Extraction de Contextes Riches en Connaissances en corpus spécialisés

Firas Hmida Emmanuel Morin Béatrice Daille
Université de Nantes, LINA UMR 6241, 2 rue de la Houssinière, BP 92208, 44322 Nantes Cedex 3
{firas.hmida, emmanuel.morin, beatrice.daille}@univ-nantes.fr

Résumé. Les banques terminologiques et les dictionnaires sont des ressources précieuses qui facilitent l'accès aux connaissances des domaines spécialisés. Ces ressources sont souvent assez pauvres et ne proposent pas toujours pour un terme à illustrer des exemples permettant d'appréhender le sens et l'usage de ce terme. Dans ce contexte, nous proposons de mettre en œuvre la notion de Contextes Riches en Connaissances (CRC) pour extraire directement de corpus spécialisés des exemples de contextes illustrant son usage. Nous définissons un cadre unifié pour exploiter tout à la fois des patrons de connaissances et des collocations avec une qualité acceptable pour une révision humaine.

Abstract.

Knowledge-Rich Contexts Extraction in Specialized Corpora
The term banks and dictionaries are valuable resources that improve access to knowledge in specialized domains. These resources are often relatively poor and do not always provide, for a given term, examples of its typicall use. In this context, we implement Knowledge-Rich Contexts (KRCs) to extract examples of contexts providing illustration of terms in specialized domain. We propose a unified framework to apply at the same time knowledge pattern and collocations with acceptable quality for human review.

Mots-clés : corpus spécialisé, CRC, patrons de connaisssances, collocations.

Keywords: specialized corpus, KRC, knowledge patterns, collocations.

1 Introduction

Les banques terminologiques et les dictionnaires sont des ressources linguistiques précieuses qui facilitent l'accès aux connaissances des domaines spécialisés. Ces ressources proposent généralement pour un terme à illustrer une définition, des exemples d'utilisation et d'autres termes en relation. Habituellement, le terme à illustrer est considéré comme « terme favori » et les autres termes reflètent des relations paradigmatiques comme la synonymie ou l'hyperonymie. La définition associée au terme favori est souvent de nature encyclopédique et les quelques exemples de contextes proposés, lorsqu'ils existent, ne couvrent qu'une partie des usages de ce terme favori (Bowker, 2011). Des travaux récents laissent entendre que les banques terminologiques actuelles n'ont pas connu d'améliorations significatives depuis les années 60 (Bowker, 2011). Elles contiennent trop peu d'informations contextuelles et les connaissances sur l'usage du terme sont assez limitées (p. ex. des exemples de collocations peuvent être indiqués mais cela n'est pas systématique). En outre, les définitions fournies par les dictionnaires comme les banques terminologiques sont généralement insuffisantes pour permettre la compréhension du terme. Les corpus spécialisés représentent un réservoir important d'informations contextuelles pour analyser le fonctionnement des termes. Cependant, tous les contextes dans lesquels les termes apparaissent ne sont pas utiles à leur compréhension. Dans ce contexte, Meyer (2001) a introduit la notion de Contextes Riches en Connaissances (CRC) désignant les contextes qui jouent un rôle prépondérant dans la compréhension des termes et renseignent sur leur fonctionnement linguistique.

Dans ce travail, nous postulons que les corpus monolingues spécialisés contiennent des contextes conceptuels et linguistiques qui peuvent être extraits automatiquement. Nous nous limitons aux corpus étudiés sans solliciter de ressources externes et mettons en œuvre deux méthodes pour identifier des CRC. La première méthode s'appuie sur la présence du terme à illustrer et l'exploitation des patrons lexicaux afin d'extraire des contextes riches en connaissances conceptuelles. Ces CRC permettent l'accès à la dimension conceptuelle du terme. Les patrons lexicaux exploités ont pour but notamment d'accéder en corpus spécialisé à la définition du terme. La seconde méthode exploite des mesures d'association pour identifier des contextes riches en connaissances linguistiques aidant ainsi à comprendre l'usage du terme. Elle repose sur

le repérage en corpus des collocations. Nous focalisons notre travail sur deux corpus spécialisés de discours scientifiques relevant du domaine de la vulcanologie en français et en anglais.

