Une méthode automatique de construction de corpus de reformulation

Ioana Buhnila

Université de Strasbourg

École doctorale des Humanités (ED520) Université de Strasbourg

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES HUMANITÉS

LiLPa UR 1339 (Linguistique, Langues, Parole)

THÈSE présentée par :

Ioana BUHNILA

soutenue le : **14 juin 2023**

pour obtenir le grade de : **Docteur de l'Université de Strasbourg**

Discipline / Spécialité : **Sciences du langage**

Informatique appliquée au langage

Une méthode automatique de construction de corpus de reformulation

THÈSE dirigée par :

Madame TODIRASCU Amalia	Professeur des Universités, Université de Strasbourg, Laboratoire LiLPa UR 1339
Monsieur TUFIȘ Dan	Professeur, Membre de l'Académie, Institut de Recherche sur l'Intelligence Artificielle de l'Académie Roumaine de Bucarest

RAPPORTEURS :

Madame CISLARU Georgeta	Professeur des Universités, Université Paris Nanterre, Laboratoire MoDyCo UMR 7114
Monsieur CONSTANT Mathieu	Professeur des Universités, Université de Lorraine, Laboratoire ATILF UMR 7118 CNRS

AUTRES MEMBRES DU JURY :

Madame ESHKOL-TARAVELLA Iris	Professeur des Universités, Université Paris Nanterre, Laboratoire MoDyCo UMR 7114
Madame BARBU-MITITELU Verginica	Chercheur, Docteur à l'Institut de Recherche sur l'Intelligence Artificielle de l'Académie Roumaine de Bucarest
Monsieur GRASS Thierry	Professeur des Universités, Université de Strasbourg, Laboratoire LiLPa UR 1339

Une méthode automatique de construction de corpus de reformulation

Thèse de Doctorat

Recherche réalisée par

Ioana BUHNILA

Sous la direction de

Amalia TODIRASCU
Professeur des Universités
LiLPa UR 1339
ED 520 Humanités
Université de Strasbourg
Strasbourg, France

Sous la codirection de

Dan TUFIȘ
Professeur Académicien Dr.
RACAI - Institut de Recherche pour l'Intelligence Artificielle « Mihai Drăgănescu » (ICIA)
Académie Roumaine, Bucarest

2018-2023

Research is formalized curiosity.

(Zora Neale Hurston)

Remerciements

Je souhaite remercier les personnes qui ont contribué à la réalisation et à l'aboutissement de ce travail de recherche.

Mes remerciements vont tout d'abord à ma directrice de thèse, Madame Amalia Todirascu, pour son encadrement soutenu, sa réactivité et ses conseils précieux concernant la rédaction scientifique et le code informatique. Je remercie également mon co-directeur de thèse, Monsieur Dan Tufis, pour ses conseils pratiques concernant les technologies d'apprentissage automatique et leur adaptation à la langue roumaine.

Je remercie les membres du jury et les rapporteurs pour avoir accepté d'évaluer ce travail de recherche. Je remercie également Madame Natalia Grabar et Madame Iris Eshkol-Taravella pour avoir fait partie du comité de suivi de thèse. Leurs conseils utiles m'ont permis d'avancer dans mon travail de recherche.

Je remercie Monsieur Rudolph Sock, directeur du laboratoire LiLPa, pour ses encouragements et pour me soutenir à présenter mes recherches lors des conférences internationales, une fois la crise sanitaire stabilisée. Je remercie les chercheurs du monde entier (France, Roumanie, Japon, Singapore, États-Unis, Corée du Sud) qui ont pris le temps d'échanger avec moi sur leur expertise et me partager leur passion pour la recherche.

Je souhaite remercier les étudiants stagiaires qui ont mené des doubles annotations sur les données et qui ont apporté des réflexions utiles pour mon travail.

Mes remerciements vont également à mes collègues du laboratoire LiLPa : Alexia, Cindy, Chang, Delphine, Emmanuelle, Erin, Igor, JiHyang, Jinwoo, Paul, Peiru, Rémi, Salomé, Seo, Seto, Simone, Thalassio. Leur encouragements, leur optimisme et leur amitié m'ont permis d'avancer dans mon travail et de me sentir moins seule avec ma thèse. J'apprécie leurs cultures riches qui m'ont permis de voyager aux quatre coins du monde à travers nos discussions. Je remercie Seto pour ses relectures minutieuses de mes articles scientifiques en anglais.

Je remercie chaleureusement mes amis (trop nombreux pour être cités ici), qui, pour beaucoup, m'ont connu en thèse et m'ont soutenu avec leur bonne humeur jusqu'à la fin. Je remercie également mes colocataires et notre chat, avec qui j'ai partagé le quotidien chargé de la rédaction de la thèse et qui ont toujours été là pour moi. Je remercie Alice pour son aide en fin de thèse lorsque mon ordinateur de travail a du subitement être réparé.

Je remercie mes parents, mon frère, ma sœur et mes amis de longue date qui m'ont soutenu moralement et ont cru en moi pendant ces longues années d'études.

Je remercie tout particulièrement mon cher Sébastien, pour sa patience, son soutien sans faille et ses encouragements constants. Je lui remercie énormément pour la relecture attentive de l'intégralité de la thèse et pour ses remarques constructives sur la rédaction du manuscrit. Je lui remercie de tout mon cœur pour son soutien moral et affectif, pour croire en moi et pour me redonner toujours la motivation à poursuivre mes rêves.

Table des matières

IV. ANALYSES, EXPÉRIENCES ET RÉSULTATS

Index des tableaux

Index des figures

Introduction

La richesse d'une langue naturelle consiste dans sa complexité notionnelle, sémantique et syntaxique, ainsi que dans la variété de ses procédés langagiers. La complexité d'une langue réside également dans la variété des moyens disponibles pour exprimer des idées similaires. Cet indicateur de la complexité d'une langue, ***la capacité de reformulation***, est également un critère important d'évaluation de l'acquisition d'une langue étrangère. Fuchs (1982 : 92) soutient qu'il y a un lien reconnu entre la capacité de ***paraphrasage*** de l'apprenant et la maîtrise linguistique. Un apprenant d'une langue étrangère doit arriver à maîtriser l'ensemble du lexique et des procédés linguistiques propres à la langue étudiée. En d'autres termes, il doit apprendre à maîtriser **l'art de la reformulation**, du *dire autrement*. Dans ce sens nous nous intéressons à la **reformulation** comme procédé linguistique de *complexification d'une langue*, mais également de *de-complexification*.

Ces deux procédés linguistiques n'interviennent pas uniquement lorsqu'il s'agit de comprendre une autre langue, mais également pour comprendre **un autre registre de langue**. La *de-complexification* d'un **registre technique ou spécialisé** est nécessaire pour la compréhension d'un lecteur non avisé. Par exemple, dans le domaine médical, un praticien doit adapter son discours pour que le patient comprenne les problèmes liés à sa maladie ou son traitement. Ce processus de **simplification** de textes spécialisés est indispensable à la **vulgarisation scientifique**. Dans notre thèse nous nous intéressons au rôle de la **reformulation** dans l'adaptation des textes en **langage technique** pour un public cible : **le grand public**.

Objectifs de la thèse

Notre travail de recherche a comme but de développer **une méthode de constitution automatique de corpus de reformulations médicales**. Nous travaillons sur la **reformulation sous-phrastique médicale**, que nous définissons comme *l'équivalence au sens large, basée sur un noyau sémantique commun, qui contribue à la vulgarisation de termes médicaux et qui ne dépasse pas le cadre d'une phrase*. Nos recherches sont menées

sur des corpus monolingues comparables du domaine de la médecine (destinés à des publics d'experts et non-spécialistes du domaine) en deux langues, français et roumain. Notre thèse s'encadre dans le domaine de recherche du **Traitement Automatique des Langues (TAL)** (en faisant appel aux techniques **d'intelligence artificielle**) et de **la linguistique**. Pour nos expériences de recherche, nous appliquons des méthodes **à base des règles d'annotation appliquées pré et post-traitement** et des méthodes **d'apprentissage automatique par réseaux de neurones**. Nous développons des **ressources utiles pour la reformulation** dans le domaine médical afin de rendre accessible à un public large des informations de santé présentées souvent avec un langage trop scientifique. Notre travail de recherche pourrait constituer une ressource pour la simplification automatique des textes, la rédaction semi-automatique des textes de vulgarisation médicale (voir même automatique avec l'avancée technologique dans le domaine de la génération automatique), avec le but d'informer le grand public de façon ciblée et adaptée.

Motivations

En linguistique et en TAL, deux notions sont utilisées pour parler de la reprise d'un texte : **la reformulation et la paraphrase**. **La reformulation** est définie comme le processus de réécriture qui a le rôle d'expliquer, reprendre ou simplifier une phrase ou un élément de la phrase. **La paraphrase**, l'équivalence basée sur un noyau sémantique commun (Fuchs, 1982), a été définie par différentes approches : énonciative et discursive (Fuchs, 1982 ; 1994), logique, sémantique (Martin, 1976) ou pragmatique (Grabar et Eshkol-Taravella, 2016b). Les reformulations et les paraphrases peuvent être de plusieurs types : lexicales, sous-phrastiques et phrastiques (Bouamor, 2012), paraphrastiques, non-paraphrastiques (Gühlich et Kotschi, 1983), etc. La notion de **reformulation est plus large que la paraphrase**, car elle inclue des techniques **d'explication** ou **d'exemplification**. Nous considérons que la reformulation est plus adaptée dans un contexte de vulgarisation de textes scientifiques, ce qui nous motive à travailler sur cette notion, plutôt que de prendre en compte uniquement les paraphrases. Pourtant, **la grande variabilité d'approches et des définitions** en linguistique rend la reformulation et la paraphrase **difficilement modélisables et exploitables en TAL**. Créer des modèles linguistiques de la reformulation ou de la paraphrase en respectant les particularités et la richesse de chaque langue et adapter ces modèles aux outils de TAL reste un grand défi.

Les travaux en **TAL** sur la construction des corpus de paraphrases (Dolan *et al.*, 2004 ; Dutrey *et al.*, 2011 ; Creutz, 2018) ou l'extraction de la paraphrase et la reformulation montrent l'intérêt croissant de la recherche sur le sujet. Diverses catégories d'approches ont été testées pour la détection de la paraphrase : par des systèmes à base de règles et traduction automatique (Ohtake et Yamamoto, 2003 ; Boumor, 2012), par traduction par pivot (Bannard et Callison-Burch, 2005 ; Pavlick *et al.*, 2015) ou neuronale (Zhou *et al.*, 2021), par les graphes syntaxiques (Chen *et al.*, 2013 ; Issa *et al.*, 2018), par apprentissage automatique statistique classique (Brockett et Dolan, 2005 ; Filice et Moschitti, 2016) et par apprentissage par réseaux de neurones (Yuan *et al.*, 2016 ; Le *et al.*, 2018). Certaines études ont été réalisées sur la paraphrase dans le domaine de la médecine également (Elhadad et Sutaria, 2007 ; Deléger et Zweigenbaum, 2009 ; Grabar et Hamon, 2015). Des ressources lexicales comme **WordNet** (Miller, 1998) pour la langue générale et **UMLS** (Bodenreider, 2004) ou **Snomed** (Spackman *et al.*, 1997 ; Donnelly, 2006) pour le domaine médical sont utiles pour l'identification automatique des paraphrases selon leur degré de synonymie (Cardon et Grabar, 2019 ; Koptient *et al.*, 2019). Par rapport à la paraphrase, la reformulation est moins abordée en TAL. Certains travaux (Grabar et Eshkol-Taravella, 2016a ; Magri, 2018) s'appuient sur l'analyse linguistique des marqueurs de la reformulation à l'aide des méthodes symboliques hybrides. Ces ressources sont disponibles pour des genres spécifiques ou registres de langues spécifiques (textes littéraires, textes transcrits de l'oral, forums de patients).

Néanmoins, les ressources lexicales et les corpus existants ne sont pas suffisants, vu la grande variété notionnelle et linguistique de la reformulation. Peu de travaux exploitent les modèles et les travaux en linguistique dans une perspective automatique. En particulier, nous proposons de construire des ressources linguistiques permettant l'identification et la génération automatique de la reformulation, sur des textes écrits du domaine médical, vu le manque de ressources pour la simplification et la vulgarisation scientifique.

Beaucoup de travaux et ressources sont disponibles pour l'anglais, mais les autres langues sont moins représentées. Nous travaillons sur des corpus comparables de textes écrits du **domaine médical** en deux langues : français et roumain. Notre choix de travailler sur un **corpus bilingue** permet de réaliser une analyse contrastive de données et d'élargir le champ de recherche sur la reformulation sous-phrastique médicale également en roumain (langue de moindre diffusion).

Nous faisons avancer les travaux réalisés sur la reformulation en français en s'inspirant des travaux disponibles sur l'anglais. Le manque de **ressources multilingues**

pour la reformulation dans des domaines de spécialité comme le domaine médical (Ganitkevitch et Callison-Burch, 2014) nous motive dans ce sens. En ce qui concerne le roumain, le travail sur la reformulation médicale est complètement novateur, car **il n'y a pas à ce jour des travaux de recherche menés sur ce sujet sur la langue roumaine**. Nous considérons qu'il est très important de travailler sur les **langues de moindre diffusion** également, même si la complexité grammaticale de ces langues et le manque de ressources rendent la tâche plus difficile. Nos recherches promeuvent la **diversité linguistique dans la recherche** et elles aident à faire progresser la recherche sur la reformulation médicale en TAL.

Notre travail de recherche représente un défi du point de vue **terminologique** également, vu la présence de **termes médicaux** dans nos corpus. **Le terme** est une unité lexicale de spécialité qui représente des connaissances spécifiques à un domaine du savoir, dans notre étude, celui de la médecine (Costa, 2005 : 84). **La terminologie médicale** se caractérise par des spécificités propres comme l'origine grecque ou latine de dénominations (comme « myocardiaque », formé avec une base latine « myo » = muscle et une base grecque « cardia » = cœur (Grabar et Hamon, 2016 : 86), l'opacité sémantique pour les non-spécialistes (« desquamation », « glomérules de néphropathie » (Buhnila, 2018 : 84)), mais aussi des combinaisons de plusieurs mots qui deviennent des termes du domaine (« touche d'essai » qui veut dire dans le langage médical « applications répétées sur une petite surface de peau pendant une dizaine de jours consécutifs »). La signification des termes scientifiques médicaux reste dans la plupart des cas obscure pour le grand public, d'où la nécessité d'une **vulgarisation** constante dans ce domaine (Grabar et Hamon, 2016). Cette richesse sémantique du langage médical et la diversité de reformulations possibles rend notre travail de recherche difficile et complexe. Notre but est d'entrecroiser la linguistique et le traitement automatique des langues pour **identifier et générer de façon automatique les reformulations de ces termes médicaux scientifiques**. Toutefois, pour l'identification automatique des termes médicaux, nous utiliserons des **terminologies reconnues** dans le domaine médical et validées par des experts du domaine, car notre thèse se concentre sur les **reformulations médicales**.

Dans cette thèse, nous analysons également **le rôle de ces reformulations médicales** dans le processus de **simplification lexicale** des textes écrits (Specia *et al.*, 2012 ; Shardlow, 2014 ; Grabar et Hamon, 2015 ; Saggion, 2017). Ce qui nous motive davantage dans notre recherche est le **nombre restreint de ressources** pour la reformulation dans le domaine médical pour la simplification de textes (Cardon, 2018 ; Cardon et Grabar, 2019). Nous considérons que les reformulations sont d'une grande

importance pour la **vulgarisation scientifique** (Vargas, 2008 ; Cardon, 2018). **Vulgariser** la médecine représente le processus de simplification lexicale et syntaxique qui a comme but de rendre les notions médicales très techniques compréhensibles en fonction du public cible. Informer correctement le grand public sur les questions médicales est une tâche qui demande un effort soutenu, vu l'innovation continue dans le domaine et les défis sanitaires constants (Pecout *et al.*, 2019).

Cette vulgarisation scientifique se réalise à l'aide des **explications** ou **définitions** de termes scientifiques monolexicaux et polylexicaux introduites par des **marqueurs de reformulation lexicaux ou grammaticaux**, de type « c'est-à-dire », « autrement dit », « encore appelé » (Antoine et Grabar, 2016 : 13), « est un », « également appelé » (Grabar et Hamon, 2015) (exemple : « un gastroentérologue spécialisé, **c'est-à-dire** un proctologue ») (Antoine et Grabar, 2016 : 6) ou voir même des **marqueurs orthotypographiques**, comme les doubles points ou les parenthèses (exemple : « des canaux galactophores **:** qui fabriquent le lait de la femme, qui sécrètent le lait » (Antoine et Grabar, 2016 : 11)). Dans cette même lignée, nous analysons les corpus en langue de spécialité pour experts et grand public pour identifier les marqueurs de reformulations spécifiques au domaine de la médecine et pour une **modélisation exploitable automatiquement**.

Des travaux ont été réalisés pour la simplification automatique de textes sur des binômes de **corpus scientifiques – corpus de vulgarisation**, mais ils restent peu nombreux. La plupart des travaux développent des ressources en langue générale sur les encyclopédies en français **Wikipédia** et **Vikidia**, la version simplifiée adaptée aux enfants, comme Brouwers *et al.* (2012) ou MUSS (Martin *et al.*, 2022). Cardon (2021) propose des ressources pour la simplification automatique dans des domaines de spécialité, notamment la médecine. Dans le cadre du projet **ClassYN** (Todirascu *et al.*, 2012) ont proposé une méthode de classification automatique des textes en fonction du public cible sur des corpus scientifiques et des corpus pour le grand public (allemand-français) dans deux domaines, informatique et médecine. Nous avons mené une étude sur les corpus médicaux en français du projet **ClassYN** et sur des corpus comparables en roumain pour l'identification automatique de la terminologie médicale et de ses équivalents en langue générale dans le cadre de notre mémoire de recherche de master (Buhnila, 2018).

Nous nous intéressons au rôle de la reformulation des termes (en particulier polylexicaux) dans la vulgarisation scientifique des contenus médicaux en fonction du **public cible**. Nous prenons en compte le lecteur cible de ces textes de vulgarisation pour

saisir les différentes formes de la reformulation médicale. Dans ce sens nous prenons en compte les variations morphologiques, lexicales, sémantiques, syntaxiques et pragmatiques qui aident à adapter la reformulation médicale à son public cible et nous proposons une ressource qui fournit des reformulations aux termes médicaux, disponible en français et en roumain. Cette ressource est extraite automatiquement et partiellement validée par plusieurs annotateurs humains.

Pour identifier et générer automatiquement des reformulations, nous faisons appel aux **méthodes d'apprentissage par réseaux de neurones**, appliquées sur les données que nous avons construit sur la base d'analyse et d'annotation de corpus. Ces méthodes viennent compléter les ressources et les outils de TAL pour la reformulation. En plus, les applications de recherche sur la reformulation médicale, la vulgarisation scientifique, la génération automatique de textes pour informer de façon ciblée et adaptée le grand public demandent des ressources exploitables. Le corpus de reformulation créé à l'issue de notre thèse pourra être intégré dans une **application de simplification de textes** dans des travaux futurs.

***

Notre thèse de doctorat est constituée de cinq **Parties**, comme suit :

- La **Partie I** présente l'état de l'art des **travaux en linguistique sur la reformulation** et de ses **diverses définitions et réalisations linguistiques** ;
- La **Partie II** présente **les approches, les tâches et les ressources proposées en TAL** pour le traitement de la **reformulation et de la paraphrase** ;
- La **Partie III** détaille notre **approche théorique sur la reformulation, nos corpus d'étude, la collecte du corpus roumain, notre méthodologie**, et **l'annotation semi-automatique des corpus** (termes et marqueurs de reformulation) **;**
- La **Partie IV** présente **les résultats de l'annotation automatique et manuelle**, les diverses **expériences** réalisées, **les analyses quantitatives et qualitatives** des données **en français** et **en roumain**, **les expériences d'apprentissage automatique par réseaux de neurones** pour la génération et l'identification des reformulations et l'évaluation du **niveau de lisibilité** des reformulations médicales ;
- La **Partie V** conclut notre travail de recherche avec un **bilan général des travaux** réalisés et introduit les **perspectives pour des recherches futures**.

I. DÉFINITION ET ÉTAT DE L'ART SUR LA REFORMULATION EN LINGUISTIQUE

1. La reformulation

Afin de construire des corpus de reformulation, nous devons identifier les reformulations dans des textes et préciser cette notion. **La notion de reformulation** elle-même a fait l'objet de nombreuses recherches et a reçu plusieurs définitions. Inkova (2020) présente l'historique des travaux en linguistique sur la reformulation depuis ses débuts dans les années '80, en se concentrant sur les travaux sur la langue française. Nous illustrons cet historique dans le **Tableau 1**. Celui-ci montre que la notion de reformulation a suscité l'intérêt des linguistes depuis déjà quatre décennies et que c'est un concept qui reste vivant et toujours intéressant à analyser dans la langue et le discours.

La reformulation représente le processus de réécriture qui a le rôle d'expliquer, simplifier ou pointer une phrase ou un syntagme. La reformulation est un processus linguistique de **transformation du discours** qui a été étudié par le prisme de plusieurs domaines tout au long de ces quatre décennies :

- les sciences du langage (Fuchs, 1982, 1994 ; Gülich et Kotschi, 1987) ;
- les études sur les interactions verbales (Roulet, 1987) ;
- la didactique (enseignement et apprentissage) (Brixhe et Specogna, 1999 ; Martinot *et al.*, 2018) ;
- la vulgarisation scientifique (Loffler-Laurian, 1984 ; Vargas, 2008) ;
- la littérature (Magri, 2018) ;
- le traitement automatique des langues (Brockett et Dolan, 2005 ; Bouamor, 2012 ; Grabar et Hamon, 2015 ; Filice et Moschitti, 2016 ; Issa *et al.*, 2018).

La notion de reformulation a été le sujet débattu et développé pour de nombreux ouvrages et chapitres d'ouvrages (Rossari, 1997 ; Le Bot *et al.*, 2008 ; Schuwer *et al.*, 2009 ; Kara, 2007 ; Kanaan, 2011 ; Martinot *et al.*, 2018). La manière dont les linguistes conçoivent et délimitent la reformulation évolue et se reflète dans des nombreux colloques, journées d'étude au niveau international (« La reformulation : à la recherche d'une frontière », « Autour de la reformulation », « Reformuler, une question de genres ? ») et numéros thématiques de revues (dont le n° 212 de la revue Langages, 2018).

Année	Pays	Auteurs & Travaux
PUBLICATIONS		
1982	France	Fuchs, 1982
1982	États-Unis	Schiffrin, 1982
1983 1987	Allemagne	Gülich et Kotschi, 1983 ; 1987
1987	Suisse	Roulet, 1987
1989	Suisse	Adam et Revaz, 1989
OUVRAGES ET CHAPITRES D'OUVRAGES		
1994	France	Fuchs, 1994
1997	Suisse	Rossari, 1997
2007	France	Kara, 2007
2008	France	Le Bot *et al.*, 2008
2009	France	Schuwer *et al.*, 2009
2011	France	Kanaan, 2011
2018	France	Martinot *et al.*, 2018
COLLOQUES, JOURNÉES D'ÉTUDE, N° THÉMATIQUES DE REVUES		
2017	Université d'Uppsala, Suède	« La reformulation : à la recherche d'une frontière »
2018	International	Revue « Langages », n° 212
2018	Université de Genève, Suisse	« Autour de la reformulation »
2019	Université de Porto, Portugal	« Reformuler, une question de genres ? »

Tableau 1. Historique des travaux sur la reformulation dans la langue française

La définition de la notion de reformulation et l'analyse de ce phénomène reste un problème complexe qui fait l'objet de plusieurs théories linguistiques que nous présentons dans la section suivante.

1.1 Différentes approches sur la reformulation

Le phénomène de la reformulation a été étudié à travers plusieurs points de vue en fonction du domaine de recherche et des différents domaines d'application, comme la linguistique interactionnelle, la linguistique de corpus ou la linguistique discursive (Eshkol-Taravella et Grabar, 2018). Nous présentons quelques approches sur la reformulation afin de mettre en avant la complexité de cette notion et montrer l'intérêt de l'aborder du point de vue de la linguistique de corpus et du TAL.

Du point de vue **linguistique et de la structure de l'énoncé**, Martinot (2003) définit la reformulation comme :

> *Tout processus de reprise d'un énoncé antérieur qui maintient, dans l'énoncé reformulé, une partie invariante à laquelle s'articule le reste de l'énoncé, partie variante par rapport à l'énoncé source.* (Martinot, 2003 : 147)

Cette définition explique le lien de synonymie qui existe entre le syntagme original (le référent, plus précisément l'élément linguistique source) et la reformulation de celui-ci, dans le cadre de deux énoncés sémantiquement liés. Inkova (2020 : 28) appelle cette définition « *extra-large* », car elle comprend « *tous les types de reformulation* » :

- les reformulations paraphrastiques ;
- les reformulations non paraphrastiques ;
- les reformulations répétitives.

Nous revenons plus en détail sur ces différents types de reformulations au **Chapitre 2. *Types des reformulations et marqueurs linguistique***.

En partant de cette définition au sens large de la reformulation, plusieurs approches spécifient cette notion. Gülich et Kotschi (1987) considèrent que les **actes de reformulation de type paraphrasage, rephrasage, correction** sont des actes verbaux utilisés pour résoudre des problèmes communicatifs. Les reformulations serviront à faire comprendre l'objectif de communication du locuteur vers l'interlocuteur cible.

Du point de vue du texte, Adam (1990) considère que la **reformulation** est un facteur qui contribue à la **textualité** parce qu'elle crée le **lien** entre les unités linguistiques (Adam, 1990 : 172). Dans ce sens, la reformulation pourrait se définir comme « un équilibre délicat entre une continuité-répétition d'une part, et une progression de l'information, d'autre part » (Adam, 1990 : 45). Par cette définition, Adam (1990) souligne que la reformulation

joue un **double rôle** : celui de répéter, reprendre une notion et, en même temps, d'en fournir des informations nouvelles.

Dans la perspective du discours didactique, Brixhe et Specogna (1999 : 12) distinguent les *autoreformulations* quand le locuteur reformule lui-même ses énoncés, et les *hétéroreformulations*, quand le locuteur reformule l'énoncé de l'interlocuteur. Dans ce scénario, la reformulation peut être *auto-initiée* (réalisée par le locuteur lui-même), ou bien *hétéro-initiée* (quand la reformulation est réalisée par le destinataire de la communication). Selon Dufour (2005), la reformulation est une opération de *re(définition)* du sens qui a lieu en discours. Dufour (2005) analyse le rôle de la reformulation comme modificateur des frontières lexicales et du rapport avec le référent (énoncé source), ce qui lui apporte le pouvoir « d'agir sur le monde ».

Plus récemment, Cislaru et Olive (2018) analysent la reformulation du point de vue des tâches de **production textuelle par ordinateur**. Ils enregistrent les jets textuels de production et de révision réalisés **par ordinateur** et ils analysent les types de reformulations produites. Concernant les **jets de production**, deux types de reformulations se démarquent : les reformulations qui gardent le même schéma syntagmatique et les reformulations qui maintiennent la structure morpholexicale. Cislaru et Olive (2018) observent que les ***révisions-reformulations*** ont lieu majoritairement au niveau phrastique.

Magri-Mourgues (2012) soutient que le **rôle de la reformulation dans le discours** écrit varie en fonction de l'approche selon laquelle la reformulation est analysée :

- **L'approche syntaxique** - la reformulation s'intéresse aux différents modes de liaisons grammaticales (morphologiques et syntaxiques) qui construisent la trame textuelle ;
- **L'approche sémantique** - analyse le rapport de sens à établir entre l'unité linguistique source (le référent) et l'unité linguistique cible (la reformulation) ;
- **L'approche pragmatique** - met en avant la motivation qui réside derrière l'utilisation de la reformulation, dans une vision interactionnelle de l'énoncé ;
- **L'approche cognitive** - établit le rapport entre l'unité reformulée et l'unité source (le référent), comme indicateur de la perception du monde réel et de sa catégorisation.

Quand on parle du **discours**, on pense également à l'intention du locuteur, au message qu'il veut transmettre et comment. Pour définir la **reformulation**, Inkova (2020 : 29) propose de prendre en compte ses fonctions dans la construction du sens dans la phrase et dans la communication. En soutient de cette approche, Inkova se rapporte à

Benveniste (2010) et appuie l'importance du **mieux dire** et de trouver une façon plus **adaptée** d'expression à travers la reformulation :

> *Les locuteurs n'utilisent pas seulement les reformulations « parce qu'ils se sont trompés » mais parce qu'ils recherchent continuellement la meilleure façon de dire et le meilleur angle d'attaque.* (Benveniste, 2010 : 87)

Nous ne pouvons pas parler du discours sans mentionner les deux grands types, **l'expression écrite** et **l'expression orale**. Est-ce que la reformulation est employée de la même façon dans les deux types de textes ? A-t-elle les mêmes fonctions ? Pennec (2006) considère qu'à l'écrit, les reformulations sont directes, c'est-à-dire que le *formulé* (segment initial) et le *reformulé* (segment dit autrement) se succèdent dans la phrase avec la fonction de « mieux dire ». À l'oral, Pennec (2006) met en avant le rôle de hiérarchisation de l'information, de structuration du discours de la reformulation. La composition de la reformulation à l'oral est moins figée, permettant d'insérer dans le discours des réflexions, opinions et précisions de la part du locuteur. Les reformulations à l'oral font l'objet des recherches d'Eshkol-Taravella et Grabar (2014 ; 2018).

Pennec (2020 : 64) développe cette vision de la **reformulation dans le discours** et considère cet acte de langage de la **méta-énonciation**, ce qui veut dire *l'énonciation qui parle de l'énonciation*. Selon Pennec (2020) la **méta-énonciation** est une sous-catégorie du **métalangage**, définit par Rey-Debove (1978 : 2) comme un sous-système d'une langue L1 qui lui permet de parler de cette langue L1.

> *il faut admettre l'importance de reconnaître à la reformulation son caractère métalinguistique, de réflexion sur le code de la langue, de retour sur la forme du dit pour en choisir une autre.* (Inkova, 2020 : 32)

Pour expliciter le code de la langue, Inkova (2020 : 39) propose dans une classification des reformulations selon le niveau **métalinguistique**. Dans ce sens, le code concerne le choix de la forme de l'énoncé et l'explication du sens d'un mot (Inkova, 2020 : 39).

1.2 Conception large / conception étroite de la reformulation

Inkova (2020) oppose ces deux concepts pour pallier les visions plus récentes sur le concept de reformulation. Si la ***conception large*** de la reformulation est représentée par la définition générale du concept comme proposée par Martinot (2003), la ***conception étroite***

limite le champ de la reformulation à la relation d'équivalence sémantique. Cette relation doit relier un segment de la phrase avec son segment reformulé (ces deux segments doivent renvoyer à la même notion ou au même référent). Cette conception étroite est présente dans les travaux de plusieurs linguistes (Gardin, 1987 ; Manzotti, 1999 ; Teston-Bonnard, 2008 ; Inkova et Guryev, 2018 ; Vassiliadou, 2020).

Vassiliadou (2020 : 82) attire l'attention sur les limites que nous devons mettre à la reformulation. Le sens large de « formuler autrement ce qui a déjà été formulé », permettrait de mettre sous la même enseigne tout type de reformulation : paraphrastique, non-paraphrastique, voir même les répétitions. Or, nous devons faire la distinction entre **la reformulation de type paraphrase** et **la reformulation de type corrective**, encore plus par rapport aux **répétitions**. Dans ce sens, Vassiliadou (2020 : 82) propose **une définition plus stricte de la reformulation** qui permettrait d'exclure de cette définition les structures beaucoup trop éloignées du sens de base (formulé / reformulé), comme la répétition.

Vassiliadou (2020 : 84) propose un critère unique qui permettrait de définir la reformulation : « reformuler signifie **formuler une deuxième fois** dans un **but d'éclaircir le sens d'un segment antérieur** ». Vassiliadou (2020) insiste sur le fait que dans la définition de la reformulation il faut donner une place importante au **segment initial qui est reformulé**. Car c'est ce premier concept qui engendre ou demande la reformulation, ce besoin **d'exprimer autrement** le sens du segment initial.

***

Ces différentes approches montrent la complexité d'interprétation du phénomène de la reformulation en linguistique et comment divers points de vue apportent plusieurs dimensions à la reformulation. Afin de mieux comprendre l'unité linguistique principale de notre travail de recherche, **la reformulation**, nous présentons les différents **types de reformulations** selon les principales théories linguistiques et la notion de **marqueur** de reformulation.

2. Types des reformulations et marqueurs linguistiques

En partant de la **conception large de la reformulation** d'Inkova (2020), nous distinguons deux grands types de reformulations : **la reformulation paraphrastique**, fondée sur l'équivalence sémantique et **la reformulation non-paraphrastique**. Cette opposition a été présentée pour la première fois par Roulet (1987) et développée par la suite par Rossari (1997). Rossari (1990, 1994) fait la distinction entre les deux types de reformulations sur les critères suivants :

- les « **reformulations paraphrastiques** » sont celles qui établissent *un rapport d'équivalence sémantique* avec la séquence source, comme dans l'exemple suivant : « <formulé> **Je suis mal** </formulé>, <marqueur> ***je veux dire*** </marqueur> <reformulé> **je suis en mauvaise forme** </reformulé> » ;

tandis que

- les « **reformulations non-paraphrastiques** » marquent *un changement de perspective énonciative*, comme dans l'exemple : « <formulé> **J'ai mal au ventre** </formulé>, <marqueur> ***finalement*** </marqueur> <reformulé> **je vais plus sortir en ville** </reformulé> » (Rossari, 1997).

Pennec (2020 : 67) divise les reformulations en trois grandes catégories en rajoutant **la paraphrase** à la liste de Rossari (1994) :

- **la paraphrase** : une reformulation qui présente **des contenus** parfaitement **équivalents** sémantiquement ;
- **la reformulation paraphrastique** : dans ce cas c'est **la forme** de l'énoncé qui est retravaillée en priorité ; cette priorité donnée à la forme peut entraîner certaines modifications de contenu sémantique ;
- **la reformulation non-paraphrastique** : dans cette situation **le contenu sémantique** est la priorité de la modification.

Nous présentons par la suite les deux types de reformulations qui se retrouvent souvent en opposition dans la littérature : la reformulation paraphrastique et la reformulation non-paraphrastique. **La paraphrase** sera présentée comme *forme* spécifique de la reformulation, ensemble avec la glose et la répétition. Nous développons davantage ses

caractéristiques et ses différents statuts (reformulation, glose) dans le **Chapitre 3.** ***Formes linguistiques de la reformulation***.

2.1 La reformulation paraphrastique

Si nous voulons définir la reformulation paraphrastique selon son acceptation au fil du temps, nous devons commencer avec les années '80 lorsque le concept commence à être étudié par les linguistes. Pour Gühlich et Kotschi (1983), la reformulation paraphrastique est caractérisée par la **conservation du sens**, comprise comme une **similarité sémantique**, allant vers l'**identification sémantique** entre le formulé et le reformulé :

> *Ce n'est pas seulement l'existence d'une équivalence sémantique entre deux énoncés qui est prise en considération, mais aussi, et surtout l'acte d'une « prédication d'identité » ; deux énoncés sont produits et enchaînés de telle manière qu'ils peuvent et doivent être compris comme « identiques ».* (Gühlich et Kotschi, 1983 : 307-308)

Ce premier principe d'**identité sémantique** est partagé par plusieurs chercheurs. En partant des deux définitions données plus haut dans l'introduction de ce chapitre, nous concluons que la notion de reformulation paraphrastique s'oriente plutôt vers une **équivalence sémantique** (Rossari, 1994) entre un formulé et une reformulé. Pennec (2020) rajoute plus tard une autre couche d'analyse : le reformulé présente des **modifications dans la forme**. Alors, la reformulation paraphrastique devient une ***presque*** **similarité sémantique**, car un **reformulé** (sous **une autre forme** lexicale, morphologique ou syntaxique) apporte inévitablement de **nouvelles nuances sémantiques**.

Nous observons également qu'il y a un changement dans la définition du concept, car pour Pennec (2020), la **paraphrase** est ce que Rossari (1994) appelait la **reformulation paraphrastique**, donc la notion qui est caractérisée principalement par l'équivalence sémantique.

Afin de bien mettre en évidence la complexité du concept, nous notons que pour Vargas (2008) la reformulation paraphrastique présente un **caractère de substitution**. Plus précisément, la reformulation paraphrastique est un « redit » qui peut être remplacé dans le texte par le « dit » (le référent reformulé). Dans ce cas, les deux unités linguistiques sont au même niveau hiérarchique, comme dans l'exemple suivant proposé par nous :

- Il a choisi de **devenir végétarien**, c'est-à-dire de **ne pas manger de la viande**.

Dans cette phrase le référent « devenir végétarien » peut être facilement remplacé par la reformulation « ne pas manger de la viande » : « Il a choisi de **ne pas manger de la viande**, c'est-à-dire de **devenir végétarien** ». La substitution est possible grâce à la similarité synonymique des deux unités. La reformulation est dans notre exemple la définition vulgarisée du référent et elle se situe sur le même plan sémantique avec le référent.

Dans la **Figure 1** nous présentons en guise de conclusion l'évolution de l'acceptation sur ce que représente la **reformulation paraphrastique** pour les linguistes.

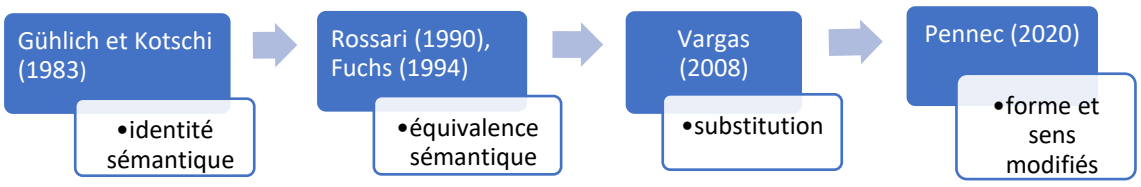

Figure 1. Acceptations et définitions de la reformulation paraphrastique. Évolutions et changements de perspective en linguistique (1983-2020)

2.2 La reformulation non-paraphrastique

Selon Rossari (1990), les reformulations non paraphrastiques expriment un **changement de perspective énonciative** dans le reformulé (Rossari, 1990 : 348). Pennec (2020) rajoute que les reformulations non-paraphrastiques sont des **reformulations correctives** ou qui apportent des **précisions** supplémentaires au segment qui est reformulé.

Pourtant, la **reformulation non-paraphrastique** ne peut pas prendre la place du référent dans le discours, car l'information transmise **change le point de vue** sur le référent. La reformulation non-paraphrastique a souvent le rôle de *rectifier* le contenu ou de donner une image plus claire sur le référent, comme dans l'exemple :

- **Le livre** est sorti hier, c'est-à-dire qu'**il sera vite dévoré par ses lecteurs**.

Nous remarquons dans notre exemple que la reformulation « il sera vite dévoré par ses lecteurs » apporte une information en plus sur le référent « livre » liée aux coordonnées

temporelles. Par conséquent, substituer les deux unités linguistiques surlignées en gras et garder le même sens de la phrase ne serait pas envisageable.

2.2.1 Reformulations non-paraphrastiques de type *description*

La reformulation non-paraphrastique a fait l'objet de multiples interprétations et elle a été analysée selon différentes approches. Pour Magri-Mourgues (2013a), les reformulations peuvent prendre la forme des « **descriptions** », surtout dans les études sur des textes littéraires. Nous remarquons que ces types de reformulations appartiennent au type de reformulation non-paraphrastique. Nous présentons les quatre types des reformulations non-paraphrastiques de type « **description** » de Magri-Mourgues (2013a) selon son analyse des récits de voyage :

- **La reformulation par addition** : le même référent est reformulé par plusieurs dénominations dans la forme de reprises anaphoriques ;
  - « [...] **une écritoire de cuivre d'un demi-pied de longueur**. Le manche de cet instrument oriental contient l'encre […]. De loin, cela peut passer pour un poignard, mais c'est l'insigne pacifique du simple lettré.[1] » (Nerval, Voyage en Orient, « Les femmes du Caire », p. 323) (Magri-Mourgues, 2013a : 2)

  Dans l'exemple ci-dessus, le référent « une écritoire de cuivre d'un demi-pied de longueur » est reformulé par l'intermédiaire de ses reprises anaphoriques sous forme de déterminants et pronoms démonstratifs « cet », « cela », « c'est ».

- **La reformulation par substitution (ou reformulation corrective)** : la reformulation a la tendance de remplacer le référent en vue d'une correction ;
  - « Nous courûmes au marché faire provision d'oranges et prendre **des glaces**, ou plutôt de la purée de neige au limon.[2] » (Th. Gautier, Voyage en Espagne, p. 226) (Magri-Mourgues, 2013a : 2)

  Cet exemple montre la reformulation corrective du référent « des glaces » qui est une reformulation ironique et simplifiante du type de dessert, « purée de neige au

[1] Les termes sont soulignés par (Magri-Mourgues 2013a : 2), à l'exception du démonstratif « cela » et de la mise en gras du référent « une écritoire de cuivre d'un demi-pied de longueur », modifications réalisées par nous.

[2] L'expression adverbiale « ou plutôt » est soulignée par (Magri-Mourgues 2013a : 2), tandis que la reformulation corrective en entier « ou plutôt de la purée de neige au limon » est soulignée par nous. Le référent « des glaces » est souligné et mis en gras par nous.

limon ». L'expression adverbiale « ou plutôt » fait la transition entre le référent et la reformulation en marquant l'alternative notionnelle corrective.

- **La reformulation par superposition (la traduction-reformulation)** : la reformulation se limite au cadre phrastique sans établir de hiérarchie entre les unités successives. Il s'agit de la **traduction inter-linguale** retrouvée dans une seule source textuelle.

  - « Ceux au contraire qui ne font que suivre la loi sans prétendre à la sagesse s'appellent *djahels*, c'est-à-dire ignorants. » (G. de Nerval, Voyage en Orient, « Druzes et maronites », p. 61).

  - « Les *cuevas*, ou habitations troglodytes, sont très nombreuses dans cette Espagne pauvre où ces hommes sobres installent comme ils peuvent des demeures rudimentaires.[3] » (A. T'sertstevens, L'Itinéraire espagnol, p. 76). (Magri-Mourgues, 2013a : 4)

  Dans les deux exemples ci-dessus provenant des récits de voyage, Magri-Mourgues (2013a) souligne l'importance de la traduction inter-linguale dans ce type de textes. Les termes inconnus dans une langue étrangère sont *traduits* par une reformulation introduite par le marqueur « c'est-à-dire » et la conjonction qui exprime l'alternative « ou ». Nous remarquons le fin sarcasme de l'auteur dans le choix des reformulations (« ignorants », « habitations troglodytes »), ce qui prouve que nous avons plutôt des reformulations subjectives que des traductions fidèles de termes inconnus au lecteur.

- **La reformulation extensive** : la reformulation repose sur les relations sémantiques lexicales, comme l'hyperonymie.

  - « C'est là qu'eût été de saison le dicton du petit Savoyard faisant l'éloge de sa gargote, et disant avec admiration qu'on y mange **cinq sortes de viandes**, à savoir : du cochon, du porc, du lard, du jambon et du salé.[4] » (G. Sand, Un hiver à Majorque, p. 138). (Magri-Mourgues, 2013a : 3)

  Nous remarquons l'ironie de la liste extensive du référent « cinq sortes de viandes » qui est reformulé à l'aide de cinq termes qui désignent le même type de viande : « du

[3] Les mots en italique appartiennent au texte original, tandis que les marqueurs « c'est-à-dire » et « ou » sont soulignés par (Magri-Mourgues, 2013a : 4). Les reformulations (« ignorants » et « habitations troglodytes ») sont mises en avant et soulignées par nous.

[4] Le verbe à l'infinitif « à savoir » est souligné par (Magri-Mourgues, 2013a : 3), tandis que la reformulation par superposition « du cochon, du porc, du lard, du jambon et du salé » est soulignée par nous. Le référent « cinq sortes de viandes » est souligné et mis en gras par nous.

cochon, du porc, du lard, du jambon et du salé ». Le verbe à l'infinitif « à savoir » à la valeur d'introducteur d'exemples.

2.2.2 Reformulation non-paraphrastique de type *intertextuelle*

Si les reformulations paraphrastiques ont une structure en trois unités (référent, marqueur, reformulation), les **reformulations intertextuelles** n'ont pas de structure bien définie. Drescher (2008) observe que les reformulations intertextuelles ressemblent plutôt à des **citations, des formes abrégées ou condensées de discours rapporté**.

Leurs marqueurs seraient les **expressions métacommunicatives** (pour emprunter la terminologie de Drescher (2008)), de type : « il parait que », « on dit que », « j'ai entendu ça », « les gens disent que ». La reformulation intertextuelle servirait à introduire une autre voix narrative dans le discours, un *savoir* présumé commun et connu[5], comme dans l'exemple ci-dessous dont le segment prend rétroactivement le statut de reformulation intertextuelle :

> « [...] il y a certaines qui ne voient pas leurs règles mais ça se manifeste d'une autre façon, par exemple les saignements, et voilà, par le nez [...] **MOI j'ai entendu** comme ça, oui, oui. » Drescher (2008 : 46)

2.2.3 Reformulation non-paraphrastique de type *intratextuelle*

Vargas (2008) soutient que la **reformulation intratextuelle** est la structure linguistique de base du texte de vulgarisation. Celle-ci aurait un « un rôle central de management du savoir et de l'information à l'écrit » (Vargas, 2008). La reformulation intratextuelle peut prendre la forme de la **dénomination,** de la **définition** ou de la **définition en mouvement inverse**. Dans ce dernier cas, la reformulation est utilisée sous forme de correction ou rectification dans un discours argumentatif (Vargas, 2008), ce qui fait le lien entre ce type de reformulation et la reformulation corrective de Magri-Mourgues (2013a). Vargas (2008) réalise son étude sur la reformulation sur le discours oral des émissions télévisées allemandes :

- « Il est très facile d'estimer le temps nécessaire à une voiture en train de rouler pour s'arrêter complètement par un freinage brusque. Et c'est pourquoi nous allons nous

[5] Remarque tirée de l'étude de Drescher (2008) sur les interactions orales entre éducateurs et public.

intéresser maintenant à ce qu'on appelle la distance de freinage. **C'est la distance dont une voiture dans une situation dangereuse a besoin pour s'arrêter.** » (*Sendung mit der Maus*, 11 mars 2006, traduit de l'allemand par Vargas) (Vargas, 2008 : 30) ;

- « Avant d'arriver *sur la route*, **ou plutôt sur la piste** avec une voiture de course, une foule de choses en plus doit être faite. » (*Sendung mit der Maus*, 18 mars 2006, traduit de l'allemand par Vargas) (Vargas, 2008 : 31)[6].

Le premier exemple est une reformulation intratextuelle de type définition, tandis que dans le deuxième exemple, le marqueur adverbial « ou plutôt » introduit une correction, donc une définition en mouvement inverse, selon la terminologie de Vargas (2008).

***

La notion de reformulation est très souvent liée à celle de **marqueurs de reformulation** typiques ou généraux, orthographiques ou complètement absents. Le prochain sous-chapitre traite justement ce concept de *marqueur de reformulation* afin de comprendre son importance et même son statut obligatoire ou optionnel dans l'identification des différents types de reformulation.

2.3 Marqueurs de reformulation

Plusieurs recherches ont été menées sur les marqueurs de reformulations qui sont basés sur le verbe « dire », comme « c'est-à-dire », « ça veut dire », « pour dire autrement », « autrement dit » (Vassiliadou, 2013a ; Steuckardt, 2018 ; Magri, 2018). Ces marqueurs peuvent être discursifs, justificatifs ou paraphrastiques. Le marqueur « c'est-à-dire », en tant que marqueur 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 » (Vassiliadou, 2013a). Vassiliadou (2013a) précise que l'emploi de « c'est-à-dire » comme **marqueur de la paraphrase d'équivalence** est plus fréquent dans des discours spécialisés.

Fuchs (2020) observe que la plupart des travaux sur la reformulation paraphrastique se sont concentrés sur l'identification de celle-ci à travers des **marqueurs paraphrastiques** de type « c'est-à-dire », « autrement dit », « en d'autres termes », « à savoir ». Un grand nombre d'études ont été dédiées aux emplois reformulatifs du marqueur prototypique

[6] Les mots en italique, en gras et surlignés appartiennent au texte original (Vargas, 2008 : 30, 31).

« c'est-à-dire » (Vassiliadou, 2004, 2013a, 2008, 2014 ; Chéria, 2010 ; Eshkol-Taravella et Grabar, 2018).

Nous avons remarqué dans les exemples donnés plus haut que la structure d'une reformulation est souvent représentée de cette manière : **formulé (S1)** (la notion – référent à reformuler) => **marqueur** de reformulation => **reformulé (S2)** (le syntagme qui constitue la reformulation du S1).

Figure 2. Structure standard d'une reformulation

La présence du marqueur de reformulation permet d'identifier la reformulation dans le discours. Cet introducteur est typique pour la reformulation paraphrastique et non-paraphrastique, mais il est absent dans le cas de la paraphrase (Pennec, 2020 : 67), *forme* de reformulation que nous présenterons dans le **Chapitre 3.**

Selon Fuchs (2020 : 50), les deux types de marqueurs de reformulation se différencient par rapport à leur fonction dans le discours et le type de reformulation :

- **le marqueur de reformulation paraphrastique** : a le rôle de mettre en équivalence le formulé avec le reformulé et exprime une homogénéité ;
- **le marqueur de reformulation non-paraphrastique** : ne sert pas à exprimer une mise en équivalence, mais plutôt une hétérogénéité de points de vue.

Chéria (2010) considère que la possibilité **d'interpréter à nouveau le contenu sémantique** du formulé est la caractéristique clé des **marqueurs de reformulations non-paraphrastiques**. Cette caractéristique joue un rôle important dans la distinction entre les deux types de marqueurs de reformulation : paraphrastiques et non-paraphrastiques.

2.4 Marqueur de reformulation : obligatoire ou optionnel ?

Si la plupart des linguistes ont été longtemps d'accord sur l'importance de la **présence** des marqueurs dans la structure d'une reformulation, des recherches plus

récentes ont été dédiées également à envisager leur **absence** dans la reformulation (Chéria, 2010 ; Vassiliadou, 2020). Cette approche analyse de critères qui peuvent servir à identifier et délimiter une reformulation, sans dépendre des marqueurs de reformulation, comme la présence des synonymes, hyperonymes ou hyponymes dans le reformulé.

Vassiliadou (2020) se questionne sur la possibilité d'insérer ou supprimer des marqueurs de reformulation et sur l'impact que cette opération peut avoir 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 être un indice de la pertinence d'une relation discursive. Pour Vassiliadou (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 serait l'indice d'une relation sémantique bien nouée entre le formulé et le reformulé.

Vassiliadou (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 analyserons cette question de la présence ou l'absence du marqueur de reformulation également lors de nos analyses concrètes sur les corpus et nous y apporterons nos observations expérimentales.

2.5 Bilan sur les types de reformulations

Nous avons réalisé une présentation exhaustive de l'état de l'art sur **les types de reformulations** en linguistique. Nous synthétisons dans la **Figure 3** les deux grands types de reformulations (paraphrastiques et non-paraphrastiques) et les différents sous-types proposés par les chercheurs en linguistique. Nous rajoutons à la classification de Rossari (1990) la **reformulation sous-phrastique.** Celle-ci se concrétise dans la **traduction intra-linguale** (traduction avec des éléments du même système linguistique) sans dépasser la longueur syntaxique d'une phrase, comme dans l'exemple suivant :

- Elle a été diagnostiquée avec **le syndrome des ovaires polykystiques**, plus précisément, **un déséquilibre hormonal**.

Dans cet exemple, le terme médical « syndrome des ovaires polykystiques » est reformulé par une paraphrase avec des mots de la langue générale, « un déséquilibre hormonal ». Cette **reformulation sous-phrastique** a le rôle de donner une reformulation rapide du terme médical scientifique. La reformulation aide à **vulgariser** la signification de ce syndrome à travers sa cause principale, en adaptant les mots employés à un public cible non-spécialiste du domaine médical. L'avantage de la reformulation sous-phrastique est l'existence du **marqueur de reformulation**, dans notre exemple, « plus précisément », qui fait le lien direct et immédiat entre le référent paraphrasé (« le syndrome des ovaires polykystiques ») et la reformulation (« un déséquilibre hormonal ») et assure la continuité du sens de la phrase.

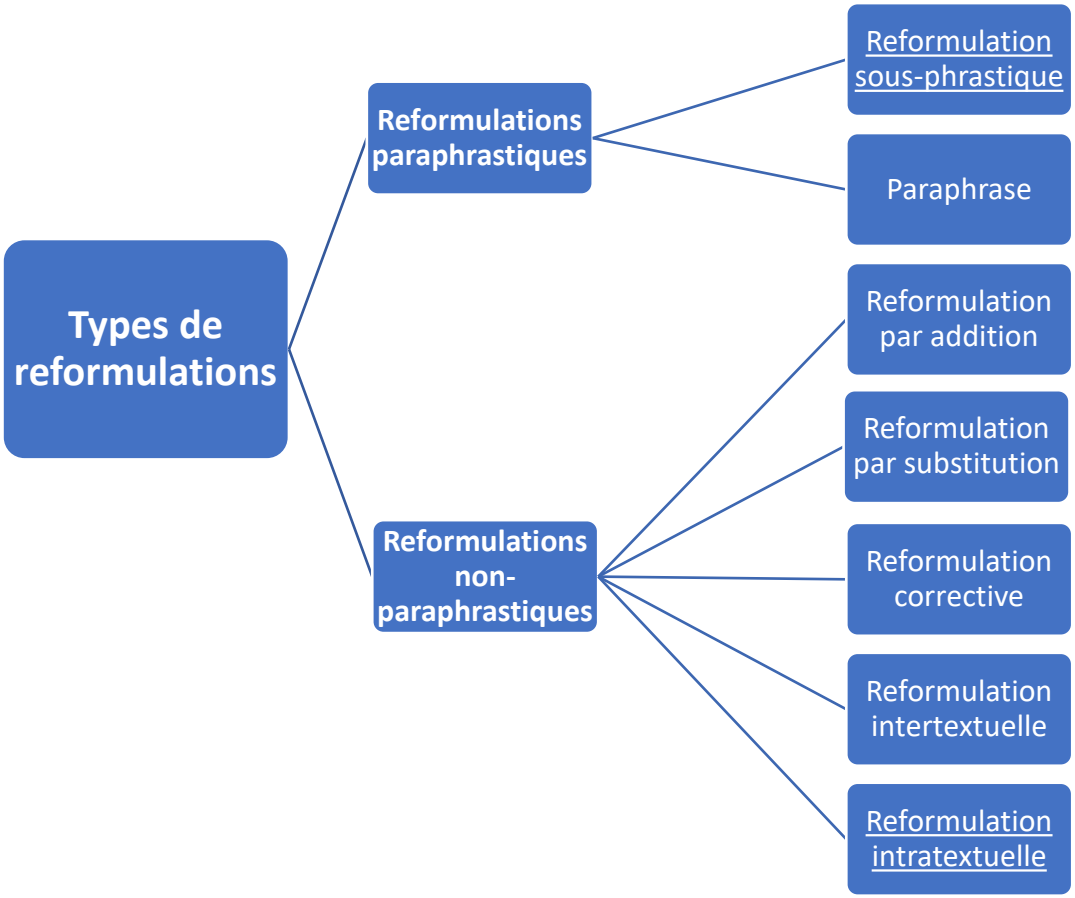

Figure 3. Les différents types de reformulations

Vu la grande variété de définitions, classifications et approches sur la reformulation, l'identification automatique de celle-ci est une opération difficile, ce qui nous motive à continuer nos recherches dans cette direction. Pour notre thèse, nous travaillons sur **les reformulations sous-phrastiques** et sur **les reformulations non-paraphrastiques de type intratextuelles** qui aident à **vulgariser le sens technique de termes médicaux**. Nous analysons les reformulations qui se trouvent dans la même phrase que le terme

médical et qui sont employées avec l'intention de *paraphraser, définir ou expliquer* le terme médical afin de rendre son sens plus accessible pour le grand public. Nous nous limitons à la taille de la phrase et nous excluons les liens de coréférence (Schnedecker et Landragin, 2014) ou les reformulations qui dépassent la portée d'une phrase.

***

Avant de construire un modèle d'exploitation automatique de la reformulation, sur la base de cette typologie de la reformulation et des marqueurs, nous analysons dans le **Chapitre 3.** les différentes **formes** que la reformulation peut prendre dans le discours.

3. Formes linguistiques de la reformulation

Les reformulations peuvent prendre différentes formes : celle de la **reprise**, de la **répétition**, de la **glose** ou de la **paraphrase** (Garcia-Debanc, 2015). La *reprise* est, selon Vion (2006), « de la pure et simple répétition d'un segment textuel aux différents degrés de ses reformulations ». Le commentaire apporté par l'énonciateur sur un mot ou une partie du discours porte le nom de *glose*. La *paraphrase* se définit comme l'équivalence basée sur un noyau sémantique commun (Fuchs, 1982). Afin de bien délimiter la reformulation qui fait l'objet de notre étude, nous analysons les définitions et l'étendue de deux principales formes de la reformulation, la glose et la paraphrase. Nous ne travaillons pas sur la reprise et la répétition, car notre objectif et d'identifier des **reformulations** qui ne sont pas sémantiquement identiques, mais qui peuvent apporter un point de vue nouveau au lecteur.

3.1 La glose

La « **glose** », définie simplement comme **commentaire sur un mot**, a été étudiée par plusieurs chercheurs (Authier-Revuz, 1995 ; Steuckardt et Niklas-Salminen, 2003 ; Steuckardt, 2005 ; Vassiliadou et Steuckardt, 2017). Pour expliquer la glose, Authier-Revuz (1995 : 6) donne l'exemple suivant : « Le sophomore, nom donné aux étudiants de seconde année, s'est fait coller ». Le commentaire « nom donné aux étudiants de seconde année » qui se trouve entre les virgules et peut-être identifié syntaxiquement comme une apposition offre une explication du terme spécialisé « sophomore ». Nous observons que la glose apparait ici comme une intervention de l'énonciateur pour clarifier une notion.

Lebaud et Ploog (2013) définissent la glose comme une reformulation d'une séquence linguistique qui garde la signification de celle-ci (Lebaud et Ploog, 2013 : 3). Les gloses sont catégorisées en deux types :

- **la glose épilinguistique** : « pratique langagière ordinaire et spontanée de tout locuteur d'une langue, sorte de rationalité pratique » ;
- **la glose métalinguistique** : « produit de la rationalité bavarde du linguiste, rationalité d'effectuation démonstrative explicite » (Lebaud et Ploog, 2013 : 9).

Nous présentons ces deux types de gloses en détail ci-dessous.

3.1.1 La glose épilinguistique

Ce type de glose aide à réduire l'incertitude sémantique et offre des équivalences considérées plus intelligibles, en employant des syntagmes de type : « Pour dire mieux, je dirai que… » ; « Pour dire les choses autrement, je dirai que… » ; « En disant ça, je veux dire que… » (Lebaud et Ploog, 2013 : 11). Ces marqueurs de subjectivité de la voix narrative, employés également dans les textes oraux comme écrits, servent à établir des équivalences interprétatives qui sont considérées plus intelligibles pour le lecteur. Leur but serait aussi d'ajouter à la construction des valeurs référentielles communes et partageables (Lebaud et Ploog, 2013 : 11).

3.1.2 La glose métalinguistique

La glose métalinguistique est contre la similarité, elle met en avant plutôt la *singularité* à l'aide de structures clivées. Ce type de glose se distingue par l'emploi de l'article défini « la » qui entraîne la préposition « à » comme dans l'exemple : « c'est la porte ouverte à toutes les dérives, tous les abus », ou par l'emploi de l'article indéfini « une » qui entraîne la préposition « sur » : « c'est une porte ouverte sur l'avenir » (Lebaud et Ploog, 2013 : 13). L'étude soutient que l'emploi de ces articles dans la structure clivée et l'ordre indiqué met en évidence l'unicité du sujet de la phrase.

> « S'agissant d'un type particulier de reformulation, nous réserverons donc le terme de glose à la reformulation d'une *séquence*, c'est-à-dire, […], d'**une petite suite de mots contextualisable et intelligible** », […]. » (Franckel (2005 : 55), cité par Lebaud et Ploog, 2013 : 12)[7]

Nous observons dans l'exemple ci-dessus que la glose métalinguistique « une petite suite de mots contextualisable et intelligible » donne le sens linguistique du terme « séquence », en mettant en avant sa spécificité et son unicité par l'emploi de l'article indéfini « une ».

Nous présentons par la suite d'autres types de glose, selon les approches de Magri (2018 ; 2013b).

[7] Le mot en italique appartient au texte original (Lebaud et Ploog, 2013 : 12), tandis que les mots surlignés et mis en gras représentent notre ajout.

3.1.3 La glose type « reformulation alternative »

Magri (2018) travaille sur un type de glose qu'elle appelle la **reformulation alternative**. Magri s'intéresse dans son étude aux **marqueurs de reformulation** qui sont formés sur le verbe « dire » utilisés dans un emploi métalinguistique. Ces marqueurs sont également appelés « marqueurs de glose » (Steuckardt, 2018). Les collocations analysées par Magri (2018) sont « c'est-à-dire », « autrement dit », « ce qui veut dire ». Magri (2018) met en avant dans ses recherches le rôle sémantique du marqueur « c'est-à-dire » qui sert à donner une *traduction* du terme reformulé. Cette *traduction* fait la transition entre l'univers inconnu par le lecteur vers le monde connu.

L'important à noter est que, selon Magri (2018), la reformulation alternative propose une substitution de l'énoncé-cible à l'énoncé-source, plaçant entre les deux énoncés une relation d'égalité.

3.1.4 La glose type « reformulation corrective »

Selon Magri (2018), la reformulation alternative qui représente l'équivalence entre deux énoncés peut être mise en opposition avec **la reformulation « corrective »** (terme proposé par Magri (2018)). Ce type de reformulation serait signalé par l'expression adverbiale « ou plutôt ». Le rôle de substitution de la reformulation évolue dans un rôle d'amélioration du sens de l'énoncé source. Pour expliquer ceci, Magri donne l'exemple suivant :

> *C'était une blonde, ou plutôt une blondine, une fraîche, toute fraîche créature qu'on devinait rose et potelée sous l'étoffe gonflée du corsage (Maupassant, Contes et nouvelles, t. 2, 1886)*. (Magri, 2018)

On observe que la reformulation « ou plutôt une blondine, une fraîche, toute fraîche créature » exprime le désir de l'énonciateur de corriger sa déclaration initiale « C'était une blonde ». Dans son étude, Magri (2018) souligne que la reformulation corrective suivie par « ou plutôt » indique une relation de synonymie ou une relation hiérarchique d'hyponymie ou d'hyperonymie, comme dans l'exemple : « C'était un petit vallon, ou plutôt une grande ondulation de terres de mauvaise qualité. » (Maupassant, Contes et nouvelles, t. 1, 1889 ; cité par Magri, 2018).

3.1.5 La « glose savante »

Magri-Mourgues (2013b) appelle des « **gloses savantes** » les reformulations introduites par des expressions formées sur le verbe « dire » comme : « autrement dit », « c'est-à-dire », « ce qui veut dire » et par la conjonction « ou ». La glose savante peut être marquée aussi au niveau orthotypographique par des signes doubles de ponctuation, des tirets ou des parenthèses. Magri-Mourgues (2013b) met en évidence les particularités de fonctionnement de ces gloses par rapport aux marqueurs introducteurs :

- **Marqueur formé sur le verbe « dire »**. Les gloses savantes introduites par ce marqueur ont une **valeur métalinguistique** et leur emploi est orienté vers le terme-cible, en suivant une orientation **du terme inconnu au terme connu**.

  > *Rendez-vous habituel des gens qui appartiennent à l'opinion modérée, et qu'on appelle cangrejos, c'est-à-dire* ***écrevisses*** *(Th. Gautier, Voyage en Espagne, p. 149)* (Magri-Mourgues, 2013b : 224) ;

- **Marqueur représenté par la conjonction « ou »** ou encore **les doubles points**. Ce type de marqueur aide à maintenir **une égalité de statut entre deux unités** (le terme commenté et la glose savante) et n'établit pas un ordre précis. Il peut adopter **les deux orientations, du terme inconnu au terme connu** et également **du terme connu au terme inconnu**.

  > *Ses bodegas, ou* ***magasins de vins****, immenses celliers aux grands toits de tuiles, aux longues murailles blanches privées de fenêtres. (Th. Gautier, Voyage en Espagne, p. 381[10])* (Magri-Mourgues, 2013b : 224) ;
  > *Un peu en avant du corps de logis des cuisines, est un charmant petit palais, entouré d'une galerie ou portique au rez-de-chaussée : c'est celui des pages ou* ***icoglans*** *du grand sérail. (A. De Lamartine, Voyage en Orient, p. 419)* (Magri-Mourgues, 2013b : 224).

Nous observons que dans le premier exemple la glose savante « magasins de vins » offre la clarification du terme « bodegas », en passant du terme inconnu au terme connu, appartenant au langage commun. Ce processus est identifié comme **vulgarisation** (transposition dans le langage général) d'un terme obscur, scientifique ou méconnu. Ces dernières qualités du terme sont jugées par le narrateur en fonction du public cible du texte écrit.

Dans le deuxième exemple, l'approche est du connu vers l'inconnu : « des pages ou **icoglans** ». Dans cette situation de discours, le narrateur du récit de voyage donne à son énoncé une orientation pédagogique, d'apport de nouvelles connaissances. Tout comme pour le premier exemple, le narrateur s'adapte en fonction d'un public cible présupposé : un passionné qui souhaite connaître des mondes inconnus.

***

Ces types de reformulations peuvent être employées avec un rôle de glose (quand l'énonciateur veut apporter une précision ou un point de vue à son énoncé), mais également avec un rôle paraphrastique. La glose savante peut être utilisée dans les textes de vulgarisation dont la voix narrative est présente, comme dans les blogs scientifiques ou les textes destinés à un public enfant. Nos corpus d'étude présentent des notions médicales de manière objective, sans intervention de la voix du narrateur, ce qui peut indiquer une absence de gloses savantes.

Nous nous intéressons par la suite à **la paraphrase** et ses possibles représentations dans le texte écrit.

3.2 La paraphrase

Le terme de **paraphrase** est défini pour la première fois dans la linguistique par Harris (1976) comme une **propriété des langues** qui permet d'établir des **relations de transformation entre des phrases d'une même langue**. Un peu plus tard, Fuchs (1982) réalise une synthèse sur les différentes approches linguistiques de la paraphrase. Fuchs (1982) conclut que la paraphrase se définit comme **l'équivalence basée sur un noyau sémantique commun**. En analysant l'hypothèse distributionnaliste de Harris (1954), Bouamor (2012) affirme que les paraphrases sont des expressions alignées qui se trouvent dans des contextes similaires.

La paraphrase est considérée comme le « résultat » de la reformulation. Eshkol-Taravella et Grabar (2018) réalisent une synthèse de différents points de vue selon lesquels la paraphrase a été étudiée :

- **la situation d'énonciation** qui donne des paraphrases linguistiques, sémantiques et pragmatiques (Culioli, 1983 ; Fuchs, 1982, 1994 ; Martin, 1976 ; Vezin, 1976) ;

- **les transformations linguistiques des segments paraphrasés** (Mel'čuk, 1988 ; Vila *et al.*, 2011 ; Bhagat et Hovy, 2013) ;
- **la taille des entités paraphrasées**, c'est-à-dire la nature syntaxique du référent (mot ou énoncé paraphrasé) (Fløttum, 1995 ; Bouamor, 2012 ; Bouamor *et al.*, 2012).

Pennec (2020 : 62) considère que les **paraphrases** sont des reformulations qui expriment **l'équivalence sémantique** au sens propre (formulé = reformulé) et que le **marqueur de reformulation** est souvent **absent** dans les paraphrases ou peut être supprimé (selon Vassiliadou, 2020).

Fuchs (1994 : 89) appelle la paraphrase « **le cas prototypique de reformulation** », dont nous retrouvons une fonction métalinguistique d'identification entre les deux segments du discours, le formulé et le reformulé. Pour analyser la paraphrase dans le discours, Fuchs (1980) propose deux approches : **l'approche énonciative** et **l'approche logique**. Nous développons ces concepts dans les sous-chapitres suivants.

3.2.1 La paraphrase selon l'approche énonciative

L'approche énonciative prend en compte la forme de l'acte d'énonciation. La paraphrase est analysée à travers son rôle de reconstruction référentielle. La paraphrase a un référent dans le texte source et se construit sur la base d'un élément de l'énonciation qui devient une valeur référentielle. Afin d'expliquer cette approche énonciative de la paraphrase, Fuchs (1980) donne l'exemple suivant :

- « Les bons résultats obtenus par l'éditeur dans la vente de l'ouvrage l'ont conduit à le diffuser massivement » paraphrasé par :
- « L'éditeur a été conduit à diffuser massivement l'ouvrage après avoir obtenu de bons résultats dans la vente de celui-ci ». (Fuchs, 1980)

Nous observons dans cet exemple que la chronologie temporelle des événements de la phrase est conservée dans la paraphrase. Par conséquent, la dimension temporelle devient la valeur référentielle, l'élément d'énonciation du texte source qui crée le lien paraphrastique.

3.2.2 La paraphrase selon l'approche logique

La deuxième approche proposée par Fuchs est *l'approche logique*. Pour l'expliquer, Fuchs fait appel à la théorie du linguiste Robert Martin qui considère que le sens linguistique de la paraphrase varie en fonction de la situation et le traitement du sens se fait d'un point de vue pragmatique. Martin (1976) étudie la paraphrase en fonction de trois variables : la signification (sens et interprétation), la double antonymie (grammaticale et lexicale) et la substitution synonymique. L'approche logique de Martin (1976) est un outil important pour la classification des paraphrases du point de vue sémantique et pragmatique.

Fuchs (1994) situe la paraphrase dans la perspective d'une linguistique de l'énonciation. La paraphrase est analysée par une **approche discursive** de l'énonciation, en mettant en avant les différents facteurs qui agissent sur la reformulation : l'objectif de la reformulation, le public cible de celle-ci et la subjectivité de l'énonciateur.

Bouamor (2012) classe les paraphrases selon les niveaux de granularité textuelle (Vanrullen, 2003), par rapport à la composition de la phrase. Dans ce sens, Bouamor (2012) identifie trois catégories de paraphrases :

- **paraphrases lexicales** - des éléments lexicaux individuels ayant le même sens ;
- **paraphrases sous-phrastiques** - des unités textuelles (segments ou fragments de texte) qui partagent le même contenu sémantique ;
- **paraphrases phrastiques** - deux phrases ou énoncés qui ont un même contenu sémantique.

Nous présentons ci-dessous plus en détail les trois types de paraphrases de Bouamor (2012).

3.2.2.1 La paraphrase lexicale

Bouamor (2012) observe que les paraphrases lexicales, des unités lexicales individuelles avec la même signification, sont souvent des **synonymes**, comme dans les exemples « marcher – se déplacer » ou « couette – couverture ». Bouamor (2012) note que la paraphrase lexicale peut dépasser le sens de la synonymie, en allant même à **l'hyperonymie**, qui représente la relation hiérarchique dont le premier terme, plus général, englobe l'extension du second terme, qui est plus spécifique (par exemple, « meuble » est hyperonyme du mot « chaise »). Nous rajoutons également **l'hyponymie**, qui représente la

relation hiérarchique inverse, dont le premier terme plus spécifique est inclus dans la classe des notions plus générales (exemple : le mot « rose » est hyponyme du mot « fleur »). Ces relations sémantiques, d'hyperonymie et d'hyponymie peuvent être considérées, selon Bouamor (2012) comme des relations paraphrastiques lexicales.

3.2.2.2 La paraphrase sous-phrastique

Bouamor (2012) considère que la **paraphrase sous-phrastique** peut être constituée également des paires de mots ou des paires de groupes de mots qu'elle identifie comme « syntagmes ou fragments textuels quelconques » (Bouamor, 2012 : 21). La taille des paraphrases sous-phrastiques peut être aussi grande que nécessaire, mais elle doit respecter la limite d'une phrase. Les éléments constituants de la paraphrase sous-phrastique doivent être en relation d'équivalence sémantique dans un contexte donné de la phrase (comme dans l'exemple de Bouamor (2012 : 21), « envisage-t-elle » a le même sens que « a-t-elle l'intention »). Ces éléments textuels peuvent se présenter également sous forme de *patrons* qui ont le rôle de lier des éléments variables de la phrase, comme dans l'exemple « X ne doute pas de Y » - « X est sûr de Y » (Bouamor, 2012 : 21). Nous remarquons que les paraphrases sous-phrastiques jouent un rôle de lien sémantique, mais que ce rôle est intrinsèquement connecté à la logique et à l'organisation syntaxique à l'intérieur de la phrase.

3.2.2.3 La paraphrase phrastique

Bouamor (2012) précise que deux phrases qui ont le même contenu sémantique peuvent être appelées des **paraphrases phrastiques** ou **paraphrases d'énoncé**, comme dans l'exemple qu'elle donne : « Elle a grondé son enfant » <=> « Elle s'est fâchée contre son enfant » (Bouamor, 2012 : 21). Bouamor (2012) souligne que la difficulté de l'utilisation des paraphrases phrastiques augmente avec le style et le registre du langage utilisé. Dans ce sens, elle donne l'exemple de phrases suivantes : « Vous n'êtes même pas en mesure de me donner ce renseignement » <=> « Tu n'es même pas fichu de me passer ce tuyau » (Bouamor, 2012 : 21). Nous observons que le deuxième énoncé utilise un langage argotique et que le sens de la phrase reste obscur si le destinataire de la phrase ne maîtrise pas suffisamment l'argot. La même situation arrive quand le locuteur utilise un langage très technique, scientifique (médical, juridique, informatique), langage que le destinataire ne connaît pas. Dans notre thèse, nous nous intéressons à cette dimension de l'incompatibilité

des registres de langue entre le locuteur et le public cible, en nous concentrant sur le langage scientifique du domaine médical et ses paraphrases en langage général.

3.3 Bilan

Les principales formes de reformulation étudiées dans la littérature sont les gloses et les paraphrases. Nous remarquons que chaque forme de reformulation s'insère à une instance différente du discours. Par exemple, la glose est adaptée au discours interne de la voix narrative, tandis que la paraphrase est plutôt liée à la sémantique du référent reformulé. Nous rajoutons les reprises (répétitions, reformulations correctives) comme formes de reformulation, mais celles-ci peuvent prendre aussi bien le corps d'une glose que celui d'une paraphrase. La **Figure 4** reprend ces trois formes de reformulation avec leurs définitions et sous-types dans un essai de synthétiser et différencier leurs caractéristiques.

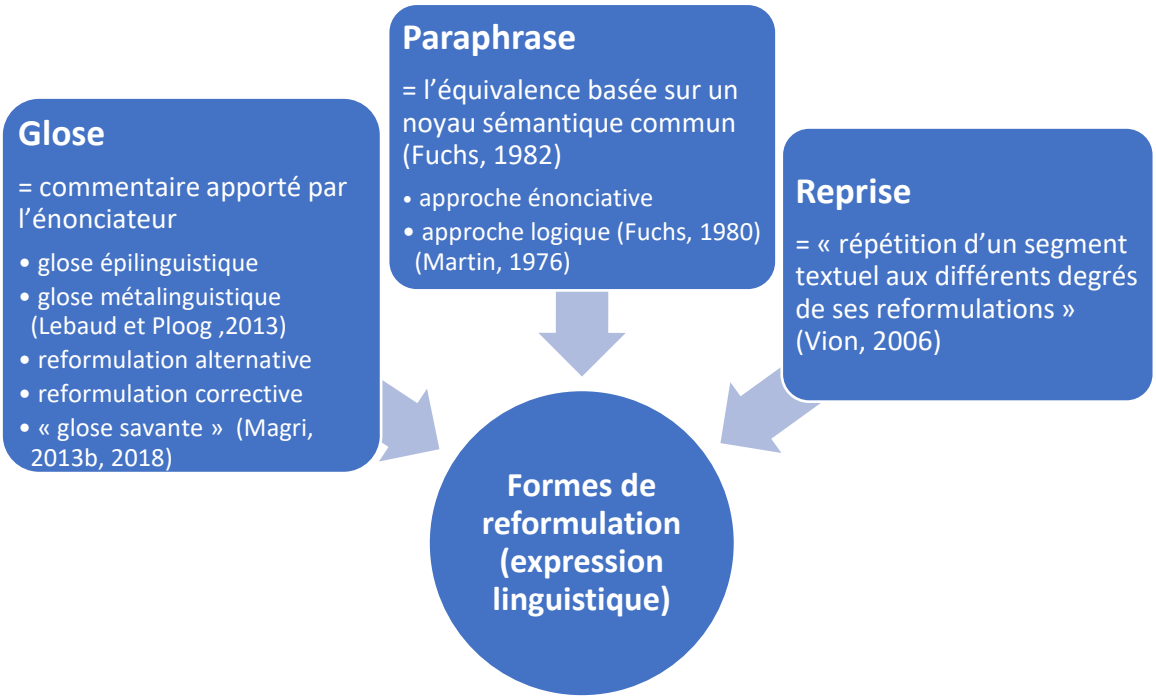

Figure 4. Formes de la reformulation

Nous constatons la variété d'approches concernant la définition et l'identification de de la glose et de la paraphrase comme formes linguistiques de la reformulation. Nous soulignons l'accent mis sur la sémantique de la paraphrase en lien avec un référent qui la précède, dans le cadre pragmatique de l'énonciation et en allant plus loin vers l'objectif

d'expliquer des notions à son public cible. Les paraphrases sont employées dans les textes de vulgarisation avec l'objectif de rendre accessible les notions médicales, ce que nous envisageons identifier également dans nos corpus.

En guise de conclusion de ces deux derniers **Chapitres (2 et 3)** sur **les types et les formes des reformulations**, nous avons regroupé toutes ces classifications dans le **Tableau 2**. Nous avons observé que dans la littérature il existe plusieurs dénominations pour le même (ou presque) type ou forme de reformulation. Dans un besoin de cohérence et clarté, nous avons indiqué dans le **Tableau 2** les relations d'équivalence totale (« **=** ») ou approximative (« **≈** ») entre les différents termes utilisés dans l'état de l'art pour désigner la reformulation dans sa diversité et complexité. Pour de raisons d'efficacité, nous avons choisi pour chaque terminologie multiple un seul terme à utiliser par la suite dans notre travail de recherche, **terme noté en gras** dans la dernière colonne intitulée « **Notre choix terminologique** ».

ÉQUIVALENCES TERMINOLOGIQUES TYPES ET FORMES DE LA REFORMULATION		
Terminologie équivalente ou semi-équivalente	**Auteurs**	**Notre choix terminologique**
Reformulation paraphrastique **=** Paraphrase	(Magri-Mourgues, 2012)	**Reformulation paraphrastique**
	(Fuchs, 1982 ; Garcia-Debanc, 2015)	
Reformulation sous-phrastique **=** Paraphrase sous-phrastique	Notre proposition	**Reformulation sous-phrastique**
	(Bouamor, 2012)	
Reformulation intratextuelle **=** Traduction intralinguale **≈** Reformulation corrective	(Vargas, 2008)	**Traduction intralinguale**
	(Jakobson, 1959)	
	(Magri-Mourgues, 2013a)	
Reformulation par superposition **=** Traduction interlinguale	(Magri-Mourgues, 2013a)	**Traduction interlinguale**
	(Jakobson, 1959)	
Reformulation intertextuelle **=** Glose	(Drescher, 2008)	**Glose**
	(Authier-Revuz, 1995 ; Steuckardt et Niklas-Salminen, 2003 ; Lebaud et Ploog, 2013 ; Vassiliadou et Steuckardt, 2017)	
Reformulation par substitution	(Magri-Mourgues, 2013a)	

= Reformulation corrective **≈** Reformulation alternative	(Magri, 2018)	**Reformulation corrective**
	(Magri, 2018)	
Reprise **=** Reformulation par addition	(Vion, 2006)	**Reprise**
	(Magri-Mourgues, 2013a)	
Glose métalinguistique **=** Glose savante	(Lebaud et Ploog, 2013)	**Glose savante**
	(Magri-Mourgues, 2013b)	

Tableau 2. Tableau d'équivalences terminologiques : types / formes de la reformulation

Les terminologies choisies reflètent nos directions de recherche. Comme nous travaillons sur la **reformulation sous-phrastique médicale**, nous avons choisi les termes de **reformulation paraphrastique**, définie comme l'équivalence basée sur un noyau sémantique commun, selon Fuchs (1982). Nous proposons le terme de **reformulation sous-phrastique** au lieu du terme de Bouamor (2012), **paraphrase sous-phrastique**, car nous travaillons sur une notion de *reformulation comprise au sens large*. Notre proposition garde la dimension de relation sémantique d'équivalence dans un contexte donné de la phrase, sans dépasser la taille de celle-ci (Bouamor, 2012), sauf pour les reformulations de type explication et exemplification. Nous développons davantage notre concept et notre position théorique dans la **Partie III** de ce travail.

Nous avons choisi les termes **traduction intralinguale** et **traduction interlinguale** de Jakobson (1959) parce qu'ils expriment le caractère de transformation lexicale et syntaxique de la reformulation dans la même langue, en gardant la valeur sémantique du contenu transformé. La **traduction interlinguale** (Jakobson, 1959) se caractérise dans notre cas par la *traduction* à travers la reformulation des notions dans des registres de langue différents (scientifique et de vulgarisation).

La **reprise** est, selon Vion (2006), « de la pure et simple répétition d'un segment textuel aux différents degrés de ses reformulations ». Nous éviterons les reprises dans notre recherche, car nous travaillons sur la **variation lexicale et syntaxique des reformulations**. La **reformulation corrective** (Magri, 2018) a le rôle d'améliorer le sens de l'énoncé source. Cette valeur pragmatique de la paraphrase peut se retrouver souvent dans les textes médicaux dans l'essai de mieux vulgariser les termes médicaux selon le public cible.

La **glose** se définit tout simplement comme commentaire sur un mot (Authier-Revuz, 1995 ; Steuckardt et Niklas-Salminen, 2003 ; Steuckardt, 2005 ; Vassiliadou et Steuckardt,

2017). Ce qui nous intéresse dans notre recherche est de bien repérer dans les textes la **glose savante** de Magri-Mourgues (2013b), si identifiable. Les gloses savantes sont des reformulations introduites par des expressions formées sur le verbe « dire » comme : « autrement dit », « c'est-à-dire », « ce qui veut dire » et par la conjonction « ou ». Au niveau orthotypographique, la glose savante peut être marquée par des signes doubles de ponctuation, des tirets ou des parenthèses. Nous testerons nos outils sur ces types de marqueurs afin de trouver des reformulations sous-phrastiques et de garder à l'écart les gloses savantes qui expriment des commentaires subjectifs du narrateur. Le but de notre recherche est d'identifier **reformulations valides de termes médicaux** et non pas des appréciations subjectives.

***

Dans la suite de notre étude sur la reformulation, nous procédons à une analyse plus fine de la reformulation en fonction de l'impact que chaque niveau linguistique (lexical, syntaxique, sémantique) a sur sa forme et son utilisation dans les textes écrits.

4. La reformulation et les différents niveaux linguistiques

En plus de différents types de reformulation et les différentes approches évoquées, nous nous rapportons également aux **fonctions de la reformulation**. Ces fonctions peuvent être lexicales, syntaxiques, sémantiques ou pragmatiques. Si la reformulation a le rôle de revenir sur un « dit » antérieur et de lui apporter des rajouts ou des modifications, ceux-ci peuvent être identifiés à plusieurs niveaux. Nous nous rapportons à ces fonctions et divers niveaux pour cadrer notre choix dans le type de reformulation recherchée dans cette étude.

4.1 La reformulation au niveau lexical

Săpoiu (2013) a étudié les **relations lexicales d'hyponymie et d'hyperonymie** dans les définitions des termes médicaux tels qu'ils apparaissent dans les dictionnaires généraux de la langue roumaine. Dans ces définitions, le terme est l'hyponyme (un mot spécifique), tandis que la définition commence par ou contient un hyperonyme (le mot générique) (Bidu-Vrânceanu, 2007). Săpoiu (2010) a proposé une classification à plusieurs niveaux des hyperonymes dans la définition d'un terme médical, en partant du plus général (« cardiomalacie : distrofie » / *cardiomalacie : dystrophie*), jusqu'à la différence la plus spécifique : déterminant (« generalizată » / *généralisé*), anatomie (« a miocardului » / *du myocarde*), processus / conséquence / cause (« prin infiltrarea fibrei musculare cu grăsimi » / *en infiltrant la fibre musculaire avec de la graisse*).

Barbu Mititelu (2011) a travaillé sur la **relation lexicale d'hyponymie / hyperonymie** appliquée sur des textes journalistiques et médicaux en identifiant automatiquement deux types de motifs : (1) « modèle d'hyponymie HYPERNYM » et (2) « modèle d'hyponymie HYPONYM HYPERNYM ». Les meilleurs résultats en termes de précision (plus de 66,66%) dans le domaine médical ont été obtenus avec les motifs suivants : GN (groupe nominal), « adică » / **qui est** (marqueur) GN (groupe nominal) ; GN « care fi » / *qui est* GN ; GN « în special » / *spécialement* GN ; GN « inclusiv » / *inclusivement* GN ; GN « precum » / *comme* GN ; GN « (în) afară de » / *en plus* GN ; GN « de obicei » / *habituellement* GN ; GN « ci (şi/doar) » / *mais (aussi/seulement)* GN.

4.2 La reformulation comme traduction intralinguale : niveau sémantique

La reformulation a lieu à l'intérieur de la même langue, par conséquent elle est intrinsèquement liée au lexique et, vu son rôle du « dire autrement », à la synonymie. La reformulation peut être donc considérée de l'ordre de la traduction intralinguale.

Ce concept de la traduction dans la même langue a été développé par Roman Jakobson (1959). Jakobson donne une définition simple, mais complète, de ce type de traduction : « La traduction intralinguale ou la reformulation consiste dans l'interprétation des signes linguistiques à travers d'autres signes de la même langue » (Jakobson, 1959). Nous utilisons d'autres mots pour expliquer les mots : c'est le principe d'une traduction, mais également de la reformulation. Nous représentons la traduction intralinguale par le niveau sémantique justifié par l'emploi des différents lexiques, mais la reformulation varie aussi en fonction de la syntaxe, de la phrase et du discours.

4.3 La reformulation au niveau syntaxique

Au niveau syntaxique, la reformulation fait partie de la structure grammaticale de la phrase et se trouve en lien direct avec la notion source. Ce lien est maintenu par les marqueurs syntaxiques. La reformulation peut être introduite par différents types de marqueurs (adverbes, expressions adverbiales, prépositions, conjonctions), mais également par des signes des ponctuation comme les virgules, les doubles points ou les points-virgules.

Afin d'identifier les reformulations dans de grands corpus des textes et les délimiter du reste du texte, nous avons besoin d'identifier d'abord leurs marqueurs syntaxiques présentés dans le **Chapitre 2.3**.

4.4 La reformulation au niveau du discours

Au niveau du discours, la reformulation reprend un élément syntagmatique qui a le rôle de **référent** de la reformulation. Ces deux syntagmes (le référent et la reformulation) peuvent se substituer l'un à l'autre parce qu'ils ont la même signification (culturelle, notionnelle ou dans la réalité extralinguistique du locuteur et de l'interlocuteur, dans le

domaine d'application). Ce lien **référent – reformulation** assure la cohésion textuelle qui s'exprime par la présence des marqueurs de cohésion tels que les chaînes de coréférence (Schnedecker et Landragin, 2014) ou les marqueurs de discours.

La délimitation du référent se réalise en fonction des marqueurs de la reformulation. Pour cela nous menons des recherches automatiques précises dans le syntagme qui précède la reformulation avec ses marqueurs (référent de type nom, verbe, etc.).

4.5 La reformulation au niveau pragmatique

Au niveau pragmatique la reformulation est analysée du point de vue interactif, afin de mettre en avant la **justification de la reformulation** et ses **enjeux communicationnels**.

Grabar et Eshkol-Taravella (2016b) étudient la **fonction pragmatique** des reformulations pour identifier l'objectif précis de celle-ci. Leur étude traite la reformulation qui s'établit à l'aide des marqueurs formés sur le verbe « dire », comme « c'est-à-dire », « je veux dire », « disons ». Ces marqueurs lient le segment reformulé du segment de la reformulation.

Ces types de **fonctions pragmatiques** ont été proposés dans la littérature par plusieurs linguistes (Gülich et Kotschi, 1987 ; Hölker, 1988 ; Beeching, 2007 ; Kanaan, 2011). Grabar et Eshkol-Taravella (2016b) réalisent une vaste typologie des possibles fonctions pragmatiques de la reformulation. Nous citons leurs exemples (extraits à partir de leur corpus d'étude) afin de mieux justifier notre choix parmi toutes ces fonctions pragmatiques.

Grabar et Eshkol-Taravella (2016b) proposent 11 catégories de fonctions pragmatiques :

- **Définition**. Le locuteur donne une définition neutre et précise avec l'objectif de faire comprendre un terme technique : « avec une ETO c'est à dire une echographie tansoesophagienne (une écho ou le palpeur est introduit dans l'estomac) » (Grabar et Eshkol-Taravella, 2016b : 5) ;
- **Explication**. Le locuteur explique quelque chose d'une manière moins formelle que dans le cas de la définition : « j'ai entendu parler (sur le net) des bêtabloquants, or il parait que céest des médicaments a vie et pour la vie, c'est-

à-dire ils ne sont efficaces que lorsquéils sont pris tous les jours »[8] (Grabar et Eshkol-Taravella, 2016b : 6) ;

- **Exemplification**. Les reformulations se présentent sous forme d'entités nommées ou d'énumérations : « 2 heures plus tard, elle a eu tous les symptomes d'un avc c'est à dire perte de parole, hémiplégie, fièvre... » (Grabar et Eshkol-Taravella, 2016b : 6) ;
- **Justification**. Le locuteur justifie des événements ou des actes à son interlocuteur : « la langue française est plus difficile disons on peut pas dire la plus difficile des langues européennes mais c'est difficile » (Grabar et Eshkol-Taravella, 2016b : 6) ;
- **Précision**. Le locuteur ajoute une information dans le but d'éclaircir : « les aînés partent eux aussi de manière moins systématique c'est-à-dire que les aînés partent pas forcément tous les ans mais souvent » (Grabar et Eshkol-Taravella, 2016b : 6) ;
- **Dénomination**. Le locuteur attribue un nom à une entité unique déjà mentionnée : « en particulier c'est l'endroit où en somme ça s'est produit le plus au début c'est-à-dire à Nanterre » (Grabar et Eshkol-Taravella, 2016b : 6) ;
- **Résultat**. Le locuteur résume ou indique une conséquence : « avec l'accent un peu de travers je veux dire l'accent » (Grabar et Eshkol-Taravella, 2016b : 6) ;
- **Correction linguistique**. Le locuteur apporte une correction : « des artisans euh hm hm hm hm hm alors c'est-à-dire artisans » (Grabar et Eshkol-Taravella, 2016b : 7) ;
- **Correction référentielle**. Le locuteur apporte une correction de lieu, de temps, ou autre : « j'habitais rue Lazare Carnot c'est à dire donc au sud de la Source » (Grabar et Eshkol-Taravella, 2016b : 7) ;
- **Paraphrase**. L'information est répétée d'une autre manière : « quelque chose de potable disons quelque chose euh de correct » (Grabar et Eshkol-Taravella, 2016b : 7) ;
- **Opposition**. L'information est reprise sous une forme négative : « elle était incapable de rien faire elle au point de vue vendeuse c'est-à-dire elle elle est pas mauvaise euh elle est agréable au point de vue clientèle elle a été incapable de passer son certificat d'études[9] » (Grabar et Eshkol-Taravella, 2016b : 7).

[8] Les fautes d'orthographe et les répétitions présentes dans les exemples font partie du texte original.
[9] Le texte a été surligné par nous.

Dans notre travail de recherche, nous nous rapportons aux cinq fonctions pragmatiques suivantes : **définition, explication, exemplification, dénomination et paraphrase**. Notre choix est justifié par le rôle que ces fonctions pragmatiques jouent dans le processus de **simplification lexicale**. Nous prenons en compte **le public cible** du texte médical (spécialiste ou grand public) et nous identifions de façon automatique les reformulations qui remplissent ces fonctions pragmatiques. Nous portons notre attention à la **fonction de dénomination** notamment pour les maladies qui portent des noms propres (comme Parkinson, Alzheimer).

D'autres fonctions, comme *la justification*, apportent plutôt un changement de perspective du discours en prenant en compte l'avis du locuteur sur le contenu paraphrasé. Nous considérons que cette fonction est présente dans plutôt dans les textes oraux, car elle exprime la *justification* du locuteur. Dans notre type de corpus, écrit, scientifique et de vulgarisation, son apparition doit être minime, faute d'absence d'emplois déictiques du locuteur (de type pronoms personnels « je », « nous »). La fonction *d'opposition* ne nous intéresse pas dans notre recherche parce que nous évaluons les résultats de reformulation positive et nous écartons les exemples de reformulation négative.

***

Dans notre travail nous analysons **le lien qui existe entre les fonctions sémantico-pragmatique et les relations lexicales** pour les paires **terme médical – reformulation**, identifiés dans des textes écrits. Nous présentons nos hypothèses de recherche dans la **Partie III** et les analyses dans la **Partie IV**.

4.6 Bilan

Dans le cas de la reformulation, sa forme et son emploi diffèrent en fonction du niveau linguistique auquel elle opère. Afin d'identifier de façon automatique les reformulations qui sont employées dans les textes écrits médicaux dans le but de la simplification lexicale, nous prenons en compte le fonctionnement des reformulations sous-phrastiques dans les niveaux linguistiques présentés. Cette analyse nous permettra de comprendre le rôle de ces reformulations dans la vulgarisation scientifique.

Figure 5. Les fonctions de la reformulation selon les niveaux linguistiques

D'autres variables très importantes pour notre recherche sont la **simplification** des textes écrits et la **vulgarisation scientifique** à travers la reformulation. Dans le **Chapitre 5,** nous présentons l'état de l'art sur la simplification lexicale et le lien avec la reformulation par rapport à un public cible spécifique.

5. La simplification lexicale par la reformulation

La simplification de textes peut se réaliser à plusieurs niveaux, en fonction de l'objectif, du public cible et du contenu textuel. Une méthode de simplification nécessite un processus de **simplification lexicale** qui représente le remplacement de termes complexes ou techniques avec des mots plus accessibles pour le public cible.

La recherche sur la simplification lexicale des textes écrits s'est développée dans deux directions :

- **l'explication des termes difficiles avec des définitions du dictionnaire**, d'autres explications en notes de bas de page ou dans un glossaire annexe

et

- **les substitutions lexicales**, procédé qui consiste à remplacer des termes considérés difficiles pour le public visé par des synonymes plus compréhensibles ou adaptés à la connaissance du lecteur cible.

Selon Siddharthan (2014), le rôle de la simplification lexicale est celui d'aider différentes catégories des lecteurs (enfants, étudiants et adultes avec un niveau faible d'alphabétisation ou tout simplement non-spécialistes du sujet présenté) à accéder plus facilement au contenu et au sens des informations des textes de spécialité (ou contenant du langage argotique ou très populaire). Un des objectifs de la simplification lexicale est d'augmenter **la lisibilité** des textes par rapport aux lecteurs cibles de celui-ci. **La lisibilité** fait référence au contenu des textes, sur deux axes : *le fond*, plus précisément aux idées exprimées dans le texte, et *la forme*, la façon d'exprimer ces contenus linguistiques (Laframboise, 1978). La lisibilité sert à identifier les parties du texte qui sont à simplifier ou à évaluer le résultat de la simplification (François, 2011).

Si la simplification lexicale peut avoir des avantages reconnus en général, il existe quand même des voix dans la recherche qui sont *contre* la simplification des textes. Siddharthan (2014) fait une synthèse de ces opinions qui se démarquent. Selon Honeyfield (1997), la simplification des textes empêche les lecteurs à acquérir les formes naturelles du langage. Honeyfield (1997) soutient également qu'un autre désavantage serait l'homogénéisation lexicale, car les informations importantes dans un texte sont souvent repérables à l'aide du vocabulaire spécialisé. Concernant les textes simplifiés pour les

lecteurs enfants, certains chercheurs soutiennent que les enfants ne les trouvent pas assez intéressantes par rapport aux textes originaux (Green et Olsen, 1986). Nous développons ces approches par la suite.

Nous considérons que la simplification lexicale joue un rôle important dans la mission de rendre un contenu plus accessible à différents types de lecteurs, que ce soit pour une visée didactique, de vulgarisation scientifique ou de partage de connaissances avec un public large. Nous développons cette position dans la section suivante.

5.1 Le rôle pédagogique de la reformulation

Magri-Mourgues (2013a) estime que la reformulation a principalement un **rôle didactique**, avec le but d'aller d'un terme inconnu vers un terme connu et vice-versa, afin d'apporter de nouvelles connaissances. Le sens du concept reformulé est gardé, l'élément qui varie est l'expression linguistique de ce concept. Le lecteur cible devient un facteur important dans le choix de la reformulation, pour compléter ses connaissances.

Le rôle pédagogique de la reformulation se fait remarquer dans l'adaptation des textes pour un public cible, par exemple pour les enfants. Green et Olsen (1986) ont mené une étude pour évaluer la pertinence de ces adaptations (textes reformulés) pour l'éducation des enfants. Ils ont analysé deux livres originaux pour les enfants et les adaptations de ces livres. L'objectif de l'étude est de savoir si les enfants préfèrent les histoires originales, non-reformulées (qui ont des phrases plus longues et complexes, mais adaptées à leur niveau scolaire) ou les textes reformulés. Les textes adaptés sont conformes avec les capacités de compréhension des enfants.

L'étude a été menée sur 58 élèves américains de la deuxième année d'école primaire (donc la langue des textes étant l'anglais). Les résultats indiquent une forte préférence pour les histoires originales par rapport aux textes reformulés. Ce résultat a été observé particulièrement chez les lecteurs avec un niveau moyen et bas d'alphabétisation. En plus, aucune différence notable de compréhension n'a pas été observée entre les textes originaux et les adaptations. Les résultats de l'étude mènent à la conclusion qu'il n'y a aucun motif valable sur le plan éducatif pour continuer à reformuler des textes autrement appropriés pour augmenter la lisibilité. Il faut pourtant prendre en compte que l'étude a été réalisée sur un nombre réduit d'enfants provenant de la même école et uniquement sur la langue anglaise. Ces résultats ne sont pas à généraliser, mais la recherche de Green et

Olsen (1986) pose des questions importantes à considérer dans la reformulation des textes pour un public enfant.

Les reformulations sont utiles également lors de l'acquisition linguistique réalisée par les enfants ou les apprenants d'une langue. Martinot (2010) étudie les reformulations réalisées par les enfants lors de l'apprentissage de leur langue maternelle selon deux critères :

- **La complexité lexicale**. Martinot (2010) identifie deux types de reformulation de prédications :
    - la *reformulation par simplification* (suppression d'un des deux prédicats ou du complément verbal de la phrase, la répétition et l'emploi d'un terme générique) ;
    - la *reformulation par complexification* (utilisation des paraphrases sémantiques, descriptives, des transformations à la forme impersonnelle ou à la forme active ou des transformations par restructuration et reprise de la métaphore).
- **La complexité syntaxique.** Ce critère concerne le choix des compléments de verbe et les variations diathèse active / passive dans la structure grammaticale de la phrase.

Ces deux types de reformulations sont complémentaires dans le processus de simplification d'un texte. Dans le cadre de notre recherche, d'une part, nous travaillons principalement sur **la reformulation par complexification** (Martinot, 2010). Nous cherchons des paraphrases, explications, définitions et synonymes qui permettent de rendre accessible le sens d'un terme spécialisé. D'autre part, nous visons également **l'identification des termes généraux (hyperonymes)** pour les termes médicaux, procédé qui est présenté comme une **reformulation par simplification** (« emploi d'un terme générique »), selon Martinot (2010).

Le rôle pédagogique de la reformulation peut se faire ressentir également dans les **textes de vulgarisation scientifique** destinés aux enfants ou au grand public adulte. Ces textes sont rédigés pour faire connaître et comprendre au public cible des notions scientifiques souvent très complexes. Dans ce sens, la reformulation et les paraphrases de celles-ci deviennent indispensables.

5.2 Le rôle de vulgarisation scientifique de la reformulation

Par rapport au rôle pédagogique, la fonction de **vulgarisation** de la reformulation se remarque par la présence des concepts techniques reformulés et par l'orientation du registre inconnu vers le connu. Suivant Vargas (2008), la **reformulation intratextuelle** est la structure de base du texte de vulgarisation pour son rôle central de management du savoir et des connaissances humaines.

Loffler-Laurian (1984) souligne *l'interdépendance* qui existe entre la vulgarisation scientifique et la reformulation : la reformulation est une **traduction monolingue** qui fait le passage d'un texte spécialisé vers un texte vulgarisé. Ainsi, les textes scientifiques sont *traduits* dans la langue commune par les remplacements des termes complexes avec leurs *traductions* dans leurs versions plus faciles à comprendre pour un lecteur non-spécialiste. Sa mission est de simplifier le langage très scientifique ou obscur en s'adaptant au lecteur cible. **Les reformulations** peuvent être employées dans le cadre de la **simplification lexicale**.

Au vu du fait que la simplification lexicale et la vulgarisation scientifique en particulier, sont réalisées par rapport à un **lecteur cible**, nous prenons en compte dans notre étude le type de public vers lequel la reformulation est destinée.

5.3 Le public cible

Reiss et Vermeer (2013) soutiennent que toute traduction de texte est réalisée pour un public cible dans la langue visée. Cette mission est considérée comme le ***skopos*** de la traduction. La théorie du skopos fait partie de ce que Vermeer appelle « l'action du traduire » (*translational action*). Reiss et Vermeer (2013) considèrent que toute forme d'activité de traduction, y compris la traduction elle-même (bilingue ou monolingue), peut être conçue comme une action qui a évidemment un objectif. Par conséquent, la définition du mot **skopos** serait **cible** ou **but** d'une traduction. La spécificité de la reformulation est qu'elle est une traduction qui a lieu entre des registres de langues différents qui appartiennent à la même langue, et non pas à deux systèmes linguistiques différents.

Au niveau du texte de reformulation, la rédaction humaine ou automatique, par le biais de l'intelligence artificielle, doit toujours tenir compte des spécificités du public cible pour que son *translatum* - pour faire appel à la terminologie de Vermeer - soit bien compris.

Nous nous intéressons également à l'objectif profond de la reformulation et au besoin d'adapter les textes aux plusieurs types de lecteurs afin de permettre une meilleure compréhension des notions et concepts techniques. Notre étude concerne le lecteur de type **grand public**, non-spécialiste du domaine médical, adulte, sans difficultés de lecture et autonome dans son apprentissage. Dans ce sens, nous développons dans la section suivante la réception des termes médicaux par ce type de lecteur.

5.4 Les termes médicaux face au grand public

Nous travaillons avec les termes médicaux, car ils posent souvent des difficultés de compréhension pour le 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). Pour comprendre la signification des termes, nous avons besoin de connaître les sciences et les techniques auxquelles elles répondent et non pas uniquement la langue dans laquelle ils sont écrits (Coseriu, 1967 : 17). 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). Des bases terminologiques de grande taille sont construites pour expliquer les termes et les notions associées, avec leur définition, source et fiabilité renseignées par les experts terminologues. En particulier, des terminologies sont disponibles notamment dans le domaine médical : UMLS (Bodenreider, 2004).

La technicité des termes médicaux est donnée par l'étymologie grecque ou latine et la composition souvent mixte de ces deux bases avec la langue moderne. Nous observons clairement la difficulté de compréhension pour les non-spécialistes avec le terme médical « cholécystectomie » qui est formé avec deux bases grecques *chole* (= bile) et *ectomy* (= ablation chirurgicale) et au milieu de celles-ci se trouve une base latine, *cystis* (= vessie) (Grabar et Hamon, 2015 : 2). Les termes médicaux peuvent être simplifiés par des synonymes de la langue générale, mais c'est parfois difficile de trouver des synonymes pour ces termes. Alors l'explication est nécessaire pour permettre la compréhension pour le grand public. Les termes médicaux sont expliqués par des reformulations et des paraphrases. Dans notre étude, nous cherchons des reformulations pour les termes monolexicaux et les termes polylexicaux, car pour ces derniers il est souvent difficile de trouver de synonymes dans la langue commune. Il est nécessaire d'identifier les termes médicaux pour ensuite chercher la reformulation possible.

Vu que nous nous intéressons à la notion de **reformulation** en premier lieu, nous utilisons **des terminologies validées et attestées** par les spécialistes du domaine médical pour identifier des termes dans nos corpus. Des outils d'identification automatique des termes, tels que **TermoStat** (Drouin, 2003) et **TTC TermSuite** (Daille *et al.*, 2011) combinent des techniques statistiques et à base de patrons lexico-syntaxiques, appliqués sur les corpus monolingues ou comparables pour la découverte de candidats terminologiques. Malgré leurs performances, ces outils peuvent donner des résultats bruités : ils proposent une liste de candidats termes, qu'il faut nettoyer pour identifier des termes du domaine. De plus, si plusieurs outils existent pour l'extraction terminologique en français, peu d'outils sont disponibles pour le roumain (à part le module BioNER intégré à la plateforme RELATE (Păiș *et al.*, 2019), que nous présentons dans la **Partie III**). Pour pallier ces problèmes, nous nous tournons vers **des bases terminologiques attestées**, multilingues ou monolingues, qui renseignent une liste de termes validée.

Ainsi, pour l'identification des termes, nous choisissons des **terminologies reconnues** dans le domaine médical. C'est le cas de **SNOMED-3.5VF** et **SNOMED International** (Côté, 1998 ; 1996) pour la langue française et une **liste de termes validés** par des médecins pour le roumain (issus de l'annotation sur le corpus **MoNERo** (Mitrofan *et al.*, 2019)).

5.5 Bilan

La reformulation est une *opération de transformation de la phrase* qui peut servir à la simplification lexicale et syntaxique des textes écrits. Dans une perspective du Traitement Automatique des Langues (TAL), notre travail est lié à la simplification lexicale automatique par l'utilisation de la reformulation et nous cherchons à adapter ce processus en fonction du public cible des textes simplifiés. Notre objectif étant de construire des corpus comparables de reformulations, nous nous intéressons aux techniques et ressources utilisées dans la simplification de textes pour nous orienter dans l'identification des reformulations de termes médicaux. Les termes du domaine médical sont souvent difficiles pour un public non-averti et nécessitent des explications. Pour l'identification de termes, nous nous appuyons sur des **terminologies médicales attestées** ou sur des **listes des termes validés** par les professionnels du domaine afin d'éviter les erreurs induites par des outils d'extraction automatique.

***

Dans la **Partie II** nous présentons **les approches en TAL** pour l'identification automatique des reformulations et des paraphrases dans de grands corpus et l'interprétation de la reformulation comme plusieurs **tâches en TAL** (désambiguïsation, traduction, simplification).

II. LA REFORMULATION EN TRAITEMENT AUTOMATIQUE DES LANGUES

1. Approches en TAL pour l'identification de la reformulation et de la paraphrase

Plusieurs travaux en TAL s'intéressent à la génération, l'extraction et la reconnaissance automatique des reformulations ou des paraphrases. **La génération de paraphrases**, qui est un problème plus large de génération de langage naturel, est le processus de création de formes alternatives du texte d'entrée. Elle trouve son application dans des domaines tels que le résumé de documents, la simplification et la traduction automatique. **L'extraction de paraphrases** implique l'identification ou la découverte de paraphrases à partir d'un grand corpus et trouve des applications dans les tâches d'extraction d'information. **La reconnaissance de paraphrases** est la tâche qui consiste à reconnaître la présence de paraphrases dans un corpus donné (Rajkumar et Chitra, 2010). La plupart des travaux en TAL s'intéressent au traitement de la paraphrase, cherchant la similarité sémantique, sauf quelques exceptions qui traitent la reformulation (Grabar et Eshkol-Taravella, 2016b). La création de corpus de paraphrases est nécessaire pour la mise en place de ces systèmes : on représente l'élément source (mot, terme, syntagme, phrase) et son équivalent reformulé.

Les travaux en TAL ont expérimenté plusieurs méthodes d'identification automatique des reformulations dans de grandes collections de textes. La reformulation étant une notion linguistique fluide (unité linguistique simple, sous-phrastique, phrastique), la première étape est de définir le type de reformulation sur lequel la recherche se réalise. Le type de reformulation détermine également la méthode automatique choisie pour obtenir les meilleurs résultats.

Dans ce chapitre nous utilisons les notions de **reformulation** et **paraphrase** de façon alternative ou conjointe, telle leur utilisation dans les études citées. Ce chapitre étant dédié aux méthodes de TAL pour l'identification des reformulations et/ou paraphrases, nous traitons les deux notions dans leur sens large (Fuchs, 1994 ; 2020). Nous analysons ces méthodes et nous prenons en compte les avantages et les limites des approches suivies.

Nous présentons les grandes catégories d'approches abordées pour identifier les reformulations à partir des corpus comparables ou des corpus parallèles : des règles définies manuellement et ressources lexicales pour l'identification des marqueurs de

reformulation, des méthodes d'apprentissage statistiques appliquées à la classification et les approches plus récentes par apprentissage profond. Souvent, les méthodes sont constituées à l'aide des approches hybrides, combinant des méthodes symboliques et statistiques. La reformulation peut être détectée dans les textes à travers des mots-clés (les marqueurs de reformulation, comme « c'est-à-dire ») ou une structure lexicale et syntaxique spécifique de type terme – marqueur – reformulation. Nous présentons ci-dessous des travaux basés sur cette méthode à base de règles définies manuellement ou automatiquement, sur la base des études linguistiques.

1.1 Méthodes à base des règles : détection de marqueurs de reformulation

Cette famille de méthodes cherche à identifier les indices linguistiques de la reformulation. On a retrouvé des structures récurrentes qui peuvent aider à son identification, telle que la structure **terme/référent – marqueur – reformulation**. Identifier les **marqueurs de reformulation** adaptés au corpus d'étude permet de mieux cibler les reformulations.

Grabar et Eshkol-Taravella (2016a) ont mené une étude sur la reformulation lexicale pour identifier des marqueurs de reformulation de type *c'est-à-dire, disons, ça veut dire* à l'aide d'un système à base de règles et des annotations manuelles. L'objectif de leur étude est de faire la différence de façon automatique entre les syntagmes qui ont le rôle de reformulations et celle qui ne l'ont pas. Pour les identifier, Grabar et Eshkol-Taravella (2016a) prennent en compte la structure syntagmatique « *S1 marker S2* », dont S1 est le référent et S2 la reformulation, les deux liées par les marqueurs de reformulation. Leur étude est menée sur le corpus oral ESLO (*Enquêtes SocioLinguistiques à Orléans*) (Eshkol-Taravella *et al.*, 2011) et sur un corpus extrait de forums de discussions. Pour distinguer les reformulations des syntagmes sans rôle de reformulation, Grabar et Eshkol-Taravella (2016a) utilisent des règles d'identification automatique comme la position du marqueur dans la phrase et la présence de certains éléments discursifs dans le contexte du marqueur (de type « donc / enfin / quoi / euh / en / hm / ouais » ou de certaines expressions comme « indépendamment »). Ces éléments discursifs ne sont pas des indices de l'absence de reformulation, mais des syntagmes des disfluences de la langue parlée, selon la théorie de Benveniste *et al.* (1990) (citée par Grabar et Eshkol-Taravella, 2016a). Il faut également délimiter le référent et sa reformulation, qui peuvent prendre des formes variées (identifiables par des patrons morphosyntaxiques).

Ces méthodes sont adaptables pour d'autres domaines ou genres, après une étude de marqueurs de reformulation spécifiques au corpus en plus de ceux étudiés dans la littérature. De plus, les référents ou les reformulations peuvent être identifiés à l'aide des patrons morphosyntaxiques.

Nous nous inspirons de cette méthode pour identifier les marqueurs les plus pertinents pour le texte médical écrit. Nous identifions des structures **terme-marqueur-reformulation**. Nous présentons nos résultats dans la **Partie V**, qui décrit les expériences d'annotation. Nous prenons également en compte la présence **des marqueurs orthotypographiques** comme les **parenthèses**.

Pour éviter le travail manuel de création de règles, d'autres méthodes appliquent **l'apprentissage automatique** pour des classifications automatiques de textes et de phrases de façon autonome, semi-supervisée ou supervisée. Nous développons par la suite le concept de classification par apprentissage automatique en TAL et ses différentes caractéristiques appliquées au domaine de la paraphrase ou de la reformulation.

1.2 Classification par apprentissage automatique

L'apprentissage automatique (en anglais, ***machine learning***), est un domaine de l'intelligence artificielle largement exploité en TAL. Un type d'application de l'apprentissage automatique est la **classification automatique** des données : on utilise un grand nombre de données comme matériel d'apprentissage pour aider un logiciel à identifier les classes auxquelles appartiennent les données. Le logiciel classe les données à l'aide des algorithmes statistiques, par exemple des arbres de décision, des arbres de type Naïve Bayes ou des fonctions linéaires, et exploitent plusieurs propriétés des données pour identifier des classes.

Nous présentons plusieurs études qui utilisent l'apprentissage automatique pour identifier les paraphrases / reformulations dans les corpus de textes. Leur objectif est de classer les paires de phrases en relation de référence. Nous nous intéressons au mélange de méthodes et aux outils nécessaires pour trouver les paraphrases ou reformulations de la manière la plus efficace.

Brockett et Dolan (2005) utilisent des corpus comparables, ressources lexicales et annotation manuelle pour extraire, par apprentissage automatique, des corpus de paraphrases monolingues à partir des corpus journalistiques comparables en anglais

(trouvés sur la toile) à l'aide d'un algorithme basé sur des machines à vecteurs de support. Les machines à vecteurs de support ou les séparateurs à vaste marge (*Support Vector Machines / SVM*) (Vapnik, 1995) sont une classe d'algorithmes d'apprentissage définis pour la discrimination de données. Ces algorithmes sont des classifieurs linéaires qui sont évalués selon la capacité de généralisation la plus grande possible[10]. Les SVM sont efficaces face aux données bruitées en grande quantité et ils se prêtent facilement à l'inclusion en masse de caractéristiques lexicales telles que les informations morphologiques et synonymiques. Parmi les propriétés utilisées par SVM, Brockett et Dolan (2005) ont choisi la similarité des segments (longueur, distance d'édition, mots voisins), la variation morphologique et les paires d'associations de mots.

Pour la création des corpus de paraphrases, ils prennent en compte les variantes morphologiques des paraphrases, les synonymes et les hyperonymes en faisant appel à la base lexicale WordNet (Miller, 1998) et aux paires de mots synonymes. Les textes ont été annotés manuellement par deux annotateurs humains afin d'identifier les paraphrases « possibles » et « sûres ». Les textes ont été par la suite alignés automatiquement avec le système de traduction automatique statistique GIZA++ (Och et Ney, 2003). Brockett et Dolan (2005) ont utilisé un algorithme de type log-likelihood (Moore, 2001) pour extraire de façon automatique des paires de mots synonymes comme « straight / consecutive » (*droit / consécutif*), « vendors / suppliers » (*vendeurs / fournisseurs*) (exemples donnés par Brockett et Dolan, 2005 : 4). Leurs expériences ont été menées sur des corpus comparables en parallèle avec un corpus de test formé par des phrases alignées qui présentent des paraphrases *sûres* et de *fausses* paraphrases. Leur travail prouve que les machines à vecteurs de support, après avoir été entraînées, peuvent aider à l'extraction des corpus de paraphrases avec presque la même efficacité même sans le recours à WordNet (Miller, 1998).

Socher *et al.* (2011) se concentrent sur l'identification de la paraphrase par des analyses sémantiques et syntaxiques. Les algorithmes transforment les mots en vecteurs et enregistrent des informations syntaxiques et sémantiques via des statistiques de co-occurrence du mot. Socher *et al.* (2011) proposent une nouvelle technique de regroupement dynamique des mots et des n-grammes qui calcule une représentation de taille fixe d'une paraphrase à partir des matrices de tailles variables. Cette technique est utilisée comme classifieur pour identifier des paraphrases synonymes. Ces méthodes sont testées sur des textes en anglais du « Corpus Microsoft Research Paraphrase (MSRP) » (Dolan *et al.*,

[10] http://wikistat.fr/

2004). Socher *et al.* (2011) utilisent 4 076 paires de paraphrases pour l'entraînement (dont 67,5% sont des paraphrases, selon des annotateurs manuels) et 1 725 paires pour le test (66,5% de paraphrases, toujours selon l'avis des annotateurs). Leur étude montre que le regroupement dynamique trouve des paraphrases exactes constituées de 2 à 5 mots, mais que pour les phrases plus longues, les mots voisins les plus proches ne sont plus dans une relation paraphrastique : « *the full bloom of their young lives* » est identifiée comme une fausse paraphrase de « *the lower bloom of their democratic lives* »[11] (Socher *et al.*, 2011 : 6).

Filice et Moschitti (2016) travaillent sur l'extraction des paraphrases sous-phrastiques en fonction des informations auxiliaires et du contexte. Ils définissent ces informations auxiliaires comme des parties de la phrase qui n'ont pas un sens précis susceptible à être paraphrasé, comme on le voit dans l'exemple : « *Although it's unclear whether Sobig was to blame* »[12] (Filice et Moschitti, 2016 : 1109). Dans cette phrase l'énoncé « *Although it's unclear whether* » ne peut pas être une paraphrase de « *Sobig was to blame* » parce que son sens n'ajoute pas de la signification sur le deuxième énoncé, donc il représente une information auxiliaire.

Pour faire cela, ils utilisent des machines à vecteurs de support (SVM) (Severyn et Moschitti, 2012), un classifieur automatique des informations auxiliaires développé dans le cadre de l'étude, l'algorithme C-SVM (Chang et Lin, 2011) et le logiciel KeLP (Carreras et Marquez, 2005)[13], une plateforme d'apprentissage fournissant la mise en œuvre d'algorithmes d'apprentissage divers[14]. Pour évaluer la qualité de leur classifieur automatique, Filice et Moschitti (2016) font recours également aux annotateurs humains pour annoter les informations auxiliaires. Le travail manuel des annotateurs est évalué à 85% avec l'indice Kappa (Cohen, 1960), une mesure d'accord commune élevée entre les deux annotateurs (rapport entre l'accord relatif et l'accord aléatoire). Le travail automatique est évalué avec les mesures statistiques de précision, rappel et F-mesure, par rapport au corpus annoté manuellement.

Le corpus utilisé est le corpus anglais « Microsoft Research Paraphrase Corpus (MSRP) » (Dolan *et al.*, 2004). Afin d'analyser les contextes de paraphrases et retrouver les informations auxiliaires, Filice et Moschitti (2016) se servent de l'analyseur syntaxique

[11] Notre traduction en français : « le plein épanouissement de leur vie de jeune », « le pauvre épanouissement de leur vie démocratique ».
[12] Notre traduction en français : « Pourtant ce n'est pas clair si Sobig était coupable ».
[13] https://github.com/SAG-KeLP
[14] http://www.kelp-ml.org/?page_id=455 (consulté le 17 mai 2020).

Stanford[15] et de deux modèles d'arbres de décision pour trouver les nœuds syntaxiques (Syntactic Tree Kernel (STK) (Collins et Duffy, 2001)) et sémantiques (pour vérifier si les phrases candidates gardent le même sens, donc elles contiennent des paraphrases) (The Partial Tree Kernel (PTK) (Moschitti, 2006)). Leurs résultats prouvent que le classifieur automatique des informations auxiliaires identifie les paraphrases avec une précision de 73,2% et une F-mesure de 71,8% sur le corpus, respectivement 78,3% et 84,5% sur le corpus de test.

Dupuch *et al.* (2013) travaillent également sur la terminologie médicale dans la perspective de l'apprentissage semi-supervisé (méthode que nous développons plus tard dans le **sous-chapitre 1.3**) pour créer des clusters de termes médicaux similaires en français et en anglais. Pour y parvenir, ils identifient les relations sémantiques et hiérarchiques entre des termes. Plus précisément, les relations visées sont les variantes morphosyntaxiques (exemple : « sténose de l'aorte, sténose aortique »), la synonymie (exemple : « tumeur gastrique, cancer gastrique ») et les relations de subsomption hiérarchique (exemple : « défaillance rénale, défaillance rénale post-opératoire ») (Dupuch *et al.*, 2013 : 64). Afin de déterminer les termes qui sont liés sémantiquement, Dupuch *et al.* (2013) utilisent la terminologie médicale MedDRA (Medical Dictionary for Regulatory Activities) (Brown *et al.*, 1999) et la base terminologique UMLS (Bodenreider, 2004). Une fois les relations identifiées, les termes sont disposés en clusters à l'aide des graphes pluridimensionnels. Le croisement des résultats dans les deux langues donne des meilleurs clusters en termes de précision : 74% (par rapport au 71,1% en anglais uniquement, respectivement 70,5% en français) et de rappel : 12,1% (par rapport au 11,8% en anglais uniquement, respectivement 8,4% en français). Pourtant, les valeurs de précision et de rappel restent faibles, ce qui montre la difficulté de la tâche.

***

Les études présentées sur le classement par apprentissage automatique statistique sont très diverses, elles mélangent plusieurs algorithmes : arbres de décision et machines à vecteurs de support (SVM). Pour utiliser ces méthodes, nous avons besoin de grands corpus pour l'entraînement, parfois des ressources lexicales (WordNet) ou terminologiques et souvent des analyseurs syntaxiques pour construire les arbres de décision ou les modèles statistiques. Des outils d'annotation syntaxique ont été développés pour le français, tels que Stanford NLP (Manning *et al.*, 2014), Mind The Gap (Coavoux et Crabbé, 2017), Talismane (Urieli et Tanguy, 2013), et le roumain (SSPR (Ion *et al.*, 2018)), ou les deux

[15] https://stanfordnlp.github.io/CoreNLP/

langues (UDPipe (Straka *et al.*, 2016), NLP-Cube (Boroș *et al.*, 2018)), mais les erreurs d'analyse (appliquées sur des textes de spécialité) peuvent générer également des erreurs dans la détection des reformulations.

De plus, notre objectif est d'identifier **une grande variété de reformulations**, non pas uniquement des paraphrases nominales, synonymes ou hyperonymes du domaine médical, de type « maladie », « trouble », mais aussi d'autres catégories de reformulations (explication, énumération, définition). Pour ces raisons, nous nous orientons vers un autre type de classement de textes par **apprentissage automatique profond** qui peut tenir compte de contextes plus larges et peut pallier l'absence des informations morphologiques ou syntaxiques.

Nous présentons par la suite l'apprentissage par réseaux de neurones et les travaux pertinents qui utilisent cette technique pour réaliser l'extraction des paraphrases ou des reformulations.

1.3 Apprentissage profond (par réseaux de neurones)

La méthode la plus récente d'apprentissage automatique est **l'apprentissage profond**. L'apprentissage par réseaux de neurones (apprentissage profond), est bien adapté aux données complexes (images, radiographies, sons), mais aussi aux données symboliques et textuelles. Les données peuvent être représentées par de nombreux attributs à valeurs réelles ou symboliques, dépendants ou non entre eux. **Les architectures neuronales** sont appropriées pour les tâches de traitement du langage en raison de leur robustesse aux entrées bruitées et de leur similitude avec les processus de pensée cognitifs humains.

Plusieurs architectures neuronales s'appliquent dans les tâches de TAL. Les **neurones artificiels** (ou *cellules*) imitent le fonctionnement d'un neurone biologique, collectant plusieurs valeurs en entrée et une valeur en sortie, calculée sur la base d'une fonction d'activation (sigmoïde, tangente ou linéaire). Les neurones sont organisés en plusieurs couches interconnectées (les liens sont associés aux poids). Une couche d'entrée recueille des données en entrée (par exemple des séquences de mots) et une couche de sortie permet d'attribuer des étiquettes aux données d'entrée (pour les tâches de classification) ou de générer des données en sortie (par exemple la traduction ou la simplification d'une séquence de mots). Lors de la phase d'entraînement, les poids sont ajustés afin d'obtenir les résultats attendus (par exemple l'étiquette qui a été attendue pour

un jeu de données d'entrée doit être la même que celle prédite par le système). L'ajustement se fait par rapport à une évaluation de l'erreur de prédiction. Plus l'erreur est importante, il faut ajuster les poids en présentant plusieurs fois (epochs) les exemples du corpus d'entraînement.

Ces architectures codent l'information linguistique dans des **contextes très larges**, sans représenter explicitement les informations morphologiques, syntaxiques ou sémantiques, ce qui permet de gérer des textes bruités, des mots inconnus ou des informations manquantes. Parmi les architectures les plus répandues appliquées en TAL, nous mentionnons les **réseaux récursifs (RNN)** et les **réseaux convolutifs (CNN)** (Schmidhuber, 2015). *Les réseaux récursifs* sont adaptés pour le traitement des données séquentielles (par exemple des séquences de mots, la génération d'un mot dépend des mots qui le précédent) : parmi les entrées d'une cellule, on retrouve la sortie de la cellule précédente (on tient compte de l'état antérieur du réseau). *Les réseaux convolutifs* sont entraînés pour modéliser des parties spécifiques de l'ensemble de données (contours d'une image, syntagmes ou patrons spécifiques pour les données textuelles). On peut ainsi spécialiser des couches du réseau pour une tâche spécifique.

L'architecture **LSTM** (*Long Short Term Memory*) est un cas particulier de réseaux récursifs, permettant de prendre en compte l'état du réseau à plusieurs moments de temps. Cette architecture est particulièrement adaptée pour traiter des séquences de données (par exemple, les séquences de mots sont associées à une série de cellules en entrée). Elle est construite sous la forme de plusieurs couches de cellules de mémoire qui sont reliées entre elles et contiennent des valeurs numériques (*vecteurs*) et qui vont prendre en considération les sorties du réseaux aux moments *Tn-1,..T1*. Si l'entrée est une séquence de mots, il est possible de prendre en compte la séquence de mots et leurs sorties pour les *n-1* mots qui précédent le mot correspondant à la cellule actuelle. Des architectures LSTM peuvent être **bidirectionnelles** (*Bidirectional Long Short-Term Memory Networks*, en français « Réseaux de Mémoire à Long-Court Terme Bidirectionnels ») : il est alors possible de prendre en compte les contextes gauche et droit simultanément.

***

Il y a deux grands types d'apprentissages par **réseaux de neurones** : **l'apprentissage supervisé** et **l'apprentissage non-supervisé**. **L'apprentissage est supervisé** quand les données d'entraînement sont annotées avec les classes ou les résultats attendus. En revanche, quand **l'apprentissage est non-supervisé**, l'apprentissage se fait de façon totalement autonome. La machine reçoit que des données brutes, sans aucun exemple de résultats attendus. L'apprentissage non-supervisé est

employé surtout pour réaliser des classements des données hétérogènes selon des caractéristiques communes.

En traitement automatique des langues, ces modèles se sont généralisés ces dernières années pour plusieurs tâches (traduction automatique, simplification, etc.), car on dispose des modèles contextuels de très grande taille. Nous devons représenter les reformulations dans leur contexte d'apparition. Dans ce sens, les **modèles de langues** qui contiennent des **vecteurs de mots** représentant le contexte sont une solution pour ne pas utiliser explicitement l'analyse syntaxique ou morphologique. Nous présentons dans le point suivant les **Transformers**, des modèles de langues permettant de représenter les mots dans leurs contextes d'apparition.

1.3.1 Modèle de langue de type Transformers

Un modèle de langue estime la probabilité qu'un mot ou une séquence de mots soit présent dans un texte. Pour interpréter correctement le sens d'un mot ou d'une séquence de mots, il faut tenir compte de ses **contextes d'occurrence**. Cette information est représentée sous forme de **vecteurs** (des valeurs numériques attribuées aux mots et séquences de mots) pour les rendre exploitables par une machine. Les premiers modèles de vecteurs contextuels tel que **word2vec** (Mikolov *et al.*, 2013) ou **GloVe** (Pennington *et al.*, 2014) représentent les mots sous forme de vecteurs représentant les cooccurrents les plus fréquents, sans faire la différence de contextes d'apparition ou de l'ordre de mots. Les **Transformers**, introduits en 2017 par des chercheurs de chez Google, (Vaswani *et al.*, 2017), sont des modèles de langues pour l'apprentissage automatique profond par des réseaux de neurones, qui proposent une représentation plus complexe de tous les contextes d'un mot.

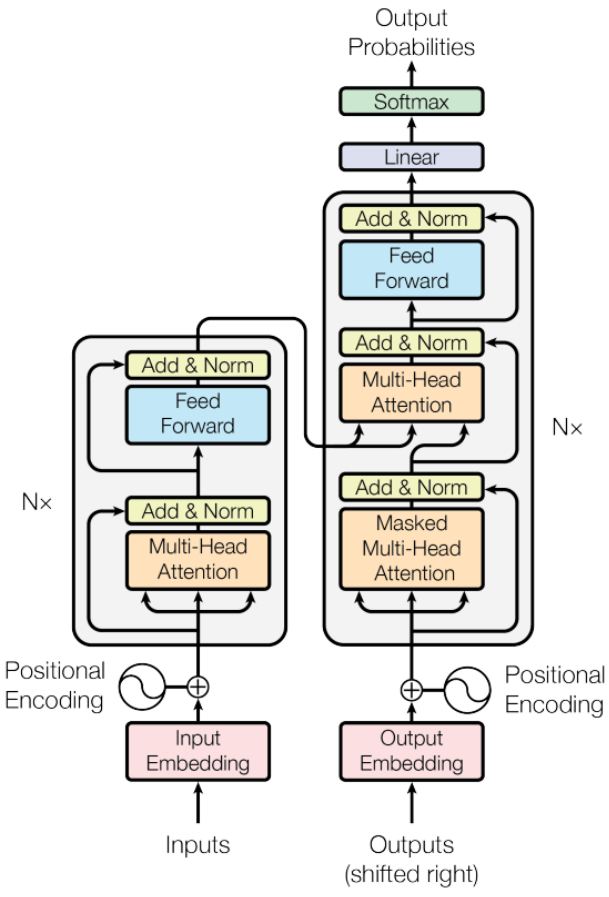

Figure 6. L'architecture du Transformer, selon les travaux de Vaswani et al. (2017)

Les **Transformers** traitent les données de manière séquentielle et utilisent des mécanismes d'auto-attention (voir **Figure 6**). Ils fonctionnent selon trois principes :

1) **Structure de type *multi-head-attention*** : apprentissage de données à l'aide de la position des mots et de plusieurs couches de règles parallèles (*multi-head*) ;
2) **Couches multiples d'auto-attention** *(self-attention)* : le Transformer enregistre des données numériques sur chaque séquence textuelle en fonction de différents sens qu'elle peut avoir dans des positions et de contextes différents et permet également l'accès à plusieurs états successifs du système. Il est possible de tenir compte des contextes précédant et suivant le mot, ainsi que d'avoir accès à plusieurs contextes simultanément ;
3) **Système de type encodeur-décodeur** : les données textuelles sont transformées en entrée en données numériques (*vecteurs*), ce qui facilite le calcul de scores de probabilité qu'une certaine *donnée de sortie* soit la plus adaptée au besoin de la tâche d'apprentissage, selon les *données d'entrée*.

À l'heure actuelle, **149 Transformers** ont été développés, selon une liste publiée sur le site de référence de la communauté internationale pour l'apprentissage automatique HuggingFace[16]. Quelques exemples des **modèles de langues de type Transformer** largement utilisés sont **BERT** (Devlin *et al*., 2019 ; Google AI) et ses dérivés français **FlauBERT** (Le *et al.*, 2020a ; CNRS) et **CamemBERT** (Martin *et al.*, 2020 ; construit conjointement par Facebook AI (actuellement Meta AI), INRIA et l'Université de Sorbonne), et d'autres comme **RoBERTa** (Liu *et al.*, 2019 ; développé par Meta AI), **T5** (Raffel *et al*., 2020 ; Google AI) et **GTP-3** (Brown *et al.*, 2020 ; OpenAI), etc.

BERT (Devlin *et al*., 2019) est un ensemble de vecteurs contextuels créés sur un corpus général et permet de représenter simultanément tous les contextes des mots sous forme d'une architecture encoder-decoder. Pour le français, il existe des variantes monolingues de cette architecture pour la langue générale en libre accès tel que **FlauBERT**[17] (Le *et al.*, 2020) et **CamemBERT**[18] (Martin *et al.*, 2020). Pour le roumain, une ressource similaire a été développée par Dumitrescu *et al.* (2020), intitulée **Romanian BERT**[19], à partir de **BERT Multilingue**[20] (Devlin *et al.*, 2019), un modèle qui contient des

[16] https://huggingface.co/docs/transformers/index, accédé le 18/11/2022 à 19h16.
[17] https://github.com/getalp/Flaubert
[18] https://github.com/pytorch/fairseq/blob/master/examples/camembert/README.md
[19] Le modèle créé par Dumitrescu et al. (2020), **Romanian BERT**, a été mis à disposition pour le roumain en 2020 : https://huggingface.co/dumitrescustefan/bert-base-romanian-cased-v1 ;
[20] https://github.com/google-research/bert/blob/master/multilingual.md

données pour plusieurs langues, dont le français et le roumain. Ces modèles ont été entraînés sur des données de la langue générale, le seul modèle existant pour le domaine médical, est, à notre connaissance, **BioBERT**[21] (Lee *et al.*, 2020). Ce modèle n'est pourtant disponible que pour l'anglais. Très récemment[22], un nouveau modèle entrainé sur des données médicales en français, **DrBERT**, a été développé par Labrak *et al.* (2023)

1.3.2 Architectures à base de réseaux de neurones pour identifier la paraphrase

La paraphrase a été abordée dans plusieurs tâches en TAL, en s'appuyant sur des architectures à base de réseau de neurones :

- **La similarité sémantique textuelle (STS),** qui mesure le degré d'équivalence dans la sémantique des textes ou phrases similaires (Agirre *et al.*, 2016) ;
- **L'identification des paraphrases (PI),** qui identifie si deux phrases ou segments ont le même sens (Dolan et Brockett, 2005 ; Xu *et al.*, 2015 ; Nie et Bansal, 2017) ;
- **La génération des paraphrases (PG)**, qui crée de nouveaux textes à partir des données d'entrée et des architectures de langues (Gupta *et al.*, 2018 ; Bowman *et al.*, 2016).

Méthodes utilisant la similarité sémantique textuelle (STS)

Un premier exemple d'architecture de réseaux de neurones de type **STS** est intitulée **PWIM** (He et Lin, 2016) (*Pairwise Word Interaction Model*, en français « Modèle d'interaction entre mots par paires »). L'architecture PWIM encode des informations sur les contextes des mots avec des scores numériques (par exemple, plus le score est élevé, plus le contexte est pertinent), à l'aide d'une architecture **Bi-LSTMs**. Le modèle **PWIM** (He et Lin, 2016) identifie la similarité entre deux phrases calculée par la similarité en cosinus, la distance euclidienne et le produit scalaire des vecteurs (Lan et Xu, 2018 : 3891), à l'aide du ConvNet, un réseau de neurones convolutifs profond à 19 couches (He et Lin, 2016 : 938-939). Pour paramétrer ce modèle, Lan et Xu (2018) utilisent un autre modèle, **GloVe** (Pennington *et*

[21] https://github.com/dmis-lab/biobert

[22] **DrBERT** (Labrak *et al.*, 2023) a été mis à disposition sur huggingface le 3 avril 2023 (https://huggingface.co/Dr-BERT).

al., 2014) (la version plus ancienne de plongement lexical, antérieure aux Transformers) qui construit des représentations vectorielles à partir de corpus[23].

Yuan *et al.* (2016) se servent des modèles d'apprentissage par réseaux de neurones pour déterminer les sens des verbes en fonction de leurs positionnements dans la phrase et les désambiguïser. D'abord, ils transforment les mots de leur corpus de textes journalistiques en **vecteurs de sens** avec l'algorithme d'apprentissage automatique supervisé **Word2Vec** (Mikolov *et al.*, 2013) (qui ne prend pas en compte l'ordre des mots). Pour déterminer l'ordre des mots dans la phrase (important pour l'identification des verbes), Yuan *et al.* (2016) utilisent une architecture neuronale de type **LSTM** (Long Short Term Memory) (Hochreiter et Schmidhuber, 1997) et un **algorithme semi-supervisé**. Yuan *et al.* (2016) identifient les mots à partir de leur contexte dans la phrase. D'une part, ce système utilise des données qui ne sont pas annotées, pour une plus grande couverture. En deuxième lieu, l'algorithme semi-supervisé annote les phrases non-traitées en fonction de leur similarité aux phrases déjà annotées. La similarité est calculée à l'aide des vecteurs de sens.

Même si cette étude ne parle pas directement des paraphrases, le travail sur la désambiguïsation sémantique des verbes peut servir à l'identification des paraphrases lexicales à partir de ces verbes synonymes.

Dans la même lignée, Le *et al.* (2018) mènent une étude pour reproduire les expériences de Yuan *et al.* (2016). Le *et al.* (2018) travaillent sur l'aspect de **la similarité sémantique des paraphrases lexicales** afin de trouver des paraphrases lexicales dans des contextes similaires. Ils utilisent le corpus « English Gigaword Fifth Edition » (Linguistic Data Consortium (LDC) (Parker *et al.*, 2011), le corpus annoté en anglais **SemCor** (Miller *et al.*, 1993), le corpus annoté anglais-chinois **OMSTI** (Taghipour et Ng, 2015) et **WordNet 3.0**.

Le *et al.* (2018) utilisent l'architecture neuronale de type **LSTM** (Hochreiter et Schmidhuber, 1997)[24]. LSTM est implémenté en prenant en compte plusieurs aspects : les lemmes qui ont une seule racine (donc un seul sens), la transmissibilité des annotations (dans le cas où le lemme du mot se retrouve plusieurs fois dans les corpus) et la fréquence des lemmes dans le corpus. Leurs expériences prouvent que l'ajout de données non-annotées et l'augmentation de la capacité des modèles diminuent la qualité des résultats. Le *et al.* (2018) obtiennent les mêmes résultats que Yuan *et al.* (2016) avec beaucoup moins

[23] Le code du modèle PWIM, tel que paramétré par Lan et Xu (2018), est libre d'accès : https://github.com/lanwuwei/SPM_toolkit/tree/master/PWIM

[24] Le *et al.* (2018) utilisent la bibliothèque Python TensorFlow 1.2.1 (Abadi *et al.*, 2015) et l'analyseur BeautifulSoup HTML pour extraire l'information de la page Web en format HTML. La tokenisation des phrases est réalisée avec les modèles pour la langue anglaise de l'outil de tokenisation Spacy 1.8.2.

de données textuelles (corpus de 1,8 milliard de tokens au lieu de 100 milliards de tokens – le corpus utilisé par Yuan *et al.* (2016)).

Ces méthodes peuvent être utiles pour la détection de paraphrases qui partagent des contexte similaires (synonymes), mais ne sont pas adaptées pour la classification des paraphrases de type explication ou définition, dont les contextes d'apparition sont très différents.

Méthodes pour l'identification des paraphrases (PI)

Pour l'identification de séquences de mots qui sont des paraphrases (**PI**), Nie et Bansal (2017) proposent **SSE**, une architecture **LSTM-RNN** bidirectionnels qui permet d'optimiser les contextes avant et après les mots. Le modèle code deux phrases d'entrée en deux vecteurs, puis utilise un classificateur sur la combinaison de vecteurs pour étiqueter le type de relation entre ces deux phrases : d'implication (*paraphrases*), de contradiction (*elles ne sont pas des paraphrases*) ou *neutre*. Comme ce modèle est développé pour traiter des phrases entières et il a été testé sur le corpus **Multi-NLI** (Williams *et al.*, 2017) qui contient des données en anglais écrit et parlé. Par conséquent, SSE n'est pas adapté à notre tâche qui implique des données sous-phrastiques (terme médical et reformulation sous-phrastique) et les langues de notre étude.

Les modèles de langues de type **Transformers** sont de plus en plus utilisées pour résoudre les tâches complexes de TAL, telle que la détection de la paraphrase. Des expériences d'identification de la paraphrase sur la langue générale en français ont été réalisées sur **FlauBERT** (Le *et al.*, 2020), avec un jeu de données multilingue PAWS-X (Yang *et al.*, 2019). Ces expériences donnent des résultats légèrement meilleurs en termes de précision que **BERT Multilingue** (Devlin *et al.*, 2019) : 89,3% versus 89,9% pour FlauBERT, sur des paraphrases qui présentent une similarité formelle forte (mots similaires ou communs).

Méthodes de génération de la paraphrase (PG)

En fonction du contexte, cette famille d'approche propose de générer un fragment de texte qui peut être équivalent sémantique. Nous nous intéressons également à l'architecture **MLM** (*Masked Language Model*, en français « Modèle de Langage Masqué ») (Wahle *et al.*, 2021), méthode destinée à la génération de paraphrases (**PG**). Cette méthode masque une partie des mots et le modèle doit « deviner » les choix de mots les plus

probables pour ces mots cachés (Wahle *et al.*, 2021 : 2). Il existe plusieurs niveaux d'implémentation des modèles pour les architectures neuronales, selon la classification de Rajkumar et Chitra (2010) :

- *char-level* : niveau du caractère ;
- *word level* : niveau du mot (unigramme) ;
- *phrasal level* : niveau du syntagme (plurigramme) ;
- *sentence level* : niveau de la phrase.

Il s'agit de masquer des caractères, des mots, des syntagmes ou des phrases. L'approche **au niveau du caractère** donne les meilleurs résultats pour la paraphrase selon les expériences de Rajkumar et Chitra (2010), **le niveau de l'unigramme** (ensemble avec la méthode MLM en cachant un terme ou mot de la paraphrase) et **le niveau du syntagme** (la paraphrase peut être également sous la forme d'un syntagme nominal indiqué comme définition ou explication). Le dernier niveau, celui de la phrase entière, ne correspond pas au but de notre étude, car notre objectif est de créer des corpus de reformulations des termes médicaux et non pas un corpus des reformulations de phrases entières, utilisable pour la détection du plagiat (Wahle *et al.*, 2021) ou la simplification de phrases (Cardon et Grabar, 2019). Cette méthode reste limitée à la génération de synonymes (mots simples ou syntagmes nominaux).

***

Le **Tableau 3** résume les différentes méthodes avec les architectures de langues disponibles.

Méthode	Définition	Architectures de langues
Similarité sémantique textuelle (STS)	Mesure le degré d'équivalence dans la sémantique des textes ou phrases similaires (Agirre *et al.*, 2016)	**PWIM** (He et Lin, 2016) (*Pairwise Word Interaction Model*, en français « Modèle d'interaction entre mots par paires »)
Identification de paraphrase (PI)	Identifie si deux phrases ou segments ont le même sens. (Dolan et Brockett, 2005 ; Xu *et al.*, 2015)	**SSE** (Nie and Bansal, 2017) (*The Shortcut-Stacked Sentence Encoder Model*, en français Le modèle d'encodeur de phrases empilées par raccourcis)
Génération des paraphrases (PG)	Crée de nouveaux textes à partir des données d'entrée et des architectures (Gupta *et al.*, 2018 ; Bowman *et al.*, 2016).	**MLM** (Wahle *et al.*, 2021) (*Masked Language Model*, en français « Modèle de Langage Masqué »)

Tableau 3. Méthodes et architectures de réseaux de neurones pour identifier les paraphrases

***

Nous observons que ces méthodes ne proposent pas une modélisation de la paraphrase et ne suivent pas une approche linguistique en particulier de la paraphrase sous-

phrastique. Certains travaux se limitent à la paraphrase lexicale seulement. D'autres traitent l'identification de la paraphrase comme un problème de classification de relations (implication, contradiction), ou se situent dans la perspective de générer la paraphrase en fonction du contexte (la plupart du temps en relation de synonymie avec le référent). Ces tâches liées à la paraphrase sont complexes à cause de la difficulté de délimitation de cette notion, la difficulté d'avoir de corpus de grande taille vérifiés et valides pour la tâche. Les résultats obtenus (en termes de précision, rappel et F-mesure) restent modestes par rapport à d'autres tâches en TAL, à cause des difficultés de délimitation de cette notion. La plupart des travaux sont réalisés sur l'anglais. Dans notre cas, il s'agit de créer des corpus qui représentent les spécificités de la reformulation sous-phrastique médicale, sans se limiter aux relations de synonymie, dans chacune de nos langues d'étude (français et roumain).

Dans le chapitre suivant nous présentons des méthodes qui interprètent la reformulation comme **tâche spécifique de TAL** de type détection de la similarité textuelle, désambiguïsation sémantique, traduction automatique et simplification automatique de textes. Ainsi, on applique les techniques adaptées pour ces tâches pour traiter la reformulation.

2. La reformulation interprétée comme tâche en TAL

Il existe différentes tâches en TAL qui ont été associées à l'identification de la reformulation et/ou de la paraphrase : la similarité textuelle pour identifier des équivalences sémantiques des paraphrases, la traduction pour identifier diverses versions d'une même reformulation en passant par une autre langue, la simplification textuelle qui a le rôle d'adapter le contenu pour exprimer plus clairement certaines notions ou phrases. Nous explorons chacune de ces tâches et nous analysons comment la reformulation (ou la paraphrase) a été modélisée pour être adaptée à ces tâches spécifiques de TAL.

2.1 Détection de la similarité textuelle à l'aide des annotations

Plusieurs travaux transposent la tâche de détection de la similarité entre phrases à l'aide d'une représentation par des **graphes syntaxiques**. Ces graphes représentent une méthode de **modélisation de la langue** en tenant compte de la syntaxe d'une langue en particulier. Nous présentons par la suite différentes études qui utilisent des graphes syntaxiques permettant d'identifier de façon automatique des **reformulations** ou des **paraphrases**.

Une étude qui fait appel aux graphes pour extraire les paraphrases est celle de Chen *et al.* (2013). Ils jugent la qualité des paraphrases selon plusieurs critères : la similarité lexicale distributionnelle, la similarité syntaxique distributionnelle et la similarité de la traduction. Leur travail de recherche se concrétise dans la création d'un graphe d'extraction des paraphrases, suivi par l'alignement des phrases avec l'outil d'alignement automatique Giza++ (Och and Ney, 2003), après avoir utilisé l'outil d'étiquetage morphologique Genia Tagger (Kim *et al.*, 2003). Le corpus utilisé pour cette étude est la deuxième version du corpus Europarl, en version bilingue danois-anglais (Koehn, 2005). Ce travail de recherche a comme but l'extraction des paraphrases en anglais, la langue danoise servant seulement comme langue de contrôle. Pour affiner l'évaluation des résultats, un test de substitution et deux annotateurs humains sont employés. Pour tester si les paraphrases sont utilisées dans certains contextes et non pas dans d'autres, Chen *et al.* (2013) utilisent comme corpus de comparaison la section « New York Times » du corpus English *Gigaword* (LDC2003T05) (Parker *et al.*, 2011). Leurs résultats prouvent que les graphes améliorés avec les critères

des similarités (lexicale, syntaxique et de traduction) donnent des paraphrases de bonne qualité en termes de précision.

D'un autre point de vue, Issa *et al.* (2018) étudient le niveau de similarité sémantique entre deux graphes syntaxiques, trait qui indique si deux phrases sont des paraphrases. La **Figure 7** ci-dessous (utilisée par Issa *et al.*, 2018) représente de façon simplifiée leur méthode de fouille de textes appelée **AMR** (*Abstract Meaning Representation*).

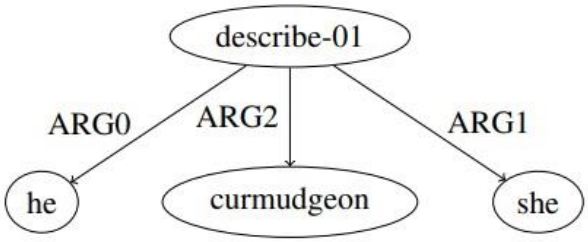

Figure 1: AMR graph for "He described her as a curmudgeon", "His description of her: curmudgeon" and "She was a curmudgeon, according to his description"

Figure 7. Modèle de graphe de type AMR (Abstract Meaning Representation) (Issa et al., 2018)

Le but de cette méthode est de convertir les phrases en graphes, en gardant uniquement les relations sémantiques entre les mots de la phrase (sans les relations grammaticales et syntaxiques). Nous observons dans la **Figure 7** que seulement les relations sémantiques sont gardées sous la forme des arguments (ARG0 /ARG1 /ARG2)[25]. Le travail de recherche est mené sur le corpus de paraphrases en anglais « Microsoft Research Paraphrase Corpus » (MSRP) (Dolan *et al.*, 2004). Pour classer les paraphrases, Issa *et al.* (2018) utilisent « PropBank » (Kingsbury et Palmer, 2002), outil qui propose des annotations sémantiques sur des ressources textuelles en anglais.

Comme méthodes de travail, Issa *et al.* (2018) utilisent le coefficient Jaccard pour déterminer le niveau de similarité sémantique des données fournies par le parseur syntaxique Stanford CoreNLP (Manning *et al.*, 2014). Pour l'évaluation, ils utilisent le corpus « Microsoft Research Paraphrase Corpus » (Dolan *et al.*, 2004) comprenant 5 801 paires de paraphrases.

***

[25] Notre traduction en français des explications d'Issa *et al.* (2018) : *« Graphe AMR pour "Il l'a décrite comme une rouspéteuse », « Sa description d'elle : rouspéteuse » et « Elle était une rouspéteuse, selon sa description ».*

Les travaux présentés dans cette section font appel aux **graphes syntaxiques** pour déterminer le niveau de **similarité sémantique** des paraphrases. Ceci n'est pas du tout paradoxal, parce que l'utilisation de différents noms, adjectifs et verbes donne la diversité lexicale et sémantique nécessaire à la création d'une paraphrase. Pour cette méthode, il est nécessaire d'identifier les parties du discours d'une phrase, les dépendances syntaxiques et les relations sémantiques pour trouver les synonymes qui peuvent constituer des paraphrases au sens similaire au segment source reformulé.

Nous considérons que la méthode à base de graphes syntaxiques n'est pas adaptée pour notre travail de recherche sur la reformulation médicale. Cette méthode demande beaucoup de données annotées et d'informations linguistiques et grammaticales correctes pour chaque langue pour réaliser une bonne analyse syntaxique, mais surtout des représentations sémantiques détaillées en format AMR. Pour le moment, nous ne connaissons pas d'outils disponibles pour les langues traitées, capable de produire un format AMR.

Dans le point suivant, nous présentons les travaux basés sur la désambiguïsation sémantique pour identifier la similarité entre les notions reformulées et leurs reformulations.

2.2 Désambiguïsation sémantique

Le concept de **désambiguïsation sémantique** représente la tâche de **déterminer automatiquement le sens correct** d'un mot cible en tenant compte du **contexte** d'apparition. Si pour l'humain, l'ambiguïté sémantique peut être déchiffrée assez facilement grâce au contexte, pour les outils de TAL, cette tâche devient très complexe (Tchechmedjiev, 2012). Afin de réaliser la simplification automatique des textes, nous devons choisir le synonyme correct qui permet de rendre plus accessible le contenu du texte original et le sens adapté. La désambiguïsation sémantique peut intervenir dans le processus de recherche de reformulations correctes, par conséquent, nous présentons quelques approches appliquées aux paraphrases.

Chaumartin (2007) a travaillé sur la désambiguïsation sémantique des paraphrases en anglais en utilisant la ressource lexicale WordNet (Miller, 1998) et des articles comparables extraits à partir de Wikipédia. Après avoir aligné ces corpus comparables, il a créé une mesure de similarité entre les verbes de ces articles (à partir des verbes de WordNet). Chaumartin (2007) a rajouté deux niveaux supplémentaires pour la désambiguïsation sémantique des verbes en plus de la hiérarchie de WordNet : une racine

commune à tous les verbes et des « catégories sémantiques » (par exemple des verbes de mouvement, verbes d'état, verbes de changement). Par la suite, ces verbes ont été recherchés dans la structure des paraphrases. On applique un algorithme de recherche automatique qui, à l'aide d'une analyse syntaxique de deux textes, arrive à identifier le plus court chemin, dans chaque graphe de dépendance, entre deux entités nommées[26]. Leurs résultats montrent une précision de 70% dans l'identification de paraphrases.

Li *et al.* (2010) ont proposé une méthode de désambiguïsation sémantique à l'aide des règles définies manuellement et des ressources lexicales pour identifier les paraphrases en anglais. Ils ont proposé trois modèles pour la désambiguïsation sémantique des mots et des expressions polylexicales. Ces modèles sont désignés pour comparer la distribution thématique d'un mot ou d'une expression cible avec les paraphrases qui ont le même sens et de choisir la plus probable. Ils construisent un modèle de sujet (ou thème) en décomposant la distribution de probabilités conditionnelles de type *mot-document* en deux distributions différentes : la distribution *mot-sujet* et la distribution *sujet-document*. Li *et al.* (2010) utilisent également WordNet (Miller, 1998) comme ressource pour l'identification des sens des mots. Leurs résultats prouvent que cette méthode peut aider à résoudre le problème de la précision de la désambiguïsation sémantique. Celle-ci peut servir à déterminer le sens des paraphrases de type expression polylexicale.

Dans le **domaine médical**, un grand nombre de travaux sur la désambiguïsation sémantique sont basés sur l'apprentissage supervisé (Liu *et al.*, 2004 ; Gaudan *et al.*, 2005 ; Leroy et Rindflesch, 2005 ; Joshi *et al.*, 2005 ; Andreopoulos *et al.*, 2008 ; Mohammad et Pedersen, 2004). Ces méthodes consistent dans l'utilisation des arbres de décision, des machines à vecteurs de support, de la classification naïve bayésienne sur des corpus d'entraînement déjà étiquetés. Les méthodes utilisent des propriétés telles que les étiquettes grammaticales, les relations sémantiques (Leroy et Rindflesch, 2005), les unigrammes, les bigrammes (Joshi *et al.*, 2005) et les relations syntaxiques et lexicales (Mohammad et Pedersen, 2004) pour la désambiguïsation. Ces méthodes sont complètement basées sur des corpus d'apprentissage qui demandent un travail d'annotation chronophage et des compétences requises dans le domaine médical.

Andreopoulos *et al.* (2008) ont utilisé la terminologie médicale UMLS (Bodenreider 2004) pour désambiguïser le sens des termes médicaux à l'aide d'un classificateur naïf

[26] Exemple des règles utilisées par Chaumartin (2007) pour chercher les verbes dans des structures paraphrastiques :
« (RIVIERE#1 riv1, couler, VILLE#1 v1) ~ (RIVIERE#1 riv1, serpenter, VILLE#1 v1) ;
(RIVIERE#1 riv1, unir, RIVIERE#1 riv2) ~ (RIVIERE#1 riv2, rejoindre, RIVIERE#1 riv1). »

bayésien. Celui-ci identifie la co-occurrence de termes médicaux dans les documents. Stevenson *et al.* (2008) utilisent les identifiants uniques de concepts médicaux d'UMLS (Bodenreider, 2004) et les termes manuellement annotés de MeSH (*Medical Subject Headings*) (Lipscomb, 2000) dans les articles du corpus MEDLINE pour construire des vecteurs de caractéristiques sémantiques des notions médicales. Ils entraînent les classificateurs basés sur les vecteurs à machine de support (SVM), les réseaux bayésiens et les modèles vectoriels. Dinh et Tamine (2010) utilisent le sens précis des concepts médicaux du thésaurus MeSH (*Medical Subject Headings*) (Lipscomb, 2000) pour désambiguïser les termes médicaux ambigus dans les documents et les requêtes afin d'en créer une indexation sémantique des textes médicaux.

***

Nous constatons que les travaux de désambiguïsation sémantique sont menés sur l'identification du sens des mots clés d'une reformulation / paraphrase ou sur des termes (dans le domaine médical). À notre connaissance, la désambiguïsation sémantique des paraphrases des termes médicaux est peu traitée et vise la détection du sens des termes spécialisés (Andreopoulos *et al.*, 2008 ; Dupuch *et al.*, 2013). L'opération de simplification lexicale des termes médicaux par des mots de la langue commune passe par la désambiguïsation sémantique et l'emploi du synonyme correct dans le contexte donné. Vu la contrainte de similarité sémantique entre le terme médical et sa reformulation, il est nécessaire d'utiliser des ressources de langue spécialisée comme UMLS (Bodenreider, 2004) pour le domaine médical. Ces ressources spécialisées sont nécessaires pour l'identification des reformulations médicales correctes.

La plupart des travaux utilisent les terminologies existantes pour l'annotation des données et la qualité des résultats dépend directement de la qualité et la pertinence des terminologies. Nous travaillons avec deux langues différentes de l'anglais, **le français et le roumain**, une langue de moindre diffusion et moins dotée en termes de ressources. Trouver pour le roumain des terminologies médicales de la même taille et qualité que celles en anglais est problématique. Comme nous réalisons notre travail de façon comparative et contrastive, il est très important de disposer des ressources lexicales et des terminologies comparables, attestés par les experts terminologues. Pour combler cette lacune, nous utilisons différentes ressources pour l'identification des termes, que nous détaillons dans le **Chapitre Méthodologie**.

***

Qui dit paraphrase et reformulation, dit **variation lexicale** dans le cadre d'une même langue. Cependant, cette variation peut également apparaitre lors de **multiples traductions**

d'une langue source vers une langue cible. Le TAL permet de créer de multiples variantes de traduction du même texte à l'aide des outils de **traduction automatique**. Nous présentons des travaux qui explorent cette tâche dans la section suivante.

2.3 Traduction automatique

La **traduction automatique** est un volet également exploité pour la recherche des paraphrases et reformulations, définies comme des variations de la langue à travers de multiples traductions. Les paraphrases, souvent phrastiques, sont identifiées de façon automatique à l'aide des logiciels qui se servent de bases de traduction, de bases terminologiques multilingues, ainsi que de corpus parallèles ou comparables de grande taille.

Dans l'identification de la paraphrase, deux types de traductions automatiques sont utilisées : la *traduction automatique statistique* et la *traduction automatique neuronale*. Les deux méthodes passent par des langues pivots et par des mécanismes de traduction et *rétrotraduction* (une deuxième traduction de la langue cible vers la langue source). Nous présentons des travaux sur l'identification ou la génération de la paraphrase qui utilisent ces deux méthodes dans les sous-chapitres suivants.

2.3.1 Traduction automatique statistique

L'étude de Bouamor (2012) utilise des corpus comparables et parallèles (Barzilay et McKeown, 2001), des méthodes de traduction automatique et des règles définies manuellement. Elle étudie la **paraphrase sous-phrastique** en français et en anglais à partir de quatre types de corpus comparables des paires d'énoncés sémantiquement liés. Les corpus d'entraînement utilisés sont de plusieurs types :

- le corpus Multitrad (Bouamor, 2010) qui contient des traductions multiples à partir d'une ou de plusieurs langues ;
- un corpus de 50 paires d'énoncés provenant de deux versions de sous-titres en français de la série américaine « Desperate Housewives »[27] ;

[27] Sous-titres librement disponibles sur http://www.opensubtitles.org

- un corpus de 50 paires de descriptions multiples de vidéos en français en provenance du corpus réalisé par crowdsourcing[28] du Chen et Dolan (2011) ;
- un corpus des 1 462 titres différents pour 100 articles de journaux, extraits avec Google News.

Bouamor (2012) utilise différentes règles d'acquisition et de classification des paraphrases sous-phrastiques. Ses méthodes sont :

- **L'alignement lexical** - Bouamor (2012) a testé cette méthode sur un corpus parallèle des paraphrases possibles obtenues à partir de traductions multiples. Elle utilise l'apprentissage statistique avec le système MOSES (Koehn *et al.*, 2007) pour réaliser l'extraction des bisegments candidats, qui à l'aide de l'outil GIZA++ (Och et Ney, 2003) aligne chaque traduction dans différents ordres.
- **Les variations entre termes** - cette méthode exploite le système FASTR (Jacquemin, 1999) qui permet de définir manuellement des métarègles pour repérer des variantes terminologiques de nature morphosyntaxique et faire appel à des ressources préexistantes telles que des familles morphologiques et sémantiques. Bouamor utilise par exemple cette métarègle : un segment qui est formé d'un nom suivi d'un adjectif peut être réécrit en un segment formé au minimum d'un verbe suivi d'un nom et d'un adjectif, uniquement si le nom et le verbe font partie de la même famille morphologique et les adjectifs sont synonymes. Un exemple de cette métarègle : « protéger de façon permanente » est identifié comme une variante du syntagme « protection constante » (Bouamor, 2012 : 75). Cette méthode sert à trouver différentes paraphrases sur un sujet donné, mais ayant une forme proche.
- **L'alignement des structures syntaxiques** - cette technique utilise la structure syntaxique des énoncés pour mettre en correspondance des segments paraphrastiques. Les énoncés sont d'abord annotés par un analyseur syntaxique. Les résultats sont pertinents surtout si les structures syntaxiques de haut niveau des deux phrases sont similaires.
- **Les traductions communes dans une langue pivot** - cette méthode exploite les équivalences de traduction obtenues à travers une langue pivot selon la méthode de Bannard et Callison-Burch (2005). Leur méthode consiste à utiliser des corpus parallèles bilingues pour créer des reformulations par traduction dans une langue suivie par une *rétrotraduction* et sélection dans la langue d'origine.

[28] Le « crowdsourcing » représente une participation collaborative (ou parallèle) des volontaires à un appel commun.

Plus précisément, un segment source est traduit dans une langue pivot, puis traduit à nouveau dans sa langue d'origine, ce qui permet d'obtenir une paraphrase candidate. Bouamor (2012) a utilisé le système de traduction automatique statistique Moses (Koehn *et al.*, 2007) et l'outil Giza++ (Och et Ney, 2003) pour aligner les phrases au niveau du mot. Après traduction et rétrotraduction, elle a sélectionné les paraphrases les plus appropriées.

- **La mesure du taux d'édition entre des séquences de mots** - cette méthode est réalisée par l'adaptation de la technique TER_p (Translation Edit Rate plus) (Snover *et al.*, 2009), originellement créée pour générer l'hypothèse de la meilleure traduction, dans une perspective monolingue, et, par conséquent, identifier des paraphrases ayant le même contenu sémantique.

À la suite de ses expériences, Bouamor (2012) construit une **typologie des paraphrases acquises** (*sûres* et *possibles*) qui décrit des cas de synonymie et qui inclut également des variations d'ordre pragmatique. Des mesures de calcul comme le rappel et la F-mesure sont utilisées pour estimer le taux de réussite de ses méthodes présentées ci-dessus. À l'issue de cette étude, une typologie bilingue des paraphrases sous-phrastiques a été créée avec des résultats comparables pour les deux langues.

Dans le cadre d'une autre étude, Bouamor *et al.* (2012) utilisent comme approche les méthodes de traduction automatique statistique et des règles définies manuellement. Bouamor *et al.* (2012) s'intéressent au cas de la validation des reformulations locales sur Wikipédia francophone. Les révisions de Wikipédia fournissent un ensemble des réécritures possibles qui peuvent contenir des paraphrases valides en contexte. Bouamor *et al.* (2012) ont évalué la qualité des classements automatiques des paraphrases à l'aide des paraphrases associées au segment dans le corpus WICOPACO (Dutrey *et al.*, 2011), des contextes construits manuellement et des paraphrases obtenues par des traductions automatiques. Ces traductions passent par une langue pivot proche du français (l'espagnol) et une autre très différente (le chinois). Les classifications réalisées à l'aide des paraphrases construites manuellement ont donné de très bons résultats, par rapport à la méthode du pivot. Les résultats sont pourtant bien meilleurs pour la langue pivot proche, l'espagnol, que pour le chinois.

Pavlick *et al.* (2015) empruntent une méthode de traduction automatique et la transposent comme tâche d'apprentissage statistique sur des corpus monolingues pour identifier les paraphrases lexicales en fonction du domaine de spécialité (la biologie). Ils se concentrent sur les différences sémantiques des termes de différents domaines de

spécialité. La langue de travail est l'anglais, ayant comme langue de contrôle le français. Pour identifier les paraphrases bilingues, Pavlick *et al.* (2015) utilisent un modèle d'algorithme bilingue de type pivot (Bannard et Callison-Burch, 2005). Les recherches sont réalisées sur le corpus de biologie de GENIA (Kim *et al.*, 2003) et sur un texte introductif d'un manuel de biologie.

Pavlick *et al.* (2015) font appel à la méthode de Moore et Lewis (2010), méthode qui identifie les expressions spécialisés par rapport à un corpus général. Cette méthode attribue un score à chaque expression du corpus général sur la base de deux modèles de langage. Ces modèles sont créés par apprentissage automatique à partir d'échantillons de texte du domaine cible et un autre du langage général. Le score est donné par la différence entre les entropies[29] croisées des deux modèles de langue. Afin de trouver les termes de spécialité (et leurs paraphrases lexicales / synonymes) dans ces expressions, Pavlick *et al.* (2015) comparent les résultats avec un corpus général composé des textes spécialisés et des sous-titres des films (Callison-Burch *et al.*, 2009) et avec un corpus général issu de Wikipédia. L'évaluation est réalisée avec les mesures de précision / rappel et par cinq évaluateurs humains. L'originalité de leur travail est la combinaison des méthodes d'apprentissage et leur application sur le corpus général en même temps que sur plusieurs sous-corpus de spécialité.

Nguyen-Son et al. (2015) identifient les paraphrases en fonction de leur rapport de synonymie avec un modèle qui cherche la similarité sémantique à base de vecteurs, une métrique nommée SimMat, combinée avec des métriques d'évaluation de la traduction automatique. La similarité sémantique est déterminée en fonction de la fréquence du terme et la co-occurrence de celui-ci dans le corpus, par rapport aux mots retrouvés dans la base WordNet (Miller, 1998) et des opérations d'édition entre les séquences. Cette méthode utilise le grand corpus de paraphrases en anglais « The Microsoft Research Paraphrase (MSRP) » (Dolan et al., 2004). Les meilleurs résultats ont été obtenus avec la métrique SimMat combinée avec huit métriques de traduction automatique (précision de 76,6% et F-mesure de 83,9%).

***

L'utilisation de la **traduction automatique statistique** (qui demande des alignements automatiques, parfois des analyseurs syntaxiques et des règles) permet d'identifier automatiquement des paraphrases dans un contexte multilingue. Pourtant, nous

[29] En linguistique, l'*entropie* représente le degré d'information discursive transmise par rapport au nombre de lexèmes de la phrase.

remarquons que la qualité des résultats obtenus dépend de la proximité étymologique des langues utilisées (Bouamor *et al.*, 2012), des ressources lexicales disponibles et du degré de comparabilité entre les textes. La méthode par traduction automatique statistique est efficace dans le cadre des langues latines grâce à la racine latine qui est la base d'une grande partie des mots (comme dans l'étude de Bouamor *et al.* (2012) sur le français et l'espagnol). Les métriques peuvent évaluer le degré de similarité entre le référent et la paraphrase par des opération d'édition (insertion, suppression, etc.)

Des études plus récentes ont exploité la **traduction neuronale** pour générer des nouvelles paraphrases à travers une ou plusieurs langues pivots. Nous présentons cette méthode dans le sous-chapitre suivant.

2.3.2 Traduction automatique neuronale

Si la traduction automatique statistique fonctionne au niveau du mot et du syntagme par l'assemblage de mots et fragments de phrases traduits et validés par des humains, **la traduction automatique neuronale** agit directement au niveau de la phrase et fonctionne sur un principe de *sémantique lexicale distributionnelle* (définir le sens des mots par rapport au contexte). Grass (2022) met en avant l'avantage de ce principe : la traduction neuronale permet de réduire les erreurs typiques de la traduction statistique, comme la synonymie et la polysémie mal traduites. Dans le même sens, certains travaux sur la paraphrase ont testé la traduction neuronale pour générer des nouvelles paraphrases à partir d'une ou plusieurs langues pivot (Narayan *et al.*, 2017 ; Sekizawa *et al.*, 2017).

Zhou *et al.* (2021) ont considéré les paraphrases comme étant des langues différentes. Ils ont utilisé le système de traduction neuronale **OpenNMT** (Klein *et al.*, 2017) et des phrases de la Bible en deux langues, anglais et français. Ils évaluent la capacité de leur modèle à générer des paraphrases phrastiques avec le score BLEU (Papineni *et al.*, 2002), algorithme qui évalue la qualité d'une traduction automatique. Zhou *et al.* (2021) concluent que l'ajout de paraphrases dans la « langue » source et cible améliore la qualité de la traduction, augmente la diversité lexicale et aide à paraphraser les mots rares (score BLEU de **57,2** sur 24 paraphrases du français vers l'anglais). Pourtant, ces résultats sont obtenus sur un nombre très réduit de paraphrases.

***

Le besoin de **corpus parallèles multilingues** limite l'utilisation de la traduction automatique pour l'extraction des paraphrases en grande quantité. Bouamor (2012) a exploité plusieurs techniques de traduction statistique pour identifier des paraphrases sous-

phrastiques en français et en anglais (alignement, traduction vers une langue pivot). Dans notre cas, il y a peu ou pas de corpus parallèles disponible en français et en roumain pour le domaine médical, pour appliquer ce type de méthodes.

La traduction automatique neuronale a été utilisé principalement sur les paraphrases phrastiques, en utilisant des langues pivot vers l'anglais (Zhou *et al.*, 2021). Si nous utilisons d'autres méthodes **d'apprentissage par réseaux de neurones** appliquées sur chaque langue, la traduction n'a pas été retenue, par manque de corpus parallèles disponibles dans les langues que nous étudions. Ainsi, nous menons des expériences à base de **corpus comparables ou généraux** (qui sont plus faciles à trouver ou constituer) et nous **générons des reformulations** en fonction du contexte représentées à l'aide de plongements lexicaux.

***

Notre étude a comme objectif final de **créer des corpus de reformulation de façon semi-automatique**. Pour développer notre méthodologie de travail, nous nous intéressons au processus de **simplification automatique des textes**. Nous menons des recherches sur la **reformulation médicale** vue comme un syntagme qui a le rôle de **simplifier** des termes médicaux difficiles à comprendre par le grand public. Dans ce sens nous présentons par la suite les tâches de simplification automatique et les travaux réalisés sur la paraphrase et la reformulation, en particulier sur les textes médicaux.

2.4 La simplification automatique

Vu que nous nous intéressons à la question de la reformulation pour la vulgarisation, **la simplification** est une tache nécessaire afin de rendre le langage technique accessible pour le grand public. Dans une perspective de **simplification automatique** des textes médicaux, nous présentons les travaux et les ressources pour la simplification automatique, lexicale et syntaxique et les possibles applications pour notre objet d'étude. En effet, la reformulation peut être considérée comme **tâche de simplification** d'un langage complexe vers un langage plus simple.

2.4.1 La simplification lexicale automatique

La **simplification lexicale** vise le remplacement des mots complexes avec leurs synonymes plus faciles à comprendre par le lecteur cible. Les travaux de recherche sur la **simplification lexicale automatique** en TAL ont exploré plusieurs dimensions : la

proposition des modèles de simplification lexicale en fonction de la taille du contexte (Ligozat *et al.*, 2013) ; la détection des mots ou fragments de textes difficiles (Grabar et Hamon, 2014) ; l'identification des éléments prédicteurs de la complexité lexicale (Gala *et al.*, 2014) ; la substitution lexicale des termes scientifiques avec des variantes simplifiées avec application dans le domaine médical (Cardon, 2018).

Afin d'aider à l'automatisation de la simplification lexicale des textes, plusieurs ressources lexicales exploitables en TAL ont été développées pour la langue française :

- **ResSyf** (Gala *et al.*, 2015), un dictionnaire qui s'applique sur le texte et qui aide à la compréhension des mots, en faisant une classification par la complexité, selon la fréquence et la variabilité lexicale. Ce dictionnaire a été conçu comme ressource pour la didactique et pour des enfants rencontrant des difficultés de lecture ;
- **MANULEX** (Lété *et al.*, 2004), une base de données lexicales qui contient les fréquences d'occurrences de mots calculées à partir d'un corpus de 54 manuels scolaires en français, catégorisées par le niveau d'étude ;
- **FLELex** (Gala *et al.*, 2014), un grand lexique pour le français langue étrangère (FLE) qui rapporte les fréquences de mots par niveau de difficulté (selon l'échelle CECR).

Cependant, il faut noter que ces outils s'adressent aux publics différents et ne correspond pas à notre public cible : MANULEX et ResSyf sont utilisés par rapport au niveau scolaire des lecteurs, tandis que FLELex est utilisé par les apprenants du français qui ont besoin de connaître des mots avec un haut niveau de difficulté pour mieux maîtriser la langue française. Des telles ressources doivent être construites pour le domaine médical, car ces ressources ne sont pas disponibles actuellement. De plus, ces approches remplacent les mots complexes avec les synonymes, ce qui est éloigné de notre objectif.

2.4.2 La simplification lexicale des termes médicaux

La variation lexicale et sémantique des termes médicaux a fait l'objet de plusieurs travaux (Grabar et Zweigenbaum, 2000 ; Zweigenbaum, 1999 ; Hahn *et al.*, 2001). Une de plus grandes difficultés rencontrées lors de la simplification des termes médicaux est de les « traduire » en langage commun. L'origine grecque et latine des termes médicaux pose des difficultés de simplification. Une méthode de simplification des termes médicaux est leur décomposition en composés néoclassiques. Par exemple, le terme « iridochoroïdite » est

décomposé en : inflammation, iris, et choroïde (Grabar et Hamon, 2016 : 6). L'analyse morphologique des termes médicaux scientifiques comme « galactose », « acromégalie » (Grabar et Hamon, 2016 : 22), suivie par l'analyse syntaxique des phrases et la fouille de textes non-spécialisés de Wikipédia du domaine médical permet d'y trouver des paraphrases. Leur méthode consiste à décomposer les termes et à traduire les composants grecs et latins en français moderne (card = cœur). Ces mots de la langue française qui correspondent à la décomposition morphologique des termes médicaux sont projetés sur le corpus pour identifier et extraire les syntagmes qui contiennent des paraphrases.

Si Grabar et Hamon (2016) ont mené leur travail sur les termes médicaux néoclassiques français monolexicaux composés de plusieurs bases (latines, grecques, françaises), nous traitons également les termes médicaux polylexicaux, en plus des termes simples. Les termes sont identifiés à l'aide de terminologies attestées. Nous pouvons considérer les paires *terme-reformulation* comme un corpus de simplification, à condition que la reformulation soit plus simple que le terme source pour le public cible. Nous chercherons d'une part à générer de nouvelles reformulations en exploitant des ressources telles que les modèles de langues et le corpus de reformulation validé manuellement, présentées dans la partie **Méthodologie**. D'autre part, nous identifions les paires de *termes-reformulation* par un système de classification automatique.

2.4.3 La simplification syntaxique automatique

La **simplification syntaxique** représente le changement de la structure grammaticale de la phrase afin de rendre les textes plus faciles à comprendre pour un public cible donné. Brouwers *et al.* (2012) ont travaillé sur la réalisation d'une typologie de simplifications syntaxiques en exploitant les corpus des révisions de Wikipédia et Vikidia (la version simplifiée de l'encyclopédie adaptée aux enfants). Minard *et al.* (2012) se sont intéressés à la simplification syntaxique de phrases dans le but d'améliorer l'extraction de relations dans un corpus de textes médicaux (relations temporelles, de lieu, âge de patients).

Une étude pionnière sur **simplification automatique de la syntaxe** des textes pour en trouver des **paraphrases** est proposée par Dras (1999). Il travaille sur le concept de « paraphrase réticente » (en anglais *reluctant paraphrase*), qui fait référence à la modification du texte pour l'adapter aux contraintes externes telles que la longueur, la lisibilité ou le style de texte. Comme Chandrasekar et Srinivas (1997), Dras utilise le formalisme de la *grammaire d'arbres adjoints* (TAG) pour représenter une phrase. Dras

utilise la programmation en nombres entiers (quand toutes les variables sont contraintes à prendre uniquement des valeurs entières) pour générer un texte qui satisfait des contraintes comme la longueur ou la lisibilité en utilisant une paraphrase minimale, donc plus facile à comprendre par les lecteurs.

Plus tard, Grabar et Hamon (2016) proposent une méthode automatique d'acquisition des **paraphrases du domaine médical en français** afin de rendre les textes médicaux trop scientifiques plus compréhensibles par le grand public. Cette étude qui se concentre sur l'étymologie néoclassique des termes médicaux emprunte une méthode fondée sur **l'analyse syntaxique** des phrases pour trouver des paraphrases qui peuvent servir à la simplification de textes. Plus précisément, la méthode employée par Grabar et Hamon (2016) est définie en quatre étapes : l'analyse morphologique de termes médicaux néoclassiques (dont la dénomination est composée des bases grecques ou latines) avec l'outil Dérif (Namer, 2009)[30], l'étiquetage et la lemmatisation du corpus avec l'outil Cordial (Laurent *et al.*, 2009), l'alignement des termes et des segments du corpus pour extraire des paraphrases.

Grabar et Hamon (2016) exploitent les termes médicaux fournis par des terminologies médicales existantes, telles que Snomed International (Côté, 1996), la partie française d'UMLS (Lindberg *et al.*, 1993) et les articles de Wikipédia. Leur étude se concentre plus sur les paraphrases qui apparaissent dans les contextes libres, sans être dans la co-occurrence du terme médical, comme dans l'exemple suivant, dont la paraphrase « une inflammation des cellules » est identifiée pour le terme technique « la cellulite » :

> « La cellulite est une infection grave qui se propage sous la peau et s'attaque aux tissus mous comme la peau elle-même et les graisses sous-jacentes.
>
> L'infection virale cause une inflammation des cellules nerveuses, conduisant à la destruction partielle ou totale du ganglion des motoneurones. » (Grabar et Hamon, 2016 : 8)

Les paraphrases sont extraites en fonction de quatre paramètres : la taille de syntagme (d'unigrammes à quadrigrammes) ; l'utilisation de formes brutes du texte pour déterminer la variation terminologique ; le taux d'alignement des termes techniques et le taux d'alignement des syntagmes syntaxiques. Grabar et Hamon (2016) identifient des paraphrases qui ont la forme des groupes nominaux (« aclasie : absence de fracture »), des

[30] Exemple d'analyse morphologique : « myocardique/A : [[[myo N*] [carde N*] NOM] ique ADJ] » (Grabar et Hamon, 2016 : 9), dont l'outil Dérif (Namer, 2009) attribue la catégorie plus probable pour chaque base du terme médical composé (N* = nom, ADJ = adjectif).

groupes prépositionnels (« périovulatoire : autour de la date d'ovulation »), des groupes participiaux (« malformatif : mal formé »), des groupes verbaux (« sinoscopie : observer les sinus maxillaires ») et de type subordonné (« agalactie : qui se caractérise par l'absence de lait ») (Grabar et Hamon, 2016 : 17).

Grabar et Hamon (2016) ont comparé leurs extractions par paramètres avec l'extraction des paraphrases identifiées en contexte définitoire à l'aide des patrons comme « est », « également appelé », « peut être défini comme ». Ils ont trouvé 2 037 définitions qui correspondent à 1 286 termes médicaux uniques, de type : « L'angiographie *est* un examen invasif »[31] (Grabar et Hamon, 2016 : 19). À la suite de ces expériences, Grabar et Hamon (2016) ont trouvé des équivalents en paraphrases, définitions et explications en langue générale pour 2 596 termes médicaux uniques.

Koptient *et al.* (2019) proposent une typologie de transformations syntaxiques pour la simplification automatique à travers la fragmentation et la réorganisation syntaxique des phrases parallèles médicales en français, de type expert et grand public. Les transformations observées sont les remplacements par des synonymes, la spécification (insertion d'informations supplémentaires), la généralisation (suppression d'informations), la pronominalisation, la substitution des adjectifs par les noms correspondants et les changements entre le singulier et le pluriel. Toutefois, ces transformations modifient le texte et il n'y a pas de lien direct entre le référent et sa reformulation.

Nisioi *et al.* (2017) explorent pour la première fois les réseaux de neurones de type LSTM (Hochreiter et Schmidhuber, 1997) et l'outil pour la traduction automatique neuronale OpenNMT (Klein *et al.*, 2017) pour *simplifier automatiquement* des phrases en anglais. Ils utilisent comme ressources d'apprentissage des paires de phrases en anglais général avec leur variante simplifiée issues de Wikipédia (Hwang *et al.*, 2015). Leurs expériences montrent que les phrases simplifiées générées ont une syntaxe plus simple, tout en gardant le contenu lexical. Plus récemment, Martin *et al.* (2022) proposent un système multilingue de simplification de phrases (MUSS) qui fonctionne par apprentissage non supervisé. MUSS utilise comme données des paraphrases phrastiques parallèles pour proposer des variantes simplifiées. L'outil a été testé sur des données générales en anglais, français et espagnol. Pourtant, ces systèmes de simplification syntaxiques (Nisioi *et al.*, 2017 ; Martin *et al.*, 2022) ne sont pas adaptés au domaine médical ou à la langue roumaine, nous ne disposons pas de corpus de taille suffisante pour appliquer ces techniques.

***

[31] Les éléments sont surlignés et mis en italique par nous.

D'autre part, nous considérons que la **simplification lexicale et syntaxique** sont indispensables dans le processus de rédaction des textes de vulgarisation. La simplification syntaxique rend le contenu scientifique de textes médicaux plus accessible au grand public à travers l'utilisation des phrases avec une structure grammaticale simple. La **simplification syntaxique**, ensemble avec la **simplification lexicale** qui, dans le domaine médical, représente le remplacement des termes techniques avec leurs synonymes ou paraphrases de la langue courante, améliorent la compréhension d'un texte scientifique.

Pour notre recherche sur les textes du domaine médical, nous cherchons la reformulation dans la même phrase, avec des variantes syntaxiques et phrastiques très variées. Nous nous inspirons des approches exploitées par Grabar et Hamon (2016) qui visent à chercher des définitions et explications des termes à base de patrons. Nous travaillons sur des données en langage scientifique et en langage simplifié et nous analysons des données qui sont déjà simplifiées (textes destinés au grand public). Notre objectif est de construire un tel corpus de simplification.

2.5 Bilan

Dans ce chapitre, nous avons passé en revue les approches d'identification de la paraphrase interprétées comme des tâches spécifiques en TAL : la désambiguïsation, la traduction ou la simplification automatique.

La désambiguïsation est utile pour identifier le sens du référent et de choisir une paraphrase adaptée. Les techniques de traduction automatique utilisent la traduction par pivot, ou font appel aux métriques. Peu de corpus parallèles sont disponibles pour le domaine médical pour les langues que nous étudions. C'est le même problème pour les méthodes de simplification automatique qui s'appuient sur des techniques de traduction. En revanche, nous adoptons des méthodes à base de patrons pour l'identification de marqueurs de reformulation suivant (Grabar et Hamon, 2016), du référent et de la reformulation.

Pour identifier les reformulations, nous avons besoin de corpus de grande taille qui contiennent des reformulations dans les langues de notre étude. Dans ce sens, nous présentons dans la section suivante les corpus de reformulations et paraphrases disponibles, leurs spécificités, ainsi que leur possible utilité pour notre travail de recherche.

3. Ressources pour le TAL : corpus de paraphrases

Le but de notre thèse est d'améliorer les méthodes existantes pour créer des corpus qui contiennent des reformulations et plus généralement de créer de nouveaux corpus comparables pour le français et le roumain. Dans ce sens, nous avons identifié les corpus disponibles pour analyser leur composition mais aussi la méthode utilisée pour leur création. La plupart des ressources identifient des paraphrases et très peu prennent en compte la notion de reformulation.

Nous avons structuré ce chapitre en fonction de la langue des textes, du caractère monolingue ou multilingue du corpus. Les corpus de paraphrases multilingues ont été groupés ensemble pour une meilleure présentation des méthodes d'identification de la paraphrase. Nous nous intéressons aussi à l'origine des langues :

- **l'anglais**, langue germanique avec une structure grammaticale plus simple et plus facilement adaptable à l'apprentissage automatique ;
- **le français** et **le roumain**, langues latines avec des structures grammaticales proches et plus complexes que celle de l'anglais.

Pour notre recherche sur les paraphrases, nous prendrons en compte les spécificités grammaticales, syntaxiques et lexicales de chacune de nos langues d'étude et du langage spécialisé du domaine de la médecine.

3.1 Corpus anglophones

Un corpus de paraphrases pour la langue anglaise très exploité dans les travaux sur l'extraction de la paraphrase est le *Corpus Microsoft Research Paraphrase* (MSRP) réalisé par Dolan *et al.* (2004) qui contient 5 801 paires de phrases reformulées. Ce corpus est constitué d'articles journalistiques recueillis à partir de milliers de sources d'information sur le web. Dolan *et al.* (2004) utilisent deux techniques : la distance d'édition des chaînes de caractères (Levenshtein, 1966) et une stratégie heuristique qui associe des phrases résumées provenant de différents articles d'actualité dans le même groupe. Ils évaluent les deux ensembles de données avec un algorithme d'alignement des mots et la métrique AER (*Alignment Error Rate*) empruntée à la traduction automatique statistique (Och et Ney,

2003). L'objectif de ce corpus est de servir à la recherche sur l'identification et la génération automatique de la paraphrase à travers des méthodes telles que la traduction (monolingue) statistique automatique et l'apprentissage automatique statistique.

Pour le domaine médical, il existe le corpus biomédical *MedMentions* (Mohan et Li, 2019), formé à partir de 4 000 résumés extraits du site PubMed[32]. *MedMentions* présente une très grande terminologie médicale comprenant plus de 3 millions de concepts médicaux. Les termes médicaux sont annotés avec les codes de la base terminologique médicale en anglais *UMLS* (Bodenreider, 2004) et sont disponibles en libre accès[33]. Pourtant, ce corpus n'identifie pas les reformulations (synonymes ou explications) de ces termes identifiés.

3.1.1 Corpus bilingue incluant l'anglais

Le corpus de paraphrases japonais-anglais (Shirai *et al.*, 2001) a été créé pour améliorer les dictionnaires japonais-anglais et également pour servir à la traduction automatique. Il contient une partie monolingue en anglais qui présente 27 000 paraphrases phrastiques (Bouamor, 2012) et 28 000 paraphrases phrastiques en japonais. Ces paraphrases ont été créées sur la base de 6 000 prédicats en japonais. Les données sont de langue générale et ont été collectées à partir des exemples de phrases proposées dans des dictionnaires japonais-anglais. L'exemple suivant prouve que ces données peuvent être exploitées également comme sources monolingues de paraphrases phrastiques :

«**J0** Kare-no kikaku-ga atatta.
His plan hit
"his plan was a success"
J1 Kare-no kikaku-ga sˆekˆo-shita.
his plan succeeded
"his plan succeeded"
E0 His plan was a success.
E1 His plan succeded.
E2 His plan was successful » (Shirai, Yamamoto et Bond, 2001 : 1)

Un autre travail similaire et plus conséquent est le *Para-Phrase DataBase PPDB* (Ganitkevitch *et al.*, 2013), une base de données en anglais et espagnol qui contient 220 millions de paires des paraphrases. Cette base de données contient 73 millions de paraphrases phrastiques, 8 millions de paraphrases lexicales, ainsi que 140 millions de

[32] https://www.ncbi.nlm.nih.gov/pubmed/ (les articles ont été recueillis pendant la période janvier 2016 - janvier 2017).
[33] https://github.com/chanzuckerberg/MedMentions

paraphrases syntaxiques (qui gardent le sens, mais aussi la structure syntaxique). Le corpus est constitué d'une collecte de plusieurs corpus en plusieurs langues, en deux étapes :

- l'extraction de paraphrases lexicales, phraséologiques et syntaxiques de grands corpus parallèles bilingues et le calcul des scores de similarité distributionnelle pour chacune des paraphrases (utilisant les n-grammes de Google (Brants et Franz, 2006 ; Lin *et al.*, 2010) et le corpus Annotated Gigaword (Napoles *et al.*, 2012) ;
- l'utilisation des langues pivots pour identifier les différences sémantiques des paraphrases par rétrotraduction en anglais et espagnol (Bannard et Callison-Burch, 2005).

3.2 Corpus multilingues : anglais et français

Une année plus tard, Ganitkevitch et Callison-Burch (2014) ont développé le « *Para-Phrase DataBase PPDB »* pour y ajouter 21 langues[34] et créer le corpus de paraphrases « The Multilingual Paraphrase Database ». Le corpus a été réalisé à partir de plusieurs corpus bilingues parallèles (Europarl-v7, textes juridiques (Koehn, 2005), JRC, textes législatifs européens (Steinberger *et al.*, 2006), OpenSubtitles (Lison et Tiedemann, 2016), GALE, CommonCrawl, Yandex) avec la méthode de la traduction par pivot (Bannard et Callison-Burch, 2005) en utilisant comme langue pivot l'anglais. Ganitkevitch et Callison-Burch (2014) cherchent trois types de paraphrases :

- **Les paraphrases lexicales** - paraphrases de mots simples ou synonymes ;
- **Les paraphrases polylexicales** - paraphrases constituées de plusieurs mots ;
- **Les paraphrases syntaxiques** - structure de paraphrases qui contiennent un marqueur syntaxique ; ce marqueur permet de substituer toute paraphrase qui a cette structure syntaxique.

Pour identifier les paraphrases, Ganitkevitch et Callison-Burch (2014) utilisent une métrique d'identification de la paraphrase spécifique à chaque langue, composée d'un ensemble de règles grammaticales, sans tenir compte du contexte. Le résultat de leurs extractions est une collection de bi-textes qui comprend plus de 100 millions de paires de

[34] Allemand, arabe, bulgare, chinois, estonien, finnois, français, grec, hongrois, italien, letton, lituanien, néerlandais, polonais, portugais, roumain, russe, slovaque, slovène, suédois et tchèque.

phrases (anglais / langues étrangères). La base de données de paraphrases multilingues est accessible gratuitement[35].

Opusparcus (*OpenSubtitlesParaphraseCorpus*) (Creutz, 2018) est un corpus de paraphrases constitué à partir de sous-titres de films et d'émissions télévisées extraits du corpus parallèle OpenSubtitles2016 (Lison et Tiedemann, 2016). Les langues prises en compte sont l'allemand, l'anglais, le finnois, le français, le russe et le suédois. Creutz (2018) utilise la méthode de la traduction par langue pivot de Bannard et Callison-Burch (2005), qui consiste à traduire un mot ou une expression d'une langue A vers une langue B et la rétrotraduction de la langue B dans la langue d'origine A. Creutz (2018) en ajoutent plusieurs langues pivots afin d'avoir des résultats plus diversifiés, et, par conséquent, un plus grand nombre des paraphrases. Le statut et la qualité des paraphrases sont jugés grâce à un nombre considérable d'annotateurs humains. L'apport nouveau de ce corpus consiste dans le registre familier utilisé dans les sous-titres. Les paraphrases identifiées à partir du registre familier peuvent devenir une source précieuse dans l'apprentissage d'une langue assistée par l'ordinateur ou pour aider les apprenants à trouver des expressions naturelles et idiomatiques dans des situations réelles de communication.

Le corpus de paraphrases PAWS-X (Yang *et al.*, 2019) est construit à partir du corpus PAWS (*Paraphrase Adversaries from Word Scrambling*) (Zhang *et al.*, 2019). PAWS-X est constitué de 23 659 paires de paraphrases traduites par des traducteurs humains de l'anglais vers six langues très différentes : français, espagnol, allemand, chinois, japonais et coréen. Les paraphrases originales en anglais du corpus PAWS proviennent de sites web Wikipedia et Quora.

3.3 Corpus francophones

Les corpus de reformulations monolingues pour la langue française sont en nombre limité, ce qui augmente encore la difficulté de notre mission. Un tel type de corpus a été constitué par Brouwers *et al.* (2012) qui ont construit un corpus parallèle à partir d'articles de Wikipédia et Vikidia dans l'objectif de simplification automatique. Tandis que Wikipédia est une encyclopédie collaborative destinée au grand public, Vikidia est conçue pour les enfants entre huit et treize ans. Bien évidemment, le langage utilisé pour constituer les articles encyclopédiques sur Vikidia est un langage simplifié, dont les termes scientifiques sont vulgarisés et traduits dans des mots du langage commun. À partir de ces deux corpus,

[35] Sur le site paraphrase.org

Wikipédia et Vikidia, Brouwers *et al.* (2012) ont établi une typologie de simplifications des termes qui posent des problèmes à la lecture en fonction du public cible, le lecteur adulte (appartenant au grand public) et les enfants. Ce corpus n'est pourtant pas libre d'accès pour le consulter pour notre travail de recherche.

3.4 Corpus multilingues : français et roumain

Pour la langue roumaine, les corpus de paraphrases ou de reformulation sont également peu nombreux. Une base multilingue de paraphrases est le corpus déjà présenté plus haut, « ***The Multilingual Paraphrase Database*** » (Ganitkevitch et Callison-Burch, 2014). Ce corpus contient des textes en 23 langues, dont le français et le roumain sont présents. Les paraphrases sont extraites par le même procédé : utiliser l'anglais comme langue pivot pour identifier les paraphrases à travers les différentes traductions vers les autres langues. Ce corpus a été créé pour aider aux recherches dans le traitement automatique de langues pour les langues moins dotées, mais les types de textes contenus dans ce corpus ne sont pas adaptés à notre situation (textes en langue générale, du domaine juridique, ou transcrites de l'oral).

Le corpus **ParaCotta** (Fikri Aji *et al.*, 2021) est un corpus synthétique des paraphrases parallèles en langue générale dans 17 langues : arabe, catalan, tchèque, allemand, anglais, espagnol, estonien, français, hindi, indonésien, italien, néerlandais, roumain, russe, suédois, vietnamien et chinois. Leur méthode repose sur des données monolingues et un système de traduction automatique neuronale pour générer des paraphrases dans les autres langues. Pourtant, ce corpus est disponible que pour la langue générale, non pas pour le domaine spécialisé.

3.5 Corpus roumains

À l'heure actuelle, un corpus monolingue de reformulations ou paraphrases en roumain (qui n'est pas inclus dans une base multilingue) n'a pas encore été créé. Pourtant, des corpus de la langue générale en roumain existent, comme **Rombac** (Ion *et al.*, 2012), **CoRoLa** (Mititelu Barbu *et al.*, 2018) et le corpus plurilingue **CURLICAT** (Váradi *et al.*, 2022) (bulgare, croate, hongrois, polonais, roumain, slovaque, slovène) qui contient les données du corpus **CoRoLa**. Il existe également des corpus du domaine médical comme : **BioRo**, (Mitrofan et Tufis, 2018), corpus biomédical en roumain et le corpus médical annoté en

termes médicaux **MoNERo** (Mitrofan *et al.*, 2019). Pourtant, les textes intégraux de ces corpus ne sont pas disponibles en libre accès pour des raisons de droits d'auteur. **GrandMed-Ro** (Buhnila, 2018) reste le seul corpus roumain des textes scientifiques et de vulgarisation existant, créé par nous-mêmes dans la cadre de notre travail de recherche de master. Ce corpus servira comme point de départ pour construire un corpus de grande taille pour l'identification et la génération des reformulations sous-phrastiques médicales, mettant en évidence les correspondances entre les mots. Nous présentons notre méthodologie en détail dans le chapitre dédié.

***

Nous remarquons l'existence d'un grand nombre de corpus de reformulations ou paraphrases en langue anglaise et un nombre plutôt restreint pour le français, ainsi que pour le roumain. Notre travail de recherche complète ce manque de corpus de reformulations pour les langues moins dotées, telles que le roumain, en proposant des ressources linguistiques annotées pour le domaine médical. Notre but est d'adapter ces ressources sur la reformulation et la paraphrase et les rendre exploitables automatiquement pour la simplification lexicale automatique et la vulgarisation scientifique des recherches et découvertes dans le domaine de la médecine.

Afin de mieux définir le concept de reformulation, respectivement de paraphrase, la **Partie III** présente les caractéristiques de la **reformulation sous-phrastique médicale** dans notre projet de recherche, nos **corpus d'étude** et notre **méthodologie** de travail.

III. NOTRE APPROCHE, MÉTHODOLOGIE ET RESSOURCES

1. La définition de la reformulation sous-phrastique médicale

La reformulation représente le processus de réécriture qui a le rôle d'expliquer, simplifier ou pointer une phrase ou un syntagme. La reformulation est un processus linguistique de transformation du discours qui a été étudié par le prisme de plusieurs domaines présentés dans la **Partie I**, **sous-chapitre 1.** La reformulation indique un lien de synonymie entre le syntagme original (le référent, plus précisément l'élément linguistique source) et la reformulation dans le cadre de deux énoncés sémantiquement liés, selon Martinot (2003). Nous élargissons ce point de vue, suivant Inkova (2020), pour traiter aussi bien les **reformulations paraphrastiques et les reformulations non-paraphrastiques**, en écartant les reformulations répétitives.

Compte tenu de la grande complexité des définitions, classifications et approches sur la reformulation, l'identification automatique de celle-ci est une opération difficile. Il s'agit d'identifier la source (**le terme médical**), de délimiter et de classer les **reformulations** (de types variés), suivant les **marqueurs** spécifiques permettant de les introduire. Pour notre recherche, 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).

Chaque type de reformulation a son rôle dans un discours de vulgarisation scientifique adressé à un public cible. Nous choisissons les *reformulations paraphrastiques* parce que nous souhaitons identifier des données équivalentes au niveau sémantique. Nous prenons en compte uniquement les **reformulations sous-phrastiques**, plus précisément

les reformulations qui ne dépassent pas la longueur d'une phrase, de type « myotonie, *c'est-à-dire* une sensation de raideur musculaire », incluant également les **paraphrases lexicales** de type synonyme comme « glucose (sucre) ». La reformulation sous-phrastique est plus facilement identifiable de façon automatique à l'aide des **marqueurs de reformulation**, tels que « signifie », « est ce qu'on appelle », « est aussi appelé », « aussi appelé », « doit être compris comme », « au sens de », « autrement dit » en français, ou, en anglais, « such as » (Hearst, 1992), « known as » (Coates-Stephens, 1991). Les marqueurs en français sont identifiés par Vassiliadou (2013a) comme indicateurs de la *paraphrase d'équivalence de type définition*. Nous cherchons également les équivalents de traduction de ces marqueurs paraphrastiques en roumain de type « înseamnă » (*signifie / veut dire* en français), « adică » (*c'est-à-dire* en français) (Barbu Mititelu, 2011 ; Săpoiu, 2013 ; Buhnila, 2018) et autres que nous cherchons à l'aide des règles définies manuellement.

Notre thèse se focalisant sur des textes écrits spécialisés du domaine médical, nous ne traitons pas la reformulation orale avec tous ses spécificités et marqueurs discursifs typiques (de type « bref », « en fait », « en gros », « comme dit », etc.). Le « mieux dire » à l'écrit dépend également du public cible de la reformulation. Dans notre recherche sur la reformulation médicale, nous nous intéressons au rôle de simplification des notions médicales difficiles à comprendre par un public non-spécialiste, à l'écrit.

Nous nous intéressons aussi à la dimension *référentielle* de la reformulation, plus précisément à la référence des descriptions définies (Fuchs, 1982). Cette référence concerne un nom précédé par un déterminant défini ou démonstratif et suivi d'une qualification, si on reste dans la limite de la phrase. Pour pouvoir créer une reformulation descriptive, nous avons besoin de connaître la réalité extralinguistique du référent (nous notons que les méthodes en traitement automatique du langage rencontrent des difficultés à identifier la réalité extralinguistique). Nos recherches seront menées sur des reformulations qui se trouvent en **relation de type signe-signe** (Rey-Debove, 1978) et qui ont comme but la **détermination du sens** (Fuchs, 1982). La relation de type signe-signe fait référence au rapport d'inclusion ou d'égalité des deux signes linguistiques. Nous nous rapportons également à **l'approche logique** (Martin, 1976) pour déchiffrer le sens et les enjeux extralinguistiques de la reformulation analysée. Ces enjeux sont le type de texte source (dans notre étude, médical), l'objectif de la reformulation et le public cible. Dans notre travail, les sources (qui renvoient au référent extralinguistique) sont les **termes médicaux scientifiques** qui ont besoin d'être reformulés et simplifiés pour faciliter la compréhension du texte médical par le grand public.

Nous prenons en compte les **spécificités du texte médical et des termes médicaux**. L'objectif de la reformulation dans ce type de texte est celui de la **simplification des concepts médicaux en fonction du public cible** pour :

- présenter dans **un langage plus simple** les informations médicales (Grabar et Hamon, 2015 ; Cardon et Grabar, 2018, Koptient *et al*., 2019) ;
- faciliter **la communication** et la **compréhension** avec les patients (Pecout *et al.*, 2019) ;
- faciliter **la compréhension des maladies** et du **monde sanitaire** pour le grand public (Pecout *et al.*, 2019 ; Koptient et Grabar, 2020).

Dans un premier temps, nous identifions des termes médicaux qui sont en relation *paraphrastique* avec leur reformulations, mais nous ne limitions pas à cette catégorie de reformulations. 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,* de type *intratextuelles,* 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** organisées sous forme de corpus de reformulations pour la **simplification des notions médicales** (Cardon, 2021 ; Grabar et Hamon, 2015 ; Grabar et Hamon, 2016).

Nous adoptons le point de vue d'Inkova (2020), en soulignant le **caractère métalinguistique de la reformulation** et de réflexion sur le code de la langue, mais en rajoutant aussi la **réflexion sur le niveau de technicité de la langue** afin *d'adapter* le dit au public cible du texte écrit. Les différents niveaux de technicité d'un texte peuvent varier entre scientifique, académique, pédagogique, d'information, médiatique et de large diffusion. Pour notre recherche nous nous intéressons au **texte de type scientifique / académique** et au **texte de vulgarisation scientifique** écrit destiné à une large diffusion auprès du grand public. Nous travaillons donc sur les reformulations qui parlent du **métalangage médical** et qui mettent en **relation d'identification** des notions médicales scientifiques avec leurs équivalents dans la langue commune (donc **traduction intra-linguale,** mais entre des registres différents de la langue).

En conclusion, nous menons notre travail de recherche sur **la reformulation sous-phrastique d'identification métalinguistique** (Bouamor, 2012 ; Fuchs, 1982). Ce type de reformulation peut prendre plusieurs formes : **contexte définitoire**, **reformulation explicative**, **propositions équationnelles** (Rey-Debove, 1978) (des unités linguistiques

équivalentes qui combinent des expressions synonymiques dans une phrase équationnelle de type « x égal x » (Jakobson, 1963)).

Selon notre proposition, **la reformulation sous-phrastique médicale** représente ***l'équivalence au sens large, basée sur un noyau sémantique commun, qui contribue à la vulgarisation de termes médicaux et qui ne dépasse pas le cadre d'une phrase.***

2. Méthodologie

Notre objectif principal est de construire des **corpus de reformulations sous-phrastiques** pour réaliser la vulgarisation scientifique dans le domaine médical, disponibles en français et en roumain. Ces corpus de reformulations doivent contenir une collection de triplets :

- **le référent source** (dans notre cas le terme médical) ;
- **les marqueurs de reformulation** permettant d'introduire et de délimiter la reformulation ;
- **la reformulation** même ;
- éventuellement **la phrase complète**.

Pour la construction de nos corpus, nous adoptons **une approche à base des règles** adaptées à nos langues d'étude, le français et le roumain, qui vise à identifier les termes médicaux et les marques linguistiques de la reformulation. Pour ce faire, nous nous servons de terminologies médicales et de listes de marqueurs. Afin de définir des règles précises pour le système à base des règles, nous réalisons une analyse contrastive des données extraites semi-automatiquement avec des **scripts Perl** (termes médicaux simples et polylexicaux, marqueurs de reformulation, reformulations sous-phrastiques) pour présélectionner les exemples et les inclure dans nos corpus.

Ensuite, nous utilisons ces corpus aussi bien pour la génération et l'identification automatique de la reformulation en utilisant **l'apprentissage automatique par réseaux de neurones**. D'une part, la génération permet d'ajouter des données au corpus de reformulation créé. D'autre part, l'identification vise à reconnaître les reformulations valides pour un terme source.

Pour ce faire, nous travaillons sur des corpus de **textes de la littérature médicale scientifique et des textes médicaux destinés au grand public** (en français : **CLEAR** (Grabar et Cardon, 2018) et **ClassYN** (Todirascu *et al.*, 2012), en roumain : **GrandMed-Ro** (Buhnila, 2018)) et **un nouveau corpus** créé par nous avec l'outil Sketch Engine (Kilgarriff *et al.*, 2014) pour le roumain. Notre méthode consiste d'abord à identifier automatiquement des termes médicaux simples et polylexicaux avec, pour le français, l'annotateur **SIFR-BioPortal** (Tchechmedjiev *et al.*, 2018). L'outil cherche de façon automatique les termes médicaux dans notre corpus à partir des terminologies médicales attestées telles que

SNOMED International (Côté, 1996) (diffusée par ASIP Santé) qui contient 150 906 concepts médicaux. Pour le roumain, nous utilisons **le corpus annoté MoNERo** (Mitrofan *et al.*, 2019) pour extraire une liste de termes médicaux (annotés avec les catégories d'UMLS (Bodenreider, 2004)) et nous cherchons à identifier cette liste de termes dans notre propre corpus.

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 de reformulation** décrits dans la littérature (Vassiliadou, 2013a ; Steuckardt, 2018 ; Magri, 2018) et identifiés selon nos observations en corpus. Nous évaluons manuellement les phrases qui contiennent les termes médicaux et respectivement les marqueurs pour identifier les reformulations correctes. Nous annotons les relations lexicales entre les reformulations correctes et les termes reformulés et les fonctions sémantico-pragmatiques des reformulations. À partir de ces résultats, nous constituons un **corpus de reformulations médicales** pour chaque langue. Nous exploitons nos corpus annotés avec des expériences de **reconnaissance et génération automatique de la reformulation** avec des architectures à base de **réseaux de neurones**.

Pour évaluer les résultats de l'annotation manuelle et de la génération automatique, nous analysons les **niveaux de difficulté et de compréhension** des **reformulations sous-phrastiques** pour déterminer leur impact sur **la simplification automatique** du texte scientifique ou de vulgarisation du domaine médical. Ces recherches peuvent servir à la génération automatique des simplifications pour un **public cible**. Informer correctement les patients, ou tout simplement le grand public, représente une nécessité pour appuyer les traitements médicaux. Dans cette perspective, la reformulation joue un rôle important vu son utilisation fréquente pour la définition et l'explicitation des notions médicales.

Pour résumer notre méthodologie, nous proposons une **méthode d'identification automatique des reformulations médicales** à partir des corpus monolingues comparables de type langage scientifique et de vulgarisation scientifique, en utilisant comme indices **les termes médicaux, les marqueurs de reformulations et leur contexte dans la phrase**.

Notre méthodologie (illustrée également dans la **Figure 8**) suit les huit étapes décrites ci-dessous :

1. Annotation automatique des termes médicaux. Pour le français, on applique les terminologies médicales SNOMED International (Côté *et al.*, 1993), SNOMED-

3.5VF (Côté, 1998), diffusées par ASIP Santé[36]. Nous adaptons le script en langage Perl[37] pour envoyer nos textes vers le service API REST de l'outil d'extraction de termes médicaux **SIFR-BioPortal** (Tchechmedjiev *et al.*, 2018). Pour le roumain, nous utilisons la liste de termes de MoNERo (Mitrofan *et al.*, 2019), pour détecter les termes médicaux simples et polylexicaux et nous sélectionnons les phrases contenant ces termes attestés ;

2. Sélection des phrases contenant également des marqueurs de reformulations étudiés dans la littérature (Fuchs, 1994 ; Grabar et Eshkol-Taravella, 2016a ; Antoine et Grabar, 2016) après vérification avec TXM sur un corpus français (pour experts et pour grand public) ;
3. Annotation manuelle par 2 annotateurs, évaluation et validation des phrases qui contiennent les termes médicaux et les marqueurs et indicateurs identifiés afin d'y trouver des reformulations médicales ;
   a. Calcul de l'accord inter-annotateur Kappa (Cohen, 1960) ;
   b. Analyse des phrases sélectionnées pour la découverte d'autres marqueurs de reformulation ;
   c. Élargissement de la liste de marqueurs de reformulation et traduction du français vers le roumain ;
   d. Application de la liste élargie de marqueurs de reformulation sur les corpus français et le corpus roumain ;
4. Classification manuelle et analyse des reformulations validées selon les relations lexicales et les fonctions sémantico-pragmatiques (Condamines, 2018 ; Săpoiu, 2013 ; Grabar et Eshkol-Taravella, 2016b) ;
5. Construction des corpus de reformulations médicales correctes en français et en roumain (jeu de données qui servira également à l'entraînement des outils de réseaux de neurones) ;
6. **a**. Expériences de génération des reformulations avec l'architecture à base de réseaux de neurones APT (*Adversarial Paraphrasing Task*) (Nighojkar et Licato, 2021). Cette étape permettra d'enrichir éventuellement le corpus de reformulations ;

   b. Expériences d'identification et de classification des reformulations avec l'architecture LSTM (*Long Short Term Memory*) (Hochreiter et Schmidhuber,

[36] https://esante.gouv.fr/terminologie-snomed-35vf
[37] Script proposé par Paul R Alexander sur la plateforme de partage de code Github et adapté par nous (https://github.com/ncbo/ncbo_rest_sample_code/blob/master/perl/annotate_text.pl)

1997). Cette expérience permet d'utiliser le corpus de reformulation pour identifier des reformulations ;

7. Évaluation des résultats par des mesures statistiques quantitatives comme la précision, le rappel, la F-mesure, le score Kappa (Cohen, 1960) et une interprétation des résultats par des analyses qualitatives ;
8. Évaluation du niveau de lisibilité des reformulations validées manuellement.

Ainsi, nous constituons des corpus de reformulations médicales qui pourront servir comme ressources textuelles exploitables pour la simplification automatique de textes, la vulgarisation scientifique, la génération automatique de textes pour informer de façon ciblée et adaptée le grand public sur des questions médicales et de santé.

Par la suite, nous présentons les corpus de textes exploités pour les deux langues de notre étude, le français et le roumain, et puis nous présentons en détail chaque étape de notre méthodologie.

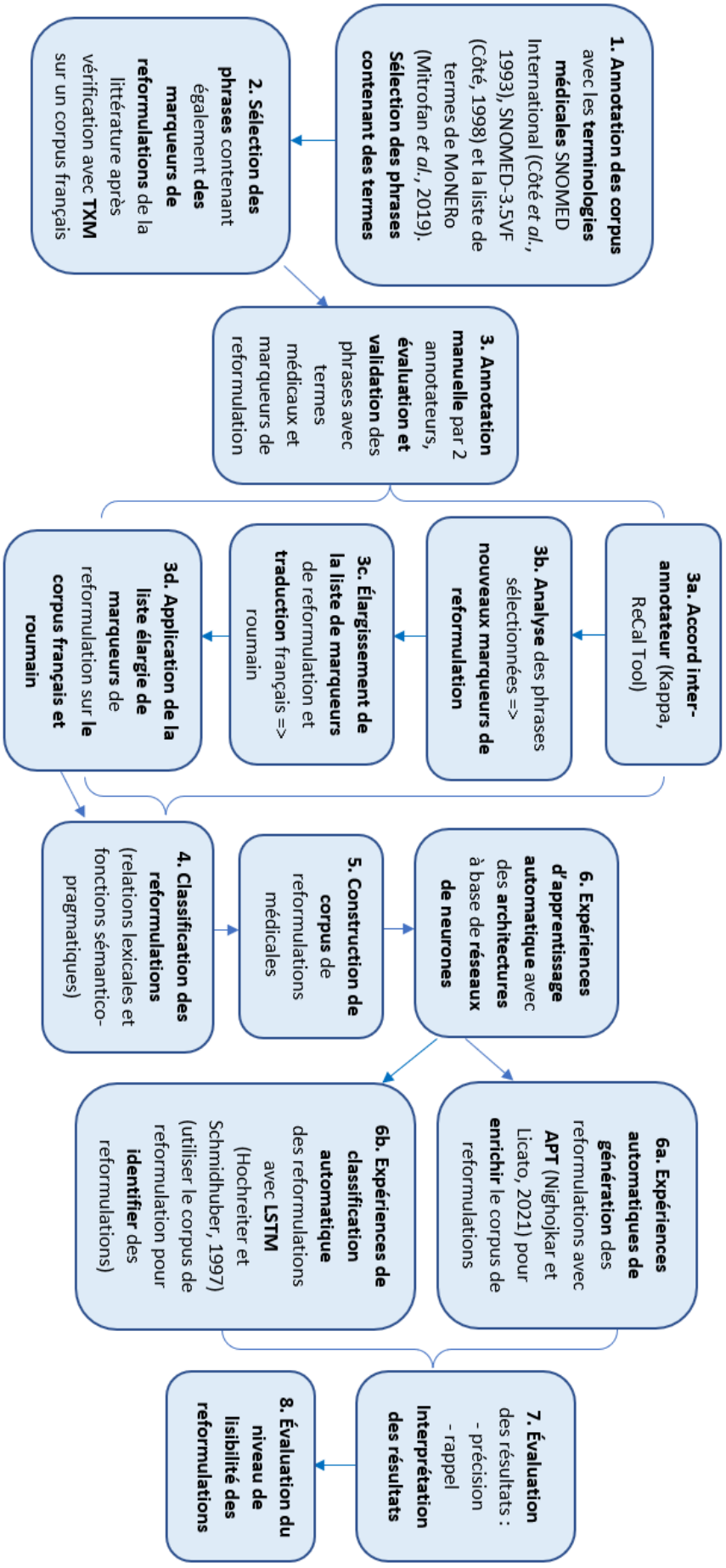

Figure 8. La méthodologie de notre travail de recherche

2.1 Corpus d'étude et collecte de données

Notre objectif est de constituer des corpus annotés de reformulations sous-phrastiques des termes médicaux en français et roumain, exploitables en TAL. Notre travail complète le nombre réduit de ressources monolingues pour le français et le roumain représentant des termes et leurs reformulations. De plus, nous proposons des méthodes d'extraction automatique de la reformulation pour les langues de moindre diffusion comme le roumain, aspect peu abordé en TAL, car la majorité des études se concentrent sur l'anglais. Dans ce sens nous souhaitons créer un modèle linguistique applicable à la langue roumaine et proposer de nouvelles règles et ressources pour le français. Ce modèle sera exploitable pour l'identification automatique de reformulations sous-phrastiques dans des textes du domaine médical.

2.1.1 Corpus français

Pour la langue française, nous utilisons la partie française du corpus déjà constitué dans le cadre du projet **ClassYN** (Todirascu *et al.*, 2012). Ce corpus comparable comprend des textes scientifiques et des textes de vulgarisation scientifique du domaine de la médecine et de l'informatique, en français et en allemand. L'objectif du projet ClassYN est de réaliser une classification automatique des textes selon leur genre textuel et le public visé (texte scientifique et texte de vulgarisation), à l'aide de propriétés syntaxiques identifiées automatiquement avec le parseur MATE (Bohnet, 2009). En ce qui concerne le corpus français du projet ClassYN (Todirascu *et al.*, 2012), le travail de nettoyage et de prétraitement du corpus a déjà été fait. Les textes ont été anonymisés manuellement et les informations non-pertinentes ont été éliminées. Les textes se présentent sur la forme de 300 paires de textes de type articles scientifiques (environ 1 million de tokens) et 300 textes de vulgarisation (environ 0,7 million de tokens) pour la partie française. Nous avons déjà travaillé avec le corpus ClassYN dans le cadre de notre travail de recherche de master (Buhnila, 2018).

Corpus ClassYN	Taille (tokens)	Taille (textes)
Scientifique	1 007 049	300
Grand public	772 374	300
Total	1 779 423	600

Tableau 4. Taille des corpus ClassYN : corpus de littérature scientifique médicale et corpus grand public

***

Nous utilisons également le corpus **CLEAR** (Grabar et Cardon, 2018), corpus comparable du domaine médical avec des textes scientifiques et des textes simplifiés. Le corpus contient des textes provenant de trois sources : encyclopédies (Wikipedia et Vikidia, sa version simplifiée adaptée aux enfants), notices de médicaments (rendus publiques par le Ministère Français de la Santé[38]) et résumés scientifiques et grand public (en collaboration avec Cochrane[39]). Le corpus CLEAR est libre d'accès pour la recherche et téléchargeable en ligne[40].

CLEAR Encyclopédie	**Articles en commun (même titre)**	**Taille (tokens)**
Wikipédia	575	2 293 078
Vikidia		197 672

Tableau 5. Taille des corpus : CLEAR à partir des encyclopédies Wikipédia et Vikidia

CLEAR Notices médicaments	**Notices en total**	**Taille (tokens)**
Scientifique	11 800	52 313 126
Grand public		33 682 889

Tableau 6. Taille des corpus : CLEAR notices des médicaments

CLEAR Cochrane	**Résumés en total**	**Résumés en commun** (même sujet)	**Taille (tokens)**
Scientifique	8 789	3 815	2 840 003
Grand public			1 515 051
Total			**4 355 054**

Tableau 7. Taille du corpus CLEAR Cochrane par type de texte (Grabar et Cardon 2018)

Parmi les trois sous-corpus, nous avons choisi **CLEAR Cochrane** pour nos expériences parce que les résumés contiennent des informations précises sur des notions médicales. Notre hypothèse est que ces textes contiennent un nombre considérable de reformulations médicales des termes et maladies présentées. La disposition du corpus Cochrane en résumés destinés à un public expert et des résumés destinés à un public non-

[38] http://base-donnees-publique.medicaments.gouv.fr/
[39] http://www.cochranelibrary.com/
[40] http://natalia. grabar.free.fr/resources.php

spécialiste nous intéresse afin de comparer les reformulations dans ces deux types de textes.

2.1.1.1 Prétraitement des corpus français

Avant de mener nos expériences sur les corpus d'étude, les textes doivent être anonymisés et uniformisés le plus possible pour avoir une taille comparable pour les deux langues. Avant le prétraitement, les corpus français présentent les caractéristiques suivantes :

- **ClassYN** (Todirascu *et al.*, 2012) : Le corpus a été anonymisé manuellement et classé en textes médicaux scientifiques et de vulgarisation, ce qui nous est très utile pour réaliser nos analyses des reformulations médicales dans les deux types de corpus ;
- **CLEAR Cochrane** (Grabar et Cardon, 2018) : Le corpus est également anonymisé manuellement. Nous réaliserons des analyses également par type de corpus, scientifique et de vulgarisation médicale.

Nous avons suivi la même procédure de prétraitement pour les deux corpus français. Nous avons découpé les textes en phrases à l'aide des caractères de fin de ligne (. ; ! ; ?) et nous avons nettoyé certains caractères spéciaux (« % », « # », « TAB »), car ils déclenchent des erreurs au lancement de l'annotation automatique des termes médicaux avec l'outil SIFR-BioPortal (Tchechmedjiev *et al.*, 2018), outil que nous présentons dans le **sous-chapitre 2.2**.

2.1.2 Corpus roumain

Pour le roumain nous partons de **l'unique corpus roumain des textes scientifiques et de vulgarisation** disponible, à notre connaissance, qui a été créé dans la cadre de notre travail de recherche de master, le corpus **GrandMed-Ro** (Buhnila, 2018). Nous agrandissons ce corpus à travers l'outil de génération des corpus **Sketch Engine** (Kilgarriff *et al.*, 2014).

Ce corpus (Buhnila, 2018) a été construit sur le modèle et la structure du corpus ClassYN (Todirascu *et al.*, 2012). Le corpus français de ClassYN nous a servi comme guide pour le choix des sources de textes médicaux (journaux scientifiques en ligne, revues médicales, sites médicaux pour le grand public), de types de textes (scientifique et de

vulgarisation) et du genre des textes (article scientifique, pages web à destination d'un public non-expert). **GrandMed-Ro** est réalisé à partir des sources en lignes. Les textes scientifiques médicaux ont été extraits à partir de sites web suivants : « Jurnalul de chirurgie »[41] (*Journal de chirurgie*), « Baza națională de date de cercetare (BNC) »[42] (*La base nationale des données de recherche*), « EMCB Educație medicală continuă »[43] (*Education médicale continue*). Les textes médicaux de vulgarisation sont issus de sites suivants : « SfatulMedicului.ro »[44] (*Le conseil du médecin*), « Doctorul zilei »[45] (*Le médecin du jour*), « Jurnal Medical »[46] (*Journal Médical*). La justification pour les choix des sources de textes scientifiques réside dans la qualité des articles et la possibilité de les télécharger en roumain, vu la tendance actuelle de publier les articles de recherche roumains directement en anglais. Le corpus scientifique est composé de 70 articles et 22 102 tokens et le corpus de textes destinés au grand public contient 261 textes avec 20 038 tokens.

Corpus GrandMed-Ro	Taille (textes)	Taille (tokens)	Taille (caractères)
Articles scientifiques	70	22 102	190 233
Textes de vulgarisation	261	20 038	190 377
Total	331	42 140	380 610

Tableau 8. Taille des corpus : GrandMed-Ro, par type de texte, scientifique et de vulgarisation

[41] Version utilisée : http://jurnaluldechirurgie.ro/desprenoi.htm (Buhnila, 2018). En 2020, une nouvelle version du site a été créée : http://jurnaluldechirurgie.ro/jurnalnou2020/
« Jurnalul de Chirurgie » est un journal universitaire publié par « l'Université de médecine et de pharmacie » de Iași, Roumanie, et représente la publication officielle du « Centre de recherche de chirurgie ouverte et laparoscopique ». Ce site accueille le premier journal chirurgical en ligne de la Roumanie (avec des publications scientifiques numérisées depuis 2005). Le site a été consulté du mois de mars à juin 2018.
[42] http://bnc.cercetaremedicala.ro/
Le site est géré par « l'Institut national des maladies infectieuses "Prof. Dr. Matei Balş" » de Bucarest, Roumanie. Le site a été consulté du mois de mars à juin 2018.
[43] https://www.emcb.ro/
Le site accueille une publication périodique en ligne du Collège des Médecins de Bucarest en collaboration avec l'Université de médecine et de pharmacie « Carol Davila » de Bucarest, Roumanie. Le site a été consulté du mois de mars à juin 2018.
[44] http://www.sfatulmedicului.ro/
Le site a été créé en 2005 et il a le rôle d'informer le grand public sur les plus connus problèmes de santé et présenter les actualités et les découvertes dans la recherche médicale dans un langage simplifié. Le site a été consulté du mois de mars à juin 2018.
[45] https://www.doctorulzilei.ro/
Le site existe depuis 2008 et il appartient à l'entreprise WHITE SPOT PRODUCTION S.R.L., située à Bucarest, Roumanie. Il propose des articles sur plusieurs thèmes : actualités, hygiène de vie, santé, médecine alternative, intimité, entretiens avec des spécialistes. Le site a été consulté du mois de mars à juin 2018.
[46] Le site exploité en 2018 (mars-juin) n'est plus accessible sur la toile ou il a changé de nom de domaine.

Afin d'avoir des résultats comparables pour nos deux langues d'étude, nous agrandissons le corpus **GrandMed-Ro** avec l'outil de génération automatique des corpus **Sketch Engine** (Kilgarriff *et al.*, 2014). Nous présentons le travail de collecte de corpus dans la section suivante.

2.1.3 Collecte du corpus roumain avec Sketch Engine

Nous utilisons la plateforme en ligne **Sketch Engine**[47] (Kilgarriff *et al.*, 2014) pour récupérer des textes médicaux de la toile et en créer des corpus médicaux en roumain. Nous choisissons la méthode de création de corpus à partir des sites web. L'accès institutionnel dont nous disposons nous permet d'extraire jusqu'à 2 000 articles et 1 million de tokens depuis un site web[48]. Nous nous intéressons aux textes de vulgarisation médicale, car la grande variété des sujets médicaux présentés dans un langage accessible représente un avantage dans la recherche des paraphrases médicales.

Lors de recherches précédentes (Buhnila, 2018), nous avons rencontré de grandes difficultés à trouver des textes ou articles scientifiques du domaine médical en roumain, en libre accès. Ceci est dû à la tendance de plus en plus conséquente des chercheurs roumains à publier leur recherche en anglais directement, pour que les articles puissent être consultés par des lecteurs internationaux. Comme nous avons besoin d'une grande quantité de textes pour nos expériences d'apprentissage automatique par réseaux de neurones, nous avons décidé de travailler uniquement sur des articles de vulgarisation pour la langue roumaine.

Nous avons choisi huit sites différents de vulgarisation médicale qui contiennent un grand nombre de textes et qui permettent l'extraction automatique de données (certains sites bloquent les outils qui font du *web scraping*[49]). Les informations sur les sites d'origine, les tailles et la date d'extraction des corpus se trouvent dans le **Tableau 9**. Nous agrandissons le corpus **GrandMed-Ro** avec **7 141 articles** qui ont une taille de **6 398 811 tokens**. Le corpus **GrandMed-Ro2** agrandi avec Sketch Engine est composé de **7 472 articles** et a une taille totale de **6 440 951 tokens**.

[47] https://www.sketchengine.eu/

[48] La limite de 2 000 articles par corpus est imposée par la plateforme Sketch Engine pour les utilisateurs qui ont un accès gratuit à travers une connexion institutionnelle (dans notre cas, à travers le partenariat entre les créateurs de la plateforme et l'Université de Strasbourg).

[49] Le *web scraping* est une technique d'extraction automatique de contenu des sites web, à l'aide d'un script, un outil ou une plateforme en ligne.

Corpus roumain	Site web	N° Articles	Taille corpus (tokens)	Date d'accès
Sites de vulgarisation médicale				
sfaturi medicale	https://sfaturimedicale.ro/	576	989 700	12/09/2021
sfatul medicului	https://www.sfatulmedicului.ro/	919	970 452	19/03/2022
doctorul zilei	https://www.doctorulzilei.ro/	1851	1 025 308	21/03/2022
romedic	https://www.romedic.ro/	769	1 027 834	25/03/2022
csid-boli	https://www.csid.ro/boli-afectiuni/	563	543 046	26/03/2022
csid-sanatate-health	https://www.csid.ro/sanatate/	1520	1 025 215	28/03/2022
Sites de cliniques privées				
cdt-babes	https://www.cdt-babes.ro/	280	198 052	26/03/2022
regina-maria	https://www.reginamaria.ro/medici	690	619 204	01/04/2022
Taille du corpus roumain extrait avec Sketch Engine		**7 141**	**6 398 811**	
Taille du corpus GrandMed-Ro original		**331**	**42 140**	01/05/2018
Taille du corpus GrandMed-Ro2		**7 472**	**6 440 951**	

Tableau 9. Les sous-corpus roumains extraits avec Sketch Engine et la taille finale du corpus GrandMed-Ro2

2.1.3.1 Prétraitement du corpus roumain

Les corpus sont constitués d'un nombre maximal de 2 000 articles téléchargés par site. Les corpus téléchargés depuis la plateforme Sketch Engine (Kilgarriff *et al.*, 2014) contiennent des métadonnées et des balises HTML et XML[50] de type **<doc>** (document) et ** ** (paragraphe) :

> *<doc id="file23932249" filename="10-boli-comune-si-tratamentul-acestora_19132" parent_folder="SfatMed" url="https://www.sfatulmedicului.ro/Educatie-pentru-sanatate/10-boli-comune-si-tratamentul-acestora_19132">*
>
> * 10 boli comune si tratamentul acestora […] </doc>*

Nous supprimons ces informations pour anonymiser les articles et réduire le bruit des résultats lors de l'exploitation automatique des corpus. Pour effacer automatiquement toutes les balises qui contiennent les métadonnées des articles, nous utilisons **l'expression régulière** *<doc id=.* ?>*. Pourtant, nous gardons également un exemplaire des versions

[50] HTML et XML sont des langages de balises qui permettent de structurer le contenu de textes sur les pages Web. HTML contient des balises prédéfinies, tandis que le langage XML permet de définir ses propres balises (par exemple, la balise <doc> est une balise XML et une balise HTML.

originales avec toutes les métadonnées. Par la suite, nous alignons les textes du corpus roumain une phrase par ligne[51] en utilisant les caractères de fin de phrase comme repère (point de fin de phrase, signe d'exclamation, signe d'interrogation).

***

Après avoir présenté les corpus français et roumain, nous présentons les étapes de notre méthodologie, les outils et les architectures choisis.

2.2 Annotation automatique des termes médicaux

L'annotation est le processus de balisage des textes (manuelle ou automatique) afin de délimiter les unités morphologiques, lexicales, syntaxiques et sémantiques de la phrase en fonction des éléments linguistiques recherchés. Par la suite, ces balises pourront être traitées de façon automatique par des scripts ou des outils d'apprentissage automatique par réseaux de neurones, méthodes que nous développons dans la suite de notre étude.

L'annotation de termes médicaux est un travail réalisé **automatiquement avec des outils** qui projettent une liste de termes sur un corpus. Nous vérifions la qualité des annotations de termes médicaux semi-automatiquement et manuellement pour une plus fine précision (pour enlever les mots de la langue courante et pour mieux identifier les termes médicaux polylexicaux). Pour identifier les termes médicaux présents dans nos corpus, nous utilisons des **terminologies médicales** en libre accès pour le français et une terminologie médicale composée par nous-mêmes pour le roumain, à partir d'un corpus annoté manuellement en termes par un expert.

L'annotation de termes est une étape cruciale dans notre travail. En particulier, nous cherchons des outils qui identifient les termes et les entités spécifiques au domaine médical, tel que **Genia Tagger** (Kim *et al.*, 2003), disponible que pour la langue anglaise. Plusieurs outils d'annotation morphosyntaxique et syntaxique sont disponibles pour le français, tel que **Stanford NLP** (Manning *et al.*, 2014) (qui annote seulement les noms de personnes, de lieu ou d'organisation, selon les modèles de langues créés sur des corpus de langue générale). Certains outils sont disponibles également pour le roumain : **TreeTagger** (Schmid, 1994),

[51] Nous avons séparé les phrases une par ligne dans l'éditeur de texte Notepad++ à l'aide des expressions régulières : nous avons remplacé «. » , « ! » et « ? » avec l'expression régulière correspondante «.\r\n » « ?\r\n » et « !\r\n », qui rajoute un retour à la ligne à la fin de chaque phrase du corpus.

UDPipe[52] (Straka *et al.*, 2016), **NLP-Cube**[53] (Boroș *et al.*, 2018), mais l'annotation en entités nommées ou termes n'est pas toujours disponible. Même si ces outils donnent de bons résultats sur l'annotation morphosyntaxique et syntaxique, la majorité ne donne pas des résultats directement exploitables pour la détection de termes médicaux, ce qui est une étape essentielle dans notre méthodologie.

Il existe des outils spécialisés pour la détection des termes, tel que **TermoStat**[54] (Drouin, 2003), ou des expressions polylexicales en français, tel que **LGTagger**[55] (Constant et Sigogne, 2011), mais ces outils ne sont pas adaptés pour le roumain. De plus, l'extraction des expressions polylexicales n'est pas directement exploitable pour la terminologie, sans validation manuelle. **NLP-Cube** ou **ATILF-LLF**[56] disposent des ressources pour le roumain, mais les résultats restent modestes, selon les études publiées (Hazem *et al.*, 2017). Pour la langue roumaine, nous avons testé **NLP-Cube** avec le service de prétraitement de texte **TEPROLIN** (Ion, 2018), qui se trouve sur la plateforme en ligne **RELATE** (Păiș *et al.*, 2019). Pourtant, nous avons rencontré des difficultés techniques d'importation d'un grand corpus de texte et d'exportation des textes annotés. Malgré les performances des outils d'extraction terminologiques ayant fait leurs preuves pour le français, les résultats obtenus peuvent contenir des erreurs. De plus, nous n'avons pas d'outil d'extraction terminologique ayant des performances similaires en roumain. Notre choix a été alors d'appliquer plutôt des ressources terminologiques validées par une vérification manuelle par les spécialistes du domaine pour l'identification des termes susceptibles d'être reformulés.

Dans notre recherche de termes médicaux, nous devons tenir compte également des particularités de chaque langue. Le français est une langue romane dont les noms varient en genre et en nombre, tandis que le roumain est une langue romane **flexionnelle**, qui présente en plus des *cas (nominatif, génitif, datif, accusatif, vocatif)*. Le roumain est la seule langue latine qui a gardé les *articles définis enclitiques* (collés à la fin du mot), qui varient en fonction du genre et du nombre. Certains noms masculins au singulier peuvent avoir une forme féminine au pluriel. Nous illustrons ces cas de figure par les exemples

[52] http://lindat.mff.cuni.cz/services/udpipe/
[53] Outil pour le découpage des phrases, tokenisation, lemmatisation, parties du discours, analyse des dépendances et reconnaissance des entités nommées, disponible pour plus de 50 langues.
[54] **TermoStat** est un outil d'identification automatique de termes par la mise en opposition de corpus spécialisés et non-spécialisés. Les langues disponibles pour TermoStat sont le français, l'anglais, l'espagnol, l'italien et le portugais. (http://termostat.ling.umontreal.ca/)
[55] **LGTagger** est un outil qui identifie les expréssions polylexicales, mais il n'est adapté que pour le français.
[56] **ATILF-LLF** est un outil pour identifier les expréssions polylexicales dévéloppé dans le cadre du projet PARSEME et adapté pour 18 langues : bulgare, espagnol, hébreu, lituanien, tchèque, maltais, roumain, slovène, allemand, grec, français, hongrois, italien, polonais, portugais, suèdois, turc, persan.

suivants pour le nom masculin **neuron** *(neurone)* et le nom neutre **calculator** *(ordinateur)*. Pour le dernier nom, le pluriel applique les articles adaptés pour le féminin :

- **neuron**, **neuronul** (nominatif, singulier, masculin) => **al neuronului / neuronului** (genitif / datif, singulier, masculin) (*neurone, le neurone => du neurone*)
- **neuronii** (nominatif, pluriel, masculin) => **ai neuronilor / neuronilor** (génitif /datif, pluriel, masculin) (*neurones, les neurones => des neurones*)
- **calculator, calculatorul** (nominatif, singulier, neutre) => **al calculatorului / calculatorului** (génitif /datif, singulier, neutre) (*ordinateur, l'ordinateur => de l'ordinateur*)
- **calculatoare, calculatoarele** (nominatif, pluriel, neutre) => **ale calculatoarelor / calculatoarelor** (génitif / datif, pluriel, neutre)

Pour répondre à ces difficultés, nous avons recherché les **termes obtenus par flexion** dans les deux langues (« pericardita constrictivă » / *péricardite constrictive ;* « antibiotiques ») et non pas les lemmes des termes (« pericardită constrictiv » / *péricardite constrictive* ; « antibiotique »), car notre objectif était de trouver des termes polylexicaux dans le discours et les organiser sous forme de corpus. Nous avons marqué les termes médicaux identifiés avec des balises XML. Cette annotation nous permet d'identifier si les termes annotés sont reformulés dans les textes. Dans notre analyse manuelle des phrases, nous marquerons tous **les termes médicaux qui sont reformulés** (mais pas tous les termes médicaux, car le but de notre étude est d'identifier uniquement les termes médicaux reformulés).

***

Les outils testés n'ont pas de performances similaires pour la détection de termes. Ainsi, nous avons développé nos propres **scripts en langage Perl** pour identifier les **termes médicaux en roumain**. Nous n'avons pas trouvé des **terminologies médicales numérisées en roumain** qui soient exploitables automatiquement, mais nous avons proposé une solution alternative, à l'aide d'une **liste de termes** validées manuellement.

Par la suite, nous présentons en détail l'annotateur automatique utilisé pour le français et la méthode appliquée pour l'annotation des termes en roumain.

2.2.1 Annotateur automatique de termes pour le français

Pour identifier les termes dans le corpus français, nous utilisons l'annotateur **SIFR-BioPortal** (*Semantic Indexing of French Biomedical Data Resources*, en français Indexation Sémantique des Ressources Françaises de Données Biomédicales) (Tchechmedjiev *et al.*,

2018). Cet outil permet d'annoter des termes scientifiques du domaine médical utilisant 37 bases terminologiques médicales pour le français[57]. Parmi ces bases, nous avons choisi les deux terminologies médicales les plus importantes, la **SNOMED International** (Côté, 1996) et la **SNOMED-3.5 VF** (Côté, 1998) afin d'identifier le plus grand nombre de termes médicaux en les cherchant dans nos textes avec l'annotateur **SIFR-BioPortal**[58].

Le Projet **SIFR** (Tchechmedjiev *et al.*, 2018) a permis la création de l'annotateur **SIFR-BioPortal** (*Semantic Indexing of French Biomedical Data Resources*) (Tchechmedjiev *et al.*, 2018) sous la forme d'une plateforme en ligne[59] facilement utilisable par un public large. Pour les spécialistes et chercheurs en TAL, le projet SIFR a créé une version utilisant API-Key avec un service web de type REST[60] pour interroger un serveur à distance. Nous avons utilisé cette option pour nos expériences d'annotation automatique de termes médicaux.

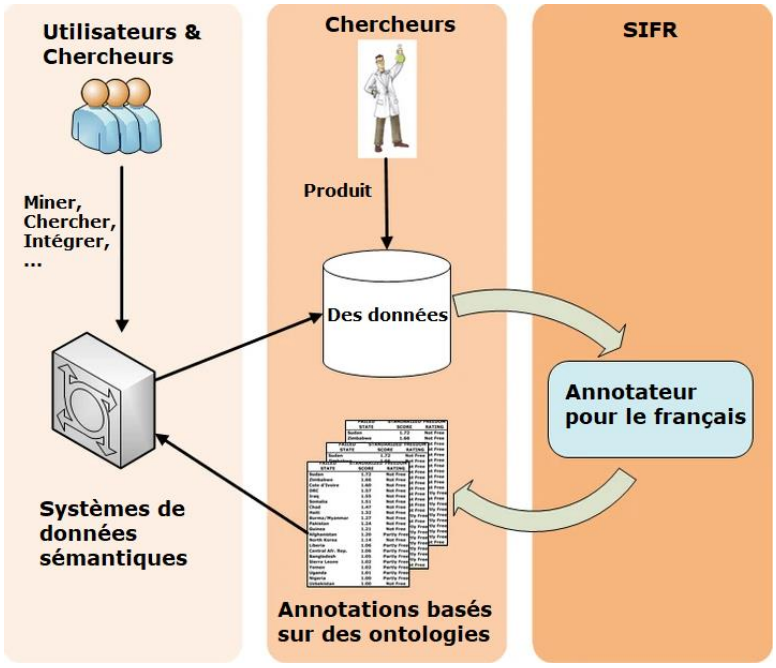

Figure 9. Fonctionnement de l'annotateur SIFR-BioPortal

Le fonctionnement de l'annotateur est expliqué dans la **Figure 9** ci-dessus. L'outil nous permet d'annoter la forme exacte du terme (non lemmatisée) dans le corpus. La configuration de l'outil est nécessaire pour obtenir des résultats d'annotation précis et évaluables, notamment la sélection de la terminologie. Nous présentons les deux

[57] Nous avons testé également d'autres annotateurs pour le français, Bio-YODIE (Gorrell *et al.*, 2018) PyMedTermino (Lamy *et al.*, 2015), mais SIFR-BioPortal est resté le plus intuitif d'utilisation.
[58] http://bioportal.lirmm.fr/annotator
[59] http://bioportal.lirmm.fr/annotator
[60] Service web REST disponible : http://data.bioportal.lirmm.fr/

terminologies sélectionnées pour l'annotation, **SNOMED-3.5VF** et **SNOMED International** (Côté, 1998 ; 1996), dans le sous-chapitre suivant.

2.2.1.1 Ressources terminologiques en français

La terminologie médicale **UMLS** (*Unified Medical Language System*, en français Système de Langage Médical Unifié) (Bodenreider, 2004) est un ensemble de fichiers et de logiciels et contient de nombreux vocabulaires, normes sanitaires et biomédicales avec le but de permettre l'interopérabilité entre les systèmes informatiques divers. Le **Métathésaurus UMLS version 2021AA** contient environ 16 millions de noms de concepts uniques et 4,4 millions de concepts provenant de 218 vocabulaires médicaux. Même si cette terminologie est en anglais, les classes et catégories proposées sont utilisées également dans d'autres langues. C'est le cas du corpus roumain annoté en entités médicales **MoNERo** (Mitrofan *et al.*, 2019), que nous présentons dans la section suivante.

La terminologie **SNOMED-3.5VF**[61] (*Systematized NOMenclature of MEDicine*, en français Nomenclature Systématisée de la Médecine, version française 3.5) contient **150 906 concepts médicaux** d'une grande variété de sous-domaines : administratif médical et traitements, agents, anatomie, diagnostics, organismes vivants, médicaments, symptômes, anatomie, maladie, procédures, substances, etc. Cette variété des sous-domaines médicaux nous permet de trouver une gamme large de termes médicaux dans nos corpus.

La terminologie **SNOMED International** (Côté, 1996) est une nomenclature médicale qui couvre tous les champs de la médecine, de la dentisterie humaine et de la médecine animale. Pour nos annotations, nous utilisons la plus récente de ses mises à jour, la **SNOMED CT** (*Clinical Terms*, en français, Termes Cliniques). Il y a des différences notables entre les mises à jour de la SNOMED-3.5VF et la SNOMED International CT. Cette dernière contient, en plus du SNOMED-3.5VF, des termes médicaux de type observations, produits de santé, thérapies, facteurs sociaux, agents pathogènes, dispositifs, événements, scores. Nous utilisons les deux terminologies pour couvrir le plus grand nombre de termes médicaux dans notre corpus d'étude.

La terminologie SNOMED International permet d'identifier des **synonymes** et la position du terme trouvé dans la phrase sans sa forme flexionnelle (le terme n'est pas lemmatisé). Dans l'exemple ci-dessous extrait du corpus français ClassYN on trouve le

[61] https://smt.esante.gouv.fr/terminologie-snomed-35vf/

terme identifié dans le texte en majuscules (HYPOXIE) et l'annotation SYN nous permet d'accéder le synonyme de ce terme trouvé par l'annotateur dans la terminologie (« diminution de l'apport d'oxygène »). Ces synonymes peuvent constituer également des reformulations.

SNOMED International CT

- diminution de l'apport d'oxygène SYN HYPOXIE 260 266

Nous faisons l'**hypothèse** de trouver des reformulations dans le contexte ***d'avant*** **ou** ***d'après*** le terme médical, dans la même phrase, ce qui nous laisse plus de **flexibilité** pour identifier un plus grand nombre des reformulations.

2.2.2 Annotation de termes pour le roumain

Un outil de traitement disponible pour le roumain est le service web **TEPROLIN** (Ion, 2018), disponible sur la plateforme en ligne **RELATE** (Păiș *et al.*, 2019). En plus des opérations de lemmatisation et d'annotation (morphologique, syntaxique, etc.), **TEPROLIN** contient un module d'identification d'entités nommées médicales, le module **BioNER**[62] (Mitrofan *et al.*, 2018 ; Boroș *et al.*, 2018) qui annote automatiquement des notions médicales en roumain, comme dans l'exemple ci-dessous :

> **Diabetul zaharat (DISO)** este un **sindrom (DISO)** caracterizat prin valori crescute ale **concentrației glucozei (CHEM)** în **sânge (ANAT)** (**hiperglicemie**) **(DISO)** și dezechilibrarea metabolismului.
>
> *(Le diabète sucré est un syndrome caractérisé par un taux de glucose élevé dans le sang (hyperglycémie) et un déséquilibre du métabolisme.)*

Dans cet exemple les notions médicales (en gras dans l'exemple) sont annotées avec différentes étiquettes, comme **DISO** – maladie, **CHEM** – chimie, **ANAT** – anatomie (catégories utilisées dans UMLS). Vu que le module **BioNER** utilise **les mêmes 14 133 termes médicaux extraits du projet MoNERo** (Mitrofan *et al.*, 2019), et que les catégories ne sont pas essentielles pour notre projet, nous avons choisi d'extraire une liste avec ces termes qui soit exploitable avec des **scripts en langage Perl** développés par nous.[63]

[62] Une version plus récente du **BioNER** (Mitrofan et Pais, 2022) a été dévéloppée en 2022 à partir du corpus **SiMoNERo** (Barbu Mititelu et Mitrofan, 2020), qui contient 17 669 entités nommées médicales.
[63] Le module **RNER** (Mitrofan et Pais, 2022) a été mis à disposition en 2022 pour l'identification des entités nommées médicales sur des textes bruts en roumain (https://github.com/racai-ai/RNER).

Le corpus **MoNERo** est extrait du corpus roumain de textes médicaux **BioRo** (Mitrofan et Tufiș, 2018) et contient des textes provenant des livres de littérature médicale scientifique (en grande partie), des articles de revues médicales scientifiques et des textes de blogs médicaux. Ce corpus contient des textes provenant de trois sous-domaines médicaux : cardiologie, diabète et endocrinologie. Le corpus **MoNERo** est annoté avec des parties du discours (catégories lexicales) et des entités nommées du domaine médical.

MoNERo comprend 154 825 tokens, tous annotés morphologiquement avec l'analyseur **TTL** (Ion, 2007 ; Mitrofan et Tufiș, 2018), selon le jeu d'étiquettes du projet **MULTEXT-EAST** (Erjavec, 2010).

2.2.2.1 Extraction d'une liste de termes médicaux en roumain

Il n'existe pas, à notre connaissance, une terminologie médicale en libre accès en roumain. Nous constituons une **liste de termes médicaux** extraits du corpus roumain pour le domaine médical **MoNERo** (Mitrofan *et al.*, 2019). Nous utilisons cette liste car les notions médicales sont annotées avec les groupes sémantiques de la base terminologique médicale **UMLS** (Bodenreider, 2004) : 1 964 pour les parties du corps (ANAT), 4 156 pour les produits chimiques et médicaments (CHEM), 6 611 pour les troubles (DISO) et 1 402 pour les procédures (PROC), en total **14 133 entités médicales**. Ces catégories ont été validées par un annotateur humain, expert du domaine médical. Le texte intégral du projet **MoNERo** n'est pas libre de droits, raison pour laquelle nous avons accès uniquement aux concepts médicaux annotés.

Nous avons utilisé les codes des catégories sémantiques qui marquent le début (**B-DISO**) et l'élément intermédiaire ou de fin du terme médical (**I-DISO**) pour extraire une terminologie médicale. Nous avons ensuite supprimé les doublons afin de n'avoir que des termes uniques. Sur les **14 133 termes médicaux extraits**, nous avons obtenu **7 528 valeurs uniques** (la même forme).

Nous avons éliminé de cette liste plusieurs types d'éléments :

- les mots identifiés comme termes : « a » (B-DISO ; Va--3s[64]) (forme du verbe « a avea », *avoir*) et le verbe « este » (*est*, forme du verbe *être*), « fie » (*soit*) (B-CHEM ; Vaip3s), car ces verbes faussent les recherches et le balisage des termes ;

Nous expériences étant déjà finies, nous envisageons à exploiter ce module dans des expériences futures sur les textes médicaux en roumain.

[64] Étiquette en format Multext qui indique verbe auxiliaire, 3e personne, singulier.

- les éléments orthotypographiques qui apparaissent dans la liste, car ils ne sont pas des termes médicaux (/ B-CHEM SLASH ; . B-ANAT PERIOD ; - B-CHEM DASH ; , B-CHEM COMMA) ;
- les termes et les abréviations en langues étrangères, principalement en anglais (classe X), tels que : speckle ; tibi-alis ; glutation ; TNF&b.alpha ;
- des abréviations (classes Y, Ya, Yc, Yn, Rc, Mc), telles que : AIRglucose ; FSIVGTT ; TNG ; AG ; Ser, etc. Pour un premier étiquetage, nous avons écarté les abréviations, car les abréviations telles que « K » pour potassium, « Ca » pour calcium créent de fausses étiquettes de termes ;
- des mots fonctionnels de type prépositions (classe Spsa) : « din » (*de*), « de » (*de*), « cu » (*avec*).

Après avoir supprimé tous ces éléments de la liste, nous avons obtenu une **terminologie médicale en roumain** constituée de **6 899 termes médicaux simples et polylexicaux uniques**. Nous avons analysé manuellement cette liste et nous avons constaté que certains termes polylexicaux sont identifiés comme tels, mais il s'agit, en réalité, de termes simples suivis de mots grammaticaux (prépositions et conjonctions) ou même de signes de ponctuation : « atrofie a » (*atrophie a*) ; « ventriculare și » (*ventriculaire et*) ; « proinsulinei și » (*proinsuline et)* ; « pancreasului . » (*du pancréas .*) ; « ruptura de » (*rupture de*) ; « sindroamelor de » (*aux syndromes de*). Par conséquent, nous avons compté ces termes comme des termes médicaux simples et non pas polylexicaux, et nous avons supprimé les mots outils.

Cette liste de termes a été utilisée pour l'annotation de termes médicaux dans notre corpus **GrandMed-Ro2**. Dans un premier temps, on sélectionne les phrases qui contiennent des termes médicaux de la terminologie extraite et nettoyée, puis nous sélectionnons les phrases qui contiennent en plus des marqueurs de reformulation présentés dans le **Tableau** 11 (**Partie III**, **sous-chapitre 2.3.2**).

***

Dans une perspective de l'automatisation de la tâche d'annotation manuelle des reformulations, nous marquons les termes avec des balises. Nous présentons cette démarche dans la section suivante.

2.2.3 Balisage des termes médicaux annotés

Pour nos expériences nous avons décidé de marquer les termes médicaux annotés par l'annotateur SIFR-BioPortal dans les textes avec **des balises**. Ce marquage nous permet d'identifier si les termes annotés avec l'outil sont reformulés dans les textes et d'évaluer la performance d'annotation de SIFR-BioPortal. Nous réalisons cette tâche avec un script en langage Perl. Nous rajoutons des **balises** de type **<t> </t>** autour des termes identifiés avec SIFR-BioPortal, ce nous permet de visualiser ces termes pendant l'annotation des reformulations :

> *L'association ticlopidine + **<t>**aspirine**</t>** par rapport aux **<t>**anticoagulants**</t>** oraux réduisait significativement le risque d'**<t><t><t>**infarctus**</t> <t>**aigu**</t></t>** du myocarde**</t>** non mortel et la revascularisation à 30 jours : critère de **<t>**jugement**</t> <t>**composite**</t>** (mortalité, **<t>**infarctus**</t>** du myocarde, revascularisation à 30 jours).*
>
> (Exemple extrait des textes scientifiques du corpus CLEAR Cochrane (Grabar et Cardon, 2018))

L'exemple donné ci-dessus montre le balisage effectué sur les termes médicaux, selon la liste de termes identifiés par SIFR-BioPortal. Nous remarquons que pour les termes médicaux polylexicaux, nous avons trois fois la balise <t>, ce qui indique des termes imbriqués, mais qui apparaissent également séparément dans la liste de termes.

Dans la liste de termes médicaux nous avons :

- infarctus aigu du myocarde
- infarctus aigu
- infarctus
- aigu

Ce qui justifie l'annotation avec les balises dans l'exemple suivant :

Lors de notre analyse manuelle des phrases, nous regardons si la phrase contient un **autre terme reformulé** qui n'est pas celui identifié par SIFR-BioPortal, comme dans l'exemple en roumain dont le terme « Mucoasa gastrica » (*La muqueuse gastrique*) n'est pas identifié par l'annotateur. Nous ajoutons ce terme à la liste des termes reformulés. Nous annotons les marqueurs identifiés automatiquement lors de nos recherches avec une paire de balises **<m> </m>**, comme dans l'exemple suivant en roumain :

Mucoasa gastrica, **<m>**mai exact**</m>** captuseala **<t>**stomacului**</t>**, este special adaptata pentru a fi protejata de acidul puternic, dar **<t>**esofagul**</t>** nu este protejat.

La muqueuse gastrique, ***<m>****c'est-à-dire****</m>*** *la paroi de* ***<t>****l'estomac****</t>****, est spécialement adaptée pour être protégée de l'acide fort, mais* ***<t>****l'œsophage****</t>*** *n'est pas protégé.*

(Exemple extrait du corpus roumain Grand-Med-Ro2)

***

Ainsi, notre **méthode** applique d'abord **une liste de termes préextraits** ou d'une **terminologie** validée manuellement pour extraire des phrases qui contiennent des termes médicaux. Une fois les termes médicaux monolexicaux et polylexicaux identifiés, nous analysons les contextes de ces termes afin d'y trouver des **marqueurs de reformulation**. Nous annotons ces **marqueurs et indicateurs de reformulations** sur la base des listes, expériences et résultats que nous détaillons dans le point suivant.

2.3 Annotation semi-automatique de marqueurs de reformulation

Les marqueurs de reformulations en français

Nous créons une liste des marqueurs de reformulation en partant de la littérature (**Tableau 10**), incluant les marqueurs :

- formés sur le verbe « dire » (*c'est-à-dire, ça veut dire / veut dire, pour dire autrement, autrement dit*) (Vassiliadou, 2013a ; Grabar et Eshkol-Taravella, 2016a ; Steuckardt, 2018 ; Magri, 2018) ;
- dérivés du verbe « désigner » ou « signifier » (Péry-Woodley et Rebeyrolle, 1998 ; Charolles et Coltier, 1986) ;
- dérivés du verbe « être » (Grabar et Hamon, 2016) suivis par des hyperonymes clés 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 ».

Nous travaillons dans cette étude également avec la notion **d'indicateur de reformulation** (Steuckardt, 2018) que nous définissons comme *des mots lexicaux ou grammaticaux qui, dans leur sémantique, renvoient au processus de vulgarisation*

(simplification, dire les choses autrement). Des **indicateurs spécifiques au domaine médical** sont les mots de type **hyperonymes généraux** (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 »). D'autres indicateurs, **« définition », « défini/e », « défini/e comme »**, placés dans le texte avant ou après le terme, annoncent la reformulation. Le choix du marqueur « définition » est justifié par la structure type d'un grand nombre d'articles qui se présente dans cette forme : « titre / terme médical » suivi par « définition » qui indique la présente d'une définition / explication ou reformulation.

Les **indicateurs grammaticaux « tel que »** et **« par exemple »** annoncent la reformulation à travers une exemplification. Dans ce cas, le terme (« antibiotiques ») devient 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. Nous évaluons manuellement les phrases pour trouver davantage de reformulations sans marqueur ni indicateur lexical, mais marquées par exemple avec des **marqueurs orthotypographiques**, comme les parenthèses ou la virgule ou les doubles points, pour y identifier des paraphrases telles que l'exemple **a.** du corpus en français QUAERO (Névéol *et al.*, 2014) :

> **a.** « De même, veuillez prévenir votre médecin si vous souffrez d'insuffisance rénale ou d'une cirrhose du foie **(**maladie du foie à un stade avancé**)**, car il faudra dans ce cas réduire la posologie de Refludan. »

La liste de marqueurs issue de ces analyses est présentée dans le **Tableau 10**.

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 10. Listes de marqueurs et d'indicateurs de reformulation établie sur la base de la littérature

Les marqueurs de reformulations en roumain

Les marqueurs de reformulation que nous recherchons pour le roumain sont inspirés des études disponibles pour le français (Péry-Woodley et Rebeyrolle, 1998 ; Chéria, 2010 ; Vassiliadou, 2016 ; Eshkol-Taravella et Grabar, 2018 ; Fuchs, 2020), des études sur le roumain (Barbu Mititelu, 2011 ; Săpoiu, 2013), des traductions à partir du français et de nos propres observations sur les corpus. Le français et le roumain sont des langues latines et elles partagent la même étymologie des termes médicaux. Par conséquent, nous avons cherché la traduction ou les équivalents roumains des marqueurs de reformulation identifiés en français, tels que « **cu alte cuvinte** » (*autrement dit*), « **înseamnă** » (*signifie*), « **este o boală** » (*est une maladie*). Nous avons trouvé d'autres marqueurs de reformulation spécifiques à la langue roumaine, de type « **sub denumirea de / sub numele de** » (*sous le nom de*), « **este cunoscut** » (*est connu*), « **cunoscut și** » (*également connu comme*), « **cum ar fi** » (*comme*), « **mai precis** » (*plus précisément*), « **este termenul** » (*est le terme*).

La liste de marqueurs en roumain issus de ces traductions et observations est présentée dans le **Tableau 11**.

Marqueurs de reformulation en roumain	Traduction en français
înseamnă / inseamna	*signifie*
mai exact	*plus précisément*
cum ar fi	*par exemple*
mai precis	*plus exactement*
este cunoscut/ cunoscută	*est connu/e*
cunoscut/ cunoscută sub numele de	*connu/e sous le nom de*
cunoscut/ cunoscuta si / cunoscută și sub denumirea de	*connu/e aussi sous le nom de*
cu alte cuvinte	*en autres mots*
définit/ă, definiție/ definitie	*défini/e, définition*
este termenul	*est le terme*
este o boală/ boala	*est une maladie*
este o afecțiune/ afectiune	*est une affection*

Tableau 11. Liste initiale de marqueurs de reformulation en roumain

Pour le roumain, nous incluons les patrons identifiés par les travaux de Barbu Mititelu (2011), de type « adică » (c'est-à-dire), « precum » (comme) et les hyperonymes généraux introducteurs de notions médicales comme « affection », « maladie », « trouble » (Săpoiu, 2013). Dans le corpus roumain **GrandMed-Ro** (Buhnila, 2018) nous identifions des également des marqueurs de type « înseamnă » (*signifie / veut dire* en français) et « adică » (*c'est-à-dire* en français) (Buhnila, 2018 : 88). Dans les exemples **b.** le terme médical « hépatite » est paraphrasé en « inflammation du foie » et dans l'exemple **c.** le terme « métastase » est paraphrasé par une longue explication (« se détachent de la tumeur initiale et migrent vers d'autres parties du corps par le sang ou le système lymphatique ») :

b. « Hepatita **înseamnă** inflamarea ficatului și un grup de infecții virale care afectează ficatul. » (*L'hépatite **signifie** une inflammation du foie et un groupe d'infections virales affectant le foie.*) (Buhnila, 2018 : 89) ;

c. « celulele anormale intră în metastază, **adică** se rup de tumoarea inițială și migrează în alte zone ale corpului, prin sistemul sanguin sau limfatic » (*les cellules anormales entrent dans la métastase, **c'est-à-dire** se détachent de la tumeur initiale et migrent vers d'autres parties du corps par le sang ou le système lymphatique*) (Buhnila, 2018 : 89).

Nous avons cherché les deux formes du mot, avec et sans **signes diacritiques** (par exemple, *este o afecțiune* / *este o afectiune*), car les textes des sites de vulgarisation médicale et les articles sur la toile sont souvent écrits sans signes diacritiques roumains (ă,

â, î, ș, ț). Cette habitude de rédaction sans signes diacritiques est spécifique au milieu numérique. Nous notons également que les exemples de mots ou phrases que nous donnons pendant notre analyse sur le corpus roumain **respectent l'orthographe des textes d'origine** (avec ou sans signes diacritiques). Par conséquent, leurs graphies *ne représentent pas* des fautes d'orthographe de notre part.

***

Nous avons réalisé une **analyse préliminaire** avec TXM sur un des corpus français pour tester les marqueurs et les indicateurs de reformulations spécifiques à chaque langue dans nos corpus médicaux. Nous menons des analyses quantitatives et qualitatives sur ces résultats afin d'améliorer davantage notre méthode et de vérifier l'efficacité des **marqueurs de reformulation** sur le corpus français **ClassYN**, que nous présentons dans la section suivante.

2.3.1 Expériences d'identification automatique des marqueurs avec TXM en français

Pour approfondir nos recherches de marqueurs, nous explorons les phrases extraites du corpus **ClassYN** (qui contiennent des termes) afin d'y trouver de nouveaux marqueurs de reformulation avec des patrons qui ont le potentiel de marquer des reformulations, comme les dérivés du verbe « être » et « désigner » (Péry-Woodley et Rebeyrolle, 1998 ; Grabar et Hamon, 2016 ; Charolles et Coltier, 1986). Nous réalisons une expérience préliminaire pour vérifier des marqueurs avec des données identifiées dans le corpus à l'aide de **TXM** (Heiden *et al.*, 2010). Le but est de trouver des variantes de ces marqueurs dans le corpus **ClassYN** (scientifique et de vulgarisation), à l'aide des requêtes **CQL**[65] précises.

Après une analyse des contextes avec l'outil TXM (Heiden *et al.*, 2010), nous identifions des marqueurs de reformulation de type, ce qui valide partiellement les listes de marqueurs établis dans la littérature :

- « **est une maladie** » : « La sarcoïdose *est une maladie* systémique chronique d'étiologie inconnue. » ;
- « **est une affection** » : « L'asthme *est une affection* pulmonaire chronique » ;
- « **désigne** » : « L'anémie *désigne* le manque de globules rouges dans le sang » ;

[65] CQL est l'acronyme de Corpus Query Language, qui est un langage d'interrogation de corpus.

- les parenthèses **()** : « La maladie de Chagas *(*infection parasitaire à Trypanosoma cruzi*)* ».

Nous présentons dans le **Tableau 12** les types de marqueurs de reformulations identifiés avec **TXM** dans 2 689 phrases (88 407 tokens), incluant termes et marqueurs du corpus, extraites de ClassYN SP[66] et 4 871 phrases (139 320 tokens) du corpus ClassYN GP[67]. L'annotation des marqueurs et la validation manuelle des reformulations nous permettent de calculer la précision des marqueurs corrects de reformulation parmi ceux identifiés avec TXM dans cette expérience. Nous remarquons que **« c'est-à-dire »** est absent du SP et présent dans **43 (0,78%)** reformulations dans GP. Parmi nos ajouts, **« défini/e »** marque **63 (0,41%)** reformulations en SP et uniquement **35 (0,43%)** en GP (probablement dû au langage informel) et **« définition » 21 (0,38%)** en SP et **52 (0,55%)** en GP. Les marqueurs « pour dire autrement », « doit être compris comme » et « au sens de » sont absents de deux corpus.

Marqueurs identifiés	**Corpus ClassYN SP**			**Corpus ClassYN GP**		
	Occ.	**Vrais marqueurs**	**P**	**Occ.**	**Vrais marqueurs**	**P**
« c'est-à-dire »	0	0	0,00	55	43	0,78
« ça/*cela/ce qui* veut dire »	2	1	0,50	25	19	0,76
« pour dire autrement »	0	0	0,00	0	0	0,00
« autrement dit »	0	0	0,00	26	7	0,26
« signifie »	7	4	0,57	136	66	0,48
« est ce qu'on appelle »	0	0	0,00	1	0	0,00
• « est aussi appelé/e » • aussi appelé • appelé	0 6	0 5	0,00 0,83	14 46 **55** (inclut les autres)	10 38 **43** (inclus les autres)	0,71 0,82 0,78
« doit être compris/e comme »	0	0	0,00	0	0	0,00
« au sens de »	0	0	0,00	0	0	0,00
« défini/e »	153 (62 + 91)	63 (20+43)	0,41	81 (50+31)	35 (15+20)	0,43
« définition »	55	21	0,38	94	52	0,55
Total	**223**	**94**	**0,42**	**473**	**265**	**0,56**

Tableau 12. Analyse par marqueurs de reformulations identifiés dans le corpus ClassYN (Occ. = occurrences ; P = précision entre les marqueurs identifiés et ceux qui introduisent une reformulation)

[66] Sélection de phrases contenant les termes, à partir du corpus ClassYN pour experts.
[67] Sélection de phrases annotées avec les termes, à partir du corpus ClassYN pour le grand public.

Nous retrouvons de façon automatique **696 termes médicaux suivis par les marqueurs donnés**, dans les phrases annotées en termes. À l'issue de notre analyse qualitative manuelle des résultats dans les deux corpus, nous trouvons dans le corpus scientifique **94** (**P 0,42 ; FRP[68] 0,106**) termes médicaux suivis des marqueurs qui permettent d'identifier correctement des reformulations et **265** (**P 0,56 ; FRP 0,190**) dans le corpus de vulgarisation, donc peu de reformulations effectives. Les marqueurs de reformulation sont plus fréquents dans le corpus de vulgarisation.

***

Nous menons une deuxième expérience avec TXM pour chercher trois structures spécifiques correspondant aux marqueurs de reformulation définis dans la littérature (Péry-Woodley et Rebeyrolle, 1998 ; Grabar et Hamon, 2016 ; Charolles et Coltier, 1986) pour mesurer leur efficacité :

- le verbe **« est »** suivi par le terme générique **« maladie »** avec la requête CQL suivante : « [word = "est"] []{0, 50} [word = "maladie.*"] within 5 » (nous permettons avoir une fenêtre de 5 mots qui peuvent s'intercaler entre les deux parties et la flexion au singulier ou au pluriel, ce qui nous donne plusieurs formes du marqueur, comme « est la maladie », « est une maladie », « est une des maladies », etc.) ;
- le verbe **« est »** suivi par **« affection »** avec la requête CQL suivante : « [word = "est"] []{0,50} [word = "affection.*"] within 5 » (idem) ;
- le verbe **« désigner »** avec ses variants de conjugaison avec la requête CQL suivante : « [word = "désigne.*"] » (nous cherchons les formes qui commencent par le préfixe verbal « désigne »).

Ces marqueurs de reformulation représentent les patrons utilisés pour les recherches, les formes exactes pouvant varier en article (défini / indéfini), conjugaison, nombre. Nous avons obtenu **534 reformulations correctes**, dont **77 (14,41%)** du corpus scientifique et **457 (85,58%)** du corpus de vulgarisation.

Marquers	**ClassYN SP**			**ClassYN GP**		
	N°	**Marq validés**	**P**	**N°**	**Marq validés**	**P**
est une maladie	80	56	0,7	498	389	0,78
est une affection	24	20	0,83	42	32	0,76
désigne	1	1	1	37	36	0,97
Total	105	**77**	**0,84**	577	**457**	**0,83**

[68] FRP représente la fréquence relative en pourcentage par rapport au nombre total de mots (88 407 pour ClassYN SP et 139 320 pour ClassYN GP).

Tableau 13. Marqueurs de reformulation trouvés avec TXM (Heiden et al., 2010) (N° : nombre d'occurrences du marqueur ; Marq validés : marqueurs validés par 2 annotateurs ; P : Précision)

Nous trouvons un nombre variable d'occurrences de marqueurs de reformulation dans les deux corpus, comme présenté dans le **Tableau 13**. Pour le corpus de littérature médicale scientifique, nous avons trouvé pour ces requêtes CQL **77** occurrences des marqueurs de reformulations avec nos patrons (précision entre les marqueurs corrects et incorrects de **0,84** et la **FRP** de **0,087**). Pour le corpus de vulgarisation, nous avons trouvé **457** marqueurs avec une précision entre les marqueurs corrects et incorrects de **0,83** et une **FRP** de **0,328**). Nous observons que la fréquence des marqueurs est plus importante dans le corpus de vulgarisation, tendance observée pour les autres marqueurs également.

***

Pour obtenir des résultats avec TXM, nous avons utilisé des expressions CQL précises pour chaque marqueur. Pourtant, ces recherches doivent être lancées successivement et nous devons exporter manuellement les données après chaque recherche. Vu que nous avons une grande quantité de données à exploiter et que nous envisageons d'élargir la liste de marqueurs, utiliser TXM **pour tous les marqueurs** peut devenir très chronophage. Dans cette perspective, nous avons décidé d'utiliser des **scripts en langage Perl** pour chercher concomitant tous les marqueurs de reformulation et exporter directement les phrases avec marqueurs. Nous évaluerons le taux d'erreur de ces scripts à l'aide d'annotations manuelles.

***

Nous avons identifié une liste de marqueurs de la littérature et issus de nos observations vérifiées sur corpus avec TXM. Nous avons d'abord appliqué cette liste sur le corpus **CLEAR**. Sur la base de l'analyse manuelle de phrases extraites de ce corpus, nous avons identifié d'autres marqueurs. Cette liste complétée a été appliquée sur le corpus **ClassYN** (liste consultable dans le bilan de la **Partie IV**, **section 1.3**) pour l'identification des phrases qui contiennent des reformulations.

***

Une autre étape est représentée par **la constitution de la liste de marqueurs de reformulation en roumain**, en partant des traductions des listes de la littérature en français et des observations sur les corpus français. Cette liste nous aide à cibler de manière plus précise les reformulations correctes. Nous présentons cette étape dans le sous-chapitre suivant.

2.3.2 Constitution de la liste de marqueurs de reformulation en roumain

Nous traduisons et nous adaptons ces marqueurs découverts en français à la langue roumaine. Nous constituons la **liste complétée avec des nouveaux marqueurs** de reformulation en roumain présentée dans le **Tableau 14** ci-dessous.

Liste complétée de marqueurs de reformulation en roumain		
N°	**Marqueurs en roumain**	**Marqueurs traduits en français**
1	denumit / ă (a) / ți (ti) / te	*appelé / e / s / es*
2	definiți /te, définit/ă, definiție/ definitie	*définis / es, défini/e, définition*
3	este un termen / este termenul	*est un terme / est le terme*
4	sunt termeni / i	*sont des / les termes*
5	sunt boli / le	*sont des / les maladies*
6	sunt afecțiuni / le (afectiuni / le)	*sont des / les affections*
7	este un sindrom / sindromul	*est un / le syndrome*
8	sunt sindromuri / le	*sont des / les syndromes*
9	sau alte	*ou autres*
10	și (si) alte	*et autres*
11	de exemplu	*par exemple*
12	spre exemplu	
13	supranumit / ă (a) / ți (ti) / te	*surnommé / e / s / es*
14	numit / ă (a) / ți (ti) / te	*appelé / e / s / es*
15	este numit / ă (a) / ți (ti) / te	*est appelé / e / s / es*
16	este numit / ă (a) / ți (ti) / te și (si)	*est aussi appelé / e / s / es*
17	sunt numiți / te	*sont appelés / es*
18	sunt numiți / te și (si)	*sont également appelés / es*
19	se numește (numeste) / numesc	*il est appelé / ils sont appelés / es*
20	se mai numește (numeste) / numesc	*il est également appelé / ils sont également appelés / es*
21	un alt nume pentru	*un autre nom pour*
22	descrie	*décrit*
23	asociat / ă (a) / ți (ti) / te	*associé / e / s / es*
24	includ / include	*inclut*
25	mai ales	*en particulier*
26	traduce (se poate traduce prin, se traduce prin, se mai traduce prin, se traduce printr-un/o)	*traduit (peut se traduire par, se traduit par, se traduit également par, se traduit par un/une)*

27	caracterizat / ă (a) / ți (ti) / te (caracterizat prin, caracterizat printr-un/o, caracterizat de, este caracterizat prin/de, fiind caracterizat)	*caractérisé / e / s / es (caractérisé par, caractérisé par un/une, caractérisé à travers, est caractérisé par un/une, étant caractérisé)*
28	constă(a) în (in)	*consiste en*
29	înseamnă / inseamna	*signifie*
30	mai exact	*plus précisément*
31	cum ar fi	*par exemple*
32	mai précis	*plus exactement*
33	este cunoscut/ cunoscută	*est connu/e*
34	cunoscut/ cunoscută sub numele de	*connu/e sous le nom de*
35	cunoscut/ cunoscuta si / cunoscută și sub denumirea de	*connu/e aussi sous le nom de*
36	cu alte cuvinte	*en autres mots*
37	définit/ă, definiție/ definitie	*défini/e, définition*
38	este termenul	*est le terme*
39	este o boală/ boala	*est une maladie*
40	este o afecțiune/ afectiune	*est une affection*

Tableau 14. Liste complétée de marqueurs en roumain traduits et adaptés des marqueurs en français

***

Une fois les termes médicaux et les marqueurs et indicateurs de reformulations repérés dans les textes, dans la même phrase, nous analysons manuellement les phrases afin d'y identifier des reformulations médicales.

2.4 Annotation manuelle de la reformulation sous-phrastique médicale

Notre méthode de construction de corpus de reformulations sous-phrastiques médicales teste de nouvelles règles d'extraction automatique des reformulations. Pour ce faire, nous menons une **analyse linguistique approfondie** de corpus et nous proposons des **modèles linguistiques** de représentation de la reformulation médicale. Dans ce sens nous analysons les termes médicaux simples et polylexicaux, les marqueurs de reformulation et la structure syntaxique de la reformulation sous-phrastique médicale. La méthode d'extraction automatique est évaluée à l'aide de plusieurs mesures statistiques

telles que la précision, le rappel, la F-mesure, Kappa (Cohen, 1960), à l'aide de la plateforme ReCal (Freelon, 2013).

Pour cette tâche, nous avons constitué un **guide détaillé d'annotation des reformulations**, consultable dans **l'Annexe 6.1**. Ce guide méthodologique permet aux annotateurs de suivre des critères clairs afin de réaliser des annotations les plus objectives possibles. Notre guide d'annotation contient des indications pour le type de données annotées, les termes médicaux, les différentes catégories de marqueurs identifiés, les types de reformulations sous-phrastiques. Nous expliquons aussi les types de relations lexicales et sémantico-pragmatiques que nous annotons entre le terme et sa reformulation (présentées dans la **section 2.5**).

L'annotateur commence par décider si la phrase contient au moins une reformulation valide. Nous avons calculé l'accord entre deux annotateurs non-spécialistes du domaine de la médecine. Nous analysons le nombre de reformulations qui ont été identifiées comme des reformulations médicales correctes par les deux annotateurs, le nombre de reformulations qui ont reçu la même étiquette de type « statut » (« oui », « oui-inv », « non »), dans les deux corpus. Nous avons également calculé le nombre de reformulations étiquetées différemment par les deux annotateurs. Nous avons procédé à une adjudication pour les doubles annotations.

Lors de l'annotation, nous identifions les reformulations qui sont plus ou moins en relation de similarité sémantique avec le terme médical, en incluant aussi les explications, les définitions et les exemplifications. Il s'agit ici de décider si la phrase contenant le terme et un marqueur de reformulation contient bien une reformulation du terme. L'extraction automatique de ces reformulations représente un défi à cause de l'énorme variation et des différentes interprétations possibles. Parfois, le terme n'est pas correctement identifié ou ce n'est pas le terme reformulé à l'aide de marqueur. Le marqueur seul peut être utilisé dans un contexte ou il n'y a pas de reformulation. Définir des règles précises pour une si grande variété de types de reformulation nécessite une recherche approfondie dans les corpus multilingues comparables. Distinguer le niveau de synonymie entre des reformulations et identifier la variation sémantique de termes médicaux est une tâche difficile, même pour les annotateurs humains.

Notre **hypothèse** est que *les corpus de vulgarisation contiennent un nombre plus important de reformulations que les corpus expert et qu'elles sont plus faciles à comprendre pour un public non-spécialiste*. Nous analysons les deux types de corpus, nous comparons les résultats et nous vérifions cette hypothèse dans la **Partie IV**. De plus, nous évaluons le

niveau de lisibilité et de compréhension d'un extrait de reformulation de chaque type de corpus par des annotateurs non-spécialistes dans le **sous-chapitre 4.**

La section suivante présente notre **analyse lexicale et sémantico-pragmatique** des reformulations médicales annotées comme correctes.

2.5 Annotation des relations lexicales et des fonctions sémantico-pragmatiques

Nous réalisons une analyse lexicale et sémantico-pragmatique des reformulations médicales identifiées manuellement par des annotateurs humains. Nous définissons **les relations lexicales** comme le *lien lexical* qui existe entre les deux segments, le terme médical et la reformulation. Nous analysons les relations lexicales de **synonymie, hyperonymie, hyponymie et méronymie**, dans le contexte du texte médical (Condamines, 2018 ; Ramadier, 2016 ; Săpoiu, 2013) et en fonction du **public cible** du texte médical (spécialiste ou grand public). Les **hyperonymes médicaux** (Săpoiu, 2013) ont un rôle important dans la classification des concepts médicaux scientifiques (« typhus des broussailles ») dans des classes plus faciles à comprendre pour le grand public, comme « maladie bactérienne » (Grabar et Hamon, 2015). Par exemple, dans le cas de **l'hyponymie**, le terme (« antibiotiques ») est **l'hyperonyme**, et la reformulation simplifie le sens du terme en utilisant des **hyponymes** (« chloramphénicol, tétracycline et doxycycline »), grâce au mot « tétracycline » qui est connu par le grand public.

Les fonctions sémantico-pragmatiques, quant à elles, représentent les *raisons* qui poussent le locuteur à utiliser la reformulation. Les définitions des relations lexicales et sémantico-pragmatiques sont inspirées de la taxinomie d'Eshkol-Taravella et Grabar (2017) et choisies par rapport à leur adaptabilité au texte écrit du domaine médical.

Dans notre travail de recherche, nous nous rapportons principalement aux cinq **fonctions pragmatiques** présentées dans Buhnila (2022a). Nous illustrons ces fonctions avec des exemples extraits du corpus CLEAR (Grabar et Cardon, 2018), où le terme médical est surligné avec deux lignes et la reformulation avec une ligne, tandis que le marqueur de reformulation apparait en italique :

- **Définition** : « La maladie de Parkinson *est une affection* invalidante du cerveau qui se caractérise par un ralentissement des mouvements, des tremblements, une rigidité musculaire et, à des stades plus avancés, une perte d'équilibre. » ;

- **Explication** : « Les transfusions répétées de globules rouges peuvent entraîner une surcharge en fer à secondaire (*c'est-à-dire* causée par les transfusions) ;
- **Paraphrase** : « Cependant, tous les essais évaluaient la " guérison microbiologique " (*autrement dit* ils ont examiné l'éradication de l'infection) et aucun n'évaluait la réduction des problèmes oculaires ou pulmonaires chez le bébé ».
- **Dénomination** : « La maladie de Crohn *est une maladie* inflammatoire chronique de l'intestin qui atteint souvent la partie inférieure de l'intestin grêle, *appelée* iléon. »
- **Exemplification** : « Des antibiotiques *(*chloramphénicol, tétracycline et doxycycline*)* sont utilisés dans le traitement de cette maladie. »

Nous analyserons le rôle de ces fonctions pragmatiques dans le processus de **simplification du terme médical**, dans le contexte de la **reformulation**. Pour la **fonction de dénomination,** nous analyserons les synonymes en langue générale (« l'intestin grêle, *appelé* iléon »), ainsi que les maladies qui portent des noms propres (exemple : « maladie de Parkinson », « maladie de Crohn »).

Nous présentons par la suite nos **hypothèses de recherche** concernant les liens entre ces fonctions sémantico-pragmatiques de la reformulation et les relations lexicales entre le référent et la reformulation.

2.5.1 Hypothèses concernant les liens entre les fonctions sémantico-pragmatiques et les relations lexicales

Dans la suite de ce travail, nous formulons **l'hypothèse que les fonctions sémantico-pragmatiques sont corrélées avec les relations lexicales,** selon Buhnila (2022a). Nous présentons quelques exemples en français et en roumain pour illustrer ces phénomènes. Les termes médicaux sont en gras, les marqueurs sont soulignés avec ligne double et les reformulations médicales sont soulignées avec ligne simple :

- *la paraphrase* (utilisée 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*) ;

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

(1b) **Glanda lenește** (hipotiroidismul) (***Glande léthargique*** *(hypothyroïdie))*

- *la dénomination* (utilisée 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.

  (2a) 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).

  (2b) **Oreionul** (parotidita epidemică) *(**Oreillons** (parotidite épidémique))*

- *la définition* (corrélée avec *l'hyperonymie* et *l'hyponymie*) : le terme est *défini*, car il est considéré comme étant trop technique et 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 » ;

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

  (3b) **Tiroidita autoimună Hashimoto** este o afecțiune în care sistemul imunitar atacă celulele tiroidei. *(**La thyroïdite auto-immune de Hashimoto** est une affection dans laquelle le système immunitaire attaque les cellules thyroïdiennes.)*

- *l'exemplification* (conjointement associée 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) ;

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

  (4b) **organele pelvine feminine**: vaginul, colul uterin, pereții uterini, trompele uterine, ovarele sau țesuturile care înconjoară uterul, vezica urinară sau chiar intestinele. *(**organes pelviens féminins** : vagin, col de l'utérus, parois utérines, trompes de Fallope, ovaires ou tissus entourant l'utérus, la vessie ou même les intestins.)*

- *l'explication* (corrélée 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 ;

  (5a) **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).

  (5b) **Hipotiroidismul** înseamna că glanda tiroidă nu poate produce cantitatea normală de hormon tiroidian. *(**L'hypothyroïdie** signifie que la glande thyroïde ne peut pas produire la quantité normale d'hormones thyroïdiennes.)*

Nous analysons ces fonctions et relations et nous testons nos hypothèses lors de l'annotation et l'analyse des reformulations identifiées dans le corpus français et, respectivement, roumain (présentées dans la **Partie IV**).

***

Les reformulations annotées et validées manuellement sont utilisées dans des tâches de TAL **d'apprentissage profond** afin de *générer* des nouvelles reformulations et de *classifier* les reformulations correctes et les reformulations incorrectes. Nous détaillons notre méthode dans la section suivante.

2.6 Approche par apprentissage profond

Nous avons présenté dans l'état de l'art (**sous-chapitre 1.3**, **Partie II**) les méthodes d'identification de la paraphrase (**PI**), la similarité sémantique textuelle (**STS**) et la génération de la paraphrase (**PG**). Nous proposons une approche basée sur l'apprentissage profond par réseaux de neurones pour **l'identification et la classification (avec LSTM) et la génération (avec une architecture APT) de la reformulation** à partir de termes médicaux et des reformulations médicales identifiées par nous et présentées en **Annexes**. Nous détaillons notre méthode et les ressources utilisées dans les sous-chapitres suivants.

2.6.1 Méthode et outils pour la génération des paraphrases

La ***génération des paraphrases (PG[69])*** consiste à créer du nouveau contenu textuel à partir des données d'entrée (dans notre cas, les termes) et des modèles de langues (Gupta *et al.*, 2018 ; Bowman *et al.*, 2016). Nous utilisons un modèle d'architecture neuronale de type **paradigme contradictoire**, qui vise à créer des ***différences lexicales et syntaxiques*** importantes entre le terme et la reformulation. Cette méthode va à l'encontre de la **similarité textuelle** (**STS**) (qui cherche la similarité formelle), dans le but de générer des reformulations les plus diverses possibles, tout en gardant **le même contenu sémantique ou similaire**. Cette méthode est la plus adaptée à notre tâche de génération de la reformulation, vu la grande diversité des reformulations médicales identifiées dans nos corpus : paraphrases, exemplifications, synonymes, de type définition et explication, voir même abréviations. En effet, les méthodes **STS** sont difficilement applicables vu la grande

[69] Abréviation du terme anglais, *Paraphrase Generation* (PG).

variation dans la structure syntaxique et les choix lexicaux des reformulations que nous avons annotées en corpus.

2.6.1.1 L'architecture neuronale APT

Pour nos expériences de **génération de reformulations à partir de termes médicaux,** nous faisons appel à **l'architecture neuronale APT** (*Adversarial Paraphrasing Task*) (Nighojkar et Licato, 2021). Cette architecture utilise un ***paradigme contradictoire*** qui consiste à générer des reformulations qui présentent, d'une part, des **contenus sémantiques équivalents**, et, d'autre part, des ***différences lexicales et syntaxiques*** par rapport au référent. Ce modèle vise à identifier **le sens général d'une phrase ou d'une expression**, non pas uniquement le sens des mots séparés. L'architecture **APT** est construite sur deux principes :

- **La similarité de sens** : cette similarité se vérifie par le fait que deux phrases qui sont *mutuellement implicites* sont sémantiquement équivalentes, et sont donc, des *paraphrases* ;
- **La dissimilarité de la structure** : mesurée avec **BLEURT** (Sellam *et al.*, 2020), un score qui évalue les textes générés automatiquement en se basant sur le modèle de langue **BERT** (Devlin *et al.*, 2019). Ce score attribue à chaque paire de phrases un score quantifiant la similarité lexicale et syntaxique des deux phrases.

Dans leur étude, Nighojkar et Licato (2021) mènent les expériences sur des paraphrases phrastiques en anglais, appartenant au langage général. Ils utilisent un score *MI (implication mutuelle)* entre deux segments, qui mesure les inférences nécessaires pour déduire le sens du premier à partir du deuxième et vice-versa, à l'aide du modèle de langue **T5-base** (Raffel *et al.*, 2020). Ainsi, ils sélectionnent des paraphrases qui ont un score d'implication mutuelle MI grand et un score BLEURT (score de similarité de phrases) réduit. Les expériences avec les corpus anglais annotés manuellement (quelques milliers de phrases et leurs paraphrases) ont donné de meilleurs résultats par rapport aux données extraites automatiquement de Twitter (d'ordre de 100 000 mots).

Nous testons ce ***modèle contradictoire d'identification de la reformulation (APT)*** sur nos **données sous-phrastiques** (terme médical *versus* reformulation), en **français et en roumain**, appartenant au **domaine médical**. Par rapport aux Nighojkar et Licato (2021), nous travaillons avec des reformulations sous-phrastiques, avec peu de similarités lexicales

ou syntaxiques. Le corpus de paires *terme-reformulation* que nous avons construit et validé manuellement servira comme jeu de données pour adapter l'outil **APT** pour générer des nouvelles reformulations à partir d'un terme présenté en entrée.

Pour lancer nos expériences, nous avons besoin des **modèles de langues** de type **Transformers**, adaptés grâce aux données annotées manuellement, plus fiables que des jeux de données de grande taille acquises automatiquement. Nous présentons le modèle de langue que nous utilisons, **T5** (Raffel *et al*., 2020), dans le sous-chapitre suivant.

2.6.1.2 Le Transformer T5

Le modèle de langue T5 (*Text-to-Text Transformer*) est **préentraîné** sur C4 (*Colossal Clean Crawled Corpus*) (Raffel *et al*., 2020), un corpus nettoyé avec une taille colossale de 7 téraoctets, extrait du corpus Web de Common Crawl. **T5** a été préentraîné pour des tâches spécifiques au Traitement Automatique des Langues (TAL), telles que :

- Jugement de l'acceptabilité d'une phrase ;
- Analyse des sentiments ;
- **Paraphrase et similarité de phrases** ;
- Inférence en langage naturel ;
- Compléter des champs manquants dans des phrases ;
- Désambiguïsation du sens des mots ;
- Réponse aux questions.

Nous utilisons ce modèle, car il a été entraîné pour la **reconnaissance de paraphrases** et pour la **génération de reformulations**.

Comme le corpus de préentraînement, C4, est un corpus de langue générale issue de la toile, nous avons besoin **d'adapter** le modèle à notre tâche spécifique et à notre domaine. Cela signifie que nous devons **entraîner le modèle avec nos données médicales (paires de termes et reformulations)** avant de lancer la génération des reformulations médicales. Cette étape d'entraînement du modèle de langue pour une tâche spécifique ou avec des données spécifiques s'appelle ***fine-tuning*** ou ***optimisation*** (Howard et Ruder, 2018). Le ***fine-tuning*** est un type d'**apprentissage par transfert** (*Transfer Learning*), et plus précisément **d'apprentissage par transfert transductif**, dont les tâches visées sont similaires (*paraphrase et similarité de phrases*), mais les données sont différentes (langue générale *versus* langage médical). Ce type d'apprentissage est utile quand nous disposons de relativement peu de données, ce qui est notre cas.

Nous évaluons la qualité des reformulations obtenues par génération automatique selon un guide d'annotation présenté dans la **Partie IV**, **section 3.2.2**.

2.6.2 Méthode et outils pour la classification automatique

Dans une autre perspective que la génération, nous nous intéressons à la **reconnaissance des reformulations** de termes médicaux. Ainsi, nous réalisons des expériences de ***classification automatique*** (voir **sous-chapitre 1.2**) afin d'identifier automatiquement les reformulations médicales correctes à l'aide d'un modèle entrainé sur nos corpus de reformulation validés manuellement. Notre méthode consiste dans une **classification binaire** (Rajkumar et Chitra, 2010) des reformulations médicales valides ou non. L'architecture choisie est de type **LSTM** (Hochreiter et Schmidhuber, 1997) en utilisant deux types de données en entrée :

- **le terme médical** ;
- **le marqueur** / **indicateur de reformulation** ensemble avec **la reformulation**.

L'objectif de cette expérience est de créer un modèle qui peut classer des nouvelles reformulations (classification binaire). Nous avons exploré l'état de l'art de plusieurs méthodes et outils (présentés dans la **Partie II**, **chapitre 1.**) afin de trouver la méthode la plus adaptée pour **les reformulations sous-phrastiques** : **les architectures neuronales de type LTSM**. Nous présentons cette architecture d'apprentissage automatique dans le point suivant.

2.6.2.1 L'architecture neuronale LTSM

L'architecture neuronale LSTM (*Long Short Term Memory*) (Hochreiter et Schmidhuber, 1997) est utilisée par des approches par **apprentissage profond** à base de **réseaux de neurones** (concepts présentés dans la **Partie III**, **sous-chapitre 1.3**). Cette architecture permet le traitement de séquences des mots et peut être adaptée pour la classification de reformulation.

Nous menons des expériences sur le corpus français et roumain, en les adaptant aux particularités de chaque langue. Dans notre cas, nous utilisons une **architecture bidirectionnelle LSTM**. L'entrée est constituée de phrases qui ont été annotées comme des reformulations valides ou sans reformulations. Ainsi les termes et les reformulations sont présentées en contexte et tous les mots (sauf les mots outils) participent à la

classification. Les entrées sont transformées à l'aide des plongements lexicaux **GloVe** (Pennington *et al.*, 2014) en vecteurs de mots (300 est la taille choisie). La sortie est binaire, indiquant que la phrase d'entrée est une vraie ou une fausse reformulation. Les données d'entraînement sont construites selon les étapes présentées dans la méthodologie : les phrases qui contiennent une double annotation automatique (termes et marqueurs) ont été annotées manuellement en fonction de leur statut (il y a au moins une reformulation valide dans la phrase).

Le réseau est constitué d'une couche d'entrée (*embeddings*), d'une couche bidirectionnelle et de 7 couches intermédiaires permettant de filtrer l'information (par exemple *Dropout* qui activent 20% de cellules des couches intermédiaires).

L'architecture est présentée dans **l'Annexe 6.2** et sur le site **github**. Nos expériences par des réseaux de neurones restent **exploratoires et novatrices**, surtout pour la langue roumaine. Nous présentons ces expériences et nos résultats dans la **Partie IV**.

***

La dernière étape de notre travail de recherche est **l'évaluation de la lisibilité des reformulations** par rapport à un public non-spécialiste. Nous présentons cette étape de la méthodologie dans la section suivante.

2.7 Évaluation du niveau de lisibilité des reformulations

Les reformulations médicales proposées automatiquement ou annotées manuellement peuvent être plus faciles ou plus difficiles à comprendre pour le grand public. Par exemple, le terme médical « antibiotiques » peut être plus simple à comprendre que son hyponyme qui fait office de reformulation, « chloramphénicol ». Même situation pour les abréviations qui font partie de la culture populaire, comme « AVC », « HIV », qui sont plus faciles à comprendre que le terme médical d'origine (*accident vasculaire cérébral*, *virus de l'immunodéficience humaine*).

Notre objectif est d'évaluer le **niveau de lisibilité** de reformulations médicales extraites à partir des corpus scientifiques et de vulgarisation du domaine médical. **La lisibilité** fait référence au contenu des textes, sur deux axes : le fond, plus précisément les idées exprimées dans le texte, et la forme, c'est-à-dire la façon d'exprimer ces contenus linguistiques (Laframboise, 1978 ; François, 2011). Nous évaluons le niveau de

compréhension des reformulations médicales obtenues par validation manuelle à l'aide de plusieurs **annotateurs humains non-spécialistes** du domaine de la médecine.

Notre **hypothèse** soutient que cette expérience d'annotation nous permettrait de **classifier** les reformulations médicales par **niveau de difficulté de compréhension** et déterminer l'impact des **relations lexicales** (*synonymie, hyperonymie, hyponymie, méronymie*) et de **fonctions sémantico-pragmatiques** (*définition, exemplification, synonymie, dénomination, explication, paraphrase*) sur la *lisibilité des reformulations médicales* (présentées dans le **sous-chapitre 2.5** de la **Partie III**). Nous présentons notre échelle de lisibilité et de compréhensions des reformulations et les résultats de l'annotation dans le **sous-chapitre 4.**

***

Nous présentons par la suite **les résultats d'annotation** de termes médicaux et des marqueurs de reformulation par corpus, nos analyses quantitatives et qualitatives des annotations manuelles, des relations lexicales et des fonctions sémantico-pragmatiques (**Chapitre 1. 2. , Partie IV**). Nous présentons également **les expériences de génération et reconnaissance automatique** des reformulations avec des architectures à base de réseaux de neurones (**Chapitre 3. , Partie IV**). Nous analysons le **niveau de lisibilité** des reformulations et nous créons un **guide de la lisibilité** des reformulations (**Chapitre 4. , Partie IV**).

IV. ANALYSES, EXPÉRIENCES ET RÉSULTATS

1. Analyses sur les corpus français

L'annotation automatique des termes médicaux et des marqueurs de reformulation identifiés dans leur contexte est présentée dans la **Partie III** et représente une étape indispensable dans la constitution d'un corpus de reformulations. Cette annotation automatique des termes médicaux est réalisée avec l'annotateur **SIFR-BioPortal** (Tchechmedjiev *et al.*, 2018), qui est croisée avec la liste de marqueurs de reformulation, présentées dans la section **2.3** de la **Partie III**. Ainsi, nous analysons **les phrases ayant une double annotation** des termes médicaux et des marqueurs de reformulation appliqués sur les **corpus français (CLEAR et ClassYN**), nos différentes expériences et nos analyses quantitatives et qualitatives. Nous réalisons une première application de la méthodologie sur le premier corpus, **CLEAR**, suivie par un élargissement de la liste des marqueurs. Nous vérifions ensuite la méthode avec la liste élargie sur le deuxième corpus, **ClassYN**.

1.1 CLEAR Cochrane

Le corpus **CLEAR Cochrane** est composé de **3 859** résumés de textes scientifiques et **3 817** résumés adaptés pour le grand public. Parmi ces résumés, 3 815 ont le même sujet, dans sa variante technique et vulgarisée. Ces dernières sont souvent plus courtes, ce qui réduit la taille du corpus à environ 1,5 million des tokens (**CLEAR GP Total**), tandis que les textes scientifiques, plus élaborés, cumulent 2,8 millions des tokens (**CLEAR SP Total**).

CLEAR Cochrane	Résumés par type de texte	Résumés en total	Résumés sur le même sujet	Taille (tokens)
CLEAR SP Total	3859	8 789	3 815	2 840 003
CLEAR GP Total	3817			1 515 051

Tableau 15. Taille des corpus : CLEAR SP Total (littérature scientifique médicale) et CLEAR GP Total (textes de vulgarisation médicale)

Le corpus CLEAR Cochrane, disponible dans les deux versions, permet de faire des comparaisons entre les résultats obtenus sur les deux types de textes. Notre **hypothèse** de départ est que les textes de vulgarisation contiennent un nombre plus important de

reformulations médicales que les textes scientifiques, puisque ces textes s'adressent au public non-spécialiste du domaine. Dans ce sens, nous réalisons des expériences et nous calculons des mesures statistiques sur chaque type de texte, scientifique et de vulgarisation, afin de quantifier et analyser les différences en termes de nombre de reformulations et de degré de lisibilité de celles-ci.

Nous évaluons les résultats de l'annotation automatique, semi-automatique et l'annotation manuelle réalisée pour identifier les reformulations dans leur contexte. Les phrases annotées et vérifiées manuellement sont incluses dans le corpus de reformulations médicales.

1.1.1 Résultats d'annotation automatique des termes médicaux

Dans le **corpus scientifique CLEAR SP Total**, l'outil SIFR-BioPortal annote un nombre de **184 446** termes médicaux, mais dont 178 728 sont des doublons. Nous identifions donc **5 718** termes uniques (sans lemmatisation)[70]. Pour le **corpus grand public, CLEAR GP Total**, nous avons 125 696 termes médicaux annotés, dont 120 450 doublons, dont **5 246** termes uniques[71].

Le **Tableau 16** met en évidence le nombre de phrases avec et sans termes médicaux dans les deux corpus. Nous remarquons que l'annotateur a identifié plus de phrases avec des termes médicaux dans le corpus grand public que dans le corpus scientifique (**73%** des phrases contiennent des termes médicaux dans le CLEAR GP Total par rapport au **59%** dans le CLEAR SP Total). À partir de ce moment, nous travaillons sur les phrases qui contiennent les termes médicaux extraits avec les terminologies, que nous appelons désormais **CLEAR SP** et **CLEAR GP**.

CLEAR Cochrane	N° total des phrases	N° de phrases sans termes médicaux	N° de phrases avec termes médicaux
CLEAR SP Total	120 089	48 507	71 585
CLEAR GP Total	63 960	16 905	46 788

Tableau 16. Phrases avec ou sans termes médicaux dans le corpus CLEAR

[70] Le temps de calcul pour l'annotation est de trois heures.
[71] Le temps de calcul pour l'annotation est de deux heures (vu la taille plus petite du corpus grand public).

1.1.1.1 Post-traitement de l'annotation de termes médicaux

Lors de notre analyse des listes de termes uniques issus lors de l'annotation, nous avons remarqué la présence des 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 :

- agents physiques ;
- diagnostics ;
- fonctions ;
- modificateurs ;
- morphologie ;
- médicaments, produits chimiques et biologiques ;
- métiers ;
- organismes vivants ;
- procédures ;
- social ;
- topographie.

Nous avons analysé toutes les classes et celle qui contient le plus grand nombre de mots de la langue générale est la liste de *modificateurs*. Nous avons extrait cette liste de 1 510 mots et nous avons délimité les numéros, les adverbes et les prépositions afin de les éliminer 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 « adjectifs »[72] 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 la sous-classe « termes relationnels » qui contient des mots qui ne sont pas essentiels à l'identification d'un terme médical, comme « après », « en plus de », « suivant », « compatible avec », etc. La liste de mots à enlever contient 286 éléments (à consulter la liste sans doublons de 160 éléments dans **l'Annexe 6.3**).

CLEAR Cochrane	Termes annotés avec SIFR-BioPortal	Termes uniques sans doublons	Termes type « modificateurs » supprimés	Termes uniques après nettoyage
CLEAR SP	184 446	5 718	140	**5 578**
CLEAR GP	125 696	5 246	146	**5 100**

Tableau 17. CLEAR Cochrane : Termes médicaux uniques après le post-traitement de l'annotation automatique

[72] Nommée « adjectifs » dans la terminologie SNOMED-3.5VF.

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 enlevés, nous avons une liste de **5 578 termes uniques** dans **le corpus scientifique** et **5 100 dans le corpus grand public (Tableau 17).**

1.1.2 Résultats de l'annotation automatique des marqueurs de reformulations

Nous avons extrait les phrases qui contiennent conjointement des termes médicaux identifiés auparavant et des marqueurs de reformulations. Cependant, nous avons choisi d'élargir notre champ de recherche de marqueurs en remplaçant « est une/la maladie/affection » avec « maladie/affection » (singulier et pluriel, les contextes ayant la structure « terme, maladie, reformulation ») et les variants du lemme « définition » et « défini ».

Dans le **Tableau 18** ci-dessous nous montrons la liste des marqueurs et indicateurs de reformulation mise à jour pour les recherches encore plus précises réalisées sur notre corpus de travail, CLEAR Cochrane.

Marqueur de reformulation		Indicateur de reformulation
c'est-à-dire • c'est-a-dire • c'est-àdire • c'està-dire • c'est à dire • c'est-à-dire		définition • definition
ça veut dire • veut dire		affection • affections
pour dire autrement		maladie • maladies
autrement dit		trouble • troubles
est une / un • affection • maladie • trouble		défini • defini • définie • definie
désigne		défini comme
ce qu'on appelle		tel que
ce que l'on appelle		ou *
est aussi appelé • aussi appelé		par exemple

doit être compris comme		
au sens de		
signifie		

Tableau 18. Liste de marqueurs et indicateurs de reformulation mise à jour

* Nous avons décidé d'enlever le marqueur de synonymie « ou », car il génère trop de bruit dans les résultats (environ 13 000 phrases).

Nous avons projeté cette liste élargie des marqueurs sur les phrases extraites. Le **Tableau 19** montre que nous avons identifié **4 687 phrases** qui contiennent simultanément des termes médicaux et des marqueurs / indicateurs de reformulation dans le corpus **CLEAR SP** (**6,54%** des phrases incluant des termes) et, respectivement, **3 980 phrases** dans le corpus **CLEAR GP** (**8,50%** des phrases contenant des termes), un total de **8 667** phrases (**7,32%** des phrases avec les termes médicaux). Nous annotons et nous analysons manuellement ces **8 667 phrases** pour trouver des reformulations médicales correctes. Nous présentons notre travail d'annotation et l'analyse des résultats obtenus dans le **sous-chapitre 1.1.3**.

CLEAR Cochrane	**N° phrases avec termes médicaux**	**N° phrases avec termes, mais sans marqueurs**	**Phrases avec termes médicaux et marqueurs**	
			N° phrases	**N° tokens**
CLEAR SP	71 585	66 899	**4 687**	**173 616**
CLEAR GP	46 788	42 814	**3 980**	**123 249**
Total	118 373	109 713	**8 667**	**296 865**

Tableau 19. CLEAR Cochrane : Phrases avec termes médicaux et marqueurs de reformulation

1.1.2.1 Une analyse préliminaire des marqueurs et indicateurs de reformulation

Nous présentons nos analyses effectuées sur un échantillon de **2 000 phrases** extraites aléatoirement, dont 1 000 phrases du corpus de textes scientifiques, CLEAR SP et 1 000 du corpus de textes pour le grand public, CLEAR GP. Cette analyse sert à tester la méthodologie et à analyser les marqueurs et indicateurs de reformulation sur un échantillon de phrases.

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 de 2000 phrases). Pour les formes morphologiques différentes (par exemple « affection » et « affections », au singulier et au pluriel), nous calculons les fréquences relatives pour chaque forme. 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 (0,764%, *fréquence relative*), suivi par « c'est-à-dire » avec 51 occurrences (0,154%) et par « signifie » avec 44 occurrences (0,143%) (dans les deux sous-corpus).

Marqueurs de reformulation	**Fréquences absolues et relatives**			
	CLEAR SP		**CLEAR GP**	
	Fréq. absolue	**Fréq. rel. %**	**Fréq. absolue**	**Fréq. rel. %**
est un/une ; sont des/un/une • affection/s • maladie/s • trouble/s	95 dont - 18 - 36 - 22	0,262 0,049 0,099 0,060	150 dont - 28 - 62 - 30	0,502 0,093 0,207 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*)	1,379	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 20. Fréquences absolues et relatives de marqueurs et indicateurs de reformulation identifiés le plus fréquents.

Parmi les marqueurs les moins fréquents se trouvent « veut dire » avec 1 seule occurrence (0,003%) dans le corpus grand public, « est aussi appelé(e) / aussi appelé(e) » avec 2 occurrences (0,005%) dans le corpus expert et seulement 1 occurrence (0,003%) dans le GP, et « désigne » avec 4 occurrences (0,012%) 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, mais ils sont mentionnés dans la littérature de spécialité.

En ce qui concerne **les indicateurs de reformulations**, les **hyperonymes du domaine médical** sont très nombreux dans les deux corpus : 1 117 (3,12%) pour « maladie », 496 (1,338%) pour « trouble » et 145 (0,31%) pour « affection », suivis par les indicateurs grammaticaux « par exemple » avec 122 occurrences (0,381%) et « tel que » avec 72 occurrences (0,224%). 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 avec *(-)* dans le **Tableau 20**). La conjonction disjonctive « ou », avec 513 occurrences (1,565%), 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 contiennent 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.

1.1.3 Annotation manuelle, évaluation et validation des reformulations

En suivant notre **guide d'annotation** manuelle (présenté dans **l'Annexe 6.1**) nous avons obtenu **4 681** phrases pour le corpus de textes scientifiques (**CLEAR SP**) et **3 980** phrases pour le corpus de textes médicaux destinés au grand public (**CLEAR GP**) qui contiennent des termes et des reformulations. Dans cette section, nous analysons les phrases avec reformulation et sans reformulation (**section 1.1.3.1**), les termes identifiés par les annotateurs humains (**section 1.1.3.2**), les marqueurs identifiées par les annotateurs humains (**section 1.1.3.3**) et l'annotation en relations lexicales et sémantiques (**sous-chapitre 1.1.3.4**).

1.1.3.1 Comparaison des annotations des phrases avec reformulations

Toutes les phrases sont analysées manuellement par les deux annotateurs et partiellement automatiquement (pour l'annotation des abréviations), avec les valeurs pour le statut (*oui*, *oui<2+>*, *oui<inv>*, *non*, selon le guide de **l'Annexe 6.1**). Nous présentons les résultats et les statistiques de ces annotations dans le **Tableau 21** et le **Tableau 22** ci-dessous.

CLEAR SP (textes scientifiques)		
Données quantitatives	**Annot 1**	**Annot 2**
Reformulations avec ***oui total (ref multiples)***	**2689**	**2767**
Reformulations avec ***oui total (phrases)***	**2471**	**2055**
Reformulations avec ***oui<inv>***	37	49
Reformulations avec ***oui+2***	207	710
Reformulations avec ***oui<inv>+2***	11	2
Reformulations avec ***non***	2210	2626
Reformulations avec les mêmes balises - ***oui***	**1059**	
Reformulations avec les mêmes balises - ***oui<inv>***	**7**	
Reformulations avec les mêmes balises - ***non***	**1560**	
Reformulations avec les mêmes balises - ***total***	**2626**	
N° reformulations **multiples**	**218** *+ 4,44%*	**712** *+13,20%*
N° d'annotations	**4899**	**5393**
N° total initial de reformulations (phrases)	**4681**	

Tableau 21. Données quantitatives sur l'annotation du corpus CLEAR SP

Pour le corpus CLEAR SP, l'annotateur 1 a identifié **2 689** reformulations correctes et l'annotateur 2 un nombre de **2 767** reformulations. Par rapport au nombre de phrases initiales, nous avons **2 471 (52,78%) phrases avec des reformulations** (annotateur 1), respectivement **2 055 (43,90%) phrases qui ont des reformulations** (annotateur 2). Les différences entre les valeurs représentent les reformulations multiples (plusieurs reformulations correctes dans la même phrase).

CLEAR GP (textes grand public)		
Données quantitatives	**Annot 1**	**Annot 2**
Reformulations avec ***oui total (ref multiples)***	**1809**	**2245**
Reformulations avec ***oui total (phrases)***	**1462**	**1600**
Reformulations avec ***oui<inv>***	114	45
Reformulations avec ***oui+2***	293	626
Reformulations avec ***oui<inv>+2***	54	19
Reformulations avec ***oui total***	**1809**	**2245**
Reformulations avec ***non***	2518	2352
Reformulations avec les mêmes balises - ***oui***	**1168**	
Reformulations avec les mêmes balises - ***oui<inv>***	**40**	
Reformulations avec les mêmes balises - ***non***	**2213**	
Reformulations avec les mêmes balises - ***total***	**3421**	
N° reformulations **multiples**	**347** *+8,01%*	**645** *+14,03%*
N° d'annotations	**4327**	**4597**
N° total initial de reformulations (phrases)	**3980**	

Tableau 22. Données quantitatives sur l'annotation du corpus CLEAR GP

Concernant les résultats d'annotation du corpus **CLEAR GP**, nous observons que l'annotateur 2 a identifié deux fois plus de reformulations multiples dans la même phrase (*oui+2*) par rapport à l'annotateur 1. Les deux annotateurs ont annoté de la même manière **3 421** phrases : **1 208 (35%)** avec *oui* et *oui<inv>*, **2 213 (65%)** avec *non*, ce qui représente **85%** du total de phrases annotées **(3 980)**. Ce pourcentage important prouve que les deux annotateurs ont suivi le guide d'annotation. Nous supposons que le travail d'identification de reformulations médicales a été plus facile vu le langage simplifié des textes destinés au grand public. Nous évaluons cette hypothèse dans le **Chapitre 4.** où nous analysons les annotations du **niveau de compréhension et de lisibilité** des reformulations par rapport au type de texte.

Nous calculons la précision et le rappel de notre méthode d'extraction de reformulation. De plus, nous calculons les accords inter-annotateur dans le sous-chapitre suivant.

1.1.3.1.1 Calcul des mesures statistiques : précision et rappel

Nous avons calculé la précision et le rappel de notre méthode de sélection des phrases avec des reformulations. **La précision** est calculée par rapport au nombre de phrases annotées manuellement avec une certaine étiquette (*oui, non, oui<inv>*, etc.) divisé par le nombre total de phrases annotées automatiquement (qui contiennent au moins un terme médical et un marqueur). Nous avons calculé **le rappel** en considérant les deux annotations comme référence, l'annotateur 1, et l'annotateur 2 respectivement, comme illustré dans la **Figure 10**, **Figure 11, Figure 12, Figure 13** et les **Tableau 23** et **Tableau 24** ci-dessous.

$$\boldsymbol{Précision} = \frac{reformulations\ en\ commun\ Annot1\ \&\ Annot2}{phrases\ extraites\ automatiquement}$$

Figure 10. Formule de calcul de la précision

$$\boldsymbol{Rappel} = \frac{reformulations\ en\ commun\ Annot1\ \&\ Annot2}{reformulations\ Annot1}$$

Figure 11. Formule1 de calcul du rappel de l'annotation

$$\boldsymbol{Rappel} = \frac{reformulations\ en\ commun\ Annot1\ \&\ Annot2}{reformulations\ Annot2}$$

Figure 12. Formule2 de calcul du rappel de l'annotation

$$\boldsymbol{Moyenne\ rappel} = \frac{(rappel\ Annot1 + rappel\ Annot2)}{2}$$

Figure 13. Formule de calcul de la moyenne du rappel

Pour le calcul de la précision et du rappel pour la balise ***oui*** nous incluons les balises ***oui*** et ***oui<inv>***, car ces phrases correspondent aux phrases initiales. En revanche, nous n'incluons pas les balises ***oui+2*** et les ***oui<inv>+2***, parce que ces reformulations sont issues de mêmes phrases (des phrases qui contiennent deux ou plusieurs reformulations correctes). Par conséquent, ces reformulations sont *en plus* du nombre initial de phrases annotées, comme le montrent les chiffres des **Tableau 21** et **Tableau 22** (le nombre initial de phrases ensemble avec le nombre total de reformulations identifiées par chaque annotateur dans ces phrases initiales).

CLEAR SP (textes scientifiques)			
Mesures statistiques (%)	**Annot 1**	**Annot 2**	**Désaccord**
Précision - ***oui***	52,78%	43,90%	8,88%
Précision - ***oui moyenne***	**48,34%**		
Précision - ***non***	47,21%	56,09%	8,88%
Précision - ***non moyenne***	**51,65%**		
Précision - moyenne générale	**56,09%**		
Rappel - ***oui***	43,14%	66,62%	23,48%
Rappel - ***oui moyenne***	**54,88%**		
Rappel - ***non***	70,58%	59,40%	11,18%
Rappel - ***non moyenne***	**64,99%**		
Rappel - moyenne générale	**59,93%**		

Tableau 23. Mesures statistiques (précision, rappel) sur les annotations du corpus CLEAR SP

Le **Tableau 23** et le **Tableau 24** illustrent les valeurs des précisions et rappels pour chaque type de texte. Les pourcentages de **désaccord** entre les deux annotateurs sont plus importants pour le corpus expert, **CLEAR SP** (moyennes, pour la précision : 8,88%, pour le rappel : 17,26%) ce qui peut être justifié par l'utilisation d'un langage scientifique et la présence d'un grand nombre de mots techniques difficiles à interpréter par les annotateurs non-spécialistes.

CLEAR GP (textes grand public)			
Mesures statistiques (%)	**Annot 1**	**Annot 2**	**Désaccord**
Précision - ***oui***	36,73%	40,20%	3,47%
Précision - ***oui moyenne***	**38,46%**		
Précision - ***non***	63,26%	59,09%	4,17%
Précision - ***non moyenne***	**61,17%**		
Précision - moyenne générale	**85,95%**		
Rappel - ***oui***	82,62%	75,50%	7,12%
Rappel - ***oui moyenne***	**79,06%**		
Rappel - ***non***	87,88%	94,09%	6,21%
Rappel - ***non moyenne***	**90,98%**		
Rappel - moyenne générale	**85,02%**		

Tableau 24. Mesures statistiques (précision, rappel) sur les annotations du corpus CLEAR GP

Nous observons que la précision a des valeurs plus importantes pour le corpus expert (**48,34%**), par rapport au corpus grand public (**38,46%**), ce qui se justifie par la présence de reformulations plus faciles à identifier lors de l'annotation, telles que les dénominations et les abréviations (ce qui n'implique pas forcément qu'elles soient plus faciles à comprendre). Le rappel bas pour le corpus **CLEAR SP** (**54,88%**) souligne la difficulté d'identification les reformulations correctes à cause du langage technique qui pose de soucis de compréhension aux annotateurs. Pour le corpus **CLEAR GP**, un rappel moyen de **79,06%** pour les reformulations correctes prouve qu'une grande partie des reformulations correctes ont été identifiés par les deux annotateurs. Notre hypothèse soutient que ces résultats sont justifiés par des phrases avec un langage plus facile à comprendre pour les deux annotateurs non-spécialistes du domaine de la médecine.

Nous continuons notre analyse avec les accords inter-annotateurs pour chaque type de texte.

1.1.3.1.2 Accord inter-annotateur

Nous avons calculé **le score inter-annotateur** de type Kappa (Cohen, 1960) sur l'intégralité des annotations sur les deux corpus avec la formule suivante :

$$\boldsymbol{Kappa} = \frac{p_r - p_e}{1 - p_e}$$

Dont $\boldsymbol{p_r}$ représente la **probabilité réelle**, c'est-à-dire le nombre total d'accords d'annotation entre annotateurs (pour les balises « oui » et « non ») et $\boldsymbol{p_e}$ représente la **probabilité attendue** sur le fait que les annotateurs soient d'accord dans l'annotation. Le calcul de la probabilité réelle des annotations est fait avec la formule :

$$\boldsymbol{p_r} = \frac{(Annot1\ \&Annot2\ balises\ \boldsymbol{oui}) + (Annot1\ \&Annot2\ balises\ \boldsymbol{non})}{N°\ total\ de\ reformulations}$$

Nous calculons la probabilité estimée ($\boldsymbol{p_e}$) avec la formule suivante :

$$\boldsymbol{p_e} = \left(Probabilité_{Annot1\ \boldsymbol{oui}} * Probabilité_{Annot2\ \boldsymbol{oui}} \right) + \left(Probabilité_{Annot1\ \boldsymbol{non}} * Probabilité_{Annot2\ \boldsymbol{non}} \right)$$

Les deux probabilités (pour les balises « oui » et « non ») sont calculées en multipliant les probabilités des deux annotateurs pour les deux balises, respectivement.

Nous utilisons l'outil **ReCal** (Freelon, 2013) qui nous permet de calculer plusieurs types de **scores inter-annotateurs** (en utilisant des valeurs ordinales, de type ordinal, intervalle et ratio) pour les deux annotations. Nous avons donné des valeurs numériques à nos annotations, par exemple 1 pour « oui », 2 pour « oui<inv> » et 3 pour « non ».

Le **score inter-annotateur Kappa** pour le corpus **CLEAR SP** est de **0,58**, qui est un **accord modéré**, et l'accord inter-annotateur pour **CLEAR GP** est de **0,64**, un accord également **modéré**, mais plus élevé (McHugh, 2012). Nous supposons que ces différences de score sont dues au niveau de technicité plus élevé du corpus expert, rendant ainsi plus difficile l'évaluation des mêmes balises pour les reformulations par les deux annotateurs. Le score plus important du corpus grand public se justifie par le fait que le langage des textes destinés au grand public soit plus facile à comprendre, et par conséquent, à annoter, pour les deux annotateurs non-spécialistes.

Données	**CLEAR SP**	**CLEAR GP**
Taille	23408 bytes	19753 bytes
N° annotateurs	2	2
N° phrases	4681	3950
N° décisions	**9362**	**7900**

Tableau 25. Données pour calculer le score inter-annotateur avec l'outil ReCal

ReCal (Freelon, 2013)	CLEAR EX	CLEAR GP
Scores inter-annotateur	**Score**	**Score**
Krippendorff's alpha (nominal)	0.571	0.687
Krippendorff's alpha (ordinal)	**0.583**	**0.644**
Krippendorff's alpha (intervalle)	0.583	0.621
Krippendorff's alpha (ratio)	0.580	0.654

Tableau 26. Score inter-annotateur avec l'outil ReCal

Nous évaluons ultérieurement la lisibilité et la compréhension des reformulations identifiées pour tester si cette hypothèse est vérifiée par des données quantifiables.

1.1.3.1.3 Adjudication entre les deux annotations

Nous avons analysé les deux annotations et nous avons remarqué certaines erreurs dans l'application du guide d'annotation. Les plus fréquentes sont :

- L'hésitation concernant **l'annotation des abréviations** comme des reformulations, de type « maladie de Parkinson (MP) » ;
- **La reprise anaphorique** d'un terme médical avec son **hyperonyme médical générique**, de type « grippe » repris avec « cette maladie ».

Nous avons décidé d'annoter ces reformulations avec ***oui*** pour le cas des abréviations et ***non*** pour le cas des reprises anaphoriques par l'hyperonyme médical sans aucune extension, de type « cette maladie ». Dans notre exemple, si le terme médical « grippe » était repris avec « maladie infectieuse », nous aurions considéré cette reformulation comme valide.

1.1.3.2 Analyse quantitative des termes médicaux reformulés

Pour le corpus **CLEAR SP**, le nombre de **termes médicaux reformulés identifiés par les deux annotateurs** est de **2 535,** dont **681 (26,86%) termes et leurs reformulations** ont été identifiés dans des phrases qui contiennent deux ou plusieurs reformulations correctes. Concernant le **corpus CLEAR GP**, nous avons une liste de **2 668 termes médicaux reformulés identifiés par les deux annotateurs, dont 901 (33,77%)** dans des contextes de **reformulations multiples** (plusieurs par phrase).

Pour avoir une liste propre de termes reformulés facilement exploitables en TAL, nous supprimons tous les déterminants[73] qui ont été annotés ensemble avec les termes médicaux. Nous commençons par la liste de **2 535** termes de **CLEAR SP** avec des expressions régulières de type « **^la\s** » pour indiquer la présence de l'article (avec minuscule et majuscule) en début de ligne (**^**) et suivi par un espace (**\s**). Nous avons identifié **1 843 déterminants**, dont les types et occurrences sont illustrés dans le **Tableau 27**.

Type	Genre	Nombre	Forme	N° occurrences	%
L'article défini	masculin	singulier	***l'***	245	13,29%
		singulier	***le***	232	12,58%
		pluriel	***les***	305	16,54%
	féminin	singulier	***la***	760	41,23%
L'article indéfini	masculin	singulier	***un***	72	3,90%
	masculin féminin	pluriel	***des***	125	6,78%
	féminin	singulier	***une***	104	5,64%
Total				**1843**	**100%**

Tableau 27. Déterminants de type article défini et indéfini des termes médicaux du corpus CLEAR SP

Parmi nos termes reformulés du corpus CLEAR SP, nous avons identifié :

- **656 (25,87%) termes médicaux simples ;**
- **1 879 (74,12%) termes médicaux polylexicaux.**

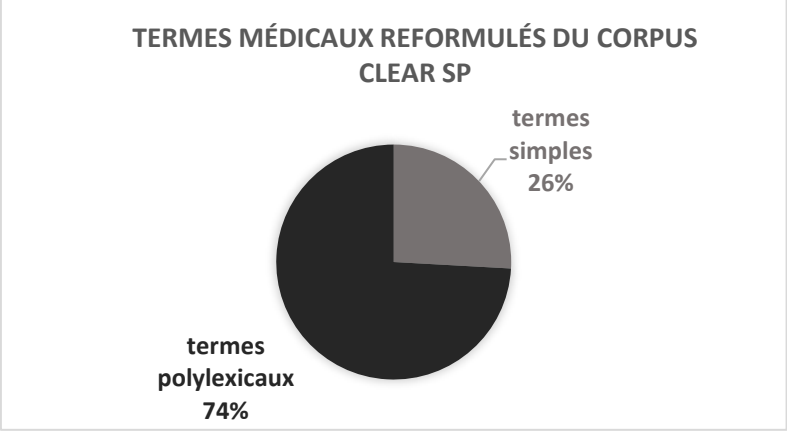

Figure 14. Types de termes médicaux reformulés extraits du corpus CLEAR SP

[73] Seuls les articles définis et indéfinis étaient annotés, d'autres types de déterminants, comme les déterminants possessifs ou démonstratifs, n'étaient pas présents dans les listes de termes.

Nous analysons la structure des **termes médicaux polylexicaux avec des expressions régulières** dans l'éditeur de texte Notepad++. Nous identifions les types suivants de termes polylexicaux :

- **629 (24,81%) termes bi-grammes** de type Nom-Adjectif (par exemple : *hystérectomie vaginale ; démence vasculaire ; embolie pulmonaire)* (expression régulière : ^\w+\s\w+$) ;
- **1 250 (49,30%) termes tri-grammes** ou **de plus de quatre tokens** de type Nom-Préposition-Nom et Nom-Adjectif-Adjectif (par exemple : *déficit cognitif léger ; sclérose en plaques ; ataxie de Freidreich*) (expression régulière : ^\w+\s\w+\s\w+).

Nous avons retrouvé **953 (33,60%) termes reformulés doublons** (qui se répètent plusieurs fois) dans notre liste, ce qui nous donne une liste de **1 582 termes médicaux *uniques* reformulés** pour le corpus **CLEAR SP**. Pourtant, nous avons besoin des toutes les occurrences de termes pour analyser les reformulations (qui peuvent être différentes). Nous présentons cette analyse de reformulations plus loin dans ce chapitre.

Un extrait de la liste de **termes simples et polylexicaux reformulés pour le corpus CLEAR SP** (sans doublons et en ordre alphabétique) se trouve dans **l'Annexe 6.4.**

****

Nous avons réalisé le même traitement pour la liste de **2 668 termes médicaux reformulés** du corpus **CLEAR GP**. Nous avons supprimé **1 793** déterminants définis et indéfinis (« la » avec 626 occurrences (34,91%) ; « l' » avec 284 (15,83%); « le » : 236 (13,16%) ; « les » : 249 (13,88%) ; « un » : 100 (5,57%) ; « une » : 123 (6,86%) ; « des » : 161 (8,97%) ; « d' » : 14 (0,78%)) pour les mêmes raisons que pour le corpus **CLEAR SP** : avoir une liste alphabétique facile à exploiter dans d'autres applications de TAL.

Parmi nos termes reformulés du corpus **CLEAR GP,** nous avons identifié :

- **826 (30,95%) termes médicaux simples ;**
- **1 842 (69,04%) termes médicaux polylexicaux.**

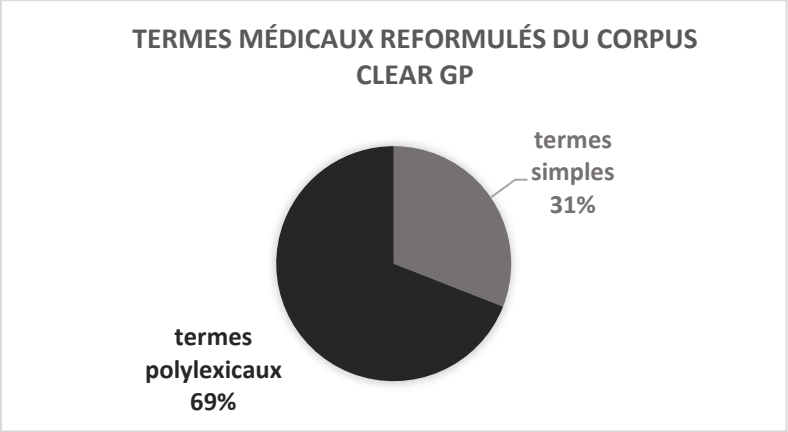

Figure 15. Types de termes médicaux reformulés extraits du corpus CLEAR GP

Nous analysons les **termes médicaux polylexicaux avec des expressions régulières**. Nous identifions plusieurs types de termes polylexicaux :

- **521 (19,52%) termes bi-grammes** de type Nom-Adjectif (par exemple : *cystite bactérienne ; neuropathie périphérique ; méralgie paresthésique)* (expression régulière : ^\w+\s\w+$) ;
- **335 (12,55%) termes tri-grammes** de type Nom-Préposition-Nom et Nom-Adjectif-Adjectif (*facteur neurotrophique ciliaire ; hémorragies digestives hautes ; maladie de Basedow*) (expression régulière : ^\w+\s\w+\s\w+$) ;
- **485 (18,17%) termes avec quatre tokens ou plus** (par exemple : *néphropathie liée au VIH ; syndrome de détresse respiratoire ; trouble déficitaire de l'attention avec hyperactivité*) (expression régulière : ^\w+\s\w+\s\w+\s\w+).

Nous avons retrouvé **988 (32,23%) termes doublons** dans notre liste, ce qui nous donne une liste de **1 680 termes médicaux *uniques* reformulés** pour le corpus **CLEAR GP**.

Un extrait de la liste de **termes simples et polylexicaux reformulés pour le corpus CLEAR GP** (sans doublons et en ordre alphabétique) se trouve dans **l'Annexe 6.5.**

****

Nous observons que la structure des termes médicaux reformulés diffère par rapport aux types de corpus :

- nous identifions plus de termes simples dans le texte grand public, **30,95%**, par rapport au **25,87%** dans le texte scientifique ;

- les termes bi-grammes de type Nom-Adjectif sont plus fréquents **(24,81%)** dans le corpus scientifique que dans le corpus grand public **(19,52%)** ;
- **49,30%** de termes polylexicaux sont complexes (tri-grammes ou plus) dans CLEAR SP par rapport au **30,72%** dans le CLEAR GP.

Ces différences prouvent que dans le corpus CLEAR GP il a une intention plus importante à reformuler même **les termes simples**, pour les rendre accessibles au grand public, tandis que dans le corpus expert nous identifions moins de termes simples qui sont reformulés. Pourtant, une plus importante partie des termes reformulés sont **complexes** (de type tri-grammes ou plus longs).

Nous continuons notre analyse avec les types de marqueurs et indicateurs de reformulation identifiés dans les deux corpus.

1.1.3.3 Analyse quantitative des marqueurs et indicateurs de reformulation

Nous analysons les marqueurs et indicateurs de reformulations et nous faisons les remarques suivantes pour le corpus **CLEAR SP** :

- **1 076 (42,44%)** reformulations apparaissent entre parenthèses **()**, de type **abréviations** (médicaments anti-inflammatoires non stéroïdiens *(AINS*)) ou **énumérations / exemplifications** (*technique d'anesthésie (anesthésie générale, sédation / analgésie, procédures de blocage régional ou paracervical (BPC))* ;
- **144 (5,68%) reformulations** marquées avec le **lemme « défini » (défini comme ; défini par ; définition)** (*La cataracte **définie comme** une perte de transparence du cristallin naturel)* ;
- **62 (2,44%)** reformulations sont délimitées par des parenthèses ensemble avec un indicateur de type ***maladie, trouble, affection*** et des marqueurs comme ***en particulier, c'est-à-dire, ou d'autres***, de type : *(un antagoniste de la nicotine (c'est-à-dire bloque l'effet de la nicotine))* ;
- **uniquement 17 (0,67%) reformulations** sont marquées à l'aide du **lemme « appelé »** *(aussi appelé, ce que l'on appelle, également appelé, généralement appelé)* (*La névralgie amyotrophiante (**également appelée** syndrome de Parsonage-Turner))*.

***

Notre analyse sur le corpus **CLEAR GP** nous permet d'extraire les observations suivantes sur les **marqueurs et indicateurs de reformulation** les plus fréquents :

- **1 023 (38,34%)** reformulations sont marquées par des parenthèses **()**, de type :
  - **abréviations** : *lupus érythémateux disséminé (LED)* ;
  - **énumérations/exemplifications** : la spasticité **(**crampes et spasmes musculaires**)** ;
  - **explication** : *hémorragie intraventriculaire **(**des saignements dans le cerveau**)*** ;
  - **paraphrase** : *chimiothérapie **(**médicaments anticancéreux**)***.
- **101 (3,78%)** reformulations sont délimitées par des parenthèses ensemble avec un indicateur de type ***maladie, trouble, affection*** et des marqueurs comme ***également connue sous le nom de, c'est-à-dire, ou d'autres***, ***comme*** de type : *La trisomie 21 **(également connue sous le nom de** syndrome de Down**)*** ;
- **44 (1,64%) reformulations** sont marquées avec le **lemme « défini » (défini comme ; défini par ; définition)** (*Un pneumothorax est **défini comme** la présence d'air dans l'espace entre les deux parois des poumons)* ;
- **31 (1,16%) reformulations** sont marquées à l'aide du **lemme « appelé »** *(aussi appelé, ce que l'on appelle, également appelé)* (*Le paludisme, **aussi appelé** malaria)*.

Nous observons que, en moyenne, **40,39%** de reformulations sont marquées par des **parenthèses**. Ceci prouve que le travail d'identification de la reformulation dans les textes médicaux expert et grand public doit viser l'analyse du *texte placé entre parenthèses*, en plus de l'identification des marqueurs lexicaux ou grammaticaux. Nous identifions deux types **d'abréviations** : les abréviations plus connues que le terme comme *AVC (accident vasculaire cérébral)*, *VIH (virus de l'immunodéficience humaine)*, et celles qui ne sont pas connues par le grand public, comme dans le cas de *lupus érythémateux disséminé (LED)*. Dans le corpus grand public, les **explications** et les **paraphrases** comme « hémorragie intraventriculaire (des saignements dans le cerveau) » et « chimiothérapie (médicaments anticancéreux) » rendent accessible le sens des termes médicaux techniques à l'aide des mots de la langue générale (*saignements, cerveau, médicaments*).

1.1.3.3.1 Nouveaux marqueurs et indicateurs de reformulations identifiés

Nous avons identifié des **nouveaux marqueurs et indicateurs de reformulation** (en plus de ceux présentés dans le **Tableau 18, sous-chapitre 1.1.2**) lors de notre analyse manuelle de phrases du corpus **CLEAR SP** et **CLEAR GP**, en total **5 203 phrases** :

- ***nouveaux indicateurs de reformulation***, comme « et autres », « et les autres », « ou autres », « ou d'autres » qui indiquent 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 », « y compris » (« L'aphasie *décrit* un trouble du langage associé à une lésion cérébrale »).

Nous notons la **fréquence élevée** sur les nouveaux marqueurs et indicateurs de reformulation médicale, dans le **Tableau 28** ci-dessous. La conjonction disjonctive **« ou »** aide à identifier **46 (0,88%) reformulations correctes** (toute seule ou dans des constructions de types ***« ou autres », « ou d'autres »*** (« schizophrénie *ou autre* maladie mentale grave »).

Nouveaux Marqueurs / Indicateurs	N° occurrences	%
parenthèses ()	2 099	**40,34%**
par exemple	232	4,45%
comme	201	3,86%
notamment	50	0,96%
ou (ou d'/autres)	46	0,88%
y compris	31	0,59%

Tableau 28. Fréquences et occurrences de nouveaux marqueurs ou indicateurs de reformulation identifiés dans le corpus CLEAR

Nous continuons notre analyse des **relations et fonctions** des reformulations médicales correctes pour vérifier les hypothèses annoncées dans la **section 2.5** et nous interprétons les annotations effectuées.

1.1.3.4 Analyse lexicale et sémantico-pragmatique des reformulations

L'approche théorique sur les relations lexicales, les fonctions sémantico-pragmatique et notre hypothèse concernant les liens qui peuvent exister entre celles-ci sont présentées en détail dans la **Partie III, Chapitre 2.5.**

Nous analysons manuellement nos données pour identifier ces différents types de relations lexicales (hyperonymie, hyponymie, synonymie, méronymie) et le lien avec la fonction sémantico-pragmatique. Nous classons les marqueurs et indicateurs par rapport aux relations et fonctions annotées, afin de confirmer ou infirmer les hypothèses présentées ci-dessus.

1.1.3.4.1 Hypothèse de recherche sur le marqueur « est un / une »

Notre première **hypothèse de recherche** concerne **le rôle du marqueur de reformulation** *« est un / une »* qui permet de mettre en évidence simultanément une :

- relation lexicale **d'hyperonymie** et
- une relation sémantico-pragmatique de **définition**.

Afin de tester la validité de cette hypothèse, nous annotons tous les marqueurs génériques « est un » et « est une » dans nos deux corpus, **CLEAR SP** et **CLEAR GP.**

Lors de nos recherches de marqueurs de reformulation, nous avons utilisé les formes *« est une affection », « est une maladie », « est un trouble »*, mais aussi les indicateurs de reformulations médicales *« affection/s », « maladie/s », « trouble/s »* séparément. Nous avons décidé de ne pas chercher le marqueur générique isolé *« est un »* pour ne pas avoir trop de résultats erronés.

Pour tester l'hypothèse, nous mettons les balises **<mdef>***est un***</mdef>** autour du marqueur générique. Nous avons choisi cette abréviation pour indiquer **m**- marqueur et **def**-définition, car notre hypothèse affirme aussi que toutes ces phrases auront à la fois :

- la **relation sémantico-pragmatique de définition**, ce qui veut dire que l'intention du locuteur et le but du texte sont de **donner une définition plus simple** au terme scientifique médical ;
  et

- la **relation lexicale d'hyperonymie**, ce qui veut dire que la **reformulation médicale est le terme générique** (*l'hyperonyme*) du terme médical, qui est à son tour le **terme spécifique** (*l'hyponyme*).

Pour extraire automatiquement les termes médicaux et les reformulations médicales de ces phrases, nous avons suivi la procédure suivante :

- nous avons introduit un élément séparateur (un TAB) avant et après le marqueur ;
- nous avons utilisé ce séparateur de colonne pour séparer automatiquement le terme médical et la reformulation qui suit le marqueur « est un/e ».

Ce traitement automatique est suivi par une validation manuelle de cette hypothèse. Chaque phrase a été évaluée et validée si elle correspond à l'hypothèse présentée. Dans le corpus de textes scientifiques, CLEAR SP, nous identifions **391** phrases qui contiennent **379** marqueurs génériques, dont **141** au masculin « est un » et **256** au féminin, « est une ». Dans le corpus de textes destinés au grand public, **CLEAR GP**, nous identifions **481** phrases qui contiennent des marqueurs génériques, dont **222** au masculin « est un » et **259** au féminin, « est une ».

Notre prochaine étape de traitement vise 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.

1.1.3.4.2 Analyse préliminaire du lien entre les marqueurs et les types de reformulations

Nous avons analysé le lien entre les marqueurs et les types de reformulation sur l'échantillon de **1 000** phrases du corpus de textes scientifiques, **CLEAR SP**, présenté dans la **section 1.1.2.1**. À la suite de cette analyse, nous avons identifié 314 reformulations médicales correctes, dont au moins un terme médical est reformulé.

Le **Tableau 29** 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 cet échantillon de 1 000 phrases du corpus scientifique sont les reformulations de type *définition*, dont **211** (**21,10%**) ont la relation lexicale *d'hyperonymie / hyponymie*, suivie par les *exemplifications*, en lien avec la relation *d'hyponymie*, avec **50** (**5%**) occurrences et la *dénomination* en lien avec la *synonymie* avec **26** (**2,6%**) occurrences. La *paraphrase*, dire les choses autrement avec des

mots plus simples, est très peu présente, avec seulement **8** (**0,8%**) occurrences. Même constat pour *l'explication*, dont nous avons identifié que **5** (**0,5%**) 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.

Type de reformulation	**Relation lexicale**	**CLEAR SP**	**% de validité des hypothèses de recherche**
paraphrase	synonymie	6 / 8	75%
dénomination		26	100%
définition	hyperonymie	211 / 223	94,61%
exemplification	hyponymie	50 / 52	96,15%
explication	méronymie	1 / 5	20%
Total		**314**	**93,63%**

Tableau 29. Reformulations validées par type de fonction sémantico-pragmatique et relation lexicale.

Le **Tableau 29** montre la répartition selon nos hypothèses de ces 314 reformulations validées manuellement. Nous observons que la plupart des reformulations correspondent aux relations lexicales selon nos hypothèses d'annotation. 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, les autres sont plutôt des *descriptions* ou des expressions de la *causalité*, comme dans l'exemple ci-dessous :

> 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.

Analyse lexicale et sémantico-pragmatique des reformulations correctes Marqueurs et indicateurs par fréquence absolue et relative						
	synonymie	**hyperony-mie**	**hyponymie**	**méronymie/ description**	**CLEAR SP**	
					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 30. Marqueurs et indicateurs par type de fonction 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 30** présente ces éléments par rapport au type de relation lexicale et fonction sémantico-pragmatique avec leur fréquence absolue et relative dans le corpus CLEAR SP. En plus de l'analyse précédente, nous observons la présence de **50** (**15,92%**[74]) occurrences du marqueur orthographique de type *parenthèses*, le plus identifié dans des fonctions *d'exemplification* (**23** ; **7,32%**) et de *définition* (**11** ; **3,50%**).

[74] du total de 314 reformulations correctes analysées.

1.1.3.4.3 Adjudication sur les phrases avec double annotation

Les deux annotations ont été mises en commun et une harmonisation des annotations différentes a été réalisée. Lorsqu'il y avait des différences dans les deux annotations (*oui* versus *non*), nous avons choisi le statut de plus adapté de la reformulation en suivant le guide d'annotation. À l'issue de cette opération, nous avons **1 854 phrases** avec des reformulations correctes pour le corpus **CLEAR SP** et **1 767** pour le corpus **CLEAR GP**. Pour ces phrases considérées comme reformulations correctes, nous avons appliqué l'adjudication pour les annotations en relation lexicale et en fonction sémantico-pragmatique.

Relation lexicale	**Fonction sémantico-pragmatique**	**CLEAR SP**		**CLEAR GP**	
		N°	**%**	**N°**	**%**
hyperonymie	définition	**857**	**46,22%**	**874**	**49,46%**
hyponymie	exemplification	**364**	**19,63%**	**466**	**26,37%**
synonymie	paraphrase	59	3,18%	118	6,67%
	dénomination	**408**	**22%**	**177**	**10,01%**
méronymie	explication	24	1,29%	71	4,01%
	définition	59	3,18%	12	0,67%
Total		**1854**	**100%**	**1767**	**100%**

Tableau 31. Paires de relations lexicales et fonctions sémantico-pragmatiques annotées dans CLEAR SP et CLEAR GP

Le corpus **CLEAR SP** a **900 (48,54%)** reformulations annotées comme *hyperonymies*, dont **857 (46,22%)** ont été annotées comme *hyperonymie-définitions* (sur **931 (50,21%)** *définitions* en total). La relation qui se trouve en deuxième place en termes de fréquence est celle de *synonymie-dénomination* avec **408 (22%)** reformulations, suivie par *hyponymie-exemplification* avec **364 (19,63%)** reformulations. Nous observons que la fonction de *dénomination* apparait uniquement en paire avec la relation de *synonymie*. Les *méronymies* sont en nombre de **89 (4,8%)**, dont **24 (1,29%)** en avec *explication*, **59 (3,18%)** avec *définition* et seulement **6 (0,32%)** avec *exemplification* (voir **Tableau 31)**. Si la paire *méronymie-explication* coïncide avec notre hypothèse de recherche, la nouvelle association ***méronymie-définition*** est un résultat nouveau. Un exemple qui représente cette nouvelle paire est :

> <terme>**Le trouble du spectre autistique**</terme> couvre une large gamme de <ref>**problèmes comportementaux et de communication**</ref>.

Le terme « trouble du spectre autistique » est défini à travers ses symptômes, qui sont des parties intégrantes de la maladie (relation *partie-tout*).

Concernant le corpus **CLEAR GP**, l'annotation manuelle des relations lexicales et fonctions sémantico-pragmatiques est réalisée sur **1 767** phrases. Notre annotation montre que la relation lexicale d'*hyperonymie* également la plus fréquente avec **908** reformulations (**51,38%**). La fonction sémantico-pragmatique de *définition* a **893** reformulations (**50,53%**). Nous observons que **874 (49,46%)** reformulations sont annotées avec *hyperonymie-définition*, ce qui constitue **96,25%** des hyperonymies. Si dans le corpus **CLEAR SP** la relation de *synonymie-dénomination* est la deuxième en termes de fréquence, dans le corpus **CLEAR GP** nous remarquons que c'est la relation *d'hyponymie-exemplification* qui est annotée dans **26,37%** de reformulations **(466).** L'annotation *synonymie-dénomination* concerne **10%** de reformulations, deux fois moins que dans le corpus scientifique. La paire *méronymie-explication* est présente uniquement dans **71 (4,01%)** reformulations et *méronymie-définition* avec uniquement **12 (0,67%)** reformulations (voir **Tableau 31**).

1.1.4 Bilan des résultats d'annotation sur le corpus CLEAR Cochrane

Nous observons que les résultats sont différents entre **CLEAR SP** et **CLEAR GP**. L'annotation du corpus scientifique permet d'identifier des reformulations dans **48,34%** des phrases (selon la moyenne de deux annotateurs), tandis que le pourcentage pour le corpus grand public est à **38,46%**.

CLEAR Cochrane	CLEAR SP	CLEAR GP
% Reformulations	**48,34%**	**38,46%**
N° total de phrases annotées	**4 687**	**3 980**
	8 667	

Tableau 32. Score de précision de reformulations annotées dans CLEAR SP et CLEAR GP

L'objectif de notre étude étant de **construire un corpus de reformulations valide** et de permettre la vérification des **hypothèses** concernant les différences entre textes de vulgarisation et scientifiques, nous avons concentré nos efforts sur une annotation manuelle et partiellement automatique (pour les abréviations) des reformulations, marqueurs de reformulations, relations lexicales et sémantico-pragmatiques.

Les nouveaux **marqueurs et indicateurs de reformulation** identifiés dans les données annotées manuellement pour le corpus **CLEAR** (*décrit, est associé à, les parenthèses, notamment, par exemple, comme*, etc.) seront utilisés dans les expériences sur le deuxième corpus français, **ClassYN**, mais ils serviront également de modèle (et leurs traductions s'ajoutent à la liste initiale de marqueurs pour le roumain) pour compléter l'annotation du corpus roumain, **GrandMed-Ro2**. L'origine latine de nos deux langues d'étude nous permet d'appliquer nos méthodes et ressources créées à partir du corpus français sur le roumain et vice-versa.

L'annotation des **relations lexicales** et des **fonctions sémantico-pragmatiques** prouve que le type de reformulation le plus fréquent dans les deux corpus, expert et grand public, reste *l'hyperonymie-définition* (**46,22%**, respectivement **49,46%**), et que la différence est notable pour la présence des reformulations de type *synonymie-dénominations* (en nombre **deux fois plus grand** dans le corpus **CLEAR SP** que le corpus **CLEAR GP**). Cela signifie que les rédacteurs de résumés scientifiques utilisent beaucoup plus de termes médicaux ou des acronymes pour *reformuler* d'autres termes médicaux (par exemple, « sclérose latérale amyotrophique (maladie du motoneurone) », « maladie du greffon contre l'hôte (MGCH) »). En revanche, dans les textes grand public, nous avons identifié **deux fois plus de paraphrases simplifiées** de termes médicaux (**6,67%** par rapport au **3,18%** dans les textes scientifiques). Ceci soutient l'objectif de textes grand public, c'est-à-dire de *simplifier les termes médicaux* avec des paraphrases de la langue commune, de type « L'entorse cervicale (ou "coup du lapin") », « placebo (traitement simulé) », « placebo (faux médicament) ». Nous observons une **nouvelle relation**, celle de la *méronymie*, utilisée pour la *définition*, qui permet d'annoter **59 (3,18%)** reformulations dans **CLEAR SP** avec uniquement **12 (0,67%)** reformulations dans **CLEAR GP**.

Nous rappelons que notre acceptation de la **reformulation médicale** comprend *tout dire autrement* qui aide à **vulgariser** le terme médical. Même si des acronymes de type « MGCH » ne sont pas plus faciles à comprendre que le terme médical, ils restent tout aussi importants pour les expériences automatiques de générations de reformulations. Lors de ces expériences, toute forme de reformulation d'un certain terme est utile pour l'apprentissage automatique de l'algorithme. Nous présentons ces expériences automatiques dans le **Chapitre 3.**

***

Dans le prochain chapitre, nous analysons les expériences et les résultats des annotations automatiques et manuelles menées sur le deuxième corpus français de notre étude, **ClassYN** (Todirascu *et al.*, 2012).

1.2 ClassYN

Nous présentons les expériences réalisées sur le corpus **ClassYN** (Todirascu *et al.*, 2012). Ce corpus comparable comprend des textes scientifiques et des textes de vulgarisation scientifique du domaine de la médecine. Nous rappelons les tailles des deux corpus dans le **Tableau 33** ci-dessous[75].

Corpus	Taille (caractères)	Taille (tokens)	Taille (textes)
ClassYN SP Total	6 903 424	1 007049	300
ClassYN GP Total	4 879 939	772 374	300

Tableau 33. Taille des corpus : ClassYN SP Total (littérature scientifique médicale) et ClassYN GP Total (textes de vulgarisation médicale)

1.2.1 Résultats de l'annotation automatique des termes médicaux

Le **Tableau 34** montre le nombre de phrases avec et sans termes médicaux identifiés dans les deux corpus, ClassYN SP Total et ClassYN GP Total. Nous observons que l'annotateur SIFR-BioPortal a identifié plus de phrases avec des termes médicaux dans le corpus grand public que dans le corpus scientifique (**73%** des phrases dans le ClassYN GP Total, par rapport au **55%** dans le ClassYN SP Total). À partir de ce moment, nous travaillons uniquement sur les **phrases extraites automatiquement** et qui contiennent des **termes médicaux** de la terminologie. Nous appelons désormais ces corpus de phrases **ClassYN SP** et **ClassYN GP**.

ClassYN	N° total des phrases	N° de phrases sans termes médicaux	N° de phrases avec termes médicaux
ClassYN SP Total	50 206	22 623	27 583
ClassYN GP Total	39 818	10 835	28 983

Tableau 34. Phrases avec ou sans termes médicaux dans le corpus ClassYN

Nous avons identifié automatiquement **77 352 termes médicaux** (dont **4 676 termes uniques**, sans lemmatisation) dans le corpus scientifique **ClassYN SP Total** et **77 284**

[75] La description détaillée du corpus **ClassYN** se trouve dans la **Partie III**, **Chapitre 2.1.1**.

termes médicaux (dont **4 153 termes uniques**, sans lemmatisation) dans le corpus **ClassYN GP Total** de textes grand public (voir la méthodologie présentée dans la **Partie III**, **sous-chapitre 2.** ; voir **Tableau 35**).

1.2.1.1 Post-traitement de l'annotation de termes médicaux

Nous éliminons les **doublons** et nous analysons manuellement la liste de résultats afin de trouver les mots qui n'ont pas le potentiel d'être des termes médicaux, mais qui appartiennent au registre de langue générale, comme « plus », « souvent », « deuxièmement ». Nous trouvons 453 mots de la langue générale pour le corpus ClassYN SP et 302 pour le corpus ClassYN GP. Nous rajoutons la liste de mots qui ne sont pas de termes médicaux issue du corpus CLEAR (disponible dans **l'Annexe 6.3**) pour nos recherches. Une fois ces mots supprimés, nous avons **4 223** termes médicaux uniques pour le corpus scientifique et **3 852** pour le corpus de vulgarisation. La précision de l'extracteur est alors de 0,90 pour ClassYN SP et 0,92 pour ClassYN GP, par rapport aux termes annotés automatiquement.

Corpus	Total termes trouvés	Termes mal identifiés	Total termes validés	Nb de doublons	Termes uniques	Termes uniques validés	P
ClassYN SP	77352	453	47389	42713	4676	4223	0,90
ClassYN GP	77284	302	58533	54380	4153	3852	0,92

Tableau 35. Termes médicaux identifiés par l'annotateur SIFR-BioPortal dans le corpus ClassYN (P : précision des termes uniques validés)

1.2.1.2 Les types des termes médicaux : simples et polylexicaux

Parmi tous les termes uniques identifiés, nous calculons le nombre de termes monolexicaux et polylexicaux. Pour cela nous utilisons l'expression régulière (\b[\w]*)(\s|(\-))(\b[\w]*) pour identifier, dans la liste de termes uniques, les termes polylexicaux dont les mots qui les composent sont séparés par des espaces (\s) ou des tirets (\-).

Nous réalisons une validation manuelle des termes polylexicaux. Nous avons remarqué la présence des candidats composés d'un article défini (« la cellule »), indéfini (« une intensité ») ou introduits par la préposition « de » (« risque de », « troubles de l ») qui sont en fait des termes simples identifiés par SIFR-BioPortal ensemble avec leurs déterminants, et non pas des termes polylexicaux. Nous annotons 39 termes incorrects dans

le corpus ClassYN SP et 53 dans le corpus ClassYN GP, ce qui nous donne une précision de **0,96** pour la validation des termes polylexicaux de deux types de corpus.

Corpus	Total termes uniques validés	Termes simples uniques	Termes polylexicaux uniques	Termes polylexicaux incorrects	Termes polylexicaux validés	P
ClassYN SP	4223	2947	1276	39	1237	0,96
ClassYN GP	3852	2438	1413	53	1360	0,96

Tableau 36. Termes médicaux monolexicaux et polylexicaux uniques identifiés par l'annotateur SIFR-BioPortal (P : précision des termes polylexicaux uniques validés)

Nous éliminons les 453 mots de la langue générale identifiés pour ClassYN SP et respectivement les 302 dans ClassYN GP avec une recherche automatique dans toutes leurs occurrences dans les listes des termes annotés par SIFR-BioPortal. Nous sommes intéressés par le nombre total de termes trouvés, même avec doublons, car cette information nous indique la prépondérance des sujets médicaux dans nos corpus. Parmi les 47 389 termes trouvés dans ClassYN SP et les 58 533 termes du ClassYN GP, nous observons **les termes qui apparaissent le plus fréquemment dans les corpus**[76] (termes et fréquences détaillés dans le **Tableau 37**.

Les termes observables dans la **Figure 16** et la **Figure 17**[77] nous montrent que les sujets médicaux abordés sont différents et que le registre langagier est en concordance avec le type de corpus. Dans le corpus scientifique, les termes médicaux les plus fréquents sont assez **techniques** (« adn », « anticorps », « cellulaire »), tandis que dans le corpus destiné au grand public les termes sont plus facilement **compréhensibles** et représentent des **questionnements médicaux populaires**, comme « cancer », « sang », « douleur ».

Les dix premiers termes les plus fréquents sont très différents dans les deux corpus, à l'exception de deux termes :

- **« maladie »**, avec une fréquence absolue (FA) de 832 (fréquence relative en pourcentage (FRP) de 0,0826) dans le corpus de littérature médicale scientifique et beaucoup plus fréquent (FA : 1977 ; FRP : 0,2560) dans le corpus de textes de vulgarisation ;

[76] Classement par fréquence des mots réalisé avec le concordancier AntConc (Anthony, 2020).
[77] Nuages des mots réalisés avec l'application web https://www.nuagesdemots.fr/

- **« virus »**, qui est presque deux fois plus présent dans le corpus scientifique que celui de vulgarisation selon les valeurs FA de 778 et 437 respectivement, mais les valeurs FRP restent très proches, 0,0773, respectivement 0,0612.

N°	ClassYN SP			ClassYN GP		
	Terme monolexical	**Nb occ**	**Fréquence relative %**	**Terme monolexical**	**Nb occ**	**Fréquence relative %**
1	maladie	832	0,0826	maladie	1977	0,2560
2	virus	778	0,0773	cancer	1237	0,1602
3	adn	761	0,0756	sang	668	0,0865
4	protéines	668	0,0663	prostate	606	0,0785
5	syndrome	554	0,0550	douleur	578	0,0748
6	expression	549	0,0545	effet	546	0,0707
7	protéine	502	0,0498	peau	510	0,0660
8	infection	492	0,0489	chirurgie	489	0,0633
9	anticorps	431	0,0428	virus	473	0,0612
10	cellulaire	404	0,0401	cardiaque	437	0,0566

Tableau 37. Termes médicaux monolexicaux les plus fréquents dans les deux corpus (Nb occ : nombre d'occurrences)

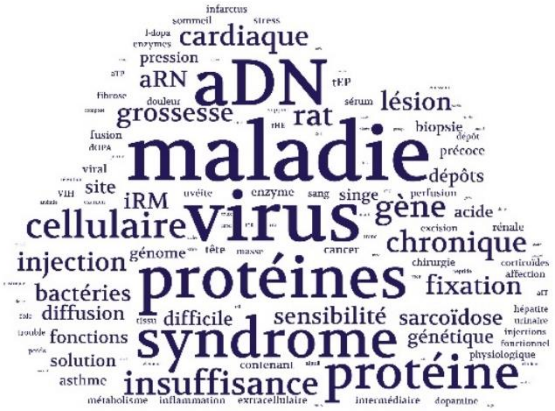

Figure 16. Nuage de termes médicaux-ClassYN SP

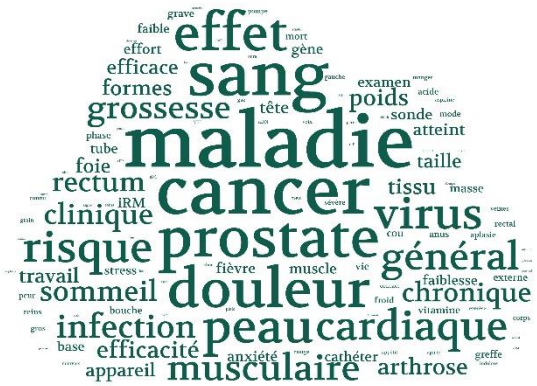

Figure 17. Nuage de termes médicaux-ClassYN GP

Concernant les **termes polylexicaux** (**Tableau 38**), nous observons que dans les textes scientifiques, le terme « l-dopa » (traitement pour la maladie de Parkinson) est le plus fréquent, suivi par « l'insuffisance cardiaque » avec une fréquence absolue (FA) de 50 (fréquence relative en pourcentage (FRP) de 0,0085) et la « fonction rénale » avec FA 47 (0,0047 FRP). Dans le corpus de vulgarisation, le terme polylexical le plus fréquent est « intervention chirurgicale », avec une FA de 284 (0,0368 FRP). Ce terme est assez accessible à la compréhension pour le grand public, tandis que le sens du deuxième terme polylexical le plus fréquent, « imagerie par résonance magnétique » est moins transparent. Nous remarquons le terme « mode de vie » avec une FRP assez importante (0,0111), ce

qui indique que les textes de vulgarisation mettent en avant le « mode de vie » en lien avec les besoins de santé.

N°	ClassYN SP			ClassYN GP		
	Terme polylexical	**Nb occ**	**Fréq relative %**	**Terme polylexical**	**Nb occ**	**Fréq relative %**
1	l-dopa	86	0,0085	intervention chirurgicale	284	0,0368
2	insuffisance cardiaque	50	0,0050	imagerie par résonance magnétique	104	0,0135
3	fonction rénale	47	0,0047	vaisseaux sanguins	103	0,0133
4	acuité visuelle	40	0,0040	examen clinique	96	0,0124
5	examen clinique	39	0,0039	médecin spécialiste	90	0,0117
6	système nerveux central	38	0,0038	mode de vie	86	0,0111
7	post-partum	37	0,0037	rapport sexuel	70	0,0091
8	embolie pulmonaire	33	0,0033	hypertension artérielle	68	0,0088
9	acides aminés	33	0,0033	anesthésie générale	64	0,0083
10	membres inférieurs	32	0,0032	anesthésie locale	57	0,0074

Tableau 38. Termes médicaux polylexicaux les plus fréquents dans les deux corpus (Nb occ : nombre d'occurrences)

Nous avons omis de cette liste les termes polylexicaux non-validés pour les raisons présentées plus haut, comme « la cellule » (FA 54, FRP 0,0053) et « une intensité » (FA 37, FRP 0,0036), dans ClassYN SP et « risque de » (le terme le plus fréquent avec une FA de 403 et une FRP de 0,0521), « de la prostate » (FA 267, FRP 0,0345), « absence de » (FA 112, FRP 0,0145), « la cellule » (FA 69, FRP 0,0089), etc. dans le corpus ClassYN GP.

1.2.1.3 Patrons morphosyntaxiques des termes médicaux

Les termes médicaux identifiés et extraits à partir des corpus ClassYN se trouvent sous la forme de plusieurs patrons. En plus de la classification binaire **termes simples / polylexicaux**, nous analysons aussi les **patrons morphosyntaxiques** des termes polylexicaux. Nous avons identifié plusieurs patrons, généralement associés aux termes du domaine : nom / adjectif, (N / Adj), nom / nom / adjectif (N / N / Adj), nom / préposition / nom (N / Prep / N), etc.

À l'aide des expressions régulières, nous avons extrait des patrons composés de 2 à 5 mots à l'aide des expressions régulières. Nous avons ensuite classé manuellement les exemples qui correspondaient aux patrons morphosyntaxiques les plus fréquents :

- **N / Adj** : expression régulière ^\w+(\s|-)\w+$; exemples : « nodule cutané », « spondylarthrite ankylosante », « rhumatisme psoriasique » ;
- **N / Adj / Adj** : expression régulière ^\w+(\s|-)(\w{4,30})(\s|-)\w+$; exemples : « cellules germinales primordiales », « noyau sous-thalamique », « système nerveux central » ;
- **N / Prep / N (de/des)** : expression régulière ^\w+\s(de|des)\s\w+$; exemples : « acétate de prednisolone », « ganglion de Gasser », « maladie de Scheuermann » ;
- **N / Prep / N (par)** : expression régulière ^\w+\s(par)\s\w+$; exemple unique « accouchement par forceps » ;
- **N / Prep / N / Adj (de/des)** : expression régulière ^\w+\s(de|des)\s\w+(\s|-)\w+$; exemples : « test de charge hydrique », « thrombophlébite des membres inférieurs ». Nous remarquons que nous trouvons **9** (**0,33%**) exemples de ce type avec des noms propres attachés par des tirets, comme « syndrome de Lennox-Gastaut ». Nous observons aussi que dans beaucoup d'exemples le déterminant défini « la » apparait, donc nous recherchons la recherche ciblée sur ce type de structure, ci-dessous ;
- **N / Prep / Det / Nom** : expression régulière ^\w+\s(de|des)\s(la)\s\w+$; exemples : « troubles de la personnalité », « dosage de la TSH », « trouble de la perception » ;
- **N / Prep / N / Adj (par)** : expression régulière ^\w+(\s|-)(par)\s\w+(\s|-)\w+$; exemples : « imagerie par résonance magnétique », « angiographie par résonance magnétique ».

	ClassYN SP			**ClassYN GP**		
Type de format	**Prep / Det**	**Nombre**	**%**	**Prep / Det**	**Nombre**	**%**
N / Adj		946	0,741	-	946	0,669
N / Adj / Adj		110	0,086	-	94	0,066
N / Prep / N	de / des	113	0,088	de / des	136	0,096
	par	1	0,000	par	0	0,000
N / Prep / N / Adj	de / des	15	0,011	de / des	67	0,047
	par	1	0,000	par	4	0,002
N / Prep / Np-Np	-	9	0,007	-	6	0,004
N / Prep / Det / Nom	de / des + la	23	0,018	de / des + la	46	0,032
N° total des termes polylexicaux	**1276**			**1413**		

Tableau 39. Termes médicaux polylexicaux par type de format morphosyntaxique

1.2.2 Résultats de l'annotation automatique des marqueurs de reformulations

Nous avons suivi la même méthode d'annotation automatique utilisée pour le corpus CLEAR, consultable dans le **sous-chapitre 2.3**. Nous avons élargi notre champ de recherche de marqueurs en remplaçant « est une/la maladie/affection » par « maladie », « maladies », « affection », « affections » et les variants du lemme « définition » et « défini ». Nous avons élargi notre liste de marqueurs et indicateurs de reformulation avec les résultats des annotations sur le corpus **CLEAR**. Les extractions sont réalisées avec des **scripts en langage Perl**.

La liste complète de marqueurs mise à jour avec les nouveaux marqueurs et indicateurs de reformulations recherchés sur les corpus **ClassYN** est composée des éléments suivants : *c'est-à-dire ; c'est à dire ; c'est a dire ; C'est-àdire ; C'est-à-dire ; c'est-à-dire ; c'està-dire ; c'est-à-dire ; ça veut dire ; pour dire autrement ; autrement dit ; signifie ; ce qu'on appelle ; ce que l'on appelle ; est aussi appelé ; aussi appelé ; doit être compris comme ; au sens de ; défini ; définie ; defini ; definie ; définis ; définies ; définition ; definition ; affection ; affections ; maladie ; maladies ; trouble ; troubles ; désigne ; par exemple ; tel que ; telle que ; tels que ; telles que ; tel qu' ; telle qu' ; tels qu' ; telles qu' ; à savoir ; appelé ; appelée ; sous la désignation de ; sous le nom de ; désigné ; désignée ; fait partie de ; en particulier ; se manifeste par ; englobe ; englobent ; inclue ; incluent ; notamment ; consiste ; se traduit par ; y compris ; couvre ; qualifié ; qualifiée*.

Nos expériences nous donnent **2 689** phrases pour le corpus **ClassYN SP** et **4 871** phrases pour le corpus **ClassYN GP**, un total de **7 560** phrases qui peuvent contenir des reformulations, sélectionnées automatiquement par l'extracteur de termes et le script qui cherche le marqueur de reformulation. Lors de notre expérience d'annotation automatique sur le premier corpus (**CLEAR**), nous avons identifié automatiquement des marqueurs de reformulations dans **7,32%** des phrases avec termes médicaux. La liste élargie nous permet de trouver des marqueurs de reformulations dans un pourcentage plus important **(13,36%)** de phrases avec des termes médicaux lors de nos annotations automatiques sur le deuxième corpus, **ClassYN**.

ClassYN	N° phrases avec termes médicaux	N° phrases avec termes, mais sans marqueurs	Phrases avec termes médicaux et marqueurs	
			N° phrases	N° tokens
ClassYN SP	27 583	24 894	**2 689**	**88 407**
ClassYN GP	28 983	24 112	**4 871**	**139 320**
Total	56 566	49 006	**7 560**	**227 727**

Tableau 40. ClassYN : Phrases avec termes médicaux et marqueurs de reformulation

Les **7 560 phrases** avec termes médicaux et marqueurs de reformulations sont analysées et annotées manuellement. Nous présentons l'analyse de cette annotation dans la section suivante.

1.2.3 Annotation manuelle, évaluation et validation des reformulations

Dans cette section, nous présentons l'évaluation de l'annotation des phrases en reformulation (selon le guide), l'annotation manuelle de termes médicaux et des marqueurs, ainsi que l'annotation des relations lexicales et des fonctions sémantico-pragmatique (suivant la même analyse pour le corpus CLEAR).

1.2.3.1 Comparaison des annotations des phrases avec reformulations

Nous avons mené des analyses quantitatives sur les phrases qui ont été identifiées comme ayant des reformulations médicales correctes par les deux annotateurs. Plus précisément, nous avons analysé les reformulations qui ont reçu la même étiquette de type « Statut » (« oui », « oui-inv », « non »), dans les deux corpus. Nous avons également calculé le nombre de reformulations étiquetées différemment par les deux annotateurs. Nous présentons les résultats et les statistiques de ces annotations dans les **Tableau 41, Tableau 42** et **Tableau 43** ci-dessous.

En analysant les résultats d'annotation pour **le corpus de textes scientifiques**, nous observons que seulement **un tiers des phrases** sont des reformulations médicales correctes. Un grand nombre de ces reformulations correctes ont été validées par les deux annotateurs (**744 ; 27,66%**), même situation pour le cas des phrases sans reformulations, annotées avec « non » (**1 913 ; 71,14%**). Les deux annotateurs identifient qu'il y a plusieurs reformulations dans une même phrase, ce qui rajoute **154** et, respectivement, **262** reformulations en plus par rapport au nombre de phrases initiales. Pourtant, nous notons le

grand écart dans l'annotation des reformulations multiples et des annotations inversées (quand la reformulation précède le terme) entre les deux annotateurs (du simple au double).

ClassYN SP (textes scientifiques)		
Données quantitatives	**Annot 1**	**Annot 2**
Reformulations avec ***oui***	603	769
Reformulations avec ***oui<inv>***	42	6
Reformulations avec ***oui+2***	154	262
Reformulations avec ***oui<inv>+2***	2	3
Reformulations avec ***oui - total***	**801**	**1040**
Reformulations avec ***non***	2045	1915
Reformulations avec balises différentes - ***total***	**404**	
Reformulations avec les mêmes balises - ***oui***	**493**	
Reformulations avec les mêmes balises - ***oui<inv>***	**3**	
Reformulations avec les mêmes balises - ***non***	**1789**	
Reformulations avec les mêmes balises - ***total***	**2285**	
N° reformulations **multiples**	**156** *+5,48%*	**265** *+8,96%*
N° d'annotations	**2846**	**2955**
N° total initial de reformulations (phrases)	**2689**	

Tableau 41. Données quantitatives sur l'annotation du corpus ClassYN SP

Concernant **le corpus de textes grand public, ClassYN GP**, la double annotation a été réalisée pour un extrait de **338 phrases**[78]. Nous observons que les annotations sont très différentes, car l'annotateur 1 a identifié **143 (42,94%)** reformulations correctes, tandis que l'annotateur 2 en a identifié **240 (72,07%)**. Ces grandes différences sont observées sur seulement 338 phrases, nous devons lancer d'autres campagnes de double annotation sur un nombre plus important de données afin de mieux juger ces écarts.

[78] Par manque de temps, le stagiaire n'a pas annoté plus de données du corpus ClassYN GP.

ClassYN GP (textes grand public ; extrait double annotation)		
Données quantitatives	**Annot 1**	**Annot 2**
Reformulations avec ***oui***	128	159
Reformulations avec ***oui<inv>***	15	81
Reformulations avec ***oui (total)***	**143**	**240**
Reformulations avec ***non***	195	98
Reformulations avec balises différentes - ***total***	**151**	
Reformulations avec les mêmes balises - ***oui***	**85**	
Reformulations avec les mêmes balises - ***oui<inv>***	**4**	
Reformulations avec les mêmes balises - ***non***	**98**	
Reformulations avec les mêmes balises - ***total***	**187**	
N° total de reformulations (phrases)	**338**	

Tableau 42. Données quantitatives sur la double annotation d'un extrait du corpus ClassYN GP

Concernant l'annotation unique réalisée sur l'intégralité de données extraites du corpus **ClassYN GP**, nous remarquons la même tendance que pour le corpus expert, seulement **un tiers des phrases (1750 ; 35,92%)** contiennent des reformulations correctes. **482** reformulations en plus du nombre de phrases se rajoutent pour arriver à un **total de 2 232 (45,82%) reformulations médicales validées** par le premier annotateur.

ClassYN GP (textes grand public)	
Données quantitatives	**Annot 1**
Reformulations avec ***oui***	**1501**
Reformulations avec ***oui<inv>***	**249**
Reformulations avec ***oui – total (phrases initiales)***	**1750**
Reformulations avec ***oui+2***	388
Reformulations avec ***oui<inv>+2***	94
Reformulations avec ***oui+2 – total (plusieurs / phrase)***	**482**
Reformulations avec ***oui/<inv> & oui+2 – total***	**2232**
Reformulations avec ***non***	3121
N° total d'annotations	**5353**
N° total initial de reformulations (phrases)	**4871**

Tableau 43. Données quantitatives sur l'annotation du corpus ClassYN GP

Nous présentons par la suite des mesures statistiques comme **la précision, le rappel** pour les deux corpus et **l'accord inter-annotateur** pour les deux corpus.

1.2.3.1.1 Calcul de mesures statistiques : précision et rappel

Nous avons calculé la précision et le rappel de la même manière que pour le corpus CLEAR. Nous rappelons que **la précision** est calculée par rapport au nombre de phrases évaluées divisé par le nombre initial total de phrases annotées automatiquement. **Le rappel** prend en considération d'abord une annotation comme référence (celle de l'annotateur 1), et puis celle de l'annotateur 2, et nous avons calculé les moyennes pour le rappel. Nous utilisons la même formule que celle utilisée sur le corpus CLEAR, et nous présentons les résultats dans le **Tableau 44** ci-dessous.

$$\boldsymbol{Précision} = \frac{reformulations\ en\ commun\ Annot1\ \&\ Annot2}{phrases\ extraites\ automatiquement}$$

Figure 18. Formule de calcul de la précision

$$\boldsymbol{Rappel} = \frac{reformulations\ en\ commun\ Annot1\ \&\ Annot2}{reformulations\ Annot1}$$

Figure 19. Formule1 de calcul du rappel de l'annotation

$$\boldsymbol{Rappel} = \frac{reformulations\ en\ commun\ Annot1\ \&\ Annot2}{reformulations\ Annot2}$$

Figure 20. Formule2 de calcul du rappel de l'annotation

$$\boldsymbol{Moyenne\ rappel} = \frac{(rappel\ Annot1 + rappel\ Annot2)}{2}$$

Figure 21. Formule de calcul de la moyenne du rappel

ClassYN SP (textes scientifiques)			
Mesures statistiques (%)	**Annot 1**	**Annot 2**	**Désaccord**
Précision - ***oui***	23,98%	28,82%	4,84%
Précision - ***oui - moyenne***	**25,50%**		
Précision - ***non***	75,93%	71,21%	4,72%
Précision - ***non - moyenne***	**73,57%**		
Précision - *moyenne générale*	**84,97%**		
Rappel - ***oui***	92,88%	71,53%	21,35%
Rappel - ***oui - moyenne***	**82,20%**		
Rappel - ***non***	93,54%	99,89%	6,35%
Rappel - ***non - moyenne***	**96,71%**		
Rappel - *moyenne générale*	**89,45%**		

Tableau 44. Mesures statistiques (précision, rappel) sur les annotations du corpus ClassYN SP

Concernant le corpus de textes scientifiques, nous remarquons une **très haute précision (84,97%) et un rappel important (89,45%)** pour l'annotation similaire (c'est-à-dire avec la même balise) entre les deux annotateurs.

ClassYN GP (textes grand public ; extrait)			
Mesures statistiques (%)	**Annot 1**	**Annot 2**	**Désaccord**
Précision - ***oui***	42,94%	72,07%	29,13%
Précision - ***oui moyenne***	**57,50%**		
Précision - ***non***	58,55%	29,42%	29,13%
Précision - ***non moyenne***	**43,98%**		
Précision - *moyenne générale*	**56,15%**		
Rappel - ***oui***	62,23%	37,08%	25,15%
Rappel - ***oui moyenne***	**49,65%**		
Rappel - ***non***	50,25%	100%	49,75%
Rappel - ***non moyenne***	**75,26%**		
Rappel - *moyenne générale*	**62,45%**		

Tableau 45. Précision et rappel sur l'extrait de 338 phrases avec double annotation de ClassYN GP

Pour le corpus de textes grand public avec annotation unique, nous observons une précision moyenne de **35,92%**, ce qui signifie que **plus d'un tiers des phrases annotées contiennent au moins une reformulation correcte.**

ClassYN GP (textes grand public)	
Mesures statistiques (%)	**Annot 1**
Précision - ***oui***	**35,92%**
Précision ***- oui+2***	***+ 9%***
Précision ***- oui total***	**45,82%**
Précision - ***non***	**64,07%**

Tableau 46. Mesures statistiques (précision) sur les annotations du corpus ClassYN GP

1.2.3.1.2 Accord inter-annotateur

Nous avons calculé **le score inter-annotateur** de type Kappa (Cohen, 1960) avec la formule suivante :

$$\boldsymbol{Kappa} = \frac{p_r - p_e}{1 - p_e}$$

Nous rappelons que $\boldsymbol{p_r}$ représente la probabilité réelle, plus précisément la probabilité observée de chaque annotation (pour les balises « oui » et « non ») et $\boldsymbol{p_e}$ représente la probabilité attendue que les annotateurs soient d'accord dans l'annotation. Le calcul du rappel des annotations est fait avec la formule :

$$\boldsymbol{p_r} = \frac{(Annot1\ \&Annot2\ balises\ \boldsymbol{oui}) + (Annot1\ \&Annot2\ balises\ \boldsymbol{non})}{N°\ total\ de\ reformulations}$$

Nous rappelons la formule de calcul pour la probabilité attendue ($\boldsymbol{p_e}$) :

$$\boldsymbol{p_e} = \left(Probabilité_{Annot1\ \boldsymbol{oui}} * Probabilité_{Annot2\ \boldsymbol{oui}} \right) + \left(Probabilité_{Annot1\ \boldsymbol{non}} * Probabilité_{Annot2\ \boldsymbol{non}} \right)$$

Les deux probabilités (pour les balises « oui » et « non ») sont calculées en multipliant les probabilités de deux annotateurs pour les deux balises. Notre calcul de l'accord Kappa pour le corpus **ClassYN SP** nous donne un accord inter-annotateur de **0,63**, ce qui est un **accord modéré** (McHugh, 2012), tandis que pour l'extrait de 338 phrases avec double annotation du corpus **ClassYN GP**, le score est bas **(0,41)**. Le nombre réduit de phrases bénéficiant d'une double annotation manuelle explique la différence par rapport au corpus CLEAR.

Lors de l'analyse des différences d'annotation pour l'extrait de phrases du corpus grand public, nous remarquons que parmi les reformulations annotées comme correctes par l'annotateur 2 se trouvent **28 (8,28%) chaines de coréférence** de type « maladie », « affection », « symptôme ». Ces hypéronymes sont erronément considérés comme des reformulations des termes médicaux, comme dans l'exemple : « Les signes de la <ref>**maladie**</ref>. Les signes évoquant le <terme>**cancer de l'estomac**</terme> sont très nombreux et variés ». Cette situation représente une compréhension erronée du guide d'annotation, car dans notre étude nous ne considérons pas ce type de chaine de référence phrastique comme une reformulation.

1.2.3.2 Analyse quantitative des termes médicaux reformulés

Le nombre de **termes médicaux reformulés** est de **1 194** pour le **corpus ClassYN SP**. Pour analyser la liste de termes reformulés, nous supprimons tous les déterminants annotés ensemble avec les termes, en nombre de **535** (**Tableau 47**).

Type	Genre	Nombre	Forme	N° occurrences	%
L'article défini	masculin	singulier	***l'***	79	14,76%
		singulier	***le***	56	10,46%
		pluriel	***les***	83	15,51%
	féminin	singulier	***la***	126	23,55%
L'article indéfini	masculin	singulier	***un***	25	4,67%
	masculin féminin	pluriel	***des***	85	15,88%
	féminin	singulier	***une***	81	15,14%
Total				**535**	**100%**

Tableau 47. Déterminants de type article défini et indéfini des termes médicaux du corpus ClassYN SP

En ce qui concerne le type de termes, nous avons identifié :

- **361 (30,23%) termes médicaux simples ;**
- **833 (69,76%) termes médicaux polylexicaux.**

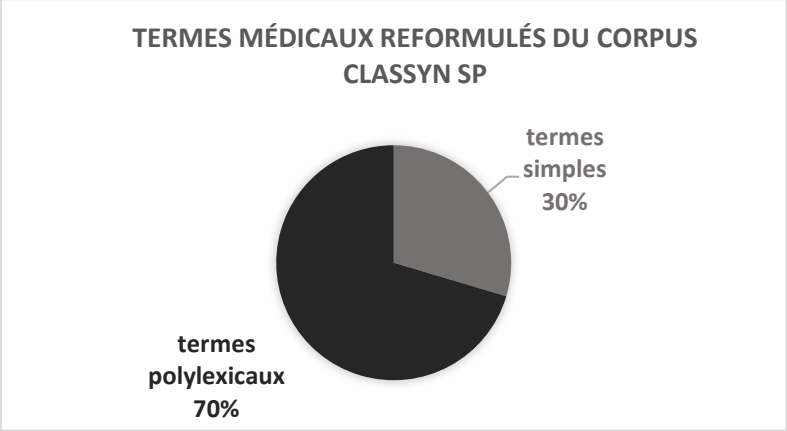

Figure 22. Types de termes médicaux reformulés extraits du corpus ClassYN SP

Nous analysons **tous les termes médicaux polylexicaux avec des expressions régulières** dans Notepad++ (indiquées entre parenthèses), en prenant en compte les phrases avec les reformulations simples et multiples. Nous identifions les types de termes polylexicaux suivants :

- **326 (16,34%) termes bi-grammes** de type Nom-Adjectif ou Nom-Nom (par exemple : *cystite bactérienne ; neuropathie périphérique ; méralgie paresthésique ; molécules sondes)* (expression régulière : ^\w+\s\w+$) ;

- **177 (14,82%) termes tri-grammes** de type Nom-Préposition-Nom et Nom-Adjectif-Adjectif (*facteur neurotrophique ciliaire ; hémorragies digestives hautes ; maladie de Basedow*) (expression régulière : ^\w+\s\w+\s\w+$) ;
- **206 (17,25%) termes avec quatre tokens ou plus** (par exemple : *néphropathie liée au VIH ; syndrome de détresse respiratoire ; trouble déficitaire de l'attention avec hyperactivité*) (expression régulière : ^\w+\s\w+\s\w+\s\w+).

Nous avons retrouvé 396 termes doublons dans nos listes, ainsi nous obtenons **798 termes médicaux uniques reformulés** pour le corpus **ClassYN SP**. Pourtant, nous avons besoin des toutes les occurrences de termes pour analyser les reformulations (qui peuvent être différentes). Nous présentons cette analyse des reformulations plus loin dans ce chapitre.

Un extrait de la liste complète de **termes reformulés pour le corpus ClassYN SP** (sans doublons et en ordre alphabétique) se trouve dans **l'Annexe 6.6**.

***

Nous avons réalisé le même traitement pour la liste finale de **2 234 termes médicaux reformulés** du corpus **ClassYN GP,** corpus de textes grand public. Nous avons supprimé un nombre de 965 déterminants définis et indéfinis (« la » avec 254 occurrences (26,32%) ; « l' » avec 183 (18,96%) ; « le » : 132 (13,67%) ; « les » : 102 (10,56%) ; « un » : 82 (8,49%) ; « une » : 143 (14,81%) ; « des » : 68 (7,04%) ; « d' » : 1 (0,10%)).

Nous notons que nous avons identifié :

- **891 (39,88%) termes médicaux simples ;**
- **1 343 (60,11%) termes médicaux polylexicaux.**

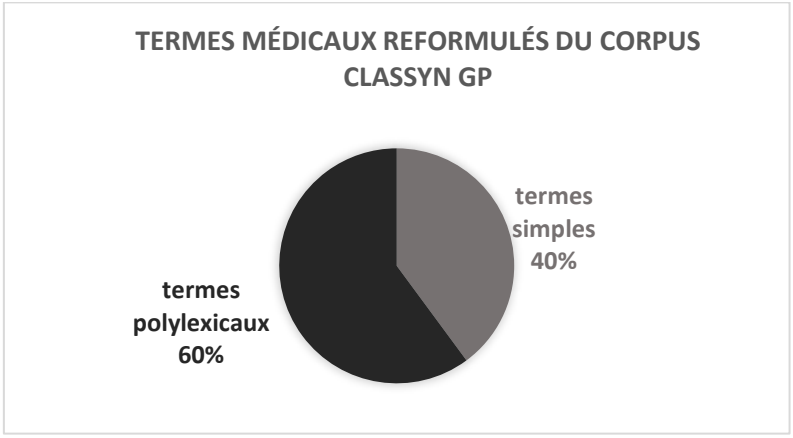

Figure 23. Types de termes médicaux reformulés extraits du corpus ClassYN GP

Parmi les termes médicaux polylexicaux nous identifions avec des expressions régulières :

- **519 (23,23%) termes bi-grammes** de type Nom-Adjectif (par exemple : *asthme paroxystique ; contractions utérines ; coliques néphrétiques)* (expression régulière : ^\w+\s\w+$) ;
- **268 (11,99%) termes tri-grammes** de type Nom-Préposition-Nom et Nom-Adjectif-Adjectif (*maladie de McArdle ; paralysie périodique hyperkaliémique ; faiblesse musculaire distale*) (expression régulière : ^\w+\s\w+\s\w+$) ;
- **278 (12,44%) termes avec quatre tokens ou plus** (par exemple : *paralysie périodique primitive hyperkaliémique* ; *neuropathie motrice héréditaire distale ; hormonothérapie du cancer de la prostate non métastatique* (expression régulière : ^\w+\s\w+\s\w+\s\w+).

Nous avons retrouvé **913 (40,33%) termes doublons** dans notre liste de termes, ce qui nous permet de retenir **1 321 termes médicaux uniques reformulés** pour le corpus **ClassYN GP**. Comme pour les autres corpus, nous gardons également toutes les occurrences de termes pour analyser les **différentes reformulations identifiées** pour le même terme médical.

Un extrait de la liste complète de **termes reformulés pour le corpus ClassYN GP** (sans doublons et en ordre alphabétique) se trouve dans **l'Annexe 6.7**.

****

Concernant les types de termes médicaux reformulés dans le corpus **ClassYN**, nous faisons les suivantes remarques :

- nous identifions beaucoup plus de termes simples dans le texte grand public, **(39,88%)**, par rapport au **30,23%** dans le texte scientifique ;
- les termes bi-grammes de type Nom-Adjectif sont plus fréquents dans le corpus grand public **(23,23%)** que dans le corpus expert **(16,34%)** ;
- les termes polylexicaux complexes (tri-grammes ou plus) sont plus reformulés dans le corpus expert **(32,07%)** que dans le corpus grand public **(24,43%)**.

Ces tendances montrent que dans le corpus **ClassYN GP** un nombre plus important de termes simples ou bi-grammes sont reformulés. En revanche, le corpus **ClassYN SP** contient un plus grand pourcentage de termes polylexicaux complexes qui sont reformulés,

ce qui prouve que le langage utilisé dans les textes scientifique est plus technique et nécessite reformulation.

Dans la prochaine section nous analysons les types de marqueurs et indicateurs de reformulation identifiés dans les deux corpus.

1.2.3.3 Analyse quantitative des marqueurs et indicateurs de reformulation

Nous analysons les marqueurs et indicateurs de reformulations qui introduisent les **776 reformulations correctes** dans le corpus **ClassYN SP** et nous obtenons les résultats suivants :

- **242 (31,18%)** reformulations sont marquées par des parenthèses uniquement **()**, sans présence d'un autre marqueur lexical, de type **abréviations** (*ADL, HCDD, TEP*) ou **énumérations/ exemplifications** (*douleur, nausées, hypoxie*) ;
- **193 (24,87%)** reformulations sont délimitées par des parenthèses ensemble avec un indicateur de type ***maladie, trouble, affection*** et des marqueurs comme ***consiste, y compris, d'autres***, de type : (*hypotension orthostatique, troubles digestifs, troubles neuropsychiques)* ;
- **68 (8,76%) reformulations** sont marquées avec le **lemme « défini » (défini comme ; défini par ; définition)** (*atteinte muqueuse **définie par** une infiltration de la muqueuse sans atteinte de la musculeuse ou de la sous-séreuse*) ;
- **65 (8,37%) reformulations** sont marquées à l'aide du **lemme « appelé »** (exemple : *dyskinésies **appelées** mouvements anormaux involontaires*).

Le marqueur prototypique de la reformulation « **c'est-à-dire** » délimite uniquement quatre reformulations dans notre corpus, ce qui est surprenant par rapport aux attentes. Nous avons identifié **55 reformulations (7,08%)** grâce à la présence des marqueurs de notre hypothèse de départ (*notamment, par exemple, tel que, en particulier, comme*, etc.), mais ils ne délimitent pas la reformulation tous seuls, ils se retrouvent souvent à l'intérieur des parenthèses.

***

Pour résumer notre analyse sur le corpus **ClassYN GP** des **marqueurs et indicateurs** de reformulation les plus fréquents, nous constatons que :

- **665 (29,76%)** reformulations sont marquées par des parenthèses uniquement **()**, de type :
    - **abréviations** : *dégénérescence maculaire liée à l'âge **(**DMLA**)* ;**

  - **énumérations/exemplifications** : troubles moteurs **(**fatigabilité, crampes, tremblement des extrémités**) ;**
  - **explication** : *jaunisse **(**peau et blanc des yeux devenant jaunes**)* **;**
  - **paraphrase** : *hirsutisme **(**augmentation de la pilosité**)* **;**
- **239 (10,69%) reformulations** sont marquées à l'aide du **lemme « appelé »** *(aussi appelé, ce que l'on appelle, également appelé)* (*L'imagerie par échographie **aussi appelée** scan échographique ou sonographie)* ;
- **151 (6,75%)** reformulations sont délimitées par des parenthèses avec un indicateur de type ***maladie, trouble, affection*** et des marqueurs comme ***se traduit par, dénommées, c'est-à-dire, d'autres***, ***comme, y compris, abrégée en,*** de type : *les ostéophytes **(parfois dénommées** "becs de perroquet"**)*** ;
- **19 (0,85%) reformulations** sont marquées avec le **lemme « défini » (défini comme ; défini par ; définition)** (*La douleur **est définie comme** la perception d'une sensation déplaisante)*.

Nous remarquons que, dans la plupart des cas, *les reformulations sont marquées par des parenthèses*, situation similaire au corpus **CLEAR**. Cependant, ce signe orthographique n'est pas toujours un marqueur de reformulation, nous avons besoin de prendre en compte aussi la présence des marqueurs lexicaux et grammaticaux à l'intérieur de la parenthèse, comme dans l'exemple : *les ostéophytes **(parfois dénommées** "becs de perroquet"**).***

1.2.3.3.1 Nouveaux marqueurs et indicateurs de reformulations identifiés

Par rapport au corpus CLEAR et la liste de marqueurs définie initialement, nous avons identifié un **grand nombre de nouveaux marqueurs et indicateurs** lors de notre processus d'annotation des reformulations. Nous les illustrons avec des exemples ci-dessous :

- **« est appelé par convention ; aussi appelé ; encore appelé » ;**

*La protéine p53 **aussi appelée** protéine gardienne du génome*
*syndrome métabolique **(encore appelé** syndrome X**)***
*l'hépatite C était **encore appelée** hépatite non-A non-B … au début des années 80*

- **« notamment » ;**

*une méningite bactérienne lymphocytaire **(notamment** tuberculose et listériose**)***

- **« d'autres / et d'autres »** (suivi par l'hyperonyme médical *trouble*) ;

et d'autres *troubles mentaux* ***comme*** *les troubles maniaques*

- **« regroupe » ;**

L'uvéite ***regroupe*** *un large ensemble de maladies inflammatoires de l'uvé*

- **« en particulier ».**

pathologies graves non thyroïdiennes ***en particulier*** *l'insuffisance rénale et la dialyse*

Nous avons identifié également d'autres marqueurs de reformulation, comme : *abrégée en ; le traduisant le plus souvent par ; se manifeste par ; consiste en ; consiste à ; classé dans ; caractérisé par ; renvoie à, représenté par ; fait partie de*.

Cette **liste élargie** sera utilisée comme source pour compléter la liste des marqueurs en roumain avec des nouvelles traductions.

1.2.3.4 Analyse lexicale et sémantico-pragmatique des reformulations

Dans le corpus **ClassYN SP**, les reformulations les plus fréquentes sont celles de type *hyponymie-exemplification* avec **326 (42,01%)** reformulations, suivies par **157 (20,23%)** reformulations de types *hyperonymie-définitions* (parmi les **228 (29,38%)** de *définitions* en total). La relation de *synonymie-dénomination* occupe la troisième place en termes de fréquence avec **149 (19,20%)** reformulations. Les *méronymies* sont en nombre de **92 (11,85%)**, dont **70 (9,02%)** avec *définition* (même situation observée dans le corpus scientifique **CLEAR**), **18 (2,31%)** avec *explication* et seulement **4 (0,51%)** avec *exemplification*.

Relation lexicale	**Fonction sémantico-pragmatique**	**ClassYN SP**		**ClassYN GP**	
		N°	**%**	**N°**	**%**
hyperonymie	définition	**157**	**20,23%**	**682**	**38,97%**
hyponymie	exemplification	**326**	**42,01%**	**300**	**17,14%**
synonymie	paraphrase	28	3,60%	206	11,77%
	dénomination	**149**	**19,20%**	184	10,51%
méronymie	explication	18	2,31%	89	5,08%
	définition	70	9,02%	**241**	**13,77%**
Total		**776**	**100%**	**1750**	**100%**

Tableau 48. Paires de relations lexicales et fonctions sémantico-pragmatiques annotées dans ClassYN SP et ClassYN GP

Notre annotation de **1 750** phrases avec des reformulations correctes du corpus **ClassYN GP** montre que **53,31%** des reformulations sont de type *définitions* **(933)**. Parmi ces définitions, les *hyperonymies-définitions* sont les plus fréquentes dans le corpus, à **38,97% (682)**. Les *hyponymies-exemplifications* sont en nombre de **300 (17,14%)**. Les reformulations de type *méronymie-définition* sont les plus fréquentes, à **13,77%** (**241** reformulations), par rapport aux autres corpus annotés (ClassYN SP : 9,02% ; CLEAR SP : 3,18% ; CLEAR GP : 0,67%). Les synonymies-paraphrases sont en nombre de **206 (11,77%)** et les *synonymies-dénominations* concerne **184** reformulations **(10,51%)**. La paire la moins fréquente reste la *méronymie-explication* avec **89** reformulations **(4,01%)**.

1.2.4 Bilan des résultats d'annotation sur le corpus ClassYN

Nous observons de grandes différences en termes de reformulations entre le corpus **ClassYN SP** et **ClassYN GP**. L'annotation du corpus scientifique permet d'identifier des reformulations médicales correctes dans seulement **25,50%** des phrases en moyenne (annotateur 1 : **23,98%**, annotateur 2 : **28,82%**), tandis que le pourcentage pour le corpus grand public pour notre annotation de données intégrales est 10% plus élevé, à **35,92%**.

ClassYN	ClassYN SP	ClassYN GP
% Reformulations	**25,50%**	**35,92%**
N° phrases annotées	**2 689**	**4 871**
Total	**7 560**	

Tableau 49. Score de précision de reformulations annotées dans ClassYN SP et ClassYN GP

Nous avons identifié également de nouveaux **marqueurs de reformulation** par rapport à la **liste élargie de marqueurs** issus du corpus **CLEAR (Tableau 18 et Tableau 28)**, de type : *est appelé par convention ; encore appelé ; d'autres ; regroupe ; en particulier ; abrégée en ; le traduisant le plus souvent par ; se manifeste par ; consiste en ; consiste à ; classé dans ; caractérisé par ; renvoie à, représenté par ; fait partie de*. Nous analysons la possibilité de traduire ces marqueurs en roumain pour identifier des reformulations médicales dans nos corpus roumains.

Concernant l'analyse **lexicale et sémantico-pragmatique** des reformulations, nous notons que les reformulations de type *hyperonymie-définition* sont les plus fréquentes dans le **corpus grand public (682 ; 38,97%)** tandis que dans le **corpus scientifique** le plus grand nombre de termes sont reformulés à travers des exemples d'hyponymes

(326 *hyponymies-exemplifications* ; **42,01%)**. Ces deux catégories représentent la grande majorité de reformulation dans les deux corpus. Les exemplifications sont privilégiées dans la littérature scientifique alors que les définitions sont plus employées dans les textes de vulgarisation. Nous observons également une particularité du corpus **ClassYN GP** : un grand nombre de reformulations de type *définitions* ne sont pas introduites par un hyperonyme (selon notre hypothèse de recherche initiale), mais par un ***méronyme*** **(241 ; 13,77%)**.

Dans le point suivant, nous présentons le bilan général des annotations sur les deux corpus français, **CLEAR Cochrane** et **ClassYN**.

1.3 Bilan général des travaux sur les corpus français

Nous avons annoté manuellement **16 227 phrases** (**8 667** du corpus **CLEAR Cochrane** et **7 560** du corpus **ClassYN**), dont nous avons identifié **8 626 reformulations médicales correctes**. Ces **8 626** reformulations ont été annotées en relations lexicales (*hyperonymie, hyponymie, synonymie, méronymie*) et en fonctions sémantico-pragmatiques (*définition, exemplification, paraphrase, dénomination, explication*).

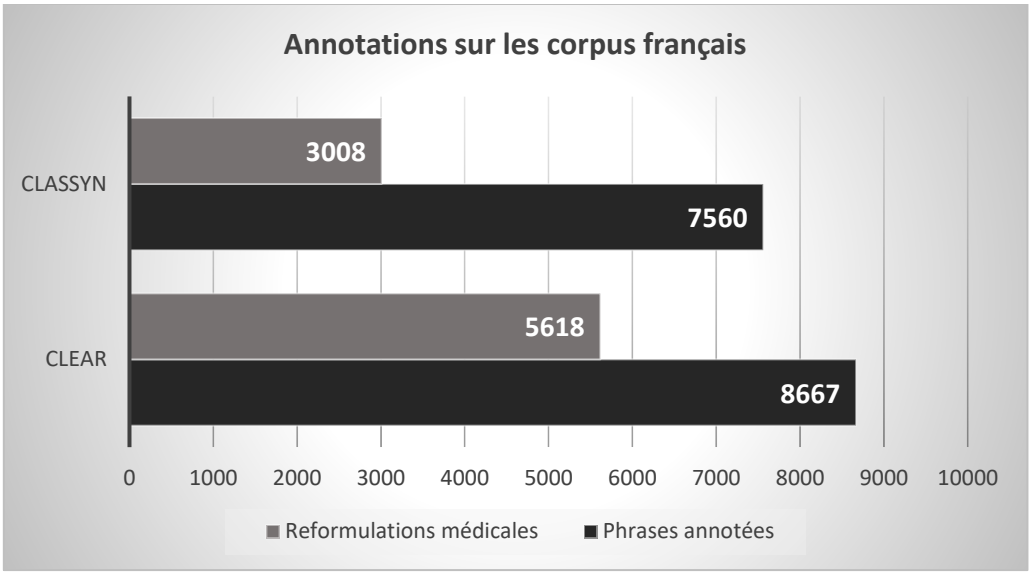

Figure 24. Données de résultats d'annotation des reformulations sur les corpus français

Les relations lexicales et les fonctions sémantico-pragmatiques nous aident à comprendre l'utilisation des reformulations dans les textes médicaux scientifiques et de vulgarisation médicale. Nous observons que pour trois sous-corpus sur quatre, les reformulations de type *définitions* introduites par des *hyperonymes* sont **les plus fréquentes** (**CLEAR SP** : **46,22%** ; **CLEAR GP** : **49,46%** ; **ClassYN GP** : **38,97%**) des reformulations. Uniquement pour le corpus **ClassYN SP,** les reformulations les plus fréquentes sont de type *exemplifications* à travers des *hyponymes* (**42,01%**).

Les reformulations de type *méronymie-définition* ont été les plus annotées dans le corpus **ClassYN GP**, à **13,77%** (**241** reformulations), par rapport aux trois autres corpus annotés (**ClassYN SP** : 9,02% ; **CLEAR SP** : 3,18% ; **CLEAR GP** : 0,67%). **ClassYN** est différent de **CLEAR**.

Sur la base de l'analyse effectuées sur nos deux corpus, nous avons identifié plusieurs **nouveaux marqueurs et indicateurs** qui soulignent la présence de la reformulation. En partant d'une liste issue de l'état de l'art (à gauche dans le **Tableau 50** ci-dessous), nous avons identifié de nouveaux marqueurs lors de nos expériences d'annotation sur les corpus français (liste à droite du **Tableau 50**). Nous utilisons cette liste élargie pour chercher des équivalents pour le corpus roumain.

N°	Liste d'origine de marqueurs / indicateurs de reformulation		N°	Liste de marqueurs / indicateurs à l'issue de nos annotations
1	c'est-à-dire		**1**	est appelé par convention
2	ça veut dire / veut dire		**2**	encore appelé
3	pour dire autrement		**3**	aussi appelé
4	autrement dit		**4**	d'autres / et d'autres / ou d'autres
5	signifie		**5**	regroupe
6	désigne		**6**	en particulier
7	ce qu'on appelle		**7**	abrégée en
8	ce que l'on appelle		**8**	le traduisant le plus souvent par
9	est aussi appelé / aussi appelé		**9**	se manifeste par
10	doit être compris/e comme		**10**	consiste en / à
11	au sens de		**11**	classé dans
12	est un / une ; sont des / un / une affection/s ; maladie/s ; trouble/s		**12**	caractérisé par
13	affection / s		**13**	renvoie à
14	maladie / s		**14**	représenté par
15	trouble / s		**15**	fait partie de
16	définition / s		**16**	y compris
17	défini / e / s / es		**17**	par exemple
18	défini / e / s / es comme		**18**	notamment
19	tel / lle / s / lles que		**19**	()
20	par exemple		**20**	comme
			21	décrit
			22	est associé à

Tableau 50. Liste de marqueurs et indicateurs de reformulation d'origine et ceux issus de nos annotations sur les corpus CLEAR et ClassYN

Nous présentons par la suite les annotations automatiques et manuelles effectuées sur la langue roumaine et nous analysons les reformulations médicales identifiées dans le corpus de vulgarisation médicale que nous avons créé, **GrandMed-Ro2**.

2. Analyses sur le corpus roumain

Un des objectifs de notre thèse est de constituer **un premier corpus médical annoté de reformulations médicales en roumain**. Cette ressource est utile pour la simplification des termes médicaux monolexicaux (un mot) et polylexicaux en roumain (deux mots et plus / expressions). Notre méthode est similaire à celle proposée pour le français : il s'agit de rechercher automatiquement les termes médicaux et les marqueurs spécifiques qui aident à les reformuler, comme dans l'exemple ci-dessous. Nous utilisons les mêmes notations employées pour le français pour mettre en avant le **terme reformulé** (<terme>colesterolului (LDL)</terme> ; *cholestérol (LDL)*), le **marqueur de reformulation** (<marq>cunoscut sub numele de</marq> ; *connu sous le nom de*) et **la reformulation médicale** identifiée (<ref>colesterol rău</ref> ; *mauvais cholestérol*).

> Grăsimile saturate și trans pot crește nivelul <terme>colesterolului (LDL)</terme> (<marq>cunoscut sub numele de</marq> <ref>colesterol rău</ref>) în sânge.
>
> *Les graisses saturées et trans peuvent augmenter le taux de <terme>cholestérol (LDL) </terme> (<marq>connu sous le nom de</marq> <ref>mauvais cholestérol</ref>) dans le sang.*

À notre connaissance, il n'y a pas de corpus de reformulations médicales disponible pour le roumain, en dehors du corpus de textes médicaux scientifiques et pour le grand public **GrandMed-Ro** de taille modeste (42 140 tokens), collecté manuellement par nous (Buhnila, 2018). Ce corpus a été utilisé pour créer un glossaire de 113 termes médicaux et leurs synonymes en langage courant. Nous avons constitué une version étendue, **GrandMed-Ro2**, avec des outils automatiques (**Partie III**, **sous-chapitre 2.1.3).**

2.1 Analyse des annotations faites sur le corpus GrandMed-Ro2

Nous présentons les analyses des annotations automatiques réalisées sur le corpus roumain **GrandMed-Ro2**. Le corpus **GrandMed-Ro2** contient **7 472 articles** et a une taille totale de **6 440 951 tokens**. Pour créer le corpus roumain de reformulations, nous avons appliqué la même méthodologie que pour les corpus français, l'annotation automatique terminologique, l'identification automatique de marqueurs et l'annotation manuelle des reformulations.

2.1.1 Post-traitement de la liste de termes

Nous avons extrait une liste de termes médicaux à partir du corpus **MoNERo** a été projetée sur le corpus avec leur annotations sémantiques (**Partie III**, **section 2.2.2.1**). Nous avons mené également une étape de **post-traitement** manuel de l'annotation automatique de termes afin d'identifier des mots qui sont des données bruitées et non pas des vrais termes médicaux. Après que ces 629 mots ont été éliminés de la liste, nous obtenons une liste finale de **6 889 termes uniques en roumain**.

TYPE DE DONNÉES	N°
Termes médicaux dans MoNERo	14 133
Termes médicaux extraits par nous	**14 133**
Termes médicaux uniques	7 528
Termes médicaux uniques après nettoyage	**6 899**
Termes médicaux simples	2 434
Termes médicaux polylexicaux	4 465

Tableau 51. Données quantitatives sur la liste de termes médicaux en roumain extraits du corpus annoté MoNERo (Mitrofan et al., 2019)

***

Nous présentons par la suite les premières expériences et analyses réalisées sur le sous-corpus « sfaturi médicales » (avis médicaux), après la double sélection des phrases. Notre analyse vise à valider la liste de marqueurs de la littérature et les traductions du français.

2.1.2 Analyse et vérification des marqueurs sur le sous-corpus « sfaturi medicale »

Nous avons suivi la même méthode que sur les corpus français : nous avons projeté en premier temps la liste des termes nettoyés sur le corpus pour sélectionner les phrases avec des termes médicaux, et en deuxième temps nous avons cherché la liste de marqueurs présentés dans le **Tableau 11** (**Partie III**, **sous-chapitre 2.3.2**) dans les phrases avec termes médicaux.

Nous avons réalisé nos premières annotations sur le **sous-corpus « sfaturi medicale »**[79] (*avis médicaux*), un corpus de **576 articles** comprenant **989 700 tokens**. Après l'extraction automatique des phrases contenant termes et marqueurs de la liste initiale de marqueurs du **Tableau** 11 (**Partie III**, **sous-chapitre 2.3.2**), nous avons obtenu **3 067 phrases** qui contiennent **5 426** occurrences de termes médicaux et **3 144** occurrences de marqueurs de reformulation.

Nous avons analysé les marqueurs de reformulation trouvés dans le corpus en termes de *fréquence absolue* (nombre d'occurrences) et de *fréquence relative* (la fréquence absolue divisée par le nombre total de tokens du corpus). Nous présentons cette analyse dans le **Tableau 52**. Nous avons considéré que les marqueurs qui se chevauchent comme « **sub denumirea de** » (*sous l'appellation de*) et « **cunoscut/ă/a sub denumirea de** » (*connu/e sous l'appellation de*) font partie du marqueur plus complexe, de type « **este cunoscut/ă/a sub denumirea de** » (*est connu/e sous l'appellation de*). Nous les avons enlevés du calcul final des occurrences (chiffres indiqués entre parenthèses avec « - »).

Nous avons ensuite classé les marqueurs les plus fréquents « este o (boală) » (*est une maladie*), « cum ar fi » (*comme*), « înseamnă » (*signifie*) et les moins fréquents : « este cunoscut sub denumirea de » (*est connu sous le nom de*), « cu alte cuvinte » (*autrement dit*) (**Figure 25**). Concernant le marqueur le plus fréquent, « este o (boală) » (*est une maladie),* parmi les **1 484** occurrences du corpus, **1 325 (89,28%)** représentent le marqueur « (o) boala/ă » (*une maladie*), et seulement **159 (10,71%)** pour la forme complète du marqueur « este o boala/ă » (*est une maladie*). Nous avons considéré ce mot comme un marqueur, car il nous a permis d'identifier des reformulations médicales telles que : « sindrom Guillain-Barré, o boală severă paralizantă » (*syndrome Guillain-Barré, une maladie paralysante sévère*) où le marqueur « o boală » (*une maladie*) apparait seul dans la structure d'une apposition, donc sans le marqueur générique « este o » (*est une*).

[79] https://sfaturimedicale.ro/, constitué le 12/09/2021.

Marqueurs en roumain	Traduction en français	N° d'occurrences	Fréquence absolue	Fréquence relative
înseamnă / inseamna	signifie	0 / 169	169	0,05375
mai exact	plus exactement	24	24	0,00763
mai precis	plus précisément	7	7	0,00222
cum ar fi	comme	1174	1174	0,37340
cu alte cuvinte	autrement dit	6	6	0,00190
este cunoscut/ă/a	est connu/e	36 / 0 / 11 *(-8)*	39	0,01240
cunoscut/ă/a și / cunoscut/ă/a si	connu/e aussi	0 / 0 / 0 / 8 / 0 / 27	35	0,01113
cunoscut/ă/a sub numele de	connu/e comme	17 / 0 / 15 1 / 0 / 4	32 5	0,01017 0,00159
este cunoscut/ă/a sub numele de	est connu/e sous le nom de			
sub denumirea de	sous l'appellation de	41 *(-8)*	33	0,01049
cunoscut/ă/a sub denumirea de	connu/e sous le nom de	0 / 0 / 8 *(-3)*	5	0,00159
este cunoscut/ă/a sub denumirea de	est connu/e sous l'appellation de	0 / 0 / 3	3	0,00095
définit/ă/a, definiție/tie	défini/e, définition	14 / 0 / 0, 0 / 5	19	0,00604
este termenul	est le terme	3	3	0,00095
este o (boală/ boala)	est une maladie	1 / 158 (2 / 1,484) *(- 159)*	1325	0,42143
este o (afecțiune/ afectiune)	est une affection	0 / 156 (0 / 413) *(- 156)*	257	0,08174
		Total	**3144**	**100%**

Tableau 52. Fréquences absolues et relatives pour les marqueurs roumains dans le sous-corpus « sfaturi medicale » (avis médicaux)

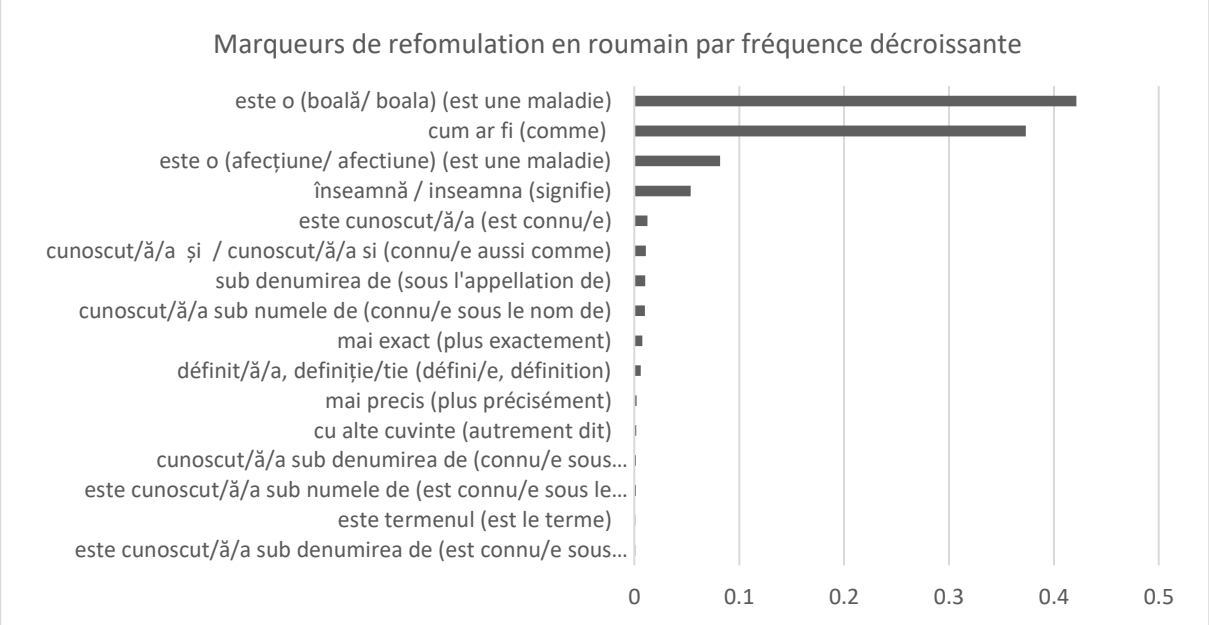

Figure 25. Classement par fréquence décroissante des marqueurs de reformulation identifiés dans le sous-corpus « sfaturi medicale » (avis médicaux)

Nous menons de recherches automatiques sur l'intégralité du corpus **GrandMed-Ro2** avec ces nouveaux marqueurs identifiés dans ce premier sous-corpus, « sfaturi médicale », en plus de la liste initiale de marqueurs pour le roumain. Lors de cette évaluation, nous identifions de *nouveaux marqueurs et de nouvelles variations de marqueurs*.

Cette nouvelle **liste élargie de marqueurs** est présentée dans le **Tableau 53**. Les marqueurs sont traduits en français par nous. Ces traductions *ne représentent pas nécessairement des marqueurs de reformulation en français*. Nous testerons leur viabilité en tant que marqueurs de la reformulation sur les corpus français dans de recherches futures.

Liste élargie de marqueurs de reformulation en roumain		
N°	**Marqueurs en roumain**	**Marqueurs traduits en français**
1	reprezintă / a	*représente*
2	adică / a	*c'est-à-dire*
3	așa-numitele / asa-numitele	*ainsi nommées*
4	așa (asa) numită (a)	*ainsi nommée*
5	denumită (a) uneori	*appelée parfois*
6	denumită (a) uneori și (si)	*appelée parfois aussi*
7	denumită (a) popular și (si)	*appelée populairement aussi*
8	popular numiți (/ti) și (si)	*populairement appelés aussi*
9	uneori numită (a)	*parfois appelée*
10	denumită (a) medical	*médicalement appelée*
11	denumită (a) sugestiv	*suggestivement appelée*
12	tehnic numit	*techniquement appelé*
13	este denumit în (in) mod obișnuit (/s)	*appelé habituellement*
14	este un proces numit	*est un processus appelé*
15	adesea numit	*souvent appelé*
16	fiind denumiți	*étant appelés*
17	este uneori numit	*est parfois appelé*
18	este adesea denumită și	*est souvent appelée aussi*
19	a fost denumită anterior și	*a été appelée antérieurement aussi*
20	o afecțiune numită	*une affection appelée*
21	este numit, de asemenea,	*est appelé également*
22	mai este numită (a) și (si)	*aussi appelée*
23	e numită și	*appelée aussi*
24	deoarece	*parce que*
25	este considerată (a)	*est considérée*

26	se referă (a) la	*se réfère à*
27	este o tulburare	*est un trouble*
28	sunt tulburări	*sont des troubles*
29	precum	*comme*
30	se caracterizează prin	*se caractérise par*
31	definește	*définit*
32	poate fi definit ca	*peut être défini comme*
33	este format din	*est constitué de*
34	este despre	*est sur*

Tableau 53. Liste élargie de marqueurs en roumain identifiés lors de l'annotation manuelle de reformulations

***

Dans le point suivant, nous présentons en détail les résultats d'annotation automatique sur chacun des sous-corpus qui constitue le corpus **GrandMed-Ro2**.

2.1.3 Résultats de l'annotation automatique des termes et des marqueurs

Le **Tableau 54** ci-dessous montre les différences entre les annotations automatiques réalisées avec des scripts en langage Perl appliquant les listes élargies de marqueurs de reformulation en roumain (présentées dans la **Partie III**, **sous-chapitre 2.3.2**). Nous menons les annotations sur chaque sous-corpus individuellement afin de garder les origines (les sites web) des reformulations. Nous illustrons le nombre total de phrases extraites par sous-corpus, dont nous mettons en avant le nombre de phrases qui contiennent les termes médicaux de notre liste de termes issue de **MoNERo**.

L'*Expérience 1* (**Exp 1**) indique le nombre de phrases qui ont des **termes médicaux et des marqueurs** selon notre **liste complétée de marqueurs** du **Tableau 14** (**Partie III, sous-chapitre 2.3.2**). Pour l'*Expérience 2* (**Exp 2**) nous rajoutons également la **liste élargie de marqueurs** des reformulation (**Tableau 53**), sur la base de l'analyse de corpus présentée dans la section précédente. Le nombre de phrases identifiées avec la liste élargie de marqueurs (*Expérience 2*) est plus grand pour la majorité des sous-corpus, dont trois sous-corpus ont une **précision** par rapport au nombre total de phrases de plus de **10%** (sous-corpus *romedic, csid-boli et regina-maria*). Néanmoins, l'analyse manuelle de phrases indique un grand nombre de *données bruitées* dans les résultats. Les *données bruitées* sont représentées par les phrases qui ne contiennent pas de reformulations. Nos observations nous montrent que ceci est dû au marqueur de type *indicateur* « boală (a) » (*maladie*) qui est cherché de manière indépendante dans les corpus.

Lors de nos premières expériences menées sur le corpus français, nous avons délibérément cherché *l'indicateur* « maladie » afin d'identifier de nouveaux marqueurs et indicateurs de reformulation. Dans une perspective d'amélioration de notre méthode de travail, nous souhaitons **affiner notre méthode d'identification de reformulations** à l'aide de **marqueurs spécifiques**. Dans ce sens, nous réalisons une troisième annotation (**Exp 3**) sur les sous-corpus roumains en supprimant « boală (a) » (*maladie*) de la *liste élargie* de marqueurs. Le nombre de phrases extraites automatiquement est moindre, mais il reste, pour la plupart des sous-corpus, plus élevé que celui de *l'Expérience 1* (**Exp 1**). Plus précisément, nous mettons en lumière les données quantitatives suivantes avec les écarts les plus grands :

- Pour le sous-corpus « **romedic** », nous avons obtenu **5 887** phrases, avec **843** phrases de plus que lors de *l'Expérience 1* ;
- Le sous-corpus « **sfatul medicului** » : **4 714** phrases, avec **704** de plus de l'**Exp 1** ;
- Pour le sous-corpus « **csid-boli** », nous avons extrait **2 685**, avec **502** phrases en plus également ;
- Pourtant, pour le sous-corpus « **regina-maria** », nous avons obtenu **3 627** phrases, avec **146** de *moins* que dans *l'Expérience 1* (**3 773** phrases). Même si nous avons un nombre plus réduit de phrases avec termes et marqueurs en éliminant le marqueur « boală (a) », nous gagnons en précision et nous limitons le travail d'annotation manuelle sur les phrases qui peuvent être des données bruitées (sans reformulations).

Corpus roumain GrandMed-Ro2	**Taille corpus (tokens)**	**N° phrases**	**N° phrases + termes - marq**	**N° phrases contenant des termes médicaux et des marqueurs Précision en pourcentage**					
				N° Exp 1	**% phr. total**	**N° Exp 2**	**% phr. total**	**N° Exp 3 - *boală***	**% phr. total**
sfaturi medicale	989 700	73 055	28 679	**5 015**	**6,86**	**6 002**	**8,21**	**5 074**	**6,94**
sfatul medicului	970 452	57 981	23 652	**4 010**	**6,91**	**5 385**	**9,28**	**4 714**	**8,13**
doctorul zilei	1 025 308	64 450	31 569	**3 611**	**5,60**	**4 870**	**7,55**	**4 044**	**6,27**
romedic	1 027 834	66 941	36 198	**5 044**	**7,53**	**6 849**	**10,23**	**5 887**	**8,79**
csid-boli	543 046	30 627	15 185	**2 183**	**7,12**	**3 169**	**10,34**	**2 685**	**8,76**
csid-sanatate	1 025 215	58 774	27 208	**3 357**	**5,71**	**4 593**	**7,81**	**3 718**	**6,32**
cdt-babes	198 052	11 731	4 957	**701**	**5,97**	**906**	**7,72**	**729**	**6,21**
regina-maria	619 204	44 895	20 971	**3 773**	**8,40**	**4 714**	**10,50**	***3 627***	***8,07***

Tableau 54. Résultats de l'annotation automatique des phrases avec les termes médicaux et les marqueurs de reformulation (liste complétée de marqueurs - Exp 1 ; liste élargie de marqueurs - Exp 2 ; liste élargie de marqueurs sans « boală » (maladie) - Exp 3)

L'étape suivante de notre travail est d'évaluer et annoter manuellement ces phrases qui ont des termes et des marqueurs de reformulation, suivant les mêmes étapes que pour le français. Nous présentons notre analyse dans le sous-chapitre suivant.

2.1.4 Annotation manuelle, évaluation et validation des reformulations

L'annotation et la validation des reformulations en roumain sont réalisées manuellement par nous uniquement, faute d'un annotateur locuteur natif du romain. Nous présentons nos expériences d'annotation sur les deux premiers sous-corpus, « sfaturi medicale » (*avis médicaux*) et « sfatul medicului » (*l'avis du médecin*). Nous avons suivi le même **guide d'annotation** que pour le français, présenté dans **l'Annexe 6.1**. Nous analysons les termes médicaux et les nouveaux marqueurs identifiés dans les reformulations médicales évaluées comme correctes.

Nous avons annoté manuellement **2 466** phrases du **sous-corpus 1** (**sfaturi medicale**) et **1 197 phrases** du **sous-corpus 2** (**sfatul medicului**), donc un total de **3 663 phrases annotées** en termes, marqueurs, reformulations, relations lexicales et fonctions sémantico-pragmatiques.

Pour le roumain, nous avons obtenu un total de **3 027 reformulations correctes,** dont **2 370 phrases (64,70%)** avec des reformulations correctes et **657 (+15,20%) reformulations multiples**, c'est-à-dire que plusieurs reformulations ont été annotées dans la même phrase. Nous avons identifié **1 889 reformulations correctes** pour le *sous-corpus 1* dont **1 585 (64,27%)** phrases avec une seule reformulation et **304 (+12,32%)** reformulations multiples. Pour le *sous-corpus 2*, nous avons annoté **1 138 reformulations correctes**, dont **785 (65,58%)** phrases avec une seule reformulation et **353 (+22,77%)** reformulations multiples. Les données quantitatives de l'annotation sont présentées en détail dans le **Tableau 55** ci-dessous.

Données quantitatives de l'annotation	**Sous-corpus GrandMed-Ro2**	
	1. *sfaturi medicale*	**2. *sfatul medicului***
Reformulations avec ***oui***	1255	727
Reformulations avec ***oui<inv>***	330	58
Reformulations avec ***oui – total (phrases initiales)***	**1585**	**785**
Reformulations avec ***oui+2***	263	322
Reformulations avec ***oui<inv>+2***	41	31
Reformulations avec ***oui+2 – total (plusieurs / phrase)***	**304**	**353**
Reformulations avec ***oui/<inv> & oui+2 – total***	**1889**	**1138**
Reformulations avec ***non***	881	412
N° total de phrases annotées par sous-corpus	**2466**	**1197**
	N° phrases	***Précision %***
N° total de reformulations avec « oui »	**2370**	**64,70%**
N° total de phrases avec « non »	**1293**	**35,29%**
Total phrases initiales	**3663**	**100%**
N° total de reformulations avec « oui » + *multiples*	**3027**	*+15,20%*
Total reformulations annotées	**4320**	***115,20%***

Tableau 55. Données quantitatives et précision de l'annotation de deux sous-corpus de GrandMed-Ro2 : « sfaturi medicale » et « sfatul medicului »

***

La section suivante présente quelques statistiques sur les termes médicaux roumains reformulés et sur les marqueurs / indicateurs de reformulation.

2.1.4.1 Analyse quantitative des termes médicaux reformulés

Lors de l'annotation de **3 663 phrases**, nous avons identifié **3 027 termes médicaux reformulés** en roumain. Nous avons supprimé les articles antéposés aux noms comme :

- L'article indéfini au masculin, singulier en roumain, « un » (*un*) avec 16 occurrences ;
- L'article indéfini au féminin, singulier « o » (*une*) avec 18 occurrences.

Nous avons identifié :

- **1 106 termes médicaux simples** (**36,53%** de termes identifiés) ;
- **1 921 termes médicaux polylexicaux** (**63,46%** du total de termes identifiés).

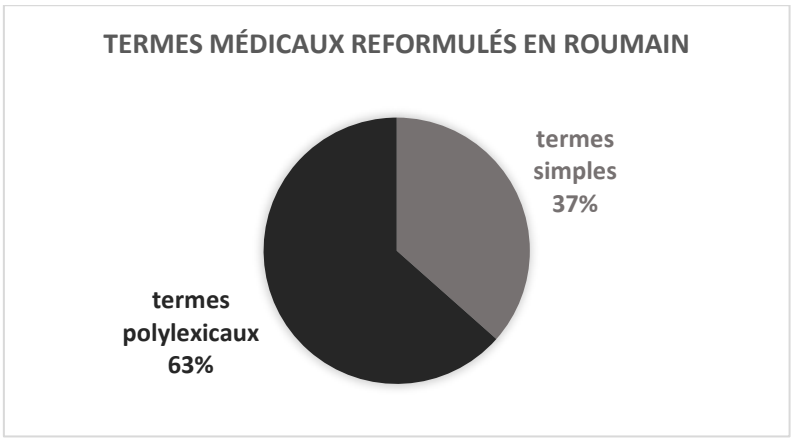

Figure 26. Types de termes médicaux reformulés extraits du corpus GrandMed-Ro2

Par la suite, nous analysons les **termes médicaux polylexicaux avec des expressions régulières** dans le logiciel Notepad++. Nous identifions plusieurs types de termes polylexicaux :

- **1 003 termes bi-grammes (52,21 % de termes polylexicaux)** de type Nom-Adjectif, comme « tulburari digestive » (*troubles digestifs*), « sarcina extopica » (*grossesse extra-utérine*), « laxative osmotice » (*laxatifs osmotiques*) (expression régulière : ^\w+\s\w+$) ;
- **418 termes tri-grammes (21,75% de termes polylexicaux)** de type Nom-Préposition-Nom et Nom-Adjectif-Adjectif, par exemple « hernie de disc » (*hernie discale*), « tromboza venoasa profunda » (*thrombose veineuse profonde*), « nefropatia cu IgA » (*néphropathie à IgA*) (expression régulière : ^\w+\s\w+\s\w+$) ;
- **500 termes avec quatre tokens ou plus** (**26,02% de termes polylexicaux)**, de type « boala pulmonara obstructiva cronica avansata » (*maladie pulmonaire obstructive chronique avancée*), « sindromul de detresa respiratorie acuta » (*le syndrome de détresse respiratoire aiguë*), « sindromul Wolff Parkinson White » (*le syndrome de Wolff Parkinson White*) (expression régulière : ^\w+\s\w+\s\w+\s\w+).

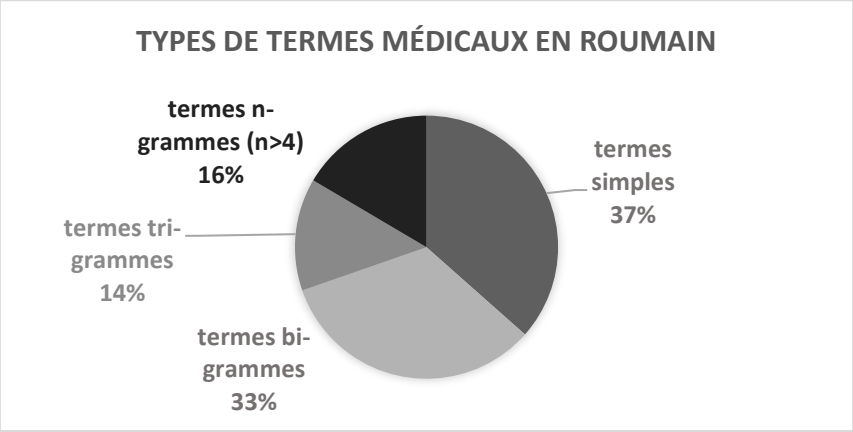

Figure 27. Sous-types de termes médicaux reformulés extraits du corpus GrandMed-Ro2

Parmi nos 3 027 termes médicaux identifiés, 1 014 sont des doublons. En supprimant les doublons, nous avons une liste finale de **2 014 termes médicaux uniques reformulés en roumain**. Nous gardons toutes les occurrences de termes doublons pour analyser les reformulations qui peuvent être différentes pour le même terme, comme pour le terme « astmul » (*l'asthme*) :

- « <terme>**Astmul**</terme> este o <ref>boala cronica ce e caracterizata prin episoade recurente de ingustare a cailor respiratorii</ref> »
  (<terme>L'asthme</terme> est une <ref>maladie chronique caractérisée par des épisodes récurrents de rétrécissement des voies respiratoires</ref>)
- « <terme>**Astmul**</terme> este o <ref>afectiune pulmonara cronica cu simptome precum, tusea, respiratie dificila si suieratoare</ref> »
  (<terme>L'asthme</terme> est une <ref>maladie pulmonaire chronique dont les symptômes sont la toux, la respiration sifflante et l'essoufflement</ref>.)
- « <terme>**Astmul**</terme> este o <ref>boala care afecteaza bronsiile</ref>. »
  (<terme>L'asthme</terme> est une <ref>maladie qui affecte les bronches</ref>.)

Ces différentes reformulations introduites par les marqueurs « este o boala » (*est une maladie*) et « este o afectiune » (*est une affection*) nous serviront pour **enrichir notre corpus de reformulations médicales** et pour **entraîner les outils d'apprentissage automatique par réseaux de neurones**, travail que nous présentons par la suite dans le **Chapitre 3.**

Un extrait de la liste complète de **termes reformulés uniques en roumain**, en ordre alphabétique, peut être consulté dans **l'Annexe 6.8**.

2.1.4.2 Analyse quantitative des marqueurs et indicateurs identifiés

Nous menons une analyse quantitative des **3 027 reformulations médicales en roumain** identifiées. Nous observons que les **marqueurs et indicateurs de reformulation** les plus fréquents sont :

- **943** reformulations (**31,15%**) sont marquées par le marqueur « **cum ar fi** » (*tel que, comme*). Ce marqueur indique une relation lexicale d'hyponymie avec une fonction sémantico-pragmatique d'exemplification, comme le prouvent les exemples suivants :
  - **«** afectiuni gastro-intestinale, **cum ar fi** ulcerul gastric si cancerul » *(maladies gastro-intestinales **telles que** l'ulcère gastrique et le cancer) ;*
  - « tulburari de neurodezvoltare, **cum ar fi** tulburarea de hiperactivitate cu deficit de atentie » *(troubles du développement neurologique **tels que** le trouble de l'hyperactivité avec déficit de l'attention).*
- **371 reformulations (12,25%) de type dénomination** sont marquées à l'aide du lemme « **numit** » (*appelé)* en tant qu'adjectif et verbe « **se numeste** » (*s'appelle*), qui fait partie de nombreux marqueurs : « **denumit »** *(connue sous le nom de)*, **« supranumit »** *(surnommé)* (toujours avec des paraphrases populaires), **« numit popular si »** *(populairement appelé)*, **« deseori numita si »** *(appelée parfois aussi)*, **« numita medical »** *(médicalement appelée)*, **« asa-numitele »** *(ainsi nommées)*, **« denumit sugestiv »** *(suggestivement appelé)* et autres que nous présentons dans la section suivante. Quelques exemples de reformulation de type dénomination :
  - Sindromul intestinului iritabil **numit si** colon spastic *(Syndrome du côlon irritable, **également appelé** côlon spastique)* ;
  - Magneziul **supranumit** « stapanul mineralelor » *(Le magnésium, **surnommé** « le maître des minéraux »)* ;
  - Mononucleoza infectioasa **denumita si** « boala sarutului » *(Mononucléose infectieuse **également connue sous le nom de** « maladie du baiser »).*
- **259 reformulations (8,55%)** marquées par « **reprezinta** » (*représente*), dont **168** (**64,86%)** sont de type *hypéronymie-définition* et le reste de type *méronymie-définition* et des *dénominations* :
  - Hiperplazia **reprezinta** o proliferare de celule necanceroase *(L'hyperplasie **représente** une prolifération de cellules non cancéreuses)* ;
  - Glicemia **reprezinta** valoarea zaharului din sange *(La glycémie **représente** la quantité de sucre dans le sang)* ;

  - Hemotoraxul **reprezinta** acumularea unei cantitati de sange intre peretele toracic si plamani *(L'hémothorax **représente** l'accumulation de sang entre la paroi thoracique et les poumons)*.
- **101 reformulations (3,33%)** de type *méronymie-définition* sont identifiées à l'aide du marqueur « **caracterizat** » (*caractérisé*) ou « **se caracterizeaza prin** » (*se caractérise par*) :
  - Diabetul zaharat **caracterizat prin** cresterea cronica a glicemiei *(Diabète sucré **caractérisé par** une élévation chronique du taux de glucose dans le sang)* ;
  - Boala Alzheimer **caracterizata prin** reducerea capacitatilor cognitive si comportamentale *(La maladie d'Alzheimer **caractérisée par** une réduction des capacités cognitives et comportementales)* ;
  - calculii simptomatici **se caracterizeaza prin** episoade de colici biliare si litiaza complicata *(les calculs symptomatiques **sont caractérisés par** des épisodes de colique biliaire et de lithiase compliquée)*.

Nous remarquons que les **parenthèses isolées ou avec un autre marqueur lexical** ne sont pas aussi fréquentes que dans les corpus français. Elles marquent **541 reformulations** correctes de **3 027 (17,87%)**, par rapport à **2 099** sur **5 198 (40,38%)** du corpus **CLEAR** (1 076 dans le corpus CLEAR SP et 1 023 dans CLEAR GP), et **907** sur **3 432 (26,42%)** reformulations correctes du corpus **ClassYN** (dont 242 dans le corpus ClassYN SP et 665 dans le corpus ClassYN GP).

2.1.4.3 Nouveaux marqueurs et indicateurs de reformulations

Lors de notre analyse manuelle de phrases, nous avons identifié d'autres **nouveaux marqueurs et indicateurs de reformulation**, comme :

- intitulata si *(intitulée aussi)* ;
- de tip / de tipul *(de type)* ;
- feluri *(types)* ;
- desemna *(désigne)* ;
- seamana clinic cu *(ressemble cliniquement à)* ;
- diversele *(les divers)* ;
- sau orice alte *(ou toute autre)* ;
- printre care si *(parmi lesquelles aussi)* ;
- semnifica (*signifie)* ;
- altfel spus *(autrement dit)* ;
- din totalul *(du total)*.

Certains marqueurs ne sont pas entièrement nouveaux par rapport à notre liste en français, comme « semnifica » (*signifie*), « altfel spus » (*autrement dit*) ou « desemna » *(désigne)*, mais ils sont nouveaux par rapport à nos recherches antérieures sur le corpus roumain.

Nous avons identifié plusieurs marqueurs de **langage métalinguistique**. Dans les textes on parle directement de la notion de *terme* ou *terme médical*, comme le montre les marqueurs suivants :

- se foloseste uneori termenul de *(on utilise parfois le terme)* ;
- este termenul care se refera la *(est le terme qui le réfère à)* ;
- este termenul medical care se refera la *(est le terme médical qui se réfère à)* ;
- asa cum este denumita in termeni medicali *(comme elle est appelée en termes médicaux)*.

Le marqueur « denumita » (*appelée*) est présent dans différentes configurations :

- asa cum este denumita afectiunea *(comme est appelée l'affection)* ;
- sau, cum mai este denumita *(ou, comme elle est encore appelée)*.

Nous avons observé que pour le marqueur « numite » (*nommées*), le terme est postposé au marqueur, avec un schéma de type « ***reformulation -> numit (nommé) -> terme médical*** » et une relation pragmatique de type **synonymie / paraphrase.**

2.1.5 Analyse lexicale et sémantico-pragmatique des reformulations

Notre approche théorique sur les relations lexicales, les fonctions sémantico-pragmatiques et notre hypothèse concernant les liens qui peuvent exister entre celles-ci sont présentées en détail dans la **Partie III**, **Chapitre 2.5.** Nous identifions les mêmes relations et fonctions que pour le français.

Nous présentons notre première analyse sur la *relation lexicale d'hyperonymie et la fonction sémantico-pragmatique de définition* dans le sous-corpus « sfaturi medicale » (*avis médicaux*). Nous les analysons selon notre **hypothèse de recherche** qui postule que ces deux éléments sont corrélés.

2.1.5.1 Analyse préliminaire sur le sous-corpus 1 « sfaturi medicale » (avis médicaux)

Nous procédons à une analyse détaillée des données, pour identifier les relations lexicales (synonymie, hyperonymie, hyponymie, méronymie) et les relations sémantico-

pragmatiques (définition, dénomination, exemplification, explication, paraphrase). Nous supposons que certains marqueurs sont corrélés à des relations spécifiques.

Une première hypothèse postule qu'une grande partie des phrases contenant le marqueur générique de reformulation « **este o / un** » (*est une / un*) contiennent des reformulations avec :

- **la relation lexicale d'hyperonymie**, ce qui signifie que la reformulation médicale est l'hyperonyme du terme médical (qui devient l'hyponyme) ;

- **la relation sémantico-pragmatique de définition**, soulignant que l'intention de la reformulation est de donner une *définition plus simple* du terme médical scientifique, c'est-à-dire de rendre le contenu notionnel du terme plus accessible.

Afin de tester la validité de cette hypothèse dans notre sous-corpus, nous avons utilisé les formes « este o afecțiune » (*est une affection*), « este o boală » (*est une maladie*), mais aussi les hyperonymes médicaux de type « afecțiune » (*affection*), « boală » (*maladie*) séparément. Nous avons décidé de ne pas rechercher le marqueur générique « este o /una ; este un » (*est une / est un*) seul afin de ne pas avoir trop de résultats erronés.

Dans ce sous-corpus de 2 466 phrases, nous avons trouvé beaucoup plus de marqueurs au féminin de type « este o » (*est une*) (365 occurrences), seulement 6 occurrences pour « este una » (*est une* ; forme féminine avec article défini enclitique), et 48 pour « este un » (*est un*), la forme masculine du déterminant. Ces marqueurs sont apparus dans 419 phrases. Nous avons ensuite analysé et annoté chaque phrase pour voir si ces marqueurs spécifiques aident à identifier les reformulations ayant le rôle sémantico-pragmatique de définition.

Nos résultats ont montré que 61 (14,55%) phrases n'étaient pas des reformulations médicales, car la phrase était incomplète : « Este o boală care afectează vederea centrală. » (*Est une maladie qui affecte la vision centrale*). Comme nous observons dans cette phrase, le terme médical reformulé n'apparait pas dans la même phrase, alors nous ne pouvons pas le considérer comme une reformulation sous-phrastique. Un autre exemple similaire est le suivant : « Aceasta este o afecțiune caracterizată de lipsa celulelor nervoase din intestinul gros. » *(Il s'agit d'une affection caractérisée par le manque de cellules nerveuses dans le gros intestin*). Ces exemples sont similaires aux cas de coréférence de type « cette maladie » que nous n'avons pas retenu comme des reformulations. Pour cette étude, nous avons conclu que les phrases commençant par « Este o/un/una » (*Est un*) ou « Aceasta este » (*C'est*) ne sont pas des reformulations, car nous ne pouvons pas identifier le terme médical qui est reformulé (il n'est pas présent dans la même phrase).

À l'issue de notre analyse manuelle des 419 phrases avec le marqueur « este o/una/un » (est une/un), nous avons obtenu **358 phrases (85%)** contenant des reformulations correctes qui confirment notre hypothèse sur la concordance entre ce marqueur, la relation lexicale *d'hyperonymie* et la fonction sémantico-pragmatique de *définition*.

2.1.5.2 Analyse sur les toutes phrases annotées du corpus GrandMed-Ro2

Parmi les **2 370** phrases qui contiennent des reformulations correctes pour le roumain (sans les reformulations multiples par phrases, en nombre de 657), **877 (37%)** ont sont de type *hyperonymie-définition*, **695** (29,32%) sont des *hyponymies-exemplifications*, **419 (17,67%)**. Nous remarquons un grand nombre de reformulations de type *synonymie-paraphrase* **(279 ; 11,77%)** ce qui prouve que les textes de vulgarisation médicale utilisent des *versions simplifiées ou populaires* du terme pour le rendre plus facile à comprendre pour le grand public. Nous notons aussi la présence des définitions introduites par des *méronymes* **(170 ; 7,17%)** et le petit nombre de *méronymies-explications* **(80 ; 3,37%)**.

Relation lexicale	**Fonction sémantico-pragmatique**	**GrandMed-Ro2**	
		N°	**%**
hyperonymie	définition	**877**	**37%**
hyponymie	exemplification	**695**	**29,32%**
synonymie	paraphrase	**279**	**11,77%**
	dénomination	140	5,90%
méronymie	explication	80	3,37%
	définition	170	7,17%
Total		**2370**	**100%**

Tableau 56. Paires de relations lexicales et fonctions sémantico-pragmatiques annotées dans le corpus roumain GrandMed-Ro2

Nous avons mené une comparaison avec un deuxième annotateur dont la langue maternelle est le roumain[80] sur **1 000 reformulations correctes.** Nous observons que les deux annotateurs ont été d'accord pour attribuer les mêmes **relations lexicales** à **60,80%** de reformulations et les mêmes **fonctions sémantico-pragmatiques**, séparément de relations, à **60,10%** de reformulations. Concernant les paires de *relations-fonctions*, les annotateurs ont attribué les mêmes pour **48,70%** de reformulations.

[80] L'annotateur est Amalia Todirascu, Professeur des Universités.

2.1.6 Bilan des analyses sur les corpus roumains

Les travaux réalisés sur le corpus roumain ont donné de meilleurs résultats en termes de précision **(65,09%)** d'identification des reformulations médicales correctes que ceux sur les corpus français. Ces résultats améliorés montrent que notre méthode d'identification des reformulations est plus précise grâce à l'utilisation d'une liste étendue de différents **marqueurs et indicateurs lexicaux de reformulation**.

Dans nos futurs travaux, nous prévoyons d'élargir également la liste des marqueurs et indicateurs pour le français (par rétrotraduction de la liste en roumain) et réaliser une **double annotation** de reformulations en roumain afin de calculer un accord annotateur de type Kappa, comme pour le corpus français. Lors des expériences futures, nous pourrions également inclure l'analyse des **chaînes de coréférence** (Schnedecker et Landragin, 2014) afin de trouver le *référent* (dans notre cas, le terme médical) de l'anaphore « aceasta » (*ceci*) dans les contextes précédents, mais cela dépassera le contexte de la phrase, donc de la reformulation sous-phrastique.

2.2 Bilan contrastif des analyses sur les corpus français et roumains

Nous avons annoté et validé manuellement un total de **19 890 phrases**, dont **16 227 phrases** du corpus français (**8 667** du **CLEAR Cochrane** et **7 560** du **ClassYN**). Nous avons également annoté **3 663 phrases** du corpus roumain, **GrandMed-Ro2**. Pour la langue française, **59%** de phrases ont une *double annotation* et *adjudication* entre annotateurs pour obtenir un accord inter-annotateur. Pour la langue roumaine, seulement **27%** des phrases ont eu une double annotation.

Notre hypothèse de départ sur la présence d'un nombre plus important des reformulations dans les textes grand public est confirmée. La seule exception est le corpus **CLEAR SP**, qui contient beaucoup de reformulations de type abréviations, ce qui fait augmenter le pourcentage de reformulations correctes à **48,34%**. Nous avons obtenu un meilleur score de précision sur l'annotation de phrases en roumain, **64,70 %** (pourtant, annotation unique, pas de double annotation) par rapport aux meilleurs scores sur les corpus français :

- **48,34%** sur le corpus **CLEAR SP** ;
- **41,77 %** sur le corpus **ClassYN GP** ;
- **38,51 %** sur le corpus **CLEAR GP** ;
- **35,96 %** sur le corpus **CLEAR** dans son intégralité ;
- **35,92%** sur le corpus **ClassYN GP** (annotation unique ; 57,50% double annotation sur un extrait) ;
- **33,63 %** sur **ClassYN** dans son intégralité ;
- **25,50%** sur le corpus **ClassYN SP**.

Cette amélioration de la précision observée sur l'annotation automatique est due à **l'agrandissement et l'affinement de la liste de marqueurs et indicateurs de reformulation en roumain**. Nos travaux réalisés sur le français, en partant des marqueurs de reformulation identifiés dans la littérature (en français et en roumain), en complétant avec nos observations sur les corpus et nos annotations manuelles des reformulations, nous ont permis de **constituer une liste finale** élargie, précise et nettement améliorée, des marqueurs et indicateurs de reformulations médicales en roumain.

Concernant **l'analyse linguistique** des données annotées, nous observons que les reformulations médicales de type *hyperonymie-définition* sont **les plus fréquentes** dans les corpus suivants, en ordre décroissant :

- **CLEAR GP** : **49,46%** ;
- **CLEAR SP** : **46,22%** ;
- **ClassYN GP** : **38,97%** ;
- **GrandMed-Ro2** : **37%**.

Nous constatons que pour le corpus **ClassYN SP,** les reformulations les plus fréquentes (**42,01%**) sont de type *hyponymie-exemplification*, qui représente la deuxième relation la plus fréquente dans les données en roumain également (**29,32%**).

Nous avons découvert une **nouvelle relation** qui n'était pas dans nos hypothèses de départ (**sous-chapitre 2.5.1**) : les reformulations de type *méronymie-définition*. Cette relation reformule le terme à travers une définition d'un élément qui fait partie du terme (relation entier (*le terme*) – partie (*la reformulation*)). