MAR-REL : une base de marqueurs de relations conceptuelles pour la détection de Contextes Riches en Connaissances

Luce Lefeuvre Anne Condamines
CLLE-ERSS, Toulouse, France

RÉSUMÉ

Les marqueurs de relation conceptuelle sont un moyen efficace de détecter automatiquement en corpus des *Contextes Riches en Connaissances*. Dans le domaine de la terminologie ou de l'ingénierie des connaissances, les *Contextes Riches en Connaissances* peuvent être utiles pour l'élaboration de ressources termino-ontologiques. Bien que la littérature concernant ce sujet soit riche, il n'existe pas de recensement systématique ni d'évaluation à grande échelle des marqueurs de relation conceptuelle. Pour ces raisons notamment, nous avons constitué une base de marqueurs pour les relations d'hyperonymie, de méronymie, et de cause, en français. Pour chacun de ces marqueurs, son taux de précision est proposé pour des corpus qui varient en fonction du domaine et du genre textuel.

ABSTRACT

MAR-REL : a conceptual relation markers database for Knowledge-Rich Contexts extraction
Markers of conceptual relation are an efficient means to automatically retrieve *Knowledge-Rich Contexts* from corpora. Those contexts can be useful when building terminological or ontological resources. Although this subject has been actively researched, there exists neither an inventory nor a broad study of the markers of conceptual relation. For those reasons, we built a French markers database for the hyperonym, the meronym and the causal relations. Each marker is associated with his accuracy rate depending on domain and text genre.

MOTS-CLÉS : Marqueur de relation conceptuelle, Contextes Riches en Connaissances, variation, domaine, genre textuel, ressources terminologiques, corpus spécialisés.

KEYWORDS: Marker of conceptual relation, knowledge-rich context, variation, domain, text genre, terminological resources, specialized corpora.

1 Introduction

En terminologie ou en ingénierie des connaissances, les linguistes s'appuient sur des corpus spécialisés pour constituer des ressources termino-ontologiques ou des réseaux de termes. Des outils informatiques peuvent aider, accompagner l'exploration de ces corpus, en détectant notamment des portions de textes supposées « riches en connaissances » car contenant des informations conceptuelles sur les termes. Ce type de portion de texte, ou *Contexte Riche en Connaissances* (Meyer, 2001) est défini comme un « *context indicating at least one item of domain knowledge that could be useful for conceptual analysis* ». Il est considéré comme pertinent car il peut constituer le point de départ de définitions, ou améliorer la connaissance d'un domaine. L'une des méthodes pour

repérer ce type de contexte consiste à définir des marqueurs de relations conceptuelles. Utilisés en ingénierie des connaissances, en lexicographie, ou en terminologie, ces éléments linguistiques permettent de repérer des relations existant entre deux termes.

Dans le cadre du projet ANR CRISTAL[1] (Contextes RIches en connaissanceS pour la TrAduction terminoLogique), nous avons constitué une base de marqueurs français permettant d'identifier en corpus spécialisé les relations d'hyperonymie, de méronymie et de cause. Afin de caractériser leur fonctionnement linguistique, nous avons mené une analyse à grande échelle en corpus variant du point de vue du domaine (Volcanologie vs. Cancer du sein) et du genre textuel (scientifique vs. vulgarisé). Les résultats obtenus à la suite de cette analyse, en particulier la précision, sont associés à chaque candidat-marqueur de la base, et constituent une indication quant à la capacité du candidat-marqueur à indiquer la relation attendue.

Nous définissons en section 2 les marqueurs de relation conceptuelle, en nous appuyant sur les travaux existants. En section 3, nous exposons notre méthodologie de recensement et d'évaluation des candidats-marqueurs de relation. Enfin, nous présentons les résultats obtenus en section 4, et discutons des suites possibles de ce travail en section 5.