2 État de l'art

De nombreuses recherches ont été menées sur les CRC dans différentes perspectives. Dans cette section, nous présentons succinctement les principaux travaux exploitables dans une perspective d'aide à la compréhension. Nous présentons tout d'abord la notion de CRC, ensuite les méthodes permettant leur extraction.

Meyer (2001) propose la notion de CRC pour désigner les contextes qui illustrent des relations entre les termes d'un domaine spécialisé. Ces relations sont souvent représentées par des unités lexico-syntaxiques appelées « *patrons de connaissances* » (PC). La phrase « *L'Olympus ci-contre est le volcan géant du système solaire* » est définie comme un CRC pour le terme *Olympus*. Dans cet exemple, le patron de connaissances *est le* explicite une relation hiérarchique entre les termes *Olympus* et *volcan*. Schumann (2012) a entrepris d'extraire des CRC à partir du Web dans le but d'enrichir une banque terminologique en langue russe. Les contextes ont tout d'abord été repérés au moyen de PC, puis ordonnés grâce à une méthode supervisée. Ce travail est similaire au nôtre. Néanmoins, dans le cadre de notre problématique, nous étudions l'identification des CRC dans un corpus spécialisé de taille modeste. Marshman (2014) a étudié la nécessité d'utiliser des ressources terminologiques mettant en évidence des CRC extraits par des PC, telles que CREATerminal [1]. Ces recherches ont également montré l'utilité des ressources enrichies par des CRC, particulièrement pour des traducteurs étudiants. Une des difficultés majeures dans le domaine des PC tient au fait qu'il n'existe aucune bibliothèque de PC qui manifesterait l'aspect cumulatif de ces travaux. Pour chaque nouvelle étude, il faut refaire une synthèse des études existantes pour établir des listes de PC. D'une part, les travaux concernant la variation dans le fonctionnement des PC sont encore récents. En effet, selon Marshman *et al.* (2008), bien que l'intérêt des PC pour repérer les CRC est indéniable, leur identification est coûteuse et l'on doit chercher à les réutiliser pour d'autres études, dans d'autres types de corpus. D'autre part, il est primordial de chercher à mesurer leur portabilité, c'est-à-dire leur degré de variabilité d'un corpus à l'autre, et donc d'un domaine à l'autre.

La notion de CRC fait écho à d'autres types de contextes tels que les définitions étudiées par Saggion (2004) et les collocations extraites par Kilgarriff *et al.* (2008). Saggion (2004) a eu recours à deux stratégies afin de repérer des définitions à partir de textes disponibles sur le Web. Il a utilisé des PC modélisant des relations de définitions, ainsi que des « termes secondaires » fournissant des connaissances spécifiques au terme en question. Saggion (2004) introduit ces termes secondaires comme les termes qui co-occurrent significativement avec le terme à illustrer dans des définitions. Un terme secondaire peut être un nom, un adjectif ou un verbe. Cette notion fait référence à celle de collocation identifiée dans un corpus de définitions. La maîtrise des collocations est une composante essentielle de la maîtrise de la langue ou d'un discours spécifique. Ceci explique l'importance accordée à cette notion pour l'aide à la compréhension. Au sens restreint, les collocations représentent des associations lexicales transparentes du point de vue de la compréhension mais qu'un locuteur non natif doit tout particulièrement apprendre à maîtriser. C'est le cas des exemples suivants : *prescrire une ordonnance*, *tenir debout*, *nuit blanche*. La notion de collocation reçoit des définitions variables selon le contexte de recherche dans lequel elle est employée. Sinclair *et al.* (1970), par exemple, définissent la collocation par la co-occurrence significative de deux items dans un contexte spécifié. En s'appuyant sur les collocations, Kilgarriff *et al.* (2008) propose GDEX (Good Dictionary Examples), un outil permettant de produire automatiquement des exemples pour des lexicographes. Cet outil a pour but de sélectionner l'exemple lexicographique le plus pertinent à partir d'un corpus de données massives. De ce point de vue, GDEX peut être considéré comme un système de filtres permettant de retenir les « bons » exemples respectant un ensemble de critères. Son fonctionnement consiste à identifier les collocations du terme en question pour associer ensuite des exemples à chaque collocation. Kilgarriff *et al.* (2008) se sont basés sur les travaux d'Atkins & Rundell (2008) pour qualifier un « bon » exemple en s'appuyant sur différents critères tels que i) la lisibilité, c'est-à-dire intelligible aux lecteurs aussi bien lexicalement que structurellement et ii) l'informativité, qui illustre des contextes typique (par exemple une collocation) et qui aide à comprendre le terme à exemplifier. Ces critères ont été mis en œuvre comme des traits positifs tels que la présence de troisième collocatif, et négatifs comme la présence de mots rares dans le contexte. Les exemples de Kilgarriff sont incontestablement des contextes riches illustrant l'usage du terme. Bien que notre objectif soit similaire à celui de Kilgarriff, nous cherchons les contextes qui fournissent des connaissances spécifiques au domaine étudié. Ces contextes permettront également au lecteur de positionner le terme par rapport à la terminologie du domaine. Ces deux types de connaissances linguistiques (collocations) et conceptuelles (PC) sont nécessaires.

