Constitution d'une base bilingue de marqueurs de relations conceptuelles pour l'élaboration de ressources termino-ontologiques

Luce Lefeuvre
CLLE-ERSS, UMR 5263
CNRS & Université Toulouse Jean-Jaurès
Toulouse, France
Luce.lefeuvre@univ-tlse2.fr

Anne Condamines
CLLE-ERSS, UMR 5263
CNRS & Université Toulouse Jean-Jaurès
Toulouse, France
Anne.condamines@univ-tlse2.fr

Résumé

Les marqueurs de relations conceptuelles sont un moyen efficace de détecter des contextes utiles à l'élaboration de ressources termino-ontologiques. De nombreux travaux existent, mais aucun recensement n'a été effectué. Nous souhaitons construire une base de marqueurs de relation pour l'hyperonymie, la méronymie et la cause, en français et en anglais. La prise en compte de la variation dans l'analyse de ces marqueurs nous permettra de caractériser leur fonctionnement.

1 Introduction

Notre étude se situe dans le cadre du projet ANR CRISTAL (Contextes RIches en connaissanceS pour la TrAduction terminoLogique) dont l'un des objectifs consiste à affiner la notion de Contextes Riches en Connaissances (Meyer, 2001) en prenant en compte différents paramètres de variation tels que la langue (français vs anglais), le domaine (oncologie vs volcanologie), le genre (scientifique vs vulgarisé) et l'utilisateur (traducteur vs terminologue). Nous adoptons ici le point de vue du terminologue. Nous nous intéressons aux relations que peuvent entretenir au moins deux termes, en considérant que ces relations sont un type de connaissance qu'il est possible de découvrir dans un corpus spécialisé. Le projet s'inscrit ainsi dans la thématique de la construction de ressources termino-ontologiques.

L'un des moyens d'accéder à ces connaissances consiste à utiliser des marqueurs de relations conceptuelles. Non ignorés de l'ingénierie des connaissances, de la lexicographie ou de la terminologie, ces éléments linguistiques n'ont pas fait l'objet d'un recensement systématique, ni d'une analyse à grande échelle.

Nous mentionnons en section 2 les travaux dans la lignée desquels nous nous situons. La section 3 décrit la méthodologie que nous avons adoptée. Nous présentons quelques résultats en section 4, et discutons des perspectives de travail en section 5.

2 Travaux antérieurs

La notion de marqueur de relation a souvent été abordée pour élaborer des réseaux de termes, que ce soit en ingénierie des connaissances, en terminologie, ou en traitement automatique des langues. Constitués d'éléments lexico-syntaxiques, typographiques ou dispositionnels (Auger et Barrière, 2008), ils peuvent être utilisés pour expliciter la relation qui unit deux termes. Cette connaissance peut être représentée par un triplet de la forme « Terme 1 - Marqueur - Terme 2 », dans lequel le marqueur précise la relation existant entre les deux termes. Par exemple, la relation d'hyperonymie (générique - spécifique) peut être indiquée par le marqueur « X est un Y + caractéristiques différentielles » (« *Le cancer est une maladie caractérisée par la prolifération incontrôlée de cellules* ») ; et la relation de méronymie (ou partie - tout) peut être indiquée par le marqueur « X être {formé/constitué} de DET Y » (« *le volcan primitif est en majorité constitué de coulées d'andésites* »). Les marqueurs étudiés concernent principalement trois relations : l'hyperonymie, la méronymie, et la cause. Considérées comme structurantes, et supposées universelles, elles apportent des éléments de connaissance sur les termes d'un domaine.

De nombreux travaux s'attachent ainsi à décrire les marqueurs de ces relations (Alarcon-Martinez, 2009 ; Hearst, 1992 ; Garcia, 1998 ; Cruse, 2002 ; Séguéla, 2001 ; Condamines et Rebeyrolle, 2000). Ces études descriptives doivent permettre d'exploiter les marqueurs de relation à l'aide d'outils dédiés, afin de détecter le plus automatiquement possible des triplets structurant les ressources termino-ontologiques.

D'autres travaux plus récents s'intéressent à la variation de ces marqueurs selon le genre textuel, le domaine, ou la langue (Condamines, 2002 ; Marshman, 2006 ; Marshman et L'Homme, 2006 ; Pearson, 1998). Ces travaux montrent que la productivité et la répartition des marqueurs varie parfois fortement d'un domaine ou d'un genre à l'autre. Ils soulignent la nécessité de prendre en compte la variation dans la description des marqueurs de relation, afin d'en étudier la « portabilité » (Marshman et L'Homme, 2006).

