Le rôle des marqueurs et indicateurs dans l'analyse lexicale et sémantico-pragmatique de reformulations médicales

Ioana Buhnila

Le rôle des marqueurs et indicateurs dans l'analyse lexicale et sémantico-pragmatique de reformulations médicales

Ioana Buhnila[1*]

[1]LiLPa UR 1339 – Linguistique, Langues, Parole, Université de Strasbourg, 67000 Strasbourg, France

Résumé. Les textes de spécialité médicale posent des problèmes de compréhension au grand public à cause de la complexité des termes médicaux. Vu l'importance croissante de rendre les recherches médicales compréhensibles pour le grand public, la reformulation et la simplification des notions techniques deviennent indispensables. L'objectif de ce travail est de présenter une méthode semi-automatique d'identification des reformulations médicales à l'aide d'une analyse lexicale et sémantico-pragmatique des marqueurs et indicateurs de reformulation. Pour automatiser l'annotation, nous vérifions l'hypothèse du lien qui existe entre les marqueurs / indicateurs de reformulation, les relations lexicales et les fonctions sémantico-pragmatiques de reformulations (par exemple, le marqueur « est un / une » indique une reformulation de type hyperonymie / définition). Nous présentons notre méthode d'annotation et d'analyse lexicale et les résultats obtenus sur un corpus des textes médicaux scientifiques écrits en français. Notre méthode servira à la création de ressources linguistiques sur la reformulation des termes de spécialité pour les vulgariser devant le grand public. Notre méthode est applicable également sur d'autres langues proches du français.

Abstract. Lexical and semantico-pragmatic analysis of markers and indicators of medical reformulations. Medical texts are difficult to understand for laypeople due to the complexity of the medical terms. Given the increasing importance of making medical research understandable by the whole society, the reformulation and simplification of technical concepts become a necessity. The aim of this work is to present a semi-automatic method for identifying medical reformulations through a lexical and semantic-pragmatic analysis of reformulation markers and indicators. To automate the annotation process, we test the hypothesis of the link that exists between reformulation markers/indicators, the lexical relations and the semantic-pragmatic functions of reformulations (e.g. the marker "is a" indicates a hypernymy/definition type of reformulation). We present our annotation method, the lexical analysis and the results obtained on written scientific medical corpus in French. This method is useful for the creation of linguistic resources on the reformulation of scientific terms. These resources can be exploited for science popularisation purposes. Our method is also applicable to other related languages.

* Ioana Buhnila : ioana.buhnila@etu.unistra.fr

1 Introduction

Les termes médicaux posent des difficultés de compréhension au grand public. *Le terme* est une unité lexicale de spécialité qui représente des connaissances spécifiques à un domaine du savoir, connaissances reconnues et partagées par les membres d'une communauté de spécialistes (Costa 2005 : 84). *Le terme* appartient à la langue de spécialité, un « sous-système » autonome de la langue qui a comme objectif la transmission de connaissances spécialisées (Contente 2005 : 456). La technicité des termes médicaux est donnée par l'étymologie grecque et / ou latine et la composition souvent mixte de ces deux bases avec la langue moderne. Des termes comme « cholecystectomie », formé avec deux bases grecques *chole* (bile) et *ectomy* (ablation chirurgicale) avec une base latine (*cystis* / vessie) au milieu, (Grabar et Hamon 2015 : 2), sont très difficiles à comprendre par un public non-spécialiste. Ces termes nécessitent un *décodage* pour les rendre compréhensibles devant le grand public. *La reformulation* permet de rendre accessible *le sens* des termes à travers des mots de la langue générale de type synonymes ou hyperonymes, ou à travers des définitions et des explications.

L'objectif de notre étude est de développer une méthode d'identification et d'annotation automatique des reformulations médicales. Le but de cette recherche est de créer un corpus de reformulations des termes médicaux monolexicaux et polylexicaux en français. Pour y parvenir, nous annotons des *termes médicaux* et des *marqueurs de reformulation*. Ces informations identifiées automatiquement nous permettent d'extraire les phrases qui ont une grande probabilité de contenir des reformulations des termes médicaux. Selon Fuchs (2020), *les marqueurs de reformulation* ont le rôle de mettre en relation d'équivalence le formulé avec le reformulé et expriment une homogénéité du sens. Nous analysons également la question de *l'absence* du marqueur de reformulation. Dans ce sens, nous cherchons dans les phrases non pas uniquement des marqueurs de reformulations, mais aussi des *indicateurs de reformulations* (Steuckardt, 2018). Nous définissons les indicateurs de reformulation comme des unités lexicales ou grammaticales qui, dans leur sémantique, renvoient au processus de reformulation (dire les choses autrement).

Les résultats de notre étude prendront la forme d'un corpus annoté sur les reformulations médicales en français. Ce corpus, disponible en format brut et annoté, pourrait servir aux chercheurs et industriels à l'exploitation des données médicales structurées (pour l'adaptation des textes spécialisés au grand public, la génération automatique des simplifications médicales, les chatbots, etc.). Informer de manière adaptée les patients, ou tout simplement le grand public, sur les maladies, les traitements médicaux et les recherches médicales est un enjeu important dans notre société.

Cet article est structuré de la façon suivante : la section 2 présente l'état de l'art, suivi par notre méthodologie, les outils et les corpus exploités (dans la section 3). Ensuite, nous réalisons l'analyse des résultats de l'annotation des termes médicaux et de l'identification des marqueurs et indicateurs de reformulation (section 4). La section 5 présente l'analyse lexicale et sémantico-pragmatique de reformulations et la discussion des hypothèses de recherche pour finir avec la conclusion et les perspectives dans la section 6.

2 Contexte

2.1 Les reformulations

La reformulation représente le processus de réécriture qui a le rôle d'expliquer, simplifier ou pointer une phrase ou un syntagme. Plusieurs recherches ont été menées au fil des années sur la reformulation (Gühlich et Kotschi, 1983 ; Fuchs, 1994 ; Rossari, 1990 ; Vassiliadou, 2013 ; Grabar et Eshkol-Taravella, 2016 ; Eshkol-Taravella et Grabar, 2017 ;

Steuckardt, 2018 ; Fuchs, 2020 ; Pennec, 2020 ; Vassiliadou, 2020). Compte tenu de la grande complexité des définitions, classifications et approches sur la reformulation, l'identification automatique de celle-ci est une opération difficile. Pour cette étude, nous prenons en compte les acceptations suivantes sur la reformulation :

- *La reformulation paraphrastique* conserve le sens et va vers une équivalence sémantique (Fuchs, 2020 ; Pennec, 2020 ; Vassiliadou, 2020) ;
- *La reformulation sous-phrastique*, une traduction intra-linguale (traduction avec des éléments du même système linguistique) qui ne dépasse pas la longueur syntaxique d'une phrase (notre proposition développée à partir de *la paraphrase sous-phrastique* de Bouamor, (2012)) ;
- *La paraphrase* exprime une équivalence basée sur un noyau sémantique commun (Fuchs, 1982 ; Bouamor, 2012 ; Kampeera, 2013 ; Pennec, 2020) ;
- *La reformulation non-paraphrastique* exprime un changement de perspective énonciative (Rossari, 1990 ; Fuchs, 1994).

Nous choisissons les *reformulations paraphrastiques* parce que nous souhaitons identifier des données équivalentes au niveau sémantique. Nous identifions de façon automatique des rapports d'équivalence entre un terme médical trop technique et difficile à comprendre et sa reformulation dans des mots plus simples. Nous cherchons des reformulations uniquement dans le cadre de la même phrase, donc des *reformulations sous-phrastiques*. Ce choix est justifié par l'objectif final de notre étude, celui de créer un corpus de phrases qui contiennent des reformulations médicales. Nous considérons que la reformulation sous-phrastique est plus facilement identifiable de façon automatique (par exemple à l'aide des marqueurs de reformulation ou des structures de n-grammes). La reformulation sous-phrastique nous intéresse pour sa dimension sémantique et la possibilité d'identifier des structures synonymiques simples et complexes, mais surtout précises au niveau notionnel, du terme reformulé.

Le besoin d'identifier des termes médicaux qui sont en relation d'équivalence sémantique avec un syntagme de la langue générale nous oriente également vers *la paraphrase*, l'exemple prototypique de ce type de relation. Le but de notre travail de recherche est de trouver le plus grand nombre possible de reformulations médicales, et pour cela nous annotons et nous analysons également *les reformulations non paraphrastiques,* dans la mesure où le reformulé donne une précision, explication, définition du terme médical (le formulé) ou exprime une cause. Nous cherchons toutes les reformulations qui peuvent servir à la vulgarisation des textes médicaux, dans le but ultérieur de créer un corpus de reformulations et une base de données utile pour la simplification des notions médicales (Cardon, 2021 ; Grabar et Hamon, 2015 ; Grabar et Hamon, 2016).

2.2 Termes médicaux

Le travail de reformulation implique également d'identifier l'élément reformulé, dans notre cas, le terme médical. Le but de la reformulation des termes médicaux est de trouver une équivalence de sens dans un registre de langue commune, adaptée à un public non-spécialiste, enfant, patients ou grand public (Brouwers et al., 2012 ; Pecout, Tran et Grabar, 2019). Plusieurs méthodes ont été exploitées dans la littérature, comme la recherche de préfixes / suffixes latins ou grecs, par exemple identifier « myocardique » à travers les bases *myo* = muscle et *carde* = cœur (Grabar et Hamon, 2016), à l'aide des ontologies médicales, comme Snomed International (Côté, 1996), ou avec des outils de détection de termes avec patrons de n-grammes (Buhnila, 2018). Si Grabar et Hamon (2016) se concentrent sur les composés néoclassiques simples (« cholécystectomie »), nous traitons également les termes médicaux polylexicaux en cherchant dans leurs contextes des

marqueurs ou indicateurs de reformulation issus de la littérature et nos propres observations sur le corpus médical. De plus, notre méthode est différente de Buhnila (2018), car nous faisons appel à l'ontologie médicale SNOMED-3.5VF (Côté, 1996) pour l'extraction des termes et nous limitons les contextes à l'aide des marqueurs et indicateurs de reformulation.

2.3 Marqueurs et indicateurs de reformulation

Une fois les termes médicaux identifiés, nous avons besoin *d'indices* pour identifier automatiquement la reformulation. Ces indices peuvent être des marqueurs ou des indicateurs de reformulation (Fuchs, 2020 ; Steuckardt, 2018). Plusieurs études sont menées sur les marqueurs de reformulations basés sur le verbe « dire », comme « c'est-à-dire », « ça veut dire », « pour dire autrement », « autrement dit » (Vassiliadou, 2013 ; Steuckardt, 2018 ; Magri, 2018). Ces marqueurs peuvent avoir un rôle discursif, justificatif ou paraphrastique. Selon Vassiliadou (2013 : 10), le marqueur « c'est-à-dire », en tant que de marqueur typique de la paraphrase, « établit des relations d'équivalence du type définition (normée et / ou naturelle), traduction, transcodage, passage d'un registre de langue à un autre ». L'emploi de « c'est-à-dire » aurait deux fonctions majeures : *l'équivalence* ou *l'identité* entre deux segments (« X et Y sont en rapport d'équivalence sémantique ») et une visée *explicative* (dont Y sert à justifier ou expliciter X) (Vassiliadou 2013 : 10). Dans notre étude nous nous intéressons à la fonction d'équivalence ou d'identité sémantique des marqueurs de reformulations basés sur le verbe « dire ».

Grabar et Eshkol-Taravella (2016) ont mené une étude sur la reformulation lexicale qui s'identifie avec des marqueurs de type « c'est-à-dire », « disons », « ça veut dire » à l'aide d'un système à base de règles et des annotations manuelles. Leur étude a l'objectif de classer automatiquement les syntagmes qui ont le rôle de reformulation et ceux qui ne l'ont pas. Pour les identifier, Grabar et Eshkol-Taravella (2016) prennent en compte la structure syntagmatique « S1 marker S2 », dont S1 est le l'élément reformulé et S2 la reformulation, les deux parties étant liées par les marqueurs de reformulation. Leur étude est menée sur deux corpus oraux de la langue générale et sur un corpus de texte de forums médicaux.