1. Interface fournissant des contextes aidant à la traduction terminologique (anglais-français) dans le domaine du cancer du sein.

3 Contribution

Dans cette section nous étudions les PC et les collocations dans une perspective d'aide à la compréhension d'un terme donné. Pour ce faire, nous mettons en œuvre deux méthodes permettant d'exploiter ces notions pour identifier des CRC à partir de corpus monolingues spécialisés.

3.1 Patrons de Connaissances pour CRC

Dans la littérature, trois relations considérées comme universelles ont été majoritairement étudiées. Il s'agit des relations d'hyperonymie, de méronymie et de cause. La relation d'hyperonymie est connue comme la plus structurante, du fait de son exploitation dans les définitions et de sa propriété de transitivité. Dans ce travail, nous mettons en œuvre l'automatisation du repérage et la portabilité des CRC à l'aide de PC d'hyperonymie. Nous exprimons ces PC sous la forme d'un triplet $(terme_1, PC\,d'hyperonymie, terme_2)$ avec $terme_1$ et $terme_2$ deux termes distincts du corpus étudié. La démarche suivie consiste à utiliser tout d'abord un outil d'extraction terminologique pour identifier les termes du corpus ; ensuite à retenir les contextes phrastiques contenant le patron $(terme\,à\,illustrer, PC\,d'hyperonymie, terme\,du\,domaine)$. Ces contextes potentiellement riches en connaissances sont considérés comme des CRC candidats. La table 1 illustre des CRC candidats associés aux termes *cendre* et *volcan*. Ces CRC candidats sont repérés grâce aux PC présentés dans la même table.

Terme à illustrer (X)	Terme du domaine (Y)	PC	CRC candidat
Cendre	produit volcanique	X_ÊTRE_LE_PRINCIPAL_Y	*Les **cendres** sont les principaux **produits volcaniques** émis par les volcans explosifs de la ceinture de feu du Pacifique.*
Volcan	Actif	X_ÊTRE_Y_LE_PLUS	*Le groupe Klyvcheskoy, dont les **volcans** sont les plus **actifs** de l'arc des Kouriles...*

TABLE 1 – Exemples de CRC candidats identifiés par des PC pour le français

3.2 Collocations pour CRC

Plusieurs mesures d'association ont été appliquées pour extraire automatiquement des collocations. Si l'Information Mutuelle spécifique (Fano, 1961) permet d'identifier des unités lexicales qui apparaissent plus souvent ensembles que séparemment, le Z-score (Berry-Rogghe, 1973) est souvent privilégié pour déterminer les collocatifs candidats d'un terme donné. Dans ce travail, nous associons à une liste de termes donnés leurs meilleurs collocatifs en nous appuyant sur la mesure Z-score puisque nous connaissons *a priori* les termes que nous souhaitons illustrer. Ces collocations serviront, par la suite, à sélectionner des CRC candidats.