2 Contextes Riches en Connaissances et marqueurs de relations conceptuelles

La notion de *Contexte Riche en Connaissances* fait souvent écho à la notion de marqueurs de relation conceptuelle, utilisés en intelligence artificielle pour l'élaboration de ressources terminologiques notamment. Constitués d'éléments lexico-syntaxiques, voire typographiques ou dispositionnels (Auger, Barrière, 2008), ils mettent en évidence le lien existant entre deux termes, et sont généralement représentés sous la forme d'un triplet « Terme 1 - Marqueur - Terme 2 », dans lequel le marqueur exprime 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]** (« *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* »). Bien que potentiellement, il existe une multitude de relations pertinentes à étudier, les marqueurs examinés concernent principalement trois relations : l'hyperonymie, la méronymie, et la cause (Marshman et al. 2012 ; Nuopponen, 2014). Considérées comme structurantes et universelles, ces trois relations apportent des éléments de connaissance sur les termes d'un domaine.

De nombreux travaux s'attachent ainsi à décrire les marqueurs de ces trois relations. Dans le domaine de la lexicologie, des recherches ont été menées autour des différentes formes de la définition (Flowerdew, 1992 ; Pearson, 1996 ; Rebeyrolle, 2000). Dans le domaine du Traitement Automatique des Langues, les travaux de Hearst (1992) ont servi de modèle pour automatiser la détection de la relation d'hyperonymie, et donc pour construire *in fine* des taxonomies ou des ontologies de manière automatique ou semi-automatique (Morin, 1999, Malaisé et al., 2004, entre autre). Les principes méthodologiques de ces travaux ont par ailleurs été repris dans de nombreux travaux s'intéressant à la mise en œuvre automatique des marqueurs de relation (Aussenac-Gilles, Séguéla, 2000 ; Condamines, Rebeyrolle, 2001, Lafourcade, Ramadier, 2016). Concernant la relation de méronymie, Girju et son équipe (2003), dans le domaine de l'extraction d'information,

[1] Projet ANR CRISTAL [ANR-12-CORD-0020], http://www.projet-cristal.org

se sont penchés sur l'expression de la méronymie en anglais et sur des méthodes efficaces pour extraire des contextes contenant cette relation. Enfin, certains chercheurs ont souligné l'intérêt d'étudier les énoncés causaux en discours et la pertinence d'intégrer la relation de cause aux ressources terminologiques (Garcia, 1998 ; Barrière, 2001).

D'autres travaux, plus rares, s'intéressent à la variation de ces marqueurs selon le genre textuel, le domaine, ou la langue (Pearson, 1998 ; Condamines, 2000, 2006, 2008 ; Marshman, 2008 ; Marshman, L'Homme, 2008). 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 » d'un corpus à un autre (Marshman, 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. Nous avons donc constitué cette base de données et avons analysé le fonctionnement de chaque candidat-marqueur dans des corpus variant en fonction du domaine et du genre.

3 Recensement et évaluation des marqueurs de relation

Notre travail s'est déroulé en 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 Recensement des candidats-marqueurs de relation

D'une manière générale, nous avons procédé de la même façon pour élaborer les listes des marqueurs des trois relations étudiées. La base de marqueurs de relation a été construite en trois phases :

1. À partir des travaux existants et dans la lignée des travaux mentionnés en section 2, nous avons relevé les marqueurs français des relations d'hyperonymie, de méronymie, et de cause,
2. Nous avons ensuite évalué la première liste obtenue, sur la base de l'introspection, afin d'y ajouter des modifications si nécessaire,
3. Enfin, nous avons uniformisé le codage linguistique des candidats-marqueurs retenus.

La liste finale résulte à la fois d'un travail introspectif et de données issues de l'observation de corpus. Le tableau suivant recense le nombre de candidats-marqueurs obtenus pour chaque relation :

Relation conceptuelle	Nombre de candidats-marqueurs recensés
Hyperonymie	33
Méronymie	95
Cause	192

TABLE 1. Nombre de candidats-marqueurs recensés par relation conceptuelle.