Vassilliadou (2020) se questionne sur la possibilité d'insérer ou supprimer des marqueurs de reformulation et quel impact cette opération a dans l'identification de la reformulation. Roulet et al. (2001 : 170) proposaient le test d'insertion pour identifier la nature d'une relation dans le discours. Ce test est considéré comme étant un indice de la pertinence d'une relation discursive. Pour Vassilliadou (2020 : 87), le test de la suppression est plus fiable que celui de l'insertion, car le fait de pouvoir maintenir une reformulation après suppression du marqueur sera l'indice d'une relation sémantique bien nouée entre le formulé et le reformulé. Vassilliadou (2020) note que, selon l'analyse d'un grand nombre de reformulations, cette possibilité de garder la reformulation après la suppression du marqueur est réalisable plutôt dans le cas des reformulations paraphrastiques. Ces reformulations présenteraient alors des relations sémantiques dites « identiques », comme la synonymie, la définition, l'hyperonymie, l'hyponymie, la traduction ou même l'étymologie.

Nous testons cette idée de l'absence de marqueur de reformulation en cherchant des *indicateurs de reformulation* spécifiques au domaine médical. Nous classifions les indicateurs de reformulation en trois types :

- *les indicateurs de langue générale* qui, par leur sémantique et leur utilisation dans le discours, renvoient à la simplification, la définition ou l'explication des notions (« définition » et « défini/e », etc.) ;
- *les indicateurs grammaticaux* qui annoncent une énumération d'hyponymes du terme médical (« tel que », « par exemple ») ;

- *les indicateurs spécifiques* au domaine médical qui sont des hyperonymes des termes médicaux (« maladie », « affection », « trouble »).

Nous évaluons manuellement les phrases pour trouver davantage de reformulations sans marqueur ni indicateur lexical (mais marquées par exemple avec des *indicateurs typographiques*, selon Steuckardt (2018), comme les parenthèses ou la virgule). Par rapport à l'étude de Grabar et Eshkol-Taravella (2016), nous analysons les reformulations des termes médicaux dans des textes médicaux écrits (articles scientifiques et de vulgarisation) afin de créer un corpus de reformulations des termes. Nous rajoutons d'autres marqueurs et indicateurs de reformulation dans notre recherche, que nous présentons par la suite dans la section 3.3.

3 Méthodologie

Nous travaillons pour cette étude sur le corpus CLEAR (Grabar et Cardon, 2018), un corpus de textes de la littérature médicale et des textes médicaux destinés au grand public en français. Notre méthode consiste d'abord à identifier de façon automatique des termes médicaux simples et polylexicaux avec l'annotateur SIFR-BioPortal (Tchechmedjiev et al., 2018). L'outil cherche de façon automatique les termes médicaux dans notre corpus à partir de l'ontologie médicale SNOMED-3.5VF (Côté, 1996) (diffusée par ASIP Santé) qui contient 150 906 concepts médicaux. Pour identifier des indicateurs spécifiques au domaine médical, nous prenons en compte le contexte du terme médical, selon la notion de « Knowledge Rich Context » proposée par Meyer (2001). Dans ce sens, nous cherchons les mots qui indiquent la présence d'un terme et des relations que ce terme peut avoir avec d'autres éléments de la phrase. Plus concrètement, nous lançons une recherche automatique sur les marqueurs qui indiquent des relations (Condamines, 2018) de type hypéronymie, hyponymie, synonymie et méronymie qui lient un terme médical à sa reformulation (Ramadier, 2016).
Dans cette étude, nous faisons l'hypothèse de trouver des reformulations dans le contexte du terme médical, dans la même phrase. Une fois les termes identifiés, nous cherchons automatiquement les marqueurs et les indicateurs de reformulation. Nous évaluons manuellement les phrases qui contiennent les termes médicaux et respectivement les marqueurs / indicateurs pour identifier les reformulations correctes. Celles-ci sont également analysées et annotées de point de vue lexical et sémantico-pragmatique, annotation que nous présentons en détail dans la section 3.4. À partir de ces résultats nous constituons un corpus de reformulations médicales qui pourrait servir comme ressource textuelle exploitable en traitement automatique de langues (TAL). Nous présentons chaque étape de la méthodologie en détail par la suite.

3.1 Données linguistiques

Nous travaillons sur le corpus CLEAR (Grabar et Cardon, 2018) et plus précisément sur le sous-corpus Cochrane. Ce corpus comparable est constitué des résumés scientifiques du domaine médical destinés aux experts et des résumés simplifiés pour le grand public. Les résumés ont été rédigés par les chercheurs de la fondation Cochrane. Grabar et Cardon (2018) ont recueilli en novembre 2017 un nombre de 8789 résumés, dont 3815 traitent en double le même concept médical : la maladie du motoneurone, l'asthme, etc. Le corpus expert contient 2 840 003 tokens et le corpus grand public 1 515 051 tokens.

Le corpus CLEAR Cochrane se présente sous la forme d'un texte expert suivi par un texte grand public sur le même thème. Afin de pouvoir exploiter et analyser les corpus

séparément pour des travaux futurs, nous avons séparé le corpus en deux sous-parties : corpus scientifique (CLEAR EX) et corpus pour le grand public (CLEAR GP). Nous partons de l'hypothèse que les textes pour les experts ont plus de termes médicaux et les textes grand public contiennent plus de synonymes, des reformulations ou explications dans un langage simplifié. Nous avons découpé les textes en phrases à l'aide des caractères de fin de ligne (. ; ! ; ?) pour afficher une phrase par ligne, dans le but de constituer un corpus avec des phrases qui contiennent des reformulations. Une fois le corpus nettoyé et aligné, nous avons procédé à l'identification automatique des termes médicaux.

CLEAR Cochrane (Grabar et Cardon, 2018)	**Résumés en total**	**Résumés sur le même thème**	**Taille (mots)**
Scientifique (EX)	8 789	3 815	2 840 003
Grand public (GP)			1 515 051

Tableau 1. Taille du corpus CLEAR Cochrane par type de texte (Grabar et Cardon, 2018).

