Federica Vezzani

Vers une méthodologie pour l'extraction et la classification automatiques des collocations terminologiques verbales en langue médicale

Résumé : Cette étude porte sur la description d'une méthodologie efficace pour le développement d'un nouveau système d'apprentissage automatique supervisé pour l'extraction et la classification de « collocations terminologiques verbales » – telles que définies dans Costa/Silva (2004) – de type semi-compositionnel et compositionnel qui se manifestent en langue de spécialité médicale. En particulier, nous décrivons les phases de prétraitement du corpus spécialisé, d'extraction automatique de base (nom) + collocatif (verbe), et de formulation de la vérité terrain pour la classification des collocations terminologiques afin d'entraîner et de valider un système automatique performant pour la langue française et optimisé sur la base du postulat théorique adopté.

Mots-clés : collocation terminologique verbale, terminologie médicale, Traitement Automatique des Langues, apprentissage automatique

Abstract: This study deals with the description of an efficient methodology for the development of a new supervised machine learning system for the extraction and classification of semi-compositional and compositional « verbal terminological collocations » – as defined in Costa/Silva (2004) – which appear in medical language. In particular, we describe the phases of preprocessing of the specialized corpus, automatic extraction of base (noun) + collocate (verb) and formulation of the ground truth for the classification of terminological collocations in order to train and validate an effective automatic system for the French language, which is optimized on the basis of the adopted theoretical premise.

Keywords: verbal terminological collocation, medical terminology, natural language processing, machine learning

https://doi.org/10.1515/9783110749854-013

1 Introduction

L'étude des unités phraséologiques qui se manifestent dans les langues de spécialité joue un rôle central en terminologie (voir, entre autres, Galinski 1990 ; Rousseau 1993 ; Candel 1995 ; Rosenbaum 2016 ; Tutin/Grossmann 2014 ; Jacques/Tutin 2018). Leur analyse est particulièrement pertinente du point de vue du terminologue, car elle permet d'acquérir les constructions linguistiques typiques d'un secteur de connaissances spécialisées donné et, ensuite, de les décrire au sein de ressources terminographiques mono ou multilingues conçues pour soutenir de multiples activités langagières comme, par exemple, la traduction spécialisée ou la rédaction technique. Dans ce contexte, les approches et les outils de Traitement Automatique des Langues (TAL) contribuent de manière significative à la tâche d'extraction automatique des unités phraséologiques contenues dans les grandes collections de textes spécialisés. À ce propos, le tout récent axe de recherche de la « Phraséologie Computationnelle » rassemble les intérêts de nombreux chercheurs appartenant à des domaines hétérogènes – lexicologie, terminologie, informatique, statistique, etc. – en ce qui concerne l'identification, la sélection et la classification automatiques (ou semi-automatiques) d'unités phraséologiques dans leur sens le plus vaste (Heid 2008 ; Constant et al. 2017 ; Pastor/Mitkov 2019 ; Pastor/Colson 2020). Dans ce panorama,

> « [o]ne of the key issues of computational phraseology is to find which algorithms are best suited for the automatic or semi-automatic extraction of phrasemes, with possible differences according to the phraseme category » (Pastor/Colson 2020, 4).

À notre avis, ce problème computationnel est un « effet secondaire » strictement dépendant de l'absence d'accord unanime sur les définitions des différentes typologies d'unités phraséologiques. En particulier, en circonscrivant le champ d'investigation au phénomène de la collocation qui fera l'objet d'étude de cet article, Seretan (2011, 131–132) propose une enquête détaillée et classe jusqu'à 21 définitions associées à ce phénomène phraséologique. Nous n'entendons pas nous interroger sur l'interprétation actuellement la plus appropriée, il faut néanmoins considérer que les systèmes automatiques sont entraînés et optimisés selon la prémisse théorique adoptée. Il s'ensuit que les performances des différents systèmes sont largement variables selon la définition retenue et difficilement comparables entre eux en termes de meilleure efficacité. Un autre aspect important à considérer est que la quasi-totalité des études de TAL qui proposent des approches ou des outils pour l'extraction automatique de collocations (tant à partir de corpus général que spécialisé) sont basées sur les propriétés – qui varient selon le théoricien de référence – de cette unité phraséologique telle qu'elle se manifeste dans la langue générale. Dans la lignée des études de Silva et al. (2004), Costa/

AU: We have shortened the running head. Please check and confirm if it is fine.

Silva (2004) et Ramos et al. (2019), nous soutenons plutôt l'idée que pour une identification automatique efficace des collocations qui peuplent les langues de spécialité, il faut distinguer entre la notion de « collocation de langue générale » et de « collocation terminologique » (voir section 2) ayant chacune ses propres spécificités qui doivent être à la base de l'optimisation du processus d'extraction.

Cette étude vise donc à proposer une méthodologie pour la conception et l'implémentation d'un nouveau système d'apprentissage automatique supervisé pour l'extraction et la classification de collocations terminologiques à partir d'un corpus spécialisé constitué de textes en français appartenant au domaine médical. En particulier, notre intérêt porte sur les « collocations terminologiques verbales » – le verbe étant de plus en plus central dans les études terminologiques récentes (voir, entre autres, De Vecchi 2021) – telles que définies dans Costa/Silva 2004, Silva et al. (2004) et Ramos et al. (2019). La méthodologie globale proposée comprend quatre étapes : 1) le prétraitement du corpus, 2) la formulation de la « vérité terrain », 3) l'entraînement et 4) la validation du système. Dans cet article, nous nous concentrons sur l'accomplissement de deux premières phases qui constitueront le point de départ pour le développement d'un nouveau système d'extraction et de classification performant sur la base du postulat théorique proposé par les auteures susmentionnées. Par ailleurs, cette étude vise à apporter une contribution en termes de reproductibilité de l'expérimentation menée et de réutilisabilité des données produites qui seront diffusées au format *open source*.

L'organisation de l'article est la suivante. La section 2 introduit la définition de *collocation terminologique* et ses propriétés syntaxico-sémantiques selon le modèle théorique adopté. La section 3 illustre un aperçu des études qui proposent différentes approches et outils pour l'extraction automatique de collocations à partir de corpus textuels. La section 4 est consacrée à la présentation de la contribution de cet article et comprend : 1) la définition de l'objet d'investigation, à savoir les collocations terminologiques verbales, 2) la description du prétraitement du corpus de textes médicaux spécialisés et 3) la formulation de la « vérité terrain » pour la classification de la compositionnalité sémantique. La section 5 définit, d'un point de vue théorique, la méthodologie que nous adopterons pour l'entraînement, la validation et l'évaluation du système. Dans la section conclusive, nous exposons nos considérations finales et nos perspectives de recherche à moyen et long terme.