Bien que la littérature sur ce sujet soit abondante, il n'existe pas de base de données recensant l'ensemble des marqueurs des relations d'hyperonymie, de méronymie et de cause, ni d'analyse systématique à grande échelle de ces marqueurs. Notre contribution sera de constituer cette base de données et d'analyser chaque candidat-marqueur afin d'en donner une description linguistique fine.

3 Méthodologie

Notre travail s'est déroulé selon deux étapes :

1) Élaboration de la liste des candidats-marqueurs en français et en anglais pour les relations d'hyperonymie, de méronymie et de cause
2) Analyse des occurrences des candidats-marqueurs français en corpus.

Nous détaillons dans la suite chacune de ces étapes.

3.1 Constitution de la base de marqueurs

La base de marqueurs de relation a été construite en deux phases :

1) Recensement des marqueurs de relation pour le français. À partir des travaux existants et dans la lignée des travaux mentionnés en section 2, nous avons fait une liste la plus exhaustive possible des marqueurs français pour trois relations : hyperonymie, méronymie, cause.

2) Élaboration de la liste des marqueurs de relation pour l'anglais (Fabre, 2014). Une première liste de marqueurs a été dressée à partir d'une étude bibliographique. Cette liste a ensuite été enrichie par la traduction de certains marqueurs de relation français. Une première validation a été effectuée en vérifiant dans le COCA corpus[1] les contextes d'apparition des nouveaux candidats-marqueurs anglais obtenus. La relecture de cette liste par une linguiste anglophone a ensuite permis de valider la liste finale.

Le tableau suivant recense le nombre de candidats-marqueurs obtenus pour chaque relation et pour chaque langue[2].

Marqueurs de relation conceptuelle	FRANÇAIS	ANGLAIS
Hyperonymie	33	35
Méronymie	95	99
Cause	192	247

Tableau 1. Nombre de candidats-marqueurs par relation et par langue.

3.2 Évaluation en corpus

La seconde étape de notre travail a concerné l'analyse à grande échelle des candidats-marqueurs en français, en prenant en compte les différents paramètres de variation que nous avons listés plus haut. Notre corpus traite ainsi de deux domaines : la volcanologie, qui appartient aux Sciences de la Terre, et l'oncologie, qui appartient aux Sciences de la Vie. Pour chacun de ces domaines, nous avons pu constituer un corpus scientifique très spécialisé et un corpus vulgarisé, en français et en anglais. Les corpus scientifiques sont constitués de textes issus de revues spécialisées, écrits par des experts à destination d'experts du domaine ou de domaines connexes. Les corpus vulgarisés sont constitués de textes issus de revues ou de sites internet de vulgarisation ; ils sont écrits par des experts du domaine et sont à destination du grand public. Les textes français ont été écrits par des auteurs francophones, et les textes anglais par des au-

[1] Davies, M. (2008-). *The Corpus of Contemporary American English: 450 million words, 1990-present.* Disponible en ligne : http://corpus.byu.edu/coca/.

[2] La liste des marqueurs français et anglais sera disponible sur le site du laboratoire CLLE-ERSS : http://w3.erss.univ-tlse2.fr/

teurs anglophones. Le tableau 2 ci-dessous synthétise ces informations.

	Oncologie	**Volcanologie**
Corpus scientifique	200 000 mots / langue	400 000 mots / langue
	2002 – 2008	1980 - 2012
Corpus vulgarisé	200 000 mots / langue	400 000 mots / langue
	2001 - 2008	1980 - 2002

Tableau 2. Constitution du corpus d'étude.

Nous avons extrait de ce corpus les contextes comportant les candidats-marqueurs recensés. Pour chaque candidat-marqueur de chaque relation, nous avons annoté le contexte comme suit :

- « Oui » : la relation est présente

« *Un dynamisme explosif, extrusif et / ou intrusif **a généré** des cônes stromboliens, des necks basaltiques* » (volcanologie, scientifique).
Le candidat-marqueur « Det X générer Det Y » lie les termes « *dynamisme explosif, extrusif et / ou intrusif* » d'une part et « *cônes stromboliens* » et « *necks basaltiques* » d'autre part par la relation de cause.

