Academic Journal of Modern Philology

ISSN 2299–7164
Vol. 13 (2021)
Special Issue
s. 75–82

Silvia Calvi[1]
Université de Vérone, Italie

Collocations terminologiques et extraction automatique : une étude pilote dans le domaine du commerce électronique

Terminological Collocations and Automatic Extraction: A Pilot Study on E-Commerce

Abstract

This article concentrates on the automatic extraction of collocations, defined by the *Explanatory and Combinatorial Lexicology* as phraseological units composed by two elements – the base and the collocate. The aim of this article is to propose a methodology to follow in order to automatically extract collocations from a terminological corpus. This method takes into account different measures: the syntactic dependences between the items of the collocation, their frequency, their tendency to co-occur (PMI) and their specificity to the e-commerce domain. After having explained the theoretical framework, the methodology is illustrated using a pilot study of the French terminology of e-commerce. In the pilot study, data were extracted from a corpus made up of e-commerce texts, which are drawn from a larger corpus called DIACOM-fr, a corpus in the process of being built at the University of Verona within the project *Digital Humanities Applied to Foreign Languages and Literatures*. Data extraction was primarily done using two tools: *Stanza* a Python natural language analysis package developed by the Stanford NLP group and *TermoStat* an automatic extractor tool developed at the Observatoire de Linguistique Sens-Texte of the University of Montreal.

Keywords: collocations, terminology, automatic extraction, e-commerce

doi:10.34616/ajmp.2021.13

1. Introduction

Le Département de Langues et Littératures étrangères de l'Université de Vérone a reçu en janvier 2018 un financement de la part du Ministère italien de l'Éducation, de l'Université et de la Recherche pour le

1 Cette étude s'insère dans mon projet de thèse sur les collocations dans la terminologie française du commerce international (directeur de thèse : Paolo Frassi, Université de Vérone; co-directeur de thèse : Patrick Drouin, Université de Montréal).

projet *Le Digital Humanities applicate alle lingue e letterature straniere* (*Les humanités numériques appliquées aux langues et littératures étrangères*). Dans le cadre de ce projet, l'équipe de recherche DIACOM[2], dont je fais partie, est en train de réaliser deux corpus en français et en espagnol dans le domaine du commerce international. Ils seront exploités pour l'extraction de termes simples et complexes pour l'alimentation d'une base de données terminologique de type *réseau lexical* (Polguère 2014 : 396–418). Aux fins de cette étude, nous analyserons uniquement la section française du projet : DIACOM-fr. Les termes, c'est-à-dire des unités lexicales caractérisant un domaine de spécialité (L'Homme 2004 : 52–82), ne sont pas le seul objet d'étude de la terminologie qui s'intéresse aussi à analyser les liens syntagmatiques et paradigmatiques que les termes nouent entre eux. Dans cet article, nous approfondirons la question des liens syntagmatiques entre les termes, plus particulièrement la question de l'extraction automatique de collocations terminologiques, qui pose des problèmes d'ordre pratique. Le fait qu'elles puissent se présenter de manière disjointe rend plus difficile leur extraction par rapport aux termes simples et aux autres unités phraséologiques, comme les locutions.

Après avoir donné une définition de *collocation* en langue générale et en langue de spécialité, nous présenterons une étude pilote menée à partir d'un sous-corpus du corpus DIACOM-fr dans le domaine du commerce électronique. Cette étude vise à proposer une méthode d'extraction semi-automatique qui permettra de repérer un échantillon représentatif des collocations du domaine d'étude, tout en mettant en relief des lacunes des extracteurs terminologiques en ce qui concerne l'extraction automatique de collocations.