2 Collocations terminologiques

La notion de *collocation terminologique* découle de la nécessité d'identifier les propriétés du phénomène de la collocation qui se manifeste dans les langues de spécialité afin de marquer une différence avec ce même phénomène apparaissant en langue générale (Silva et al. 2004 ; Costa/Silva 2004, Ramos et al. 2019). En général, la collocation peut être définie comme

> « [. . .] une association entre morphèmes lexicaux présentant, entre eux, un phénomène d'attraction motivé par un ensemble d'affinités combinatoires » (Silva et al. 2004, 347).

Les deux éléments qui composent la collocation sont nommés par Hausmann (1979) comme *base* et *collocatif* et sont liés par une attraction selon laquelle, le collocatif est sélectionné pour exprimer un sens spécifique en fonction de la base donnée (ou *mot-clé* selon la dénomination de Mel'čuk et al. 1995). En effet, comme l'affirme L'Homme (1998, 516), « le sens d'un des éléments, la base ou le mot clé, demeure inchangé ; en revanche, le collocatif acquiert un nouveau sens au sein du groupement ». Quelques exemples de ce phénomène en langue générale sont rassemblés dans L'Homme (1998) et sont : *déception* (base) *amère* (collocatif), *éperdument* (base) *amoureux* (collocatif), *soulever* (collocatif) *l'enthousiasme* (base).

Analysées dans un contexte terminologique, les collocations prennent des spécificités sémantiques et morphosyntaxiques précises. D'un point de vue méta-terminologique, la dénomination de « collocation terminologique » a été introduite dans Silva et al. (2004, 352) en tant que néonyme moins porteur d'ambigüité que la dénomination de « combinaison lexicale spécialisée » précédemment introduite par L'Homme (1998). En particulier, la « collocation terminologique » a été définie dans Silva et al. (2004, 352) en tant que regroupement :

a) de deux lexèmes (base et collocatif), l'un possédant le statut de terme simple ou complexe et l'autre celui de non-terme (par exemple « optimiser la dose ») ; ou
b) de deux lexèmes (base et collocatif), dont les deux peuvent avoir le statut des termes simples ou complexes (par exemple « diagnostiquer des masses pancréatiques »).

Les auteures exposent ensuite les spécificités au niveau 1) sémantique, 2) morpho-syntaxique, 3) lexicographique/terminographique du phénomène de la collocation qui se manifeste dans la langue générale et dans la langue de spécialité (Silva et al. 2004, 353–354). En particulier, au niveau sémantique, la collocation de langue générale a un sens semi-compositionnel puisque la base A conserve son sens tandis que le collocatif B prend un sens C dans le groupe A+B (voir l'exemple susmentionné de « déception amère »). La collocation terminologique peut avoir et

un sens semi-compositionnel (A+B[C])[1] et un sens totalement compositionnel (A +B[B])[2] à partir de la somme des sens de ses composants. Des exemples de ces deux cas de figure sont fournis par les auteures elles-mêmes et sont, respectivement, « image en anneau » et « image cardiaque ».

En outre, la définition de la notion de *collocation terminologique* permet d'introduire la question relative à la distinction entre l'entité-collocation et l'entité-terme complexe et d'en tracer, au moins théoriquement, les frontières. En effet, alors que le terme complexe se présente comme une unité multilexémique ayant pour fonction de dénommer un et un seul concept (par exemple « néoplasie mammaire »), la totalité de la construction morphosyntaxique qui constitue la collocation terminologique ne renvoie pas à un seul concept et assume donc le statut de non-terme (par exemple « administrer un médicament ») (Silva et al. 2004). À ce propos, Pitar (2020, 94–95) remarque les deux critères suivants faisant la distinction entre ces deux éléments :

– le terme complexe est un signifiant qui recouvre un seul signifié, c'est-à-dire une seule notion, tandis que la collocation constitue une combinaison préférentielle d'un terme avec des déterminants ;
– le terme appartient au vocabulaire, tandis que la collocation, construite autour d'un terme noyau, appartient surtout au discours.

Cependant, comme l'indiquent Silva et al. (2004, 352–353), dans la pratique ces contours sont labiles car le statut de collocation terminologique – quel que soit le type de termes et/ou non-termes qui la composent (verbes, noms, adjectifs ou adverbes) – peut évoluer dans le temps et, si la relation de dépendance entre les deux éléments devient solide, la construction peut prendre le statut de terme complexe (un exemple de ce phénomène fourni par les mêmes auteures est « prendre la fièvre »). Enfin, comme le remarquent Meyer/Mackintosh (1994), L'Homme/Jia (2015) et L'Homme (2017) ce problème est particulièrement évident pour les groupes composés d'un terme modifié par un adjectif ou un syntagme prépositionnel (par exemple « logiciel de modélisation »), ouvrant ainsi une série de défis considérables pour l'accomplissement de la tâche d'extraction automatique.