3.2 Annotation automatique des termes médicaux

Pour notre étude nous identifions les termes médicaux présents dans nos données avec l'annotateur SIFR-BioPortal (Tchechmedjiev et al., 2018) dans sa version française, utilisé pour nos expériences à l'aide d'un script en Perl. Cet outil met à disposition 28 terminologies et ontologies médicales en français. Nous utilisons l'ontologie SNOMED-3.5VF qui contient une grande variété des notions médicales : administratif médical et traitements, agents, anatomie, diagnostics, organismes vivants, médicaments, symptômes, anatomie, maladie, procédures, substances, etc. Cette variété de concepts médicaux et la recherche par lemme nous permet de trouver une gamme large de termes médicaux dans le corpus CLEAR, dans la partie scientifique et vulgarisée également.

3.3 Annotation semi-automatique des marqueurs et indicateurs de reformulation

Nous créons une liste des **marqueurs de reformulation** en partant de la littérature, comme les marqueurs :

- formés sur le verbe « dire » (*c'est-à-dire, ça veut dire / veut dire, pour dire autrement, autrement dit*) (Vassiliadou, 2013 ; Grabar et Eshkol-Taravella, 2016 ; Steuckardt, 2018 ; Magri, 2018) ;
- dérivés du verbe « désigner » ou « signifier » (Péry-Woodley et Rebeyrolle, 1998 ; Charolles et Coltier, 1986) ;
- dérivés du verbe « être » avec ses différentes formes morphologiques, « est une/e/des », « sont un/une/des » (Meyer, 2001 ; Grabar et Hamon, 2016) suivis par des hyperonymes du domaine médical comme « maladie », « affection » et « trouble » ;
- observés dans nos corpus, comme ceux formés sur le verbe « appeler » (*qu'on appelle ce, que l'on appelle, est aussi appelé / aussi appelé)* et d'autres, de type « doit être compris comme », « au sens de ».

Ces marqueurs de reformulation sont indépendants du domaine (à l'exception des hypéronymes médicaux) et peuvent indiquer des différents types de relations entre le terme médical et sa reformulation, relations que nous développons par la suite dans la section 3.4.

Nous travaillons dans cette étude avec la notion d'**indicateur de reformulation**. Steuckardt (2018) considère que les *indicateurs de reformulation* sont les indices

prosodiques (la saillance, la mélodie) et les marques typographiques (la virgule, les points de suspension, les pauses à l'oral). Nous incluons dans cette catégorie les unités qui ne sont pas considérées comme *marqueurs typiques* de la reformulation dans la littérature, par exemple la locution conjonctive « c'est-à-dire » qui est considérée un marqueur typique de la reformulation (Vassiliadou, 2013 ; Grabar et Eshkol-Taravella, 2016 ; Steuckardt, 2018) (cf. liste présentée en début de cette section).

Nous incluons dans la liste des indicateurs de reformulation les hyperonymes de langue générale (Săpoiu, 2013), comme « affection », « maladie », « trouble ». Ces hyperonymes aident à classer des notions médicales très techniques (« typhus des broussailles ») dans des catégories plus faciles à comprendre pour le grand public (« maladie bactérienne ») (Grabar et Hamon, 2015). Les indicateurs lexicaux définitoires, comme « définition », « défini/e », « défini/e comme », placés dans le texte avant ou après le terme, peuvent annoncer la reformulation d'un terme. Les indicateurs grammaticaux « tel que » et « par exemple » annoncent la reformulation à travers une exemplification. Dans ce cas, le terme (« antibiotiques ») est l'hyperonyme et la reformulation simplifie le sens du terme à l'aide d'hyponymes : (« chloramphénicol, tétracycline et doxycycline »). L'indicateur grammatical « ou » qui exprime l'altérité peut introduire des synonymes ou paraphrases du terme. Dans nos recherches automatiques, « ou » a été recherché en dernier pour réduire le nombre de résultats erronés, vu la grande diversité d'utilisation de cette conjonction disjonctive dans la langue et le discours.

Marqueurs de reformulation	Indicateurs de reformulation
c'est-à-dire	affection / s
ça veut dire / veut dire	maladie / s
pour dire autrement	trouble / s
autrement dit	définition / s
signifie	défini / e / s / es
désigne	défini / e / s / es comme
ce qu'on appelle	tel / lle / s / lles que
ce que l'on appelle	par exemple
est aussi appelé / aussi appelé	ou
doit être compris/e comme	
au sens de	
est un / une ; sont des / un / une • affection/s • maladie/s • trouble/s	

Tableau 2. Listes de marqueurs et d'indicateurs de reformulation.

Les phrases qui contiennent des termes médicaux et des marqueurs et / ou indicateurs de reformulations sont extraites. Ces phrases sont par la suite annotées semi-automatiquement et analysées manuellement du point de vue lexical et sémantico-pragmatique selon la typologie ci-dessous.

3.4 Annotation des reformulations médicales

3.4.1 Fonctions sémantico-pragmatiques correspondant aux relations lexicales

Nous annotons les termes, les reformulations, les relations lexicales et les fonctions sémantico-pragmatiques. *Les relations lexicales* montrent le lien lexical entre les deux segments, le terme médical et la reformulation. *Les fonctions sémantico-pragmatiques*

représentent les raisons qui poussent le locuteur à utiliser la reformulation. Les définitions des relations lexicales et des fonctions sémantico-pragmatiques sont inspirées de la taxinomie d'Eshkol-Taravella et Grabar (2017) et adaptées par nous au type de texte écrit du domaine médical. Pour faciliter la lecture de nos exemples, nous surlignons les termes médicaux, nous présentons les marqueurs ou indicateurs en italique et les reformulations en gras. Dans cette étude, nous testons l'hypothèse que les fonctions sémantico-pragmatiques sont corrélées avec les relations lexicales, comme suit :

- *la paraphrase* (en lien avec *la synonymie*) : le sens du terme est exprimé dans la reformulation avec d'autres mots dans le but de simplifier le terme, tout en gardant une relation lexicale d'équivalence sémantique (la *synonymie*).