- « Non » : le candidat-marqueur n'indique pas la relation attendue

« *Mais notre but **est un** autre volcan **très** actif et dangereux* » (volcanologie, vulgarisation)
Le candidat-marqueur testé « Y être DET X très Adj » n'indique pas la relation d'hyperonymie attendue entre « *but* » et « *volcan* ».

- « Plutôt oui » : le candidat-marqueur exprime la relation conjointement avec un autre élément.

« *Trop de repos ou un manque d'activité peuvent **diminuer** l'oxygénation des tissus musculaires* » (oncologie, vulgarisation)
La nominalisation « oxygénation » associée au candidat-marqueur « diminuer » nous permet d'interpréter la relation comme causale. Deux éléments du triplet sont ainsi présents.

- « Plutôt non »: la relation est difficile à interpréter ; ou alors les éléments en relation ne nous intéressent pas dans l'optique de construire des ressources termino-ontologiques (ce ne sont pas des termes du domaine par exemple).

« *Cette découverte **motive** son élection à l'Académie des sciences* » Relation de cause (volcanologie, vulgarisation)
Il ne nous semble pas pertinent d'intégrer les éléments en relation à une ressource terminologique liée au domaine de la volcanologie.

- « Indéterminé »: nous ne pouvons évaluer la relation (par manque d'indices linguistiques ou par manque de connaissances sur le domaine).

« *Hormones hypophysaires : Ce sont des hormones **sécrétées par** l'hypophyse, glande cérébrale située juste sous le cerveau* » (oncologie, vulgarisation)
Les candidats-termes « *hormones* » et « *hypophyse* » peuvent être reliés par une relation de cause ou une relation de fonction. Aucun indice linguistique ne nous permet de statuer pour l'une ou l'autre des relations.

Environ 10000 contextes ont été annotés selon ces critères.

4 Résultats

Comptabilisant ensemble les « oui » et « plutôt oui », nous avons effectué deux types de calculs : la fréquence d'apparition des candidats marqueurs dans les corpus, et la productivité de chaque candidat-marqueur. Cette productivité correspond au pourcentage des énoncés contenant un candidat marqueur pouvant être interprétés comme contenant la relation attendue.

Nous avons ainsi pu mettre au jour quelques phénomènes de variation liés au domaine ou au genre textuel que nous présentons ici.

4.1 Influence du genre textuel

Les candidats-marqueurs de la relation de méronymie sont organisés selon différentes catégories, qui peuvent préciser par exemple : le type de liaison que les parties d'un ensemble entretiennent (fusion, jonction, inclusion), le type même des parties, si ces parties sont organisées ou non (organisation, non organisation), si elles proviennent de la décomposition d'objets, si elles correspondent à l'expression d'un lieu. Plusieurs candidats-marqueurs n'apparaissant pas du tout dans les corpus, nous avons choisi d'observer la façon dont les occurrences des candidats-marqueurs sont réparties à travers les catégories plutôt que de les comparer de façon isolée.

Catégories de regroupement	**Occ. VULGARISATION**	**Occ. SCIENT**	TOTAL
Inclusion	71	109	180
Non-organisation	37	3	40
Organisation	12	10	22
Types de parties	28	28	56
Lieu	38	40	78
Parties de même genre	29	20	49
TOTAL	215	210	425

Tableau 3. Répartition des occurrences totales de certains candidats-marqueurs de la relation de méronymie par catégorie.

Le tableau 3 ci-dessus présente la répartition des occurrences des candidats-marqueurs selon certaines catégories. On remarque que dans les catégories « Inclusion » et « Non-organisation », les occurrences ne sont pas réparties de manière équilibrée. Les candidats-marqueurs exprimant l'inclusion d'une partie dans une autre sont plus fréquents dans le corpus scientifique. Les candidats-marqueurs indiquant que les parties ne sont pas organisées entre elles sont plus fréquents dans le corpus vulgarisé. Un Chi-test[3] ($p \leq 0{,}001$) a confirmé la différence des deux corpus par rapport aux catégories des candidats-marqueurs.