3.2 Évaluation en corpus

La seconde phase de notre travail a concerné l'analyse à grande échelle des candidats-marqueurs, 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 le cancer du sein chez la femme, qui appartient aux Sciences du Vivant. Ces deux domaines scientifiques sont très spécialisés et font l'objet de beaucoup de vulgarisation. Nous avons donc pu constituer pour chaque domaine une partie scientifique et une partie vulgarisée. 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 ou par des auteurs ayant des connaissances avancées dans le domaine et sont à destination du grand public. Le tableau 2 ci-dessous synthétise ces informations.

	Cancer du sein	Volcanologie
Corpus scientifique	200 000 mots	400 000 mots
	2002 – 2008	1980 - 2012
Corpus vulgarisé	200 000 mots	400 000 mots
	2001 - 2008	1980 - 2002

TABLE 2. Constitution du corpus

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* » (cancer du sein, 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* » (cancer du sein, 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 9000 contextes ont été annotés selon ces critères.

4 Variabilité du fonctionnement des candidats-marqueurs

Comptabilisant ensemble les « oui » et « plutôt oui », nous avons calculé la précision de chaque candidat-marqueur. Étant donné le nombre d'occurrences total d'un candidat-marqueur, nous avons calculé le ratio du nombre de cas dans lesquels la relation était présente sur le nombre d'occurrence total. Ce rapport, exprimé en pourcentage, correspond à la formule suivante :

$$Précision = \left(\frac{\text{Nombre d'occurrences dans lesquelles la relation est présente}}{\text{Nombre d'occurrences total}}\right) \times 100$$

Nous présentons dans la suite deux types de résultats : une analyse quantitative comparant les taux de précision des candidats-marqueurs de méronymie selon le domaine, et une analyse qualitative de la variation selon le domaine également. À partir de ces observations, nous proposons quelques pistes de réutilisabilité des candidats-marqueurs.

4.1 Analyse quantitative

Le calcul des taux de précision des candidats-marqueurs nous a permis de comparer la répartition des candidats-marqueurs au sein des sous-corpus, ainsi que leur taux de précision. Deux hypothèses ont émergé : le domaine influence les types de relations présentes en corpus ; tandis que le genre textuel influence l'apparition de certains candidats-marqueurs. À travers l'exemple des relations de méronymie, nous illustrons les phénomènes de variation observés.

Le tableau ci-dessous présente la répartition et la précision des candidats-marqueurs de méronymie selon les sous-relations auxquelles ils appartiennent et selon le domaine :

	VOLCANOLOGIE		CANCER	
Sous-relation	Nb Occ	Précision	Nb Occ	Précision
1- Réunion compacte	1	100%	na	na
2- Fusion	10	90%	1	**100%**
3- Décomposition	52	90%	6	50%
4- Non-organisation	18	83%	22	**100%**
5- Inclusion	187	80%	123	60%
6- Rapprochement	4	75%	11	0%
7- Organisation	20	75%	2	50%
8- Expression du Lieu / Temps	126	71%	52	65%
9- Type de parties	40	70%	10	60%
10- Parties de genre différent	14	64%	na	na
11- Jonction	21	57%	9	**67%**
12- Parties de même genre	22	41%	27	**70%**
Total	515	75%	263	63%

TABLE 3. Nombre d'occurrences et précision des candidats-marqueurs par sous-relation selon le domaine.

D'une manière générale les candidats-marqueurs de la relation de méronymie ont un taux de précision plus élevé dans le domaine de la volcanologie (75%) que dans le domaine du cancer du sein (63%). Deux cas sont possibles quant à l'influence du domaine sur la précision selon les sous-relations : soit le taux de précision des candidats-marqueurs est supérieur dans le domaine de la volcanologie, soit il est supérieur dans le domaine du cancer du sein :

- Les candidats-marqueurs des sous-relations *Parties de même genre*, *Jonction*, *Fusion*, et *Non-Organisation* (numérotées 2, 4, 11, 12) ont des taux de précision plus élevés dans le domaine du cancer du sein
- Les candidats-marqueurs des sous-relations *Parties de genre différent, Rapprochement*, *Réunion compacte*, *Inclusion, Organisation, Décomposition, Type de parties, Expression du lieu/temps* ont des taux de précision plus élevés dans le domaine de la volcanologie.