Les mesures d'association peuvent également être combinées à des analyses linguistiques telles que l'analyse syntaxique (Fellbaum, 1998). Ces analyses, jouant un rôle de filtre, permettent d'affiner la qualité des collocations obtenues et de les classer selon leurs catégories grammaticales. Evert & Krenn (2005) montrent qu'il faut distinguer les catégories syntaxiques des collocations avant d'appliquer une mesure d'association. Nous retenons alors deux catégories de collocations nominales dans lesquelles la base est un terme à illustrer : $(terme, nom)$ et $(terme, adjectif)$. Ces collocations ont été identifiées dans Josselin-Leray *et al.* (2014) comme pertinentes dans un exercice d'aide à la compréhension.

Après avoir filtré les mots outils dans le corpus, nous avons repéré les collocations constituées de deux mots pleins dans une fenêtre bigramme : un mot avant ou un mot après la base (sans compter les mots vides) en respectant les catégories syntaxiques étudiées. Afin d'extraire les CRC candidats nous avons suivi les deux étapes suivantes :

1. Identifier pour un terme à illustrer ses collocatifs en fonction de sa catégorie syntaxique et les ordonner selon le Z-score ;
2. Parcourir les collocatifs de chaque terme à illustrer et retenir le premier (selon le Z-score) qui procure au moins un contexte phrastique lisible, tel que défini par Kilgarriff *et al.* (2008). Ici, nous nous sommes limités à un seul collocatif en vue d'obtenir un nombre acceptable de CRC candidats pour une évaluation humaine.

Les CRC candidats de la table 2 sont identifiés par des collocations dans lesquelles la base est le terme à illustrer.

Terme à illustrer	Collocatif	CRC candidat
Gaz	carbonique	*Ce **gaz carbonique** qui, transformé par les plantes, a donné de l'oxygène, indispensable à la vie.*
Gas	dissolved	***Gas dissolved** in the molten rock expanded and literally blew the volcano apart...*
Cendre	retombée	*Les explosions phréatiques se font plus violentes qu'en 1792, et deux ou trois d'entre elles provoquent des **retombées de cendres** sur les villes du prêcheur.*
Cendre	retombée	*Veaucoup d'habitants du prêcheur et de ses environs viennent se réfugier à Saint-Pierre, épargnée par les **retombées de cendres**.*

TABLE 2 – Exemples de CRC candidats identifiés par des collocations

4 Expériences et résultats

Dans cette section nous décrivons les différentes ressources mobilisées pour nos expériences et présentons ensuite les résultats obtenus en appliquant les deux méthodes précédentes.

4.1 Ressources linguistiques

Les expériences ont été réalisées sur deux corpus français et anglais relevant du domaine de la vulcanologie. Il s'agit d'un corpus comparable constitué par Amélie Josselin-Leray de CLLE-ERSS. Il est composé de documents scientifiques contenant environ 400 000 mots par langue, obtenus grâce à une recherche thématique à partir de journaux et magazines tels que *Le Monde*, *Sciences et avenir*, *Sciences et Vie*... L'ensemble des documents ont été nettoyés et normalisés à travers les traitements suivants réalisés par la plateforme TermSuite [2] : segmentation en occurrences de formes, étiquetage morphosyntaxique, lemmatisation et extraction terminologique. En ce qui concerne l'extraction terminologique, nous nous sommes limités aux termes simples et complexes qui apparaissent au moins 5 fois dans chaque corpus. Ces termes sont nécessaires à l'exploitation des PC. À ce niveau, nous nous appuyons sur une liste de PC relatifs à la relation d'hyperonymie : 33 en français (Rebeyrolle & Tanguy, 2000) (*cf.* table 3 pour un extrait) et 34 en anglais (Séguéla, 2001; Marshman *et al.*, 2012). Ces PC permettent d'extraire des CRC comportant éventuellement des connaissances définitoires qui illustrent les termes en question.