2. Modèle théorique

2.1. La collocation en langue générale

Nous avons adopté la définition de *collocation* proposée par la *Lexicologie Explicative et Combinatoire* (LEC) (Mel'čuk, Clas, Polguère 1995 ; Mel'čuk 1993, 2008, 2013). En tant que phrasème, la collocation est une unité multilexémique non libre contrainte sur l'axe paradigmatique : il est impossible de remplacer ses composantes par des unités lexicales synonymiques. Sur l'axe syntagmatique la collocation se distingue d'une autre unité phraséologique, la locution : les locutions ne sont pas libres sur l'axe syntagmatique, tandis que les collocations se caractérisent par un certain degré de liberté, qui se manifeste, entre autres, par la possibilité pour le locuteur d'ajouter du matériau linguistique dans cette unité phraséologique (p.ex. *atteindre les clients*, *atteindre beaucoup de clients*). Cela dépend du fait que ce phrasème est composé de deux éléments distincts. Pour sa part, les composantes de la locution perdent leur autonomie, elles représentent une seule lexie. Les deux éléments de la collocation – la base et le collocatif – ont un statut sémantique différent : la base est autonome et choisie librement par le locuteur, alors que le collocatif est choisi en fonction de la base et du sens que le locuteur veut exprimer. Par exemple, dans la collocation *gravement malade* le locuteur choisit librement la base *malade* et s'il désire exprimer l'intensité le collocatif *gravement* s'imposera. La *LEC* distingue deux types de collocations : les collocations standards

2 Pour plus d'informations sur le projet DIACOM nous renvoyons le lecteur au site : https://dh.dlls.univr.it/it/progetti/patrimonio-linguistico-culturale/#diacom [consulté le 20/07/2021].

et les collocations non standards, formalisées respectivement par des fonctions lexicales standards et non standards. C'est la notion de *fonction lexicale* qui permet de formaliser et de classer les collocations. La définition de *fonction lexicale* s'appuie sur celle de *fonction mathématique*, *f(x)=y*, les deux variables *x* et *y* correspondent à deux lexies et la fonction *f* indique le lien syntagmatique ou paradigmatique qui les relie. Dans l'exemple de collocation *gravement malade* que nous venons de présenter, la fonction en question est la fonction syntagmatique qui indique l'intensité, c'est-à-dire Magn, et les deux lexies sont *malade* et *gravement*. Nous pouvons donc la représenter de façon formelle comme suit : Magn (malade) = gravement (Mel'čuk, Clas, Polguère 1995 ; Wanner 1996).

Les fonctions lexicales standards représentent un lien récurrent dans la plupart des langues du monde ; dans notre exemple, le sens d'intensité est en effet récurrent dans plusieurs langues : *gravement malade* (fr.), *gravemente malato* (it.). Cependant, pour certaines collocations le lien sémantique entre la base et le collocatif n'est pas aussi systématique (p.ex. *année bissextile*) : ces collocations sont formalisées à travers des fonctions lexicales non standards (p.ex. année ayant 366 jours : *année bissextile*). Comme Mel'čuk, Clas et Polguère (1995) l'ont affirmé, la frontière entre les collocations non-standards et les locutions au sens plus transparent est très fine. Notre équipe de recherche vient de démontrer (Frassi *et al.* 2020) que dans les langues de spécialité, les collocations non-standards « pour plusieurs raisons liées à l'application des restrictions syntactico-sémantiques particulières, […] s'apparentent davantage aux locutions faibles qu'aux collocations » (Frassi *et al.* 2020 : 331). Même si les collocations non-standards/locutions faibles sont très répandues dans les langues de spécialité, nous ne les approfondirons pas dans la présente étude, puisque nous nous intéressons ici uniquement aux collocations standards pour lesquelles la *LEC* a déjà formalisé des fonctions lexicales pouvant exprimer le sens qui relie les bases aux collocatifs.

2.2. La collocation en langue de spécialité

Les textes des langues de spécialité contiennent autant de collocations que les textes de langue générale. En effet, à côté des termes simples (p.ex. *site*) et dérivés (p.ex. *dépersonnalisation*), on retrouve un pourcentage élevé de termes complexes (p.ex. *bouche-à-oreille*, *commerce électronique*) qui correspondent bien souvent aux phrasèmes que nous venons de décrire, à savoir les locutions (p.ex. *marge arrière*) et les collocations (p.ex. *client fidèle*). Dans l'exemple *client fidèle*, la distinction entre les deux éléments qui composent la collocation est évidente : *client*, terme et base de la collocation, s'associe à une autre unité lexicale, le collocatif *fidèle*, pour exprimer le sens de bonté. En adaptant la définition de *collocation* de la LEC aux langues de spécialité, nous définissons dans la présente étude une collocation comme un groupe de deux ou plusieurs lexies dont au moins une lexie est un terme du domaine d'étude. De plus, la collocation doit pouvoir être formalisée en ayant recours à une fonction lexicale standard.

3. Le corpus DIACOM-fr et l'étude pilote sur le commerce électronique

Le corpus du commerce international DIACOM-fr en voie de constitution à l'Université de Vérone se compose à l'heure actuelle de 583 textes en langue française (environ dix millions de *word tokens*). Nous avons adopté trois critères pour la sélection des textes : 1) un critère chronologique, 2) un critère thématique et 3) un critère textuel que nous présentons dans le tableau 1.

Tableau 1- Les critères de constitution du corpus DIACOM-fr

Critère chronologique	Critère thématique	Critère textuel
Trois périodes clés de l'histoire du commerce international : 1) 1850–1914 (la deuxième révolution industrielle) ; 2) 1945–1970 (le boom économique) ; 3) 1985–2020 (le développement du marketing et la naissance du commerce électronique).	Trois sous-domaines : 1) Macroéconomie et économie internationale (Politique commerciale et relations internationales, pays d'étude, aspects sociaux) ; 2) Secteur (Produit, secteur) ; 3) Type d'activité dans les entreprises (management, marketing, commerce électronique, logistique, droit).	Quatre types de textes : 1) Textes scientifiques ; 2) Textes institutionnels ; 3) Articles de presse spécialisée ; 4) Documentation d'entreprise.

L'étude pilote que nous présentons ici ne prend en compte qu'une seule partie de ce corpus : les textes allant de 1985 à 2020 qui appartiennent au sous-domaine *type d'activité dans les entreprises*. Ce sous-corpus contient notamment des textes portant sur le commerce électronique. Il s'agit de 32 documents (961 611 *word tokens*) de nature différente : textes scientifiques et académiques (40,62%), textes institutionnels (34,38%) et articles de la presse spécialisée (25%).

4. Essais d'extraction automatique

Avant de soumettre notre corpus aux extracteurs automatiques, nous avons converti les *.pdf* en *.txt*. Pour réduire le bruit dans les résultats, nous avons décidé de supprimer de manière manuelle toutes les parties qui n'avaient aucun intérêt linguistique, comme les frontispices, les index, les titres, les notes en pied de page, les grilles, les bibliographies, *etc*. Dans les paragraphes suivants, nous décrivons la méthode que nous avons adoptée pour repérer les collocations du domaine d'étude. D'abord, nous avons employé un extracteur de termes disponible en ligne – *Termostat* – dont les résultats comportaient plusieurs lacunes pour l'extraction automatique de collocations terminologiques. Les failles illustrées dans §4.1 nous ont obligés à adopter une autre méthode d'extraction automatique qui prend en compte quatre indices pour le repérage de collocations : la dépendance syntaxique, la fréquence, la tendance à la cooccurrence (calculée à partir du score *PMI*, *pointwise mutual information*) et la spécificité au domaine d'étude.