(1) Orthophonie versus placebo *ou* **absence d'intervention** pour le traitement des troubles de la parole dans la maladie de Parkinson.

- *la dénomination* (en lien avec *la synonymie*) : le terme est reformulé à l'aide d'un autre nom (ou terme), en gardant une relation lexicale d'équivalence sémantique (la *synonymie*), mais sans l'intention d'expliciter ou simplifier le terme reformulé.

(2) Cependant, on ignore si ces médicaments sont bénéfiques chez les personnes atteintes de broncho-pneumopathie chronique obstructive (BPCO, *c'est-à-dire* **bronchite chronique** *ou* **emphysème**, ou les deux).

- *la définition* (en lien avec *l'hyperonymie* et *l'hyponymie*) : le terme est *défini*, car il est considéré comme étant trop technique ou spécialisé et donc, difficile à comprendre, à travers un mot / syntagme générique (*hyperonymie*) (3) ou spécifique (*hyponymie*) (4). Nous rajoutons notre analyse de marqueurs et indicateurs : « est un/e » et les indicateurs hyperonymiques « affection », « maladie », « trouble ».

(3) Le typhus des broussailles *est une* **maladie bactérienne** prévalente dans les régions de l'Asie et du Pacifique.

- *l'exemplification* (en lien avec *l'hyponymie*) : la reformulation est constituée d'exemples qui aident à illustrer le sens du terme à travers plusieurs entités du même type (des sous-types spécifiques).

(4) Des antibiotiques *(***chloramphénicol, tétracycline et doxycycline***)* sont utilisés dans le traitement de cette maladie.

- *l'explication* (en lien avec *la méronymie*) : le terme est suivi par une situation ou une procédure en particulier et la reformulation donne une explication en apportant des détails en plus sur une partie / composante.

(5) Le programme de réadaptation devait avoir été multidisciplinaire, (*c'est-à-dire* **comprendre une consultation médicale associée à une intervention psychologique, sociale ou professionnelle, soit une combinaison de celles-ci**).

Nous analysons manuellement nos données pour identifier ces différents types de relations (hyperonymie, hyponymie, synonymie, méronymie et d'autres, par exemple, cause) et nous classons les marqueurs et indicateurs par rapport aux relations et fonctions marquées, afin de confirmer ou infirmer les hypothèses présentées ci-dessus.

4 Analyse des résultats d'annotation

4.1 Analyse sémantique de l'annotation des termes

Lors de notre analyse de listes de termes uniques issus lors de l'annotation, nous avons remarqué la présence de mots de la langue commune comme « après », « trois », « une durée ». Après avoir analysé la composition de la base terminologique SNOMED-3.5VF, nous avons observé qu'elle est constituée de plusieurs classes de termes. Nous avons analysé toutes les classes et nous avons observé que celle qui contient le plus de mots de la langue générale est la liste de *modificateurs*. Nous avons extrait cette liste de 1510 mots et avons délimité les numéros, les adverbes et les prépositions afin de les supprimer de manière automatique de notre liste de termes annotés.

La classe de modificateurs contient des sous-classes. Nous avons décidé de garder la sous-classe « nom-adjectif » dans son intégralité parce qu'elle contient des termes à usage médical assez techniques, comme « en rémission », « phase précoce », « stade intermédiaire » dont le sens peut être inconnu au grand public. Nous avons supprimé également la sous-classe « termes relationnels » qui contient des mots qui ne sont essentiels à l'identification d'un terme médical, comme « après », « par suite de », « suivant », « délimité par », « contrôlé par », etc. La liste de mots à supprimer contient 286 éléments.

CLEAR Cochrane (Grabar et Cardon, 2018)	Termes annotés avec SIFR-BioPortal	Termes uniques sans doublons	Termes type « modificateurs » enlevés	Termes uniques après nettoyage
Scientifique (EX)	184 446	5718	140	**5578**
Grand public (GP)	125 696	5246	146	**5100**

Tableau 3. Termes médicaux annotés par type de corpus.

Nous avons gardé tous les modificateurs importants dans la structure de termes médicaux, comme : « néoplasie récidivante », « tumeur récidivante », « chevauchement néoplasique », etc. Nous utilisons un script afin d'extraire tous les mots de cette liste qui peuvent se retrouver dans nos listes de termes médicaux uniques annotés avec SIFR-BioPortal. Une fois ces mots supprimés, nous avons une liste de 5578 termes uniques simples et polylexicaux dans le corpus scientifique et 5100 dans le corpus grand public.

Une fois les listes nettoyées, nous extrayons automatiquement avec nos propres scripts les phrases qui contiennent à la fois les termes médicaux annotés par SIFR-BioPortal et des occurrences de marqueurs ou indicateurs de reformulation. Nous adaptons notre script pour identifier toutes les formes morphologiques et pour annoter automatiquement les termes médicaux et les marqueurs / indicateurs. Nous obtenons 4687 phrases pour le corpus de textes scientifiques (CLEAR EX) et 3975 phrases pour le corpus de textes médicaux pour le grand public (CLEAR GP).

CLEAR Cochrane (Grabar et Cardon, 2018)	N° total des phrases avec termes médicaux	N° de phrases avec termes, mais sans marqueurs / indicateurs	N° de phrases avec termes médicaux et marqueurs / indicateurs
Scientifique (EX)	71 585	66 899	**4687**
Grand public (GP)	46 788	42 814	**3975**

Tableau 4. Phrases qui contiennent des termes médicaux et des marqueurs / indicateurs de reformulation.

Notre prochaine étape de traitement est l'analyse des marqueurs et indicateurs de reformulation afin de tester nos théories sur le lien qui peut exister entre ces unités, les relations lexicales et les fonctions sémantico-pragmatiques des reformulations.