La tendance générale se vérifie donc au niveau des sous-relations puisque 8 sur 12 d'entre elles ont des candidats-marqueurs dont les taux de précision plus élevés dans le domaine de la volcanologie. Dans la suite, nous illustrons certains des phénomènes de variation observés, en établissant un lien avec les thématiques traitées dans les corpus d'étude.

4.2 Analyse qualitative

Nous présentons dans la suite la variation d'un point de vue linguistique, en nous penchant plus particulièrement sur les candidats-marqueurs ayant un taux de précision plus élevé dans le domaine de la volcanologie. Les thèmes abordés dans ce corpus concernent les types de volcanisme, mais également les roches, les laves, le magma. La composition, mais également la décomposition des éléments chimiques ou des roches lors d'un phénomène volcanique sont souvent étudiées. Le volcan

est souvent décrit comme un *édifice* ayant une *structure*, au sein de laquelle sont organisées différentes parties. Cela peut ainsi expliquer la précision plus élevée des sous-relations énoncées plus haut dans le domaine de la volcanologie. Par exemple, les candidats-marqueurs exprimant la *Décomposition* permettent de détecter des contextes comme le (1) ci-dessous :

(1) Le magma basique, en contact avec le magma acide plus froid, ***se fragmente en*** *micropillows (ou pillows) et en flammes. (volcanologie, scientifique)*

Le contexte (1) illustre l'expression de la décomposition d'un tout, *le magma basique*, en parties, les *micropillows* et les f*lammes*. Le marqueur testé ici est **[DET X se fragmenter en DET Y]**. Le terme *fragmentation* est présent parmi les candidats-termes du domaine de la volcanologie, mais pas parmi ceux du cancer du sein. Cela est également une indication sur la présence de tels marqueurs dans le corpus de volcanologie. Il est intéressant de noter ici que certains éléments lexicaux présents dans les marqueurs sont des candidats-termes du domaine. Autrement dit, l'un des éléments du candidat-marqueur, un verbe ou une nominalisation, est à la fois candidat-terme du domaine et candidat-marqueur. Dans ce cas, les marqueurs sont moins soumis à la polysémie, ce qui permet d'interpréter la relation attendue.

5 Conclusion et perspectives

La construction de ressources terminologiques ou ontologiques est l'un des champs explorés par le TAL ou l'ingénierie des connaissances. L'intérêt d'utiliser des marqueurs de relation pour détecter de manière automatique ou semi-automatique des Contextes Riches en Connaissances a de nombreuses fois été souligné. Notre contribution est d'avoir constitué la base MAR-REL[2], qui rassemble les principaux candidats-marqueurs français des relations d'hyperonymie, de méronymie et de cause. Chacun des candidats-marqueurs a été évalué en corpus variant selon deux domaines (volcanologie et cancer du sein) et deux genres textuels (scientifique vs vulgarisé).

La caractérisation du fonctionnement linguistique des marqueurs, en termes de stabilité ou de variation, permet d'émettre l'hypothèse qu'il est possible de réutiliser certains résultats pour des corpus ayant des caractéristiques proches de ceux de notre étude. Ainsi par exemple, le marqueur de méronymie [DET X se fragmenter en DET Y] pourrait donner des résultats intéressants dans des corpus liés à l'urbanisme ou dont les thématiques sont liées à la description de structures.