4.1. *TermoStat* et l'extraction de collocations terminologiques

L'extracteur de termes que nous avons choisi est *TermoStat*, outil réalisé au sein de l'Observatoire de Linguistique Sens-Texte de l'Université de Montréal (Drouin 2003). Cet extracteur met en opposition des corpus spécialisés et non-spécialisés pour extraire la terminologie. L'extraction automatique de termes simples et de termes complexes nous a fourni 7794 candidats termes dont les patrons syntaxiques les plus récurrents sont :

- Nom + adjectif : 2707 candidats termes (35%)
- Nom + préposition + nom : 1882 candidats termes (24%)
- Nom : 1575 candidats termes (20%)

- D'autres patrons syntaxiques : 1630 (21%)

TermoStat n'est pas conçu en tant qu'extracteur de collocations, c'est pour cela que nous avons évalué si cet outil nous permet de repérer automatiquement un échantillon représentatif de collocations. Nous n'avons analysé que les premiers 300 résultats triés par *TermoStat* en ordre de *spécificité*. Ce calcul a été proposé par Lafon (1980) et permet d'analyser le vocabulaire spécifique à un corpus d'analyse par rapport à un corpus de référence. Dans notre cas, le corpus DIACOM-fr du commerce électronique 1990–2019 a été opposé à un corpus de français langue générale composé d'articles de journaux tirés du quotidien *Le Monde* 2002. Cette comparaison permet à l'extracteur de repérer les candidats termes, c'est-à-dire les candidats dont la fréquence est significativement plus élevée dans le corpus d'analyse par rapport à celle du corpus de référence en langue générale. Nous avons ensuite passé en revue les résultats pour les filtrer de manière manuelle afin de retenir uniquement les collocations terminologiques du domaine du commerce électronique. Premièrement, nous avons supprimé 172 candidats termes qui n'appartiennent pas au domaine d'étude (p.ex. *échelle, législateur, règles spéciales, justice procédurale*). Deuxièmement, nous avons distingué les termes simples (76) des termes complexes (52). L'analyse qualitative des termes complexes confirme que *TermoStat* n'est pas conçu pour l'extraction automatique de collocations terminologiques. En effet, l'extraction automatique donne beaucoup plus de locutions que de collocations : l'analyse que nous avons menée a montré que seulement 4 termes complexes sur 52 rentrent dans la catégorie de *collocation* (*téléchargement illégal, pratique illicite, établissement stable, contrat conclu*). De plus, l'extraction porte uniquement sur les syntagmes nominaux. Pour l'extraction de collocations à collocatif verbal, il faut employer une autre fonction de *TermoStat* : la fonction *Bigrammes* qui extrait les paires de mots composées d'un verbe et d'un nom (sujet ou objet du verbe). Pour notre étude, nous avons analysé 300 bigrammes. Nous avons d'abord supprimé les bigrammes qui n'appartiennent pas au domaine d'étude : 172 bigrammes (p.ex. *jouer rôle, poser problème, poser question, résoudre un problème*). De 128 bigrammes appartenant au domaine du commerce électronique, 56 peuvent être classés comme collocations (p.ex. *neutraliser dépersonnalisation, exercer activité, conclure contrat*) (Mel'čuk 1993, Wanner 1996).

Sans approfondir la question du classement des collocations standards, ce qui est évident lors de l'analyse de ces résultats c'est que l'outil *TermoStat* ne permet pas de repérer de manière exhaustive les collocations terminologiques. Pour les collocations nominales, l'extracteur ne les reconnaît pas toujours puisqu'elles sont bien souvent composées d'un élément qui appartient à la langue générale (p.ex. *fidèle* dans la collocation *client fidèle*). Pour les collocations verbales, la fonction *Bigrammes* donne des résultats intéressants, mais elle présente les résultats sans considérer la spécificité des termes, ce qui entraine beaucoup de bruit dans les résultats obtenus. Ces résultats sont quand même surprenants puisque l'outil a été conçu pour l'extraction de termes.