4.2 Analyse des marqueurs et indicateurs de reformulation

Nous présentons les résultats obtenus sur un échantillon de 2000 phrases, dont 1000 phrases du corpus de textes scientifiques, CLEAR EX (avec 36 182 tokens) et 1000 du corpus de textes pour le grand public, CLEAR GP (avec 29 863 tokens). Notre méthode consiste à chercher dans chaque sous-corpus les marqueurs et indicateurs de reformulation et d'analyser leur fréquence absolue (le nombre d'occurrences) et leur fréquence relative en pourcentage (par rapport au nombre total de tokens du corpus analysé). Pour les formes morphologiques différentes (par exemple « affection » et « affections », au singulier et au pluriel), nous calculons les fréquences relatives pour les deux formes. Nous remarquons que, parmi les marqueurs, les plus fréquents sont ceux formés avec le verbe « être » dans la structure « est /une maladie / affection / trouble » avec 245 occurrences, suivi par « c'est-à-dire » avec 51 occurrences et par « signifie » avec 44 occurrences. Parmi les marqueurs le moins fréquents se trouvent « veut dire » avec une seule occurrence dans le corpus grand public, « est aussi appelé(e)/aussi appelé(e) » avec 2 occurrences dans le corpus expert et seulement une dans le GP et « désigne » avec 4 occurrences dans le GP. Les marqueurs de reformulation « pour dire autrement », « autrement dit », « ce qu'on appelle / ce que l'on appelle », « doit être compris comme » et « au sens de » n'ont pas été retrouvés dans les corpus d'étude.

Marqueurs de reformulation	**Fréquences absolues et relatives**			
	CLEAR EX		**CLEAR GP**	
	Fréq. absolue	**Fréq. rel. %**	**Fréq. absolue**	**Fréq. rel. %**
est un/une ; sont des/un/une	95 dont	0.262	150 dont	0.502
• affection/s	- 18	0.049	- 28	0.093
• maladie/s	- 36	0.099	- 62	0.207
• trouble/s	- 22	0.060	- 30	0.100
c'est-à-dire / c'est à dire	26	0.071	25	0.083
signifie	11	0.030	34	0.113
Indicateurs de reformulation				
affection/s	76/44(s) (*-18*)	0.160	69/9(s) (*-24*)	0.150
maladie/s	535/153(s) (*-36*)	0.013	582/120(s) (*-62*)	1.741
trouble/s	274/215(s) (*-22*)	0.696	222/162(s) (*-30*)	0.642
défini/e/s/es	66/30(e) (*-21*)	0.124	18/6(e) (*-5*)	0.043
défini/e/s/es comme	21	0.058	5	0.016
définition/s	11	0.030	3	0.010
tel / lle / s / lles que	28	0.077	44	0.147
par exemple	46	0.127	76	0.254
ou	260	0.718	253	0.847
Total	**1431**	**3.955**	**1480**	**4.955**

Tableau 5. Fréquences absolues et relatives de marqueurs et indicateurs de reformulation identifiés le plus fréquents.

En ce qui concerne les indicateurs de reformulations, les hyperonymes du domaine médical sont très nombreux : 1117 pour « maladie », 496 pour « trouble » et 145 pour

« affection », suivis par les indicateurs grammaticaux « par exemple » avec 122 occurrences et « tel que » avec 72 occurrences. Pour que les calculs soient corrects, nous avons enlevé les occurrences d'hyperonymes du domaine quand ils font partie du marqueur « est une maladie / affection / trouble » (représenté en italique et entre parenthèses dans le Tableau 5). La conjonction disjonctive « ou », avec 513 occurrences, a fait l'objet d'une recherche automatique afin d'analyser son impact dans la reformulation. Nous remarquons que dans le corpus pour le grand public, les marqueurs et indicateurs ont une fréquence relative plus élevée (4.955%) que dans le corpus de textes scientifiques (3.955%), ce qui soutient l'hypothèse que les textes de vulgarisation peuvent contenir un nombre plus grand de reformulations. La prochaine étape de notre expérience, l'analyse et l'évaluation manuelle des phrases, nous permet de déterminer le pourcentage de marqueurs et d'indicateurs qui aident à identifier des reformulations correctes.

5 Analyse lexicale et sémantico-pragmatique des reformulations

Nous présentons les résultats des analyses réalisées sur l'échantillon de 1000 phrases du corpus de textes scientifiques, CLEAR EX. Suite à cette analyse, nous avons identifié 314 reformulations médicales correctes, dont au moins un terme médical est reformulé. Le Tableau 5 ci-dessous est un classement des reformulations correctes par type de relation lexicale en rapport avec la relation sémantico-pragmatique de la reformulation. Nous observons que les plus fréquentes reformulations dans le corpus scientifique sont les reformulations de type *définition*, dont 211 ont la relation lexicale *d'hyperonymie / hyponymie*, suivies par les *exemplifications*, en lien avec la relation *d'hyponymie*, avec 50 occurrences et la *dénomination* en lien avec la *synonymie* avec 26 occurrences. La *paraphrase*, dire les choses autrement avec des mots plus simples, est très peu présente, avec seulement 8 occurrences. Même constat pour *l'explication*, dont nous avons identifié que 5 occurrences. Cette observation peut s'expliquer par le type de texte dont les phrases font partie, le texte scientifique. Ce type de texte fait rarement usage des paraphrases simplifiées et des explications pour rendre accessible le sens d'un terme médical, vu qu'il est destiné à un public expert. Dans des prochaines études nous analyserons le corpus de textes médicaux destinés au grand public, car nous faisons l'hypothèse d'y trouver un nombre plus grand d'occurrences pour ces deux types de reformulations, la paraphrase et l'explication.

Le Tableau 5 montre la répartition selon nos hypothèses de ces 314 reformulations validées manuellement. Nous observons que la plupart des reformulations correspondent aux relations lexicales selon les hypothèses d'annotation présentées dans la section 3.4. Le plus faible pourcentage de correspondance (20%) est celui des *explications* en lien avec la *méronymie* car nous avons trouvé une seule explication de ce type (5), les autres sont plutôt des *descriptions* ou des expressions de la *causalité* (8).

(8) Une élévation significative de la pression artérielle peut être dangereuse *(par exemple*, **conduire à des accidents vasculaires cérébraux***)*, mais il existe peu d'informations sur la façon de prévenir ou de traiter l'hypertension du post-partum.