Patrons de connaissances (FR)
X_ÊTRE_UN_Y
X_ÊTRE_UNE_SORTE_DE_Y
X_ÊTRE_LE_Y_LE_PLUS
X_ÊTRE_AUTRES_Y
Y_ET_ADVERBE_DE_SPECIFICATION_X

TABLE 3 – Exemples de PC exploités pour le français (X est un terme et Y son hyperonyme)

Enfin, les termes que nous cherchons à illustrer avec nos deux approches et qui sont caractéristiques du domaine de la vulcanologie sont présentés dans la table 4. Ces termes ont été sélectionnés par des linguistes pour des expériences portant sur l'aide à la traduction.

4.2 Résultats des patrons de connaissances

La table 5 présente les résultats obtenus après projection des PC d'hyperonymie sur les corpus de vulcanologie en français et en anglais afin d'illustrer les termes de la table 4 rappelés en colonne *# Termes à illustrer*. Nous désignons par *# Termes*

2. https://logiciels.lina.univ-nantes.fr/redmine/projects/termsuite

Corpus	Termes à illustrer
Français	*basalte*, *cendre*, *cratère*, *cône*, *débris*, *dégazage*, *dôme*, *fontaine*, *gaz*, *jaillir*, *lave*, *magma*, *phase*, *roche*, *scorie*, *téphra*, *vacuole*, *volcan*, *vésicule*, *éruption*
Anglais	*basalt*, *blobs*, *cinder*, *cone*, *eruption*, *fountaining*, *gas*, *layers*, *scoria*, *softball*, *spongelike*, *vesicles*

TABLE 4 – Liste des termes à illustrer

extraits le nombre de termes intégrant des PC avec au moins un hyperonyme, et par *# CRC candidats* le nombre de CRC associés aux termes extraits. *# CRCC* indique le nombre de CRC Conceptuels. Ce sont des CRC candidats contenant le terme à illustrer ainsi que d'autres termes du domaine étudié. Des relations conceptuelles doivent être explicitées à travers un lien sémantique entre le terme visé et les autres termes présents dans le CRC candidat.

Le CRC candidat « *Les **cendres** sont les principaux **produits volcaniques** émis par les volcans explosifs de la ceinture de feu du Pacifique* » ($cf.$ table 1) explicite une relation d'hyperonymie définissant le terme *cendre*. Il s'agit alors d'un CRCC. En ce qui concerne le deuxième cas « *Le groupe Klyvcheskoy, dont les **volcans** sont les plus **actifs** de l'arc des Kouriles* », la relation d'hyperonymie est invalide. En effet, le terme *actif* n'est pas un hyperonyme de *volcan*, d'où nous considérons ce CRC candidat comme non intéressant. Même s'ils peuvent contenir d'autres relations sémantiques intéressantes, les CRC candidats ne sont retenus que lorsque la relation d'hyperonymie est valide.

La colonne *# Termes extraits* montre que les PC sont présents dans les deux corpus spécialisés même si uniquement 33 à 40 % des termes à illustrer sont retrouvés. Cependant, la qualité des CRC obtenus après validation est relativement bonne. Ces résultats sont finalement assez conformes à l'état de l'art (Morin, 1999), à savoir un faible rappel des patrons de connaissances au bénéfice de la précision.

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats	# CRCC (P.)
Français	20	8	21	17 (80,95 %)
Anglais	12	4	14	10 (71,42 %)

TABLE 5 – Résultats de la projection des PC après une validation manuelle

4.3 Résultats des collocations

En ce qui concerne l'évaluation de la méthode basée sur les collocations, chaque contexte identifié a été évalué par un annotateur natif. Nous désignons par un contexte intéressant, un contexte révélant des connaissances conceptuelles ou linguistiques et aidant à comprendre le terme à illustrer. Nous distinguons, en plus des CRCC, les contextes riches en connaissances linguistiques (CRCL). Un CRCL représente un contexte contenant seulement le terme à illustrer et son collocatif, à l'exception d'autres termes du domaine auxquels le terme à illustrer est conceptuellement lié. Un CRCL doit être grammaticalement bien formé.