La catégorie « Inclusion » comporte des candidats-marqueurs comme « X {comprendre/ abriter/comporter/compter/inclure/intégrer} DET Y », ou « Y (être) {classé/classifié/catalogué/ rangé/placé/inclus/étiqueté/catégorisé/groupé} dans DET X ». Leur fréquence plus importante en corpus scientifique peut être due à deux facteurs. Le premier concerne la notion d'inclusion elle-même, qui peut être difficile à appréhender, et que l'on retrouve souvent dans les domaines des mathématiques, de la logique, de la biologie, de la minéralogie. L'autre facteur concerne les éléments en relation dans ces structures. Dans la plupart des contextes contenant ces candidats-marqueurs, les éléments en relation sont des candidats-termes : « acte chirurgical » et « curage axillaire », « complexe volcanique » et « cratère » par exemple. Si l'on ne connait pas la signification de ces termes, un effort de compréhension est nécessaire pour saisir le lien de méronymie qu'il peut exister. On peut ainsi émettre l'hypothèse que l'apparition de ces candidats-marqueurs est liée à une volonté des auteurs, experts de leur domaine, de s'adresser à leurs pairs, sans avoir à détailler à la fois la relation d'inclusion et la spécificité des termes en relation.

La catégorie « Non-organisation » comporte quant à elle des candidats-marqueurs comme « X {être/résulter/de/issu de} DET {tas/amas/ ramassis/masse/accumulation/entassement} de (DET) Y » ou « {tas/amas/ramassis/masse/ accumulation/entassement} de (DET) Y {dans/en/pour former /pour constituer/donner} (DET) X ». La présence d'éléments du lexique comme « tas » ou « accumulation » rend ces structures facilement compréhensibles. Elles ne fournissent pas d'information précise sur les liens que peuvent entretenir les parties. Assez générales et peu spécialisées, elles peuvent être comprises par tous les lecteurs ; quand bien même les éléments en relation seraient des candidats-termes comme « lave » et « dôme » ou « cellules » et « ganglions lymphatiques » par exemple. Ce manque de précision peut expliquer la très forte fréquence d'apparition de ces candidats-marqueurs en corpus vulgarisé. Les auteurs ne peuvent en effet pas détailler toutes les connaissances d'un domaine spécialisé.

Finalement, il semblerait que le genre textuel ait une influence à plusieurs niveaux : au niveau des catégories de la relation de méronymie, au niveau des candidats-marqueurs eux-mêmes, au niveau des éléments en relation.

4.2 Influence du domaine

Le fonctionnement des candidats-marqueurs de cause semble varier de manière significative en fonction du domaine (figure 1).

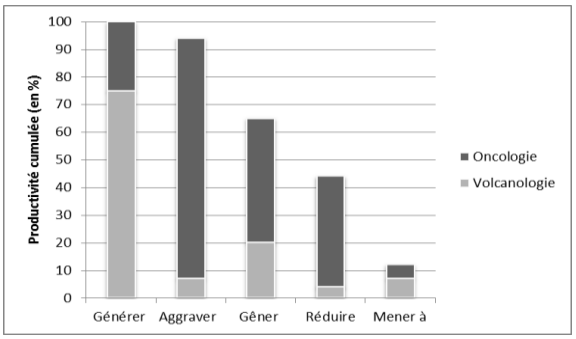

Figure 1. Répartition de quelques candidats-marqueurs de cause selon le domaine.

Dans le domaine de l'oncologie, les candidats-marqueurs de cause les plus représentés (*aggra-*

[3] Je remercie sincèrement Basilio Calderone, membre de CLLE-ERSS pour son aide.

ver, *gêner*, *réduire*, *diminuer*) appartiennent aux catégories /influencer/ et /gêner/, que l'on peut paraphraser par « X cause une influence/une gêne sur Y ». Dans le domaine de la volcanologie, les candidats-marqueurs les plus représentés (*générer*, *mener à*, mais aussi *déclencher*, *créer*, *engendrer*) sont liés à la catégorie /créer/, qui indique qu'un phénomène ou une situation X est la cause de l'existence d'un phénomène ou d'une situation Y. Les objectifs distincts des deux domaines peuvent expliquer ces différences. L'oncologie, et la médecine plus généralement, a pour but de soigner, d'étudier le développement des maladies, de décrire des symptômes, des effets secondaires liés aux traitements. En objet des candidats-marqueurs présents, on retrouve des éléments du lexique comme "séquelles", "dépression", "lymphœdème", "cancer", qui sont liés aux symptômes, aux diagnostics, aux traitements du cancer. La volcanologie a pour objectif d'étudier l'origine ainsi que les mécanismes du volcanisme. Elle s'intéresse à la création des volcans, mais également à ce qu'ils produisent, ce qui va de concert avec la catégorie /créer/ de la relation de cause. On retrouve ainsi en objet des candidats-marqueurs de cause présents des éléments lexicaux qui désignent les produits des volcans : "cendres", "lahars", ou qui concernent la typologie des volcans : "structures", "cônes". Dans les deux cas, il semble bien que ce soit le domaine qui ait une influence sur l'apparition des candidats-marqueurs de cause.