1 Avec la forme graphique B[C] nous nous référons au collocatif B qui prend un sens C dans le groupe A+B.

2 Avec la forme graphique B[B] nous nous référons au collocatif B qui maintient son sens B dans le groupe A+B.

3 Extraction (semi-)automatique des collocations : approches et outils

Dans la littérature, il existe de nombreuses méthodes pour extraire automatiquement des collocations à partir de corpus. En général, les systèmes automatiques reposent sur trois grandes familles d'approches : statistiques, linguistiques et hybrides (Seretan 2011). L'enquête de Kolesnikova (2016) illustre un panorama exhaustif des stratégies adoptées dans les systèmes de TAL pour mesurer la force d'attraction entre deux unités linguistiques qui constituent potentiellement les deux éléments clés caractérisant la collocation, à savoir la base et le collocatif. Les mesures dites d'association (ou de cohésion) reposent sur 1) des formules statistiques, telles que *Log-likelihood* (Dunning 1993), *Mutual Information* (Church/Hanks 1990), *T-score* (Church et al. 1991) etc., 2) des méthodes basées sur des règles formulées *ad hoc* sur les spécificités de chaque langue et 3) la combinaison des deux méthodes précédentes afin de bénéficier des avantages découlant des deux stratégies. Comme nous l'avons décrit dans l'introduction, les performances des systèmes d'extraction automatique sont fortement dépendantes du postulat théorique adopté. Par exemple, en prenant comme point de départ la théorie Sens-Texte dans le cadre de la Lexicologie Explicative et Combinatoire (Mel'čuk et al. 1995), des études récentes proposent des systèmes d'identification et de classification automatiques des collocations à partir des fonctions lexicales (Mel'čuk 1996). En ce sens, l'étude d'Anke et al. (2019) introduit le nouveau jeu de données Lex-FunC dans lequel les collocations sont catégorisées selon les fonctions lexicales et présente des expériences afin d'évaluer les meilleures représentations vectorielles pour la classification des collocations. Les résultats obtenus prouvent qu'à l'heure actuelle la classification par fonctions lexicales pose de plus grands défis que les représentations vectorielles syntaxiques et sémantiques traditionnelles utilisées en littérature. Toujours dans cette optique, l'étude de Kolesnikova et al. (2020) illustre des expériences afin d'identifier automatiquement les fonctions lexicales verbales en appliquant des techniques d'apprentissage automatique supervisé sur un ensemble de données de collocations verbe-nom en espagnol. Les auteurs ont testé 68 algorithmes d'apprentissage supervisé et évalué leurs performances sur la détection automatique de fonctions lexicales en obtenant de meilleurs résultats que l'état de l'art en termes de précision et de rappel.

Les études illustrées jusqu'à présent se concentrent sur le processus d'extraction de collocations à partir de corpus de langue générale. Si l'on regarde les études menées sur des corpus spécialisés représentatifs d'une langue de spécialité donnée, nous pouvons constater qu'un des principaux problèmes qui se pose est la difficulté de distinguer automatiquement entre collocations et termes complexes. Comme

nous l'avons mentionné auparavant, ce problème largement connu dans la littérature constitue un défi de longue date pour les terminologues car l'identification manuelle des frontières entre l'entité-terme et l'entité-collocation est une tâche difficile qui varie considérablement selon l'interprétation donnée. L'automatisation de ce processus rend le défi encore plus insidieux. Parmi les principales contributions en ce sens figurent les travaux de Maniez qui visent à proposer des modèles d'acquisition (semi-) automatique pour la classification des « candidats termes » et des « candidates collocations » (à collocatif adjectival et verbal) à partir de corpus spécialisés pour le domaine médical (Maniez 2001 ; 2002a ; 2002b). Un autre aspect important à considérer en ce sens est introduit dans l'étude de Roche/Prince (2008) et concerne la notion de « pertinence » des collocations. Les auteurs évaluent que, bien que du point de vue syntaxique l'extraction – dans leur cas – des syntagmes nominaux soit *a priori* compatible avec la définition de Mel'čuk de base et collocatif, du point de vue sémantique et pragmatique ces candidates collocations doivent être validées par une forme de pertinence. Or, cette dernière est totalement dépendante de la tâche à réaliser, de la situation en vigueur, du genre textuel et du domaine considéré. La variabilité de la pertinence constitue donc une valeur à prendre en considération dans l'évaluation des systèmes d'extraction automatique. La contribution récente de L'Homme/Azoulay (2020) s'inscrit dans cette perspective en présentant une analyse comparative du comportement collocationnel de 15 items lexicaux analysés dans un corpus général et un corpus spécialisé sur le thème de l'environnement. En particulier, les items lexicaux sélectionnés sont des noms alors que les collocatifs sont de type verbal, adjectival et nominal. Dans ce contexte, l'étude montre que les corpus ont un impact significatif sur la tâche d'identification automatique des collocations et pourraient influencer ce que les lexicographes et les terminologues considèrent comme des collocations pertinentes.

Nous concluons cet aperçu par la présentation de quelques outils spécifiquement conçus et mis à la disposition pour l'extraction de collocations à partir de corpus spécialisés. En particulier, l'étude de Seretan et al. (2004) décrit un outil d'extraction de collocations à collocatif nominal, adjectival et verbal conçu en tant qu'aide à la traduction spécialisée et qui peut être intégré dans un environnement de traduction réel. L'outil Colex présenté par Orliac (2006) est plutôt un système hybride d'extraction des collocations conçu afin d'extraire les collocations de type verbe + nom à partir de corpus spécialisés et basé sur trois relations grammaticales modélisées par les « fonctions lexicales verbales » (Orliac 2006, 269) : sujet + verbe, verbe + premier complément et verbe + deuxième complément. Un autre outil est TermeX (Delač et al. 2009), basé sur l'utilisation de mesures d'association lexicale pour extraire automatiquement les collocations de tout type. Il offre à l'utilisateur une variété de mesures d'association parmi lesquelles choisir, ainsi que la possibilité de sélectionner manuellement les collocations les plus pertinentes parmi celles

extraites automatiquement. Enfin, bien qu'ils n'aient pas été spécifiquement conçus pour l'extraction de collocations, des outils tels que TermoStat (Drouin 2003), BiTermEx (Planas 2012), TermEvaluator (Inkpen 2016) et Sketch Engine (Kilgarriff et al. 2014) offrent de nombreuses fonctionnalités qui permettent d'obtenir des listes de combinaisons lexicales récurrentes qui se prêtent à une validation manuelle par le terminologue quant à leur statut de termes et/ou de collocations pertinents pour l'objet d'étude.

4 Conception et implémentation du système d'apprentissage automatique supervisé

La méthodologie proposée dans cette contribution constitue le point de départ pour la conception et l'implémentation d'un nouveau système d'apprentissage automatique supervisé pour l'extraction et la classification efficace de « collocations terminologiques » telles que définies dans la section 2 de cet article. En particulier, notre intérêt est tourné vers les collocations terminologiques à collocatif verbal telles que déterminées dans Costa/Silva (2004) et Ramos et al. (2019) qui se manifestent dans des textes médicaux (voir sous-section 4.1) et qui sont constituées :

1. d'une base A de type nominal ayant le statut de terme (simple ou complexe) spécialisé dans le domaine médical dont le sens demeure inchangé ; et
2. d'un collocatif B de type verbal ayant le statut de terme (simple ou complexe) spécialisé ou de non-terme de langue générale pouvant acquérir un sens B[C] ou conserver son sens B[B] originel.