Type de reformulation	Relation lexicale	CLEAR EX	% de validité des hypothèses de recherche
paraphrase	synonymie	6 / 8	75%
dénomination		26	100%
définition	hyperonymie hyponymie	211 / 223	94,61%
exemplification		50 / 52	96,15%

explication	méronymie	1 / 5	20%
Total		**314**	**93,63%**

Tableau 5. Reformulations validées par type sémantico-pragmatique et relation lexicale.

La prochaine étape de notre analyse est de tester l'hypothèse selon laquelle certains marqueurs ou indicateurs seront utilisés plus fréquemment pour certains types de reformulations. Le Tableau 6 ci-dessous présente ces éléments par rapport au type de relation lexicale et fonction sémantico-pragmatique avec leur fréquence absolue et relative dans le corpus CLEAR EX. En plus de l'analyse précédente, nous observons la présence avec 50 occurrences du marqueur orthographique de type *parenthèses*, le plus identifié dans des fonctions *d'exemplification* (23 occurrences) et de *définition* (11 occurrences).

Analyse lexicale et sémantico-pragmatique des reformulations correctes Marqueurs et indicateurs par fréquence absolue et relative						
	synonymie	**hyperony-mie**	**hyponymie**	**méronymie/ description**	**CLEAR EX**	
					Fréq. absolue	**Fréq. relative**
paraphrase	- () - c.à.d				6 1	0.016 0.002
dénomina-tion	- () + maladie /trouble - / - ou - aussi appelé - c'est-à-dire - également nommée				8 7 5 2 2 1	0.022 0.019 0.013 0.005 0.005 0.002
définition		- est un/e/la/l'+ sont un/une/des • maladie/s • affection/s • trouble/s • problème, agent, etc. - et/ou les/autres/d'autres - défini/e comme - () , - c'est-à-dire - définition			85+12 29 16 11 7 2	0.268 0.080 0.044 0.030 0.019 0.005
exemplifi-cation			- () - par exemple - tel que + maladie / trouble - c'est-à-dire - notamment - y compris		23 14 9 3 1 1	0.063 0.038 0.024 0.008 0.002 0.002
explication				- c'est-à-dire - ()	3 2	0.008 0.005

Tableau 6. Marqueurs et indicateurs par relation sémantico-pragmatique lexicale.

Nous avons identifié également de nouveaux marqueurs et indicateurs de reformulation lors de notre analyse manuelle des phrases :

- *nouveaux indicateurs de reformulation*, comme « et autres », « et les autres », « ou autres », « ou d'autres » dans des relations lexicales *d'hyperonymie*, avec la fonction sémantico-pragmatique de *dénomination* (« la mucoviscidose et autres maladies génétiques » ; « la démence et les autres troubles cognitifs ») ;

- *nouveaux indicateurs de type hyperonymes spécifiques au domaine médical*, comme « problème de santé publique », « agent », « symptôme » (« L'agressivité *est un problème de santé publique* majeur directement associé à plusieurs troubles mentaux. ») ;

- *nouveaux marqueurs* : « décrit », « est associé à » (« L'aphasie *décrit* un trouble du langage associé à une lésion cérébrale »).

La conjonction disjonctive « ou » aide à identifier 32 reformulations correctes (toute seule ou dans des constructions de types « ou autres », « ou d'autres » (« schizophrénie *ou autre* maladie mentale grave »). Les hyperonymes génériques médicaux font parties de 236 reformulations médicales correctes. Notre méthode d'identification des reformulations à une précision de 31.40% pour le corpus expert et de 48.80% pour le corpus grand public.

6 Conclusion et perspectives

Cette étude montre une méthodologie d'identification de reformulations médicales dans des corpus médicaux en français. Nos annotations automatiques permettent de cibler les phrases qui ont le plus grand potentiel de contenir des reformulations, à travers l'identification des termes médicaux et des marqueurs et indicateurs de reformulation. Notre méthode est transposable à d'autres domaines scientifiques, car les marqueurs et indicateurs de reformulation font partie du langage général et aident à l'identification des relations entre des termes (contexte définitoire, hypéronymie, méronymie ou synonymie). La méthode est applicable à d'autres langues latines proches du français, comme le roumain (Buhnila, 2021, article à paraître).

Nos contributions sont l'analyse lexicale et sémantico-pragmatique des 314 reformulations correctes identifiées et des marqueurs / indicateurs qui met en lumière les différents procédés linguistiques (hyperonymie, hyponymie, synonymie) utiles dans la construction des reformulations. Ces résultats feront partie d'un corpus de reformulations médicales qui sera agrandit avec l'analyse de la totalité des phrases extraites (en nombre de 8 662 des deux corpus, EX et GP). Nous envisageons d'augmenter la précision de notre méthode en exploitant également les abréviations pour faciliter l'identification de nouvelles reformulations. Le corpus de reformulations servira, dans nos futurs travaux, à créer des corpus de test et d'entrainement des outils d'apprentissage automatique avec des réseaux de neurones (Nighojkar et Licato, 2021) pour la recherche automatique des reformulations médicales.

Références bibliographiques

Bouamor, H. (2012). Etude de la paraphrase sous-phrastique en traitement automatique des langues. *Université Paris Sud - Paris XI.* Français. ⟨NNT : 2012PA112100⟩. ⟨tel-00717702⟩.

Bowker, L. et Pearson, J. (2002). Working with Specialized Language. A Practical Guide to Using *Corpora. Londres, New York: Routledge.*

Brouwers, L., Bernhard D., Ligozat A.-L. et François T. (2012). Simplification syntaxique de phrases pour le français (syntactic simplification for french sentences) [in French]. *In Actes de la conférence conjointe JEP-TALN-RECITAL 2012, 2 : TALN :211–224.* Grenoble, France : ATALA/AFCP.

Buhnila, I. (2018). Simplification lexicale entre les textes scientifiques et les textes de vulgarisation du domaine de la médecine. *Mémoire de Master, Université de Strasbourg,* Strasbourg, France.

Buhnila, I. (2021). Building a Corpus of Medical Paraphrases in Romanian. *In Proceedings of the The 16th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2021* (à paraître).