4.2. Dépendance syntaxique, fréquence, PMI et spécificité

Les lacunes que nous venons de présenter nous ont obligés à développer une autre méthode d'extraction automatique visant à obtenir des résultats plus satisfaisants tant du point de vue quantitatif que qualitatif. La méthode que nous proposons naît de la combinaison des résultats obtenus par deux outils *Stanza* et *TermoStat*. *Stanza* est une librairie Python d'analyse de la langue naturelle, développée par le Stanford NLP Group. La librairie contient des outils permettant la reconnaissance des parties du discours et des dépendances syntaxiques (Qi *et al.* 2020). Cet outil est devenu le point de départ de notre extraction au-

tomatique : nous avons extrait les combinaisons de deux mots dont le patron syntaxique correspondait aux types de collocations étudiées par la LEC. En particulier nous avons extrait les verbes et leurs dépendants (sujet-objet), les adverbes en relation avec un verbe et les adjectifs en relation de dépendance avec les noms. Pour ces paires de mots nous avons retenu la fréquence du mot 1, la fréquence du mot 2 et la fréquence de la cooccurrence des mots 1 et 2. Nous avons aussi calculé le PMI, c'est-à-dire le *pointwise mutual information* (Church & Hanks 1990), une mesure statistique qui quantifie l'écart entre la probabilité de la cooccurrence d'une paire de mots, en analysant leur distribution conjointe et leurs distributions individuelles. Cet indice est bien souvent employé en traitement automatique de la langue pour repérer des associations de mots fréquentes : une combinaison de lexies dont la valeur PMI est élevée (> 0) aura plus de possibilités d'être une collocation qu'une combinaison au PMI négatif. Cette extraction, à laquelle nous n'avons appliqué aucun type de filtre statistique, a conduit à un nombre très élevé de combinaisons (21.415) qui relèvent à la fois de la langue générale que de la langue de spécialité. Ce nombre élevé de paires de mots nous a amenés à combiner cette extraction aux résultats donnés par *TermoStat*, qui nous a permis d'observer la spécificité des mots (Lafon 1980). L'idée derrière la fusion des résultats des deux outils est d'exploiter le concept de spécialité et d'ainsi écarter les paires qui relèvent de la langue générale.

Dans le but d'évaluer cette méthode, nous n'avons analysé que les paires qui présentent plus de 50 cooccurrences dans notre corpus (353 paires). Nous avons supprimé automatiquement la seule combinaison dont la valeur de PMI était négatif (*de tout*) et toutes les paires (40) où les deux mots ne sont pas spécifiques au domaine d'étude, c'est-à-dire qui présentent une spécificité de 0 (p.ex. *reconnaître juridiquement*, *dire proprement*). Nous avons ensuite passé en revue manuellement les résultats triés par fréquence de cooccurrence du mot 1 et du mot 2 pour les filtrer de manière manuelle. Nous avons supprimé toutes les paires (98) qui n'appartiennent pas au domaine d'étude (p.ex. *autre part*, *jouer un rôle*) et les paires (156) ne correspondant pas à la définition de *collocation* que nous avons adoptée, à savoir les locutions (p.ex. *commerce électronique*) et les syntagmes libres (p.ex. *loi française*). Ce filtrage manuel a conduit à un total de 58 collocations terminologiques. Les collocations extraites présentent plusieurs patrons syntaxiques : Nom + Adjectif (p.ex. *effet négatif*, *pratique illicite*), Verbe + Nom (p.ex. *exercer une activité*, *conclure un contrat*), Nom + Verbe (p.ex. *tableau présente*, *Internet permet*), Verbe + Adverbe (p.ex. *appliquer aisément*). Nous soulignons que cette méthode qui s'appuie sur les dépendances syntaxiques permet de surmonter le problème de la disjonction des collocations qui cause des difficultés dans l'extraction automatique de certains outils comme *TermoStat*. Après avoir filtré les résultats, nous avons calculé la *précision* de cette méthode, c'est-à-dire le nombre de bons candidats par rapport au nombre total des candidats extraits[3]. La précision obtenue est de 16,43% (58 collocations valides sur 353), un pourcentage plus élevé que celui obtenu à l'aide de *TermoStat* qui est de 10% (60 bons candidats sur 600). De plus, l'extraction par dépendance syntaxique permet d'extraire un échantillon plus représentatif de plusieurs collocations ayant des patrons syntaxiques différents par rapport à ceux extraits par *TermoStat*.

Dans cette étude pilote, nous n'avons pas extrait de combinaisons de type Nom + Préposition + Nom, toutefois nous prévoyons implémenter ce patron dans notre méthodologie.

3 Pour l'instant notre évaluation ne concerne que la précision, mais pour des études ultérieures nous prévoyons prendre en considération une autre mesure : le *rappel*, c'est-à-dire le nombre de bons candidats extraits par rapport au nombre de toutes les collocations présentes dans le corpus.

5. Conclusions

Bien conscients de l'importance des relations syntagmatiques entre les termes, nous avons essayé de présenter une méthode d'extraction semi-automatique de collocations terminologiques. La méthode que nous avons illustrée est née de l'analyse des lacunes des résultats obtenus à partir d'un extracteur terminologique. Ces faiblesses nous ont permis de développer une nouvelle méthode d'extraction automatique qui prend en compte plusieurs indices et qui permet d'avoir des résultats plus satisfaisants en termes de *précision*. Cette méthode sera exploitée pour une étude plus exhaustive des collocations terminologiques à partir de notre corpus DIACOM-fr, dans le but ultime de décrire les collocations dans une base de données terminologique de type *réseau lexical* que nous sommes en train de réaliser à l'Université de Vérone dans le cadre du projet *Le Digital Humanities applicate alle lingue e letterature straniere*.

Références

Church, Kenneth, Patrick Hanks (1990) "Word Associations Norms, Mutual Information, and Lexicography." [In:] *Computational linguistics*. Vol. 16/1; 22–29.

Drouin, Patrick (2003) "Term Extraction Using Non-technical Corpora as a Point of Leverage." [In:] *Terminology*. Vol. 9/1; 99–115.

Frassi, Paolo, Silvia Calvi, John Humbley (2020) "Fouille de textes et repérage d'unités phraséologiques." [In:] Catherine Brune, Christophe Roche (eds.) *TOTh 2019 Terminologie & Ontologie*. Chambéry: Presses Universitaires Savoie Mont Blanc; 321–338.

L'Homme, Marie-Claude (2004) *La terminologie : principes et techniques*. Montréal : Les Presses universitaires de Montréal.

Lafon, Pierre (1980) "Sur la variabilité de la fréquence des formes dans un corpus." [In:] *MOTS*. Vol. 1; 128–165.

Mel'čuk Igor (1993) "La phraséologie et son rôle dans l'enseignement-apprentissage d'une langue étrangère." [In:] *ELA*. Vol. 92; 82–113.

Mel'čuk, Igor (2008) "Phraséologie dans la langue et dans le dictionnaire." [In:] *Repères & Applications* VI; 187–200.

Mel'čuk, Igor (2013) "Tout ce que nous voulions savoir sur les phrasèmes mais …." [In:] *Cahiers de lexicologie*. Vol. 102/1; 129–149.

Mel'čuk, Igor, André Clas, Alain Polguère (1995) *Introduction à la lexicologie explicative et combinatoire*. Paris: Duculot.

Polguère, Alain (2014) "From Writing Dictionaries to Weaving Lexical Networks." [In:] *International Journal of Lexicography*. Vol. 27/4; 396–418.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, Christopher D. Manning (2020) "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." [In:] *Association for Computational Linguistics (ACL) System Demonstrations*.

Wanner, Leo (ed.) (1996) *Lexical Functions in Lexicography and Natural Language Processing*. Amsterdam/Philadelphia: John Benjamins Publishing.

Sitographie

DIACOM: https://dh.dlls.univr.it/it/progetti/patrimonio-linguistico-culturale/#diacom [consulté le 20/07/2021].