Les tables 6 et 7 présentent les résultats obtenus pour les collocations respectivement de type $(terme, adjectif)$ et $(terme, nom)$. Dans ces tables, *# Termes extraits* est le nombre de termes ayant une collocation fournissant des contextes, *CRCC* et *CRCL* représentent quant à eux les pourcentages des connaissances respectivement conceptuelles et linguistiques illustrées par les contextes repérés grâce à des collocations. *Non CRC* est le pourcentage de contextes qui ne sont pas intéressants par rapport à la compréhension des termes visés. Nous focalisons notre analyse des résultats sur le type des connaissances illustrées ainsi que la cohérence entre les résultats des corpus étudiés. Nous pouvons néanmoins constater que la qualité des CRC obtenue par cette approche est en dessous de la seule utilisation des patrons de connaissances.

Dans le cas du corpus français de vulcanologie, les contextes identifiés par les collocations de type $(terme, adjectif)$ contiennent plus souvent des connaissances conceptuelles que linguistiques. Par exemple, la phrase *Ce **gaz carbonique** qui, transformé par les plantes, a donné de l'oxygène, indispensable à la vie* ($cf.$ table 2) contient une relation sémantique qui peut être traitée comme une relation de cause. Dans d'autres cas, les collocatifs de cette catégorie peuvent révéler des connaissances conceptuelles quand il s'agit de participe passé ou présent. En effet, ce type de collocatif, mettant en jeu un verbe conjugué, traduit éventuellement un lien sémantique entre le terme en question et d'autres termes du domaine. Nous parlons alors d'un contexte conceptuel. Dans le corpus anglais de vulcanologie, le contexte ***Gas dissolved** in the molten rock expanded and literally blew the volcano apart* ($cf.$ table 2), considéré comme CRCC, le terme *gas* est illustré par son collocatif *dissolved*. Ces résultats semblent être cohérents avec ceux des corpus anglais de vulcanologie dont les contextes ont été évalués par des linguistes.

Les collocations de type $(terme,\ nom)$ favorisent l'illustration des connaissances conceptuelles et linguistiques de façon mitigée dans les deux corpus étudiés : dans le corpus vulcanologie français 31,33 % des contextes sont conceptuels et 27,71 % sont linguistiques. En effet, ces collocations peuvent informer sur l'usage du terme, notamment en présence de prépositions comme dans le cas de *retombée de cendre* (base : *cendre* et collocatif : *retombée*). Les résultats sont mis en évidence dans les deux corpus vulcanologies.

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats	CRCC	CRCL	Non CRC
Français	20	18	74	45,95 %	12,16 %	41,81 %
Anglais	12	10	41	41,46 %	14,63 %	43,90 %

TABLE 6 – Évaluation manuelle des contextes extraits par la collocation de type $(terme\ adjectif)$

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats	CRCC	CRCL	Non CRC
Français	20	18	83	31,33 %	27,71 %	40,96 %
Anglais	12	10	43	41,86 %	30,23 %	27,91 %

TABLE 7 – Évaluation manuelle des contextes extraits par la collocation de type $(terme,\ nom)$

4.4 Synthèse

Si nous combinons maintenant les résultats obtenus par les deux précédentes approches, nous pouvons constater à la lecture de la table 8 que nous sommes en mesure de proposer des CRC pour l'ensemble des termes de la liste de référence. La qualité de ces CRC est autour de 70 %, ce qui reste acceptable. Il n'y a que deux termes anglais *spongelike* et *softball* pour lesquels nous ne proposons pas de CRC pertinents.

Il serait en outre intéressant de pouvoir ordonner ces derniers pour ne conserver que les plus intéressants. Dans un premier temps, il semble pertinent de proposer en premier lieu les CRC issus des patrons de connaissances puis ceux issus des collocations.