L'enquête que nous présentons porte donc sur les collocations terminologiques verbales semi-compositionnelles A+B[C] et compositionnelles A+B[B]. D'un point de vue syntaxique, les deux éléments peuvent être liés par les relations grammaticales suivantes : 1) sujet A + verbe B[C] ou B[B], 2) verbe B[C] ou B[B] + complément d'objet direct (COD) A, et 3) verbe B[C] ou B[B] + complément d'objet indirect (COI) A.

Afin de décrire le développement du futur système, les deux sections suivantes seront consacrées à la description des phases de prétraitement du corpus et de formulation de la « vérité terrain » servant à jeter les bases des processus d'entraînement et de validation du système.

4.1 Prétraitement du corpus spécialisé

Une fois l'objet d'investigation défini, nous décrivons dans ces paragraphes la méthodologie de mise en forme du corpus à partir duquel le système d'apprentissage automatique supervisé sera entraîné et validé. La première phase de notre *pipeline* porte sur le prétraitement de la collection de textes et implique la démarche méthodologique suivante :

[1] Collection des textes
[2] Tokenisation
[3] Lemmatisation
[4] Étiquetage morphosyntaxique
[5] Analyse des dépendances
[6] Isolement de la base nominale
[7] Extraction de candidats collocatifs verbaux

Les textes qui composent notre corpus ont été extraits à partir du Manuel MSD disponible en ligne.[3] Cette ressource rassemble des articles de nature médicale sur des sujets hétérogènes et constitue une source de référence à la fois pour la consultation de connaissances spécialisées par des médecins experts et pour la vulgarisation de ces connaissances à un public non expert (Bullers 2016). Dans une perspective de variation diastratique de la langue, le manuel permet la consultation de deux versions d'un même sujet médical rédigées, d'un côté, pour les professionnels de santé et, de l'autre, pour le grand public afin d'en faciliter la compréhension. De plus, la quasi-totalité des articles est disponible en 11 langues différentes.

À partir de cette ressource, nous avons extrait toutes les pages HTML des articles en français – les sujets médicaux étant classés par ordre alphabétique de A à Z – destinés aux professionnels de santé. L'extraction des pages a été menée de manière semi-automatique grâce à l'utilisation du paquet *rvest*[4] du langage de programmation R. La collection téléchargée comprend un ensemble de 9.102 documents pour un total de 9.636.383 occurrences. Par la suite, sur la base de l'approche R *Tidyverse*,[5] nous avons procédé aux phases d'importation [1], de tokenisation [2], de lemmatisation [3], d'étiquetage morphosyntaxique [4], et d'analyse des dépendances [5], en utilisant le paquet *udpipe*[6] pour l'annotation du texte. Ce paquet fait

3 https://www.msdmanuals.com/fr/accueil.
4 https://cran.r-project.org/web/packages/udpipe/index.html.
5 https://www.tidyverse.org.
6 https://cran.r-project.org/web/packages/udpipe/index.html.

partie du projet *Universal Dependencies* (UD)[7] et permet de choisir, pour chaque langue, différents systèmes d'annotation entraînés sur des corpus spécifiquement sélectionnés. Dans cette expérimentation, nous avons choisi le modèle français *Sequoia*[8] entraîné également sur des collections de documents de l'Agence européenne des médicaments (Popa-Fabre et al. 2020). Le Tableau 1 résume un extrait de texte annoté via le *pipeline* adopté et illustre la sortie de UDpipe en format tabulaire où chaque ligne correspond à l'annotation d'un jeton spécifique.

AU: The cross references to figures, tables and equations are highlighted for the author/editor to check and confirm its correct placement. These highlights will be removed in the next stage. Please make changes if necessary.

Tableau 1 : Exemple de sortie de l'annotateur UDpipe.

	Token	Lemma	Upos	Head_token_id	Dep_rel
1	Un	un	DET	2	det
2	abcès	abcès	NOUN	5	nsubj
3	est	être	AUX	5	cop
4	une	un	DET	5	det
5	poche	poche	NOUN	0	root
6	de	de	ADP	7	case
7	pus	pus	NOUN	5	nmod
8	provoquée	provoquer	VERB	5	acl
9	. . .	. . .	. . .	. . .	. . .

Après avoir collecté et annoté le corpus spécialisé, nous avons exploré les deux dernières étapes de la démarche proposée pour la phase de prétraitement, à savoir l'isolement de la base nominale [6] et l'extraction de candidats collocatifs verbaux [7]. L'extraction des termes médicaux de type nominal constituant la base des candidates collocations terminologiques a été réalisée grâce au thésaurus *Medical Subject Headings* (MeSH)[9] qui a joué le rôle de ressource centrale pour la cartographie des termes contenus dans le corpus préalablement constitué. MeSH est un vocabulaire contrôlé de la *National Library of Medicine* (NLM)[10] des États-Unis et il est généralement utilisé par la NLM pour indexer et cataloguer les articles biomédicaux de la base de données PubMed[11] de manière uniforme et cohérente afin de soutenir la recherche d'informations dans le domaine médical (Coletti et al. 2001). La version française du vocabulaire MeSH est fournie gratuitement par l'Institut national de la santé et de la recherche médicale (INSERM) et

7 https://universaldependencies.org.
8 https://deep-sequoia.inria.fr.
9 https://meshb.nlm.nih.gov/search.
10 https://www.nlm.nih.gov.
11 https://pubmed.ncbi.nlm.nih.gov.

comprend au total 122.819 termes simples et complexes.[12] Pour cette première étude, nous nous sommes concentrés sur l'extraction des termes MeSH simples contenus dans notre corpus et, sur la base de la fonction *Term frequency – Document Frequency* TF-DF (Salton/Yang 1973), nous avons extrait un total de 11.908 termes médicaux de type nominal. Les termes sont classés sur la base de la moyenne entre les fréquences d'occurrences dans les documents et le nombre de documents de la collection dans lesquels les occurrences apparaissent. De cette liste, nous avons extrait automatiquement les verbes associés aux termes grâce à l'analyse de dépendance réalisée précédemment avec UDpipe. Pour chaque terme simple MeSH, nous avons récupéré les occurrences étiquetées *nsubj*, *obj*, et *iobj* – respectivement sujet, COD et COI – et extrait le verbe qui, dans l'arbre syntaxique, est associé au terme. Ce *pipeline* nous a permis d'extraire les bases et les candidats collocatifs qui font l'objet de l'annotation sémantique manuelle suivante.