Compte-tenu de la taille de nos corpus, les résultats obtenus nécessitent cependant d'être nuancés. Un travail sur des corpus variés pourrait permettre de mener une autre étude contrastive, qui fournirait une description linguistique encore plus étendue des marqueurs de relation. Enfin, dans l'objectif de donner une perspective plurilingue à cette base, le même travail est en cours pour l'anglais et pour l'italien. Une étude des marqueurs de relation en espagnol est également envisagée.

[2] La base est constituée de deux parties : les marqueurs décrits de manière lexico-syntaxique, et leur traduction en langage UIMA (effectuée par Lingua&Machina). Nous indiquerons l'adresse à laquelle l'ensemble de la base MAR-REL sera disponible lors de la Conférence si notre communication est acceptée.

Références

AUGER A., BARRIÈRE C. (2008). Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology* 14, 1–19.

AUSSENAC-GILLES N., SÉGUÉLA P., (2000). Les relations sémantiques : du linguistique au formel. *Cahiers de Grammaire, « Sémantique et Corpus »* 25, 175–198.

BARRIERE C. (2001). Investigating the causal relation in informative texts, *Terminology* 7, 135–154.

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

CONDAMINES A. (2006). Avec et l'expression de la méronymie: l'importance du genre textuel. in G. Kleiber, C. Schnedecker et A. Thyssen (Eds) : *La relation «Partie - Tout».* Leuven : Peeters, 633–650.

CONDAMINES A. (2008). Taking genre into account when analysing conceptual relation patterns. *Corpora* 8, 115–140.

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*, Eyrolles, 225-241

FLOWERDEW J. (1992). Definitions in Science Lectures. *Applied Linguistics* 13, 203–221.

GARCIA D. (1998). *Analyse automatique des textes pour l'organisation causale des actions : réalisation du système informatique COATIS.* Thèse de Doctorat en Informatique. Université Paris 4.

GIRJU R., BADULESCU A., MOLDOVAN D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*-Volume 1, Association for Computational Linguistics,1–8.

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

Lafourcade M., Ramadier, L. (2016) Semantic Relation Extraction with Semantic Patterns Experiment on Radiology Reports. 10th edition of the Language Resources and Evaluation Conference (LREC 2016), 23-28 May 2016, Portorož (Slovenia), 6 p.

MALAISÉ V., ZWEIGENBAUM P., BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires en corpus pour l'aide à la construction d'ontologie. Actes de *TALN'04*, *Conférence sur le Traitement Automatique des Langues Naturelles*, 269–278.

MARSHMAN E. (2008). Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts. *Terminology* 14, 124–151.

MARSHMAN E., L'HOMME M.-C. (2006). Portabilité de la relation causale : étude sur deux corpus spécialisés, *Corpus et dictionnaires de langues de spécialité* : *Actes des Journées du CRTT*, 87-110.

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

MARSHMAN E., GARIÉPY JL., HARMS C (2012). Helping language professionals relate to terms: Terminological relations and termbases. *The Journal of Specialised Translation* 18, 30-56.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. in Bourigault, D., Jacquemin, C., L'Homme, M.-C. (Eds) : *Recent Advances in Computational Terminology*, Amsterdam/Philadephia : John Benjamins, 279–302.

MORIN E. (1999). *Extraction de liens sémantiques entre termes à partir de corpus de textes techniques*. Thèse de doctorat en Informatique, Université de Nantes.

NUOPPONEN A. (2014). Tangled Web of Concept Relations. Concept relations for ISO 1087-1 and ISO 704. *Terminology and Knowledge Engineering 2014*, 1-10.

PEARSON J. (1998). *Terms in context*. Amsterdam/Philadephia : John Benjamins.

PEARSON J. (2000). Une tentative d'exploitation bi-directionnelle d'un corpus bilingue. *Cahiers de Grammaire, « Sémantique et Corpus »* 25, 53–69.

REBEYROLLE J. (2000). *Forme et fonction de la définition en discours*. Thèse de doctorat en Sciences du Langage, Université Toulouse 2.