Corpus	# Termes à illustrer	# Termes extraits	# CRC candidats (sans doublons)	# CRC valides (P.)
Français	20	20	143	100 (69,93 %)
Anglais	12	10	97	67 (69,07 %)

TABLE 8 – Tableau récapitulatif de la combinaison des deux méthodes

5 Conclusion et perspectives

Dans ce travail, nous avons proposé de mettre en œuvre la notion de Contextes Riches en Connaissances pour extraire directement de corpus des exemples illustrant le fonctionnement des termes. Ces CRC, qui sont extraits de corpus en s'appuyant sur des patrons de connaissances et des collocations, permettent d'accéder tout à la fois aux connaissances linguistiques et conceptuelles. L'originalité de notre approche est de considérer l'ensemble des CRC disponibles à la différences des travaux existants qui se restreignent soit à des patrons de connaissances pour extraire des définitions (Marshman, 2014) soit à des collocations pour extraire des exemples (Kilgarriff *et al.*, 2008). La complémentarité des deux approches mises en œuvre permet d'obtenir des CRC variés avec une qualité acceptable pour une révision humaine. Néanmoins, il serait intéressant de pouvoir réduire le nombre proposé de CRC en cherchant à proposer systématiquement pour un terme à illustrer un exemple de CRC linguistique et un autre de CRC conceptuel. Pour ce faire, il sera nécessaire de pouvoir associer à chacun de ces CRC un score de confiance qui pourrait être fonction du patron de connaissances déclenché dans un cas et de l'ordonnancement global des collocations dans l'autre cas.

Remerciements

Ce travail qui s'inscrit dans le cadre du projet CRISTAL `www.projet-cristal.org` a bénéficié d'une aide de l'Agence National de la Recherche portant la référence ANR-12-CORD-0020.

Références

ATKINS B. S. & RUNDELL M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

BERRY-ROGGHE G. (1973). The computation of collocations and their relevance in lexical studies. *The Computer and Literary Studies*, p. 103–112.

BOWKER L. (2011). Off the record and on the fly : Examining the impact of corpora on terminographic practice in the context of translation. *Corpus-based Translation Studies : Research and Applications. London/New York : Continuum*, p. 211–236.

EVERT S. & KRENN B. (2005). Using small random samples for the manual evaluation of statistical association measures. *Computer Speech & Language*, **19**(4), 450–466.

FANO R. M. (1961). *Transmission of Information : A Statistical Theory of Communication*. MIT Press.

FELLBAUM C. (1998). *WordNet : An electronic lexical database*. MIT Press.

JOSSELIN-LERAY A., FABRE C., REBEYROLLE J., PICTON A. & PLANAS E. (2014). Good Contexts for Translators - A First Account of the Cristal Project. In *Proceedings of the XVI EURALEX International Congress*, p. 631–645, Bolzano, Italy.

KILGARRIFF A., RYCHLÝ P., HUSÁK M., RUNDELL M. & MCADAM K. (2008). GDEX : Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX International Congress*, p. 425–432, Barcelona.

MARSHMAN E. (2014). Enriching terminology resources with knowledge-rich contexts : A case study. *Terminology*, **20**(2), 225–249.

MARSHMAN E., GARIÉPY J. L. & HARMS C. (2012). Helping language professionals relate to terms : Terminological relations and termbases. *JoSTrans*, **18**.

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**(2), 141–172.

MEYER I. (2001). Extracting knowledge-rich contexts for terminography - A conceptual and methodological framework. In B. DIDIER, J. CHRISTIAN & M.-C. L'HOMME, Eds., *Recent Advances in Computational Terminology*, p. 279–302.

MORIN E. (1999). Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues (TAL)*, **40**(1), 143–166.

REBEYROLLE J. & TANGUY L. (2000). Repérage automatique de structures linguistiques en corpus : le cas des énoncés définitoires. *Cahiers de grammaire*, **25**, 153–174.

SAGGION H. (2004). Identifying Definitions in Text Collections for Question Answering. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, p. 1927–1930.

SCHUMANN A.-K. (2012). Towards the Automated Enrichment of Multilingual Terminology Databases with Knowledge-Rich Contexts–Experiments with Russian EuroTermBank Data. In *Proceedings of the 2nd Workshop on the Creation, Harmonization and Application of Terminology Resources (CHAT'12)*, p. 27–34.

SÉGUÉLA P. (2001). Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques. *Thèse en Informatique, Université Toulouse 3*.

SINCLAIR J. M., JONES S. & DALEY R. (1970). *English Lexical Studies. Final Report of O.S.T.I. Programme C/LP/08.*