5 Perspectives

Les premiers résultats nous ont permis de valider nos hypothèses sur l'influence du genre et/ou du domaine sur le fonctionnement des marqueurs de relation. Nous souhaitons pour la suite mener des analyses plus fines, afin de mettre en évidence des fonctionnements propres à chaque sous-corpus en lien avec la nature de sa variation. Cela nous permettra de mettre au point des catégories de fonctionnement des marqueurs de relation en fonction du domaine et du genre. Nous pourrons ainsi dresser une typologie des marqueurs de relation, indiquant les cas dans lesquels les marqueurs sont productifs : dans tous les corpus, dans le domaine de la volcanologie, dans le genre vulgarisé, etc.

Le second aspect que nous souhaitons développer concerne l'amélioration de la productivité des marqueurs. Pour cela, nous souhaitons utiliser différentes ressources externes pour contraindre le co-texte. Ces ressources, de type lexical, nous permettront à la fois de sélectionner et de filtrer les contextes extraits. L'utilisation de la liste des candidats-termes ainsi que celle des nominalisations déverbales nous permettront par exemple de sélectionner des triplets complets. Le lexique transdisciplinaire scientifique pourra nous permettre de filtrer certains contextes n'apportant pas de connaissances spécifiques sur le domaine.

Enfin, il serait intéressant de projeter des couples de termes dont on connaît la relation afin de pouvoir découvrir des marqueurs spécifiques au domaine.

Références

Alarcon Martinez, R. (2009). *Descripción y evaluación de un sistema basado en reglas para la extracción automática de contextos definitorios*. Thèse de doctorat (non publiée) de l'Université Pompeu Fabra (discipline Sciences du Langage), Barcelone.

Auger, A., & Barrière, C. (2008). Pattern based approaches to semantic relation extraction: a state-of-the-art. *Terminology*, 14(1), 1-19.

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

Condamines, A., & Rebeyrolle, J. (2000). Construction d'une base de connaissances terminologiques à partir de textes : expérimentation et définition d'une méthode. In J. Charlet, M. Zacklad, G. Kassel, D. Bourigault, (eds.), *Ingénierie des Connaissances, évolutions récentes et nouveaux défis* (pp. 225-242). Paris: Eyrolles.

Cruse, A. (2002). Hyponymy and its Varieties. In R. Green, C.A. Bean, & S.-H Myaeng (eds.), *The semantics of relationships* (pp. 3-22). Dordrecht/Boston/London, Kluwer Academic Publishers.

Fabre, L. (2014). *Élaboration d'une liste de marqueurs de relations conceptuelles en anglais*. Rapport de stage de Master 2 (discipline Linguistique Anglaise) au sein du laboratoire CLLE-ERSS, Université Toulouse – Jean Jaurès, Toulouse.

Garcia, D. (1998). *Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis*. Thèse de doctorat de l'Université Paris IV - Sorbonne (discipline Informatique), Paris.

Hearst, M. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes.

Marshman, E. (2006). *Lexical Knowledge Patterns for the Semi-automatic Extraction of Cause-effect and Association Relations from Medical Texts: A Comparative Analysis of English and French*. Thèse de doctorat de l'Université de Montréal (discipline Traduction), Montréal.

Marshman, E., & L'Homme, M.-C. (2006). Portabilité des marqueurs de la relation causale : étude sur deux corpus spécialisés. In F. Maniez, P. Dury, N. Arlin & C. Rougemont (eds.), *Corpus et dictionnaires de langues de spécialité*. *Actes des Journées du CRTT 2006* (pp. 87-110), Nantes.

Meyer, I. (2001). Extracting Knowledge-rich Contexts for Terminography: A Conceptual and methodological Framework. In D. Bourigault, M.C. L'Homme & C. Jacquemin (eds.), *Recent Advances in Computational Terminology* (pp. 279-302). Amsterdam/Philadelphia: John Benjamins.

Pearson, J. (1996). The Expression of Definition in Specialized Texts: A Corpus-based Analysis. In M. Gellerstam et al. (eds.), *Proceedings of the Seventh Euralex International Congress* (pp. 817–824), Göteborg.

Séguéla, P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de doctorat de l'Université Paul Sabatier (discipline Informatique), Toulouse.