4.2 Formulation de la « vérité terrain »

Dans l'apprentissage supervisé, un système automatique apprend à effectuer sa tâche de manière optimale (extraction, classification, etc.) à travers un entraînement basé sur des données soigneusement étiquetées et jugées véridiques selon la tâche à accomplir. En ce sens, la deuxième phase de la méthodologie proposée ici comprend la formulation de la « vérité terrain » (Søgaard et al. 2014) qui constituera le modèle sur lequel le futur système d'apprentissage automatique supervisé sera entraîné pour l'extraction et la classification des collocations terminologiques à collocatif verbal.

À partir des données extraites dans la phase de prétraitement du texte (noms et verbes), nous basons leur annotation sur la notion de compositionnalité du sens de la candidate collocation. Comme nous l'avons décrit précédemment, la compositionnalité d'une collocation terminologique est dictée par le collocatif qui peut varier de sens (B[C]) – collocation semi-compositionnelle – ou conserver son sens usuel (B[B]) – collocation compositionnelle – tout en restant le sens de la base A inchangé. À cet égard, nous proposons une annotation manuelle des verbes associés aux termes médicaux extraits en termes de variation de sens. Reprenant la notion de « pertinence » introduite dans la section 2 (Roche/Prince 2008), notre objectif est de marquer comme :

– « pertinents » les verbes qui changent de sens B[C],
– « non-pertinents » les verbes qui gardent leur sens B[B].

12 http://mesh.inserm.fr/FrenchMesh/.

Or, afin de définir une méthodologie rigoureuse pour l'annotation de la « pertinence », il est d'abord nécessaire de s'interroger sur ce que l'on entend par variation de sens du collocatif. Reprenant l'affirmation selon laquelle « le sens d'un des éléments, la base ou le mot clé, demeure inchangé ; en revanche, le collocatif acquiert un nouveau sens au sein du groupement » (L'Homme 1998, 516), nous nous demandons par rapport à quoi le sens de la base demeure inchangé et par rapport à quoi le collocatif acquiert un nouveau sens. Les deux questions sont au cœur des deux cas de figure de collocation terminologique dans lesquels, alors que la base A reste inchangée, le collocatif peut à la fois prendre un nouveau sens et conserver son sens originel. À cet égard, pour définir un critère procédural clair pour l'annotation de la variation de sens du collocatif, nous nous appuyons sur la convention suivante : « [. . .] it is a lexicographic convention that the most frequent and therefore typical sense is given the highest rank in the list of senses and is put first in the dictionary entry » (Gelbukh/Kolesnikova 2012, 62).

Tous les verbes extraits – ayant le statut de termes ou de non-termes – font donc l'objet d'une analyse manuelle minutieuse afin de vérifier si le sens exprimé par rapport à la base nominale – ayant le statut de terme simple – occupe la première place dans les ressources lexicographiques pour la langue française. En particulier, si le sens du collocatif B exprimé en fonction de la base A correspond à la première définition répertoriée dans la ressource lexicographique (B[B]), la collocation sera étiquetée comme compositionnelle ; tandis que si le sens du collocatif B exprimé en fonction de la base A correspond à une définition qui se positionne à un niveau inférieur de la liste[13] (B[C]), la collocation sera étiquetée comme semi-compositionnelle. Afin de garantir une certaine cohérence sur le jugement de pertinence, notre démarche repose sur la consultation croisée de deux ressources lexicographiques disponibles en ligne, respectivement Le Robert[14] et Le Larousse.[15] De cette façon, nous comparons les données des deux sources en ce qui concerne la position des définitions (sens) associées au collocatif verbal. Le choix de s'appuyer sur des ressources lexicographiques – et non terminographiques – découle de la volonté d'identifier tous les sens attribués au terme analysé et pas seulement le sens spécifique que ce terme peut avoir dans la langue de spécialité étudiée. En ce sens, consulter une ressource terminographique risquerait de nous faire perdre les informations relatives à toutes les définitions qui, de manière transversale, peuvent être associées au terme quel que soit le domaine de travail.

13 Par « niveau inférieur » nous entendons toutes les définitions qui se placent après la première de la liste.

14 https://dictionnaire.lerobert.com.

15 https://www.larousse.fr.

Les travaux d'étiquetage sont en cours et l'ensemble des annotations sera librement mis à la disposition sur la plateforme GitHub[16] pour permettre la réutilisation de ces données. Le Tableau 2 illustre un extrait du processus d'annotation des verbes les plus fréquemment associés à trois termes médicaux contenus dans le corpus : « médicament », « chirurgie » et « maladie ».

Tableau 2 : Formulation de la vérité terrain.

Nom (N)	Verbe (V)	Pertinence – Sens (V)	Type de collocation
médicament	administrer	pertinent – B[C]	semi-compositionnelle
médicament	prescrire	non-pertinent – B[B]	compositionnelle
médicament	prendre	pertinent – B[C]	semi-compositionnelle
chirurgie	subir	pertinent – B[C]	semi-compositionnelle
chirurgie	réaliser	non-pertinent – B[B]	compositionnelle
maladie	contracter	pertinent – B[C]	semi-compositionnelle
maladie	dépister	pertinent – B[C]	semi-compositionnelle
maladie	traiter	pertinent – B[C]	semi-compositionnelle
maladie	enrayer	pertinent – B[C]	semi-compositionnelle

Tous les termes de type nominal de ce tableau – première colonne – jouent le rôle d'*obj* (COD) par rapport aux verbes extraits de la deuxième colonne. La troisième colonne contient le jugement de pertinence basé sur la variation (B[C]) ou non-variation (B[B]) du sens du collocatif, telle qu'identifiée à partir des deux ressources précitées. Cette annotation permet de classer, dans la quatrième colonne, le type de collocation terminologique verbale comme semi-compositionnelle ou entièrement compositionnelle conformément au postulat théorique adopté.

5 Entraînement, validation et évaluation du système

La collection des données annotées résultant du processus précédent constitue donc la vérité terrain sur laquelle s'appuieront l'entraînement et la validation du futur système d'apprentissage supervisé. Dans cette section, nous présentons la méthodologie que nous adopterons une fois la formulation de la vérité terrain terminée. En particulier, les prochaines étapes de notre démarche portent sur :

16 https://github.com/gmdn.

1. la représentation mathématique des collocations ;
2. le choix du modèle de classification à entraîner ;
3. la validation de ses performances.

Tout d'abord, nous évaluerons la forme mathématique la plus appropriée pour la représentation des collocations terminologiques verbales sur la base de l'état de l'art actuel. En général, cette représentation est performée au moyen de vecteurs numériques qui sont utilisés dans un espace vectoriel pour calculer les « distances » entre les unités linguistiques : l'hypothèse est que plus les vecteurs représentant ces unités dans cet espace sont proches, plus leurs sens seront proches en langue naturelle. Les valeurs de ces vecteurs peuvent changer en fonction de l'approche statistique choisie, par exemple binaire (1 ou 0) ou TF-IDF, ou en fonction de l'entraînement basé sur une approche neuronale, par exemple à travers la méthode des « plongements de mots » afin de capturer de façon optimale la cooccurrence d'unités linguistiques dans des contextes similaires (Manning/Schutze 1999).

La deuxième étape concerne le choix du modèle de classification supervisé qui sera utilisé pour prédire l'étiquette d'une collocation. Dans notre cas, nous utiliserons l'étiquette « positif » pour les collocations terminologiques verbales semi-compositionnelles et l'étiquette « négatif » pour les collocations terminologiques verbales compositionnelles. Loin d'entrer dans les détails des différents types de méthodes et d'algorithmes, l'idée principale est que, étant donné un ensemble de données étiqueté, un modèle supervisé est entraîné sur un sous-ensemble de cet ensemble de données pour apprendre les caractéristiques d'une classe d'objets (dans notre cas les caractéristiques des deux classes de collocations terminologiques verbales semi-compositionnelles et compositionnelles) sur la base de leurs représentations vectorielles. Par exemple, dans Kolesnikova/Gelbukh (2020), les auteurs ont entraîné, sur la base de 50 % de l'ensemble de données disponibles, six classificateurs différents, tels que les « machines à vecteurs de support », le « perceptron multicouche », les « k plus proches voisins », l'« arbre de décision », les « forêts aléatoires » et « Ada Boost ». En termes de reproductibilité de cette expérience, l'un de nos objectifs futurs est d'évaluer lequel parmi ces modèles est le plus performant pour la classification de notre objet d'étude.

Dans la troisième étape, la partie restante de l'ensemble de données sera ensuite utilisée pour valider les performances du système de classification. En général, étant donné un classificateur binaire, c'est-à-dire un classificateur qui prédit si une collocation est soit semi-compositionnelle soit compositionnelle, la sortie peut être 1) correcte si la prédiction correspond à la vraie étiquette (la vérité terrain) de l'objet – par exemple « vrais positifs » ou « vrais négatifs » respectivement pour les collocations qui appartiennent à la classe semi-compositionnelle ou à la classe compositionnelle – ou 2) incorrecte lorsque la prédiction ne corres-

pond pas à la vraie étiquette (« faux négatifs » ou « faux positifs »). Par exemple, nous aurons un « faux négatif » lorsque nous attribuons l'étiquette « négatif » à une collocation terminologique verbale semi-compositionnelle, alors que nous aurons un « faux positif » lorsque nous attribuons l'étiquette « positif » à une collocation terminologique verbale compositionnelle. Ces quatre valeurs – « vrai positif », « vrai négatif », « faux négatif », « faux positif » – sont combinées de différentes manières pour former la base de l'évaluation et de la comparaison des différents classificateurs. Dans l'étude de Kolesnikova/Gelbukh (2020), 50% de l'ensemble des données qui restent est utilisé pour tester les performances des classificateurs à l'aide de quatre mesures d'évaluation : le rappel, la précision, le F1 et l'exactitude. Dans ce cas également, nous prendrons ces mesures d'évaluation pour tester la performance de notre système de classification supervisé.

Il faut préciser que, dans la toute récente littérature (Kolesnikova/Gelbukh 2020 ; Shwartz/Dagan 2019 ; Anke et al. 2021), ces trois étapes sont généralement performées pour la tâche de classification des fonctions lexicales de type verbe-nom à partir de grands corpus de langue générale. En ce sens, notre travail se distingue de ces études en ce que la méthodologie proposée ici vise plutôt à apporter une contribution en termes de classification de la compositionnalité sémantique des collocations terminologiques verbales qui se manifestent en langue de spécialité médicale. De plus, la description détaillée de cette méthodologie vise à apporter une contribution en termes de reproductibilité de l'expérimentation et de réutilisabilité de la collecte de données annotées qui sera mise à la disposition au format *open source*.

6 Conclusions et perspectives

Cette étude est consacrée à la description du cadre méthodologique pour la conception et l'implémentation d'un système d'apprentissage supervisé pour l'extraction et la classification automatiques de collocations terminologiques verbales semi-compositionnelles et compositionnelles. En particulier, après avoir défini notre objet d'investigation, nous avons décrit les phases de prétraitement du texte et de formulation de la vérité terrain pour l'étiquetage des données sur lesquelles s'appuieront l'entraînement et la validation du futur système. Bien que ce dernier processus soit toujours en cours, nous pouvons avancer quelques considérations préliminaires en termes d'efficacité de la performance du système. Tout d'abord, nous soulignons sa dépendance à l'égard 1) du postulat théorique adopté, 2) de la langue d'étude et 3) de la langue de spécialité analysée. Le système sera en effet à considérer comme – plus ou moins – performant selon la définition re-

tenue de collocation terminologique verbale (Silva et al. 2004) qui se manifeste en langue de spécialité médicale française ayant un haut niveau de spécialisation. Par conséquent, l'entraînement et la validation d'un système automatique avec des prémisses théoriques différentes nécessitera un ensemble de données étiquetées différemment conformément à ces mêmes postulats théoriques.

En ce qui concerne nos perspectives à moyen terme, nous entendons achever la phase de formulation de la vérité terrain non seulement en considérant les collocatifs associés aux termes simples cartographiés par le thésaurus MeSH mais aussi en intégrant les collocatifs associés aux termes complexes constituant la base de la collocation terminologique et ayant le rôle de sujet, COD et COI. De plus, au sein du jeu de données annoté, nous prévoyons également l'intégration d'informations relatives à la fonction lexicale exprimée par le collocatif verbal en fonction de sa base nominale. Enfin, après le développement du système soigneusement entraîné et validé (section 4), nous visons à mener des expérimentations à long terme afin de le tester en termes de variation diastratique de la langue médicale. En particulier, nous nous intéressons à comparer le phénomène de la collocation terminologique verbale qui se manifeste dans les textes médicaux destinés aux experts et dans les textes médicaux destinés aux non-experts en utilisant comme corpus de validation la collection d'articles offerte par le manuel MSD à l'usage du grand public.

Bibliographie

Anke, Luis Espinosa/Codina-Filbá, Joan/Wanner, Leo, *Evaluating language models for the retrieval and categorization of lexical collocations*, in : Merlo, Paola/Tiedemann, Jorg/Tsarfaty, Reut (edd.), *Proceedings of the 16th conference of the European chapter of the association for computational linguistics. Main volume*, Association for Computational Linguistics, 2021, 1406–1417, DOI: 10.18653/v1/2021.eacl-main.120.

Anke, Luis Espinosa/Schockaert, Steven/Wanner, Leo, *Collocation classification with unsupervised relation vectors*, in : Korhonen, Anna/Traum, David/Màrquez, Lluís (edd.), *Proceedings of the 57th annual meeting of the association for computational linguistics*, Association for Computational Linguistics, 2019, 5765–5772, DOI: 10.18653/v1/P19-1576.

Bullers, Krystal, *Merck manuals*, Journal of the medical library association 104:4 (2016), 369–371, DOI: 10.3163/1536-5050.104.4.028.

Candel, Danielle, *Locutions en langues de spécialité*, Cahiers du français contemporain 2 (1995), 151–173.

Church, Kenneth/Hanks, Patrick, *Word association norms, mutual information, and lexicography*, Computational linguistics 16:1 (1990), 22–29.

Church, Kenneth/Gale, William/Hanks, Patrick/Hindle, Donald, *Using statistics in lexical analysis*, in : Zernik, Uri (ed.), *Lexical acquisition. Exploiting on-line resources to build a lexicon*, Hillsdale (NJ), Lawrence Erlbaum, 1991, 115–164.

Coletti, Margaret H./Bleich, Howard L., *Medical subject headings used to search the biomedical literature*, Journal of the American medical informatics association 8:4 (2001), 317–323.

Constant, Matthieu/Eryiğit, Gülsen/Monti, Johanna/van Der Plas, Lonneke/Ramisch, Carlos/Rosner, Michael/Todirascu, Amalia, *Multiword expression processing. A survey*, Computational linguistics 43:4 (2017), 837–892.

Costa, Rute/ Silva,Raquel, *The verb in the terminological collocations. Contribution to the development of a morphological analyser. MorphoCom*, in : Lino, Maria Teresa/ Xavier,Maria Francisca/ Ferreira, Fátima/ Costa,Rute/ Silva,Raquel (edd.), *Proceedings of the fourth international conference on language resources and evaluation (LREC'04)*, European Language Resources Association (ELRA), 2004, 1531–1534, http://www.lrec-conf.org/proceedings/lrec2004/pdf/795.pdf [dernière consultation : 25.01.2022].

Delač, Davor/ Krleza,Zoran/ Snajder,Jan/Dalbelo Basic, Bojana/ Saric,Frane, *TermeX. A tool for collocation extraction*, in : Gelbukh, Alexander (ed.), *Computational linguistics and intelligent text processing*, Berlin/ Heidelberg,Springer, 2009, 149–157.

De Vecchi, Dardo, *D'une terminologie de quoi à une terminologie de et pour qui? Le rôle des verbes en entreprise*, in : Delavigne, Valérie/De Vecchi, Dard (edd.), *Termes en discours. Entreprises et organisations*, Paris, Presses Sorbonne Nouvelle, 2021, 69–82.

Drouin, Patrick, *Term extraction using non-technical corpora as a point of leverage*, Terminology 9:1 (2003), 99–115.

Dunning, Ted E., *Accurate methods for the statistics of surprise and coincidence*, Computational linguistics 19:1 (1993), 61–74.

Galinski, Christian, *Terminology and phraseology*, Terminology science and research 1 (1990), 70–86.

Gelbukh, Alexander/Kolesnikova, Olga, *Semantic analysis of verbal collocations with lexical functions*, Berlin/Heidelberg, Springer, 2013.

Jacques, Marie-Paule/Tutin, Agnès (edd.), *Lexique transversal et formules discursives des sciences humaines*, Londres, ISTE Éditions, 2018.

Hausmann, Franz Josef, *Un dictionnaire des collocations est-il possible?*, Travaux de linguistique et de littérature Strasbourg 17:1 (1979), 187–195.

Heid, Ulrich, *Computational phraseology. An overview*, in : Granger, Sylviane/ Meunier,Fanny (edd.), *Phraseology. An interdisciplinary perspective*, Amsterdam/ Philadelphia,John Benjamins Publishing, 2008, 337–360.

Inkpen, Diana/Paribakht, Sima T./Faez, Farahnaz/Amjadian, Ehsan, *Term evaluator. A tool for terminology annotation and evaluation*, International journal of computational linguistics and applications, 7:2 (2016), 145–165.

Kilgarriff, Adam/ Baisa,Vít/ Bušta,Jan/ Jakubíček,Miloš/ Kovář,Vojtěch/ Michelfeit,Jan/ Rychlý,Pavel/ Suchomel,Vít, *The Sketch Engine. Ten years on*, Lexicography 1 (2014), 7–36.

Kolesnikova, Olga, *Survey of word co-occurrence measures for collocation detection*, Computación y sistemas 20:3 (2016), 327–344.

Kolesnikova, Olga/Gelbukh, Alexander, *A study of lexical function detection with word2vec and supervised machine learning*, Journal of intelligent & fuzzy systems 39:2 (2020), 1993–2001.

Kolesnikova, Olga/ Gelbukh,Alexander/Chanona Hernández, Liliana, *Automatic detection of semantic classes of verb-noun collocations*, Computación y sistemas 24:1 (2020), 141–150.

L'Homme, Marie-Claude, *Caractérisation des combinaisons lexicales spécialisées par rapport aux collocations de langue générale*, in : Fontenelle, Thierry/Hiligsmann, Philippe/Michiels, Archibald/ Moulin, André/Theissen, Siegfried (edd.), *EURALEX'98 Proceedings. Papers submitted to the eighth EURALEX international congress on lexicography in Liège, Belgium*, Liège, University of Liège, 1998, 513–522.

L'Homme, Marie-Claude/Jia, Zhichao, *Classement des combinaisons lexicales spécialisées à base nominale dans un dictionnaire d'informatique*, Cahiers de lexicologie 106 (2015), 229–251.

L'Homme, Marie-Claude, *Combinatoire spécialisée. Trois perspectives et des enseignements pour la terminologie*, TTR. Traduction, terminologie, rédaction 30:1–2 (2017), 215–241.

L'Homme, Marie-Claude/Azoulay, Daphnée, *Collecting collocations from general and specialised corpora*, in : Pastor, Gloria Corpas/Colson, Jean-Pierre (edd.), *Computational phraseology*, Amsterdam/Philadelphia, John Benjamins, 2020, 151–175.

Maniez, François, *Extraction d'une phraséologie bilingue en langue de spécialité. Corpus parallèles et corpus comparables*, Meta. Journal des traducteurs/Meta. Translators' journal 46:3 (2001), 552–563.

Maniez, François, *Distinguer les termes des collocations. Étude sur corpus du patron « adjectif–nom » en anglais médical*, in : Pierrel, Jean-Marie (ed.), *Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles. Posters*, Nancy, ATALA, 2002, 344–349 (=2002a).

Maniez, François, *Un modèle d'extraction des collocations en langue de spécialité*, ASp. la revue du GERAS 35–36 (2002), 35–47 (=2002b).

Manning, Christopher/Schutze, Hinrich, *Foundations of statistical natural language processing*, Cambridge (MA), MIT press, 1999.

Mel'čuk, Igor A, *Lexical functions. A tool for the description of lexical relations in a lexicon*, in : Wanner, Leo (ed.), *Lexical functions in lexicography and natural language processing*, Amsterdam/ Philadelphia, John Benjamins, 1996, 37–102.

Mel'čuk, Igor A./Clas, André/Polguère, Alain, *Introduction à la lexicologie explicative et combinatoire*, Louvain-la-Neuve, Duculot, 1995.

Meyer, Ingrid/Mackintosh, Kristen, *Phraseme analysis and concept analysis. Exploring a symbiotic relationship in the specialized lexicon*, in : Martin, Willy/Meijs, Willem/Moerland, Margreet/ten Pas, Elsemiek/van Sterkenburg, Piet (edd.), *EURALEX'94 Proceedings. Papers submitted to the 6th EURALEX international congress on lexicography in Amsterdam, The Netherlands*, Amsterdam, Vrije Universiteit, 1994, 339–348.

Orliac, Brigitte, *Colex. un outil d'extraction de collocations spécialisées basé sur les fonctions lexicales*, Terminology 12:2 (2006), 261–280.

Pastor, Gloria Corpas/Mitkov, Ruslan (edd), *Computational and corpus-based phraseology. Third international conference. Europhras 2019. Malaga, Spain, September 25–27, 2019, proceedings*, Cham, Springer Nature, 2019.

Pastor, Gloria Corpas/Colson,Jean-Pierre (edd.), *Computational Phraseology*, Amsterdam/ Philadelphia, John Benjamins, 2020.

Pitar, Mariana, *Sur la cohésion interne des termes complexes*, Buletinul stiintific al universitatii politehnica din Timisoara, Seria limbi moderne 19 (2020), 92–103.

Planas, Emmanuel, *BiTermEx. Un prototype d'extraction de mots composés à partir de documents comparables via la méthode compositionnelle*, in : Antoniadis, Georges/Blanchon, Hervé/Sérasset, Gilles (edd.), *Proceedings of the joint conference JEP-TALN-RECITAL 2012*, vol. 2, Grenoble, ATALA/ AFCP, 2012, 415–422.

Popa-Fabre, Murielle/Ortiz Suárez, Pedro Javier/Sagot, Benoît/De la Clergerie, Éric, *French contextualized word-embeddings with a sip of CaBeRnet. A new French balanced reference corpus*, in : Bański, Piotr/Barbaresi, Adrien/Clematide, Simon/Kupietz, Marc/Lüngen, Harald/Pisetta, Ines (edd.), *Proceedings of the 8th workshop on challenges in the management of large corpora*, European Language Ressources Association, 2020, 15–23, https://aclanthology.org/2020.cmlc-1.3 [dernière consultation : 25.01.2022].

Ramos, Margarida/ Costa,Costa/ Roche,Christophe, *Dealing with specialised co-text in text mining. Verbal terminological collocations*, in : Brune, Catherine/ Roche,Christophe (edd.), *TOTh 2019. Terminologie & ontologie. Théories et Applications*, Chambéry, Presses Universitaires Savoie Mont Blanc, 2019, 339–362.

Roche, Mathieu/Prince, Violaine, *Evaluation et détermination de la pertinence pour des syntagmes candidats à la collocation*, in : Heiden, Serge/Pincemin, Bénédicte (edd.) *JADT'08. Journées internationales d'analyse statistique des données textuelles*, Lyon, Presses universitaires de Lyon, 2008, 1009–1020.

Rosenbaum Franková, Lucia (2016), *Phrasèmes spécialisés dans les textes économiques*, Cahiers de lexicologie 108 (2016), 43–57.

Rousseau, Louis-Jean, *Terminologie et phraséologie, deux composantes indissociables des langues de spécialités*, Terminologies nouvelles 10 (1993), 9–12.

Salton, Gerard/ Yang,Chung-Shu, *On the specification of term values in automatic indexing*, Journal of documentation 29:4 (1973), 351–372.

Seretan, Violeta, *Syntax-based collocation extraction*, Dordrecht, Springer Science & Business Media, 2011.

Seretan, Violeta/Nerima, Luka/Wehrli, Eric, *A tool for multi-word collocation extraction and visualization in multilingual corpora*, in : Williams, Geoffrey/Vessier, Sandra (edd.), *Proceedings of the 11th EURALEX international congress*, Lorient, Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, 2004, 755–766.

Shwartz, Vered/ Dagan,Ido, *Still a pain in the neck. Evaluating text representations on lexical composition*, Transactions of the Association For Computational Linguistics 7 (2019), 403–419.

Silva, Raquel/Costa, Rute/Ferreira, Fátima, *Entre langue générale et langue de spécialité une question de collocations*, Études de linguistique appliquée 3 (2004), 347–359.

Søgaard, Anders/ Planl,Barbara/ Hovy,Dirk, *Selection bias, label bias, and bias in ground truth*, in : Liu, Qun/ Xia,Fei (edd.), *Proceedings of COLING 2014. 25th international conference on computational linguistics. Tutorial abstracts*, Dublin City University/Association for Computational Linguistics, 2014, 11–13, https://aclanthology.org/C14-3005.pdf [dernière consultation : 25.01.2022].

Tutin, Agnès/Grossmann, Francis (edd.), *L'écrit scientifique. Du lexique au discours*. *Autour de Scientext*, Rennes, Presse universitaire de Rennes, 2014.