approach to a Terminological Knowledge Base (CKTB): Method and Results. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME Eds. *Recent Advances in Computational Terminology*. p. 127–148. Amsterdam/Philadelphia.

FELIU, J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. Thèse de doctorat, Universitat Pompeu Fabra.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING-92*. p. 539-545. Nantes.

GARCIA, D. (1997). Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes. In *Actes des deuxièmes rencontres — Terminologie et Intelligence Artificielle, TIA '97*. p. 7–26. Toulouse.

JACQUES M.-P. & AUSSENAC-GILLES N. (2006). Variabilité des performances des outils de TAL et genre textuel. *T.A.L.* 47(1), p. 11-32.

MALAISÉ, V., P. ZWEIGENBAUM & B. BACHIMONT. (2005). Mining defining contexts to help structuring differential ontologies. *Terminology*. 11(1), p. 21–53.

MARSHMAN E. (2007). *Lexical Knowledge Patters for Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts: A Comparative Analysis of English and French*. Doctoral thesis, Département de linguistique et de traduction, Université de Montréal, Montreal, Canada.

MARSHMAN, E. & M.C. L'HOMME. (2006). Disambiguating lexical markers of cause and effect using actantial structures and actant classes. In Picht, H. Ed. *Modern approaches to terminological theories and applications. Proceedings of the 15th European Symposium on Language for Special Purposes, LSP 2005*. p. 261–285. Bern.

MARSHMAN E., L'HOMME M.-C. & SURTEES V. (2008). Portability of cause-effect relation markers across specialized domains and text genres: A comparative evaluation. *Corpora*. 3(2), p.141-172.

MARSHMAN E., L'HOMME M.-C. & SURTEES V. (2008a). Verbal Markers of Cause-Effect Relations across Corpora. In B. NISTRUP MADSEN & H. ERDMAN THOMSEN Eds. *Managing Ontologies and Lexical Resources. Proceedings of the 8th International Conference on Terminology and Knowledge Engineering, TKE 2008*. p. 159-173. Copenhagen.

MARSHMAN E., MEYER I. & MORGAN T. (2002). French patterns for expressing concept relations. *Terminology*. 8(1), p. 1-29.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In D. BOURIGAULT, C. JACQUEMIN & M.-C. L'HOMME Eds. *Recent Advances in Computational Terminology*. p. 279-302. Amsterdam/Philadelphia.

MEYER I., MACKINTOSH K., BARRIÈRE C. & MORGAN T. (1999). Conceptual sampling for terminographical corpus analysis. In *Proceedings of Terminology and Knowledge Engineering TKE '99*. p. 256-267. Innsbruck.

NUOPPONEN, A. (1994). Causal Relations in Terminological Knowledge Representation. *Terminology Science and Research* 5(1). p. 36-44.

NUOPPONEN, A. (2005). Concept relations: An update of a concept relation classification. In B.N. Madsen & H.E. Thomsen Eds. *Terminology and Content Development: Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*. p. 127–138. Copenhagen.

PEARSON J. (1998). *Terms in Context*. Amsterdam/Philadelphia.

SÉGUÉLA P. (1999). Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés. *Terminologies nouvelles*. 19, p. 52-60.