Cardon, R. (2021). Simplification automatique de textes techniques et spécialisés. *Informatique et langage [cs.CL]. Université de Lille.* Français. ⟨NNT : 2021LILUH007⟩. ⟨tel-03343769v2⟩.

Charolles, M. et Coltier, D. (1986). Le contrôle de la compréhension dans une activité rédactionnelle : l'exemple des reformulations paraphrastiques. *Pratiques 49 (1):* 51-66. https://doi.org/10.3406/prati.1986.2450.

Condamines, A. (2018). Nouvelles perspectives pour la terminologie textuelle. *J. Altmanova; M. Centrella; K.E. Russo. Terminology and Discourse, Peter Lang*, 1-13. 978-3-0343-2415-1. ff10.3726/978-3-0343-2414-4ff. ffhalshs-01899150f.

Contente, M. (2005). Termes et textes : la construction du sens dans la terminologie médicale. *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction,* 453-65. Bruxelles, Belgique.

Costa, R. (2005). Texte, terme et contexte. *Actes des septièmes Journées scientifiques du réseau de chercheurs Lexicologie Terminologie Traduction*, 79-88. Bruxelles, Belgique.

Côté, R. (1996). Répertoire d'anatomopathologie de la SNOMED internationale, v3.4. *Université de Sherbrooke,* Sherbrooke, Québec.

Eshkol-Taravella, I. et Grabar, N. (2017). Taxinomie dans les paraphrases du point de vue de la linguistique de corpus, *Syntaxe et Sémantique*, vol. 18, no. 1, 149-184.

Fuchs, C. (1982). *La Paraphrase*. PUF, Paris, 184 pages.

Fuchs, C. (1994). *Paraphrase et énonciation*. Editions OPHRYS, 185 pages.

Fuchs, C. (2020). Paraphrase et reformulation : un chassé-croisé entre deux notions. *Autour de la reformulation, 36, Droz, Coll. Recherches et Rencontres*, 978-2-600-06051-6, 41-55.

Grabar, N., et Cardon R. (2018). CLEAR - Simple Corpus for Medical French. *In Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA).* Tilburg, the Netherlands: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-7002, 3–9.

Grabar, N., et Eshkol-Taravella, I. (2016). Disambiguation of occurrences of reformulation markers c'est-à-dire, disons, ça veut dire. *JADT 2016 : 13ème Journées internationales d'Analyse statistique des Données Textuelles*, 1-13. Nice, France.

Grabar, N., et Hamon T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. *22ème Traitement Automatique des Langues Naturelles, 14.* Caen, France.

Grabar, N., et Hamon, T. (2016). Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *Traitement Automatique des Langues, Varia, 57 (1)*, 85-109.

Gühlich E. et Kotschi T. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de Linguistique française 5*, 305-351.

Kampeera, W. (2013). Analyse linguistique et formalisation pour le traitement automatique de la paraphrase. *Linguistique. Université de Franche-Comté*. Français. ⟨NNT : 2013BESA1011⟩. ⟨tel-01288926⟩.

Lindberg D., Humphreys B., et Mccray A. (1993). The Unified Medical Language System, *Methods Inf Med, vol. 32, no 4,* 281-291.

Magri, V. (2018). Marqueurs de reformulation : exploration outillée et contrastive dans deux corpus narratifs. *Langages N° 212 (4)*, 35-50.

Meyer, I. (2001). Extracting Knowledge-Rich Contexts for Terminography: A conceptual and methodological framework. *Dans D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Éds), Recent Advances in Computational Terminology* (pp. 279-302). Amsterdam: John Benjamins.

Namer F., (2009). Morphologie, Lexique et TAL : l'analyseur DériF. *TIC et Sciences cognitives, Hermes Sciences Publishing*, London.

Nighojkar, A. et Licato, J. (2021). Improving Paraphrase Detection with the Adversarial Paraphrasing Task. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7106–7116, Online. Association for Computational Linguistics.

Pecout, A, Tran, T. M., et Grabar, N. (2019). Améliorer la diffusion de l'information sur la maladie d'Alzheimer : étude pilote sur la simplification de textes médicaux. *Ela. Etudes de linguistique appliquée N° 195 (3):* 325-41.

Pennec, B. (2020). Les reformulations : des formes méta-énonciatives par excellence. Spécificités et introducteurs. *Autour de la reformulation, 36, Droz*, *Coll. Recherches et Rencontres*, 57-75.

Péry-Woodley M. et Rebeyrolle J. (1998). Domain and genre in sublanguage text: definitional microtexts in three corpora, *LREC,* 987-992.

Ramadier, L. (2016). Indexation et apprentissage de termes et de relations à partir de comptes rendus de radiologie. *Informatique. Université Montpellier*, Français. ⟨NNT : 2016MONTT298⟩. ⟨tel-01479769v2⟩.

Rossari, C. (1990). Projet pour une typologie des opérations de reformulation. *Cahiers de linguistique française 11*, 345-359.

Roulet, E., Filliettaz L. et Grobet A. (2001). Un modèle et un instrument d'analyse de l'organisation du discours. Bern/Berlin/Bruxelles/New York/Oxford/Wien, *Peter Lang.*

Săpoiu, C. (2013). Hiponimia în terminologia medicală. Modalităţi de abordare în semantică şi lexicografie. *Piteşti, Editura Trend*, 199 pages.

Steuckardt, A. (2018). Les marqueurs de reformulation formés sur dire : exploration outillée. *Langages N° 212 (4)*, 17-34.

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Zevio, S. et Jonquet, C. (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. *BMC bioinformatics, 19(1)*, 405.

Vassiliadou, H. (2013). C'est-à-dire (que) : embrayeur d'énonciation. *Semen. Revue de sémio-linguistique des textes et discours, no 36 (octobre)*. 1-14. http://journals.openedition.org/semen/9684. https://doi.org/10.4000/semen.9684.

Vassiliadou, H. (2020). Peut-on aborder la notion de "reformulation" autrement que par la typologie des marqueurs ? Pour une analyse sémasiologique et onomasiologique. Olga Inkova. *Autour de la Reformulation, Droz*, 978-2-600-06051-6, 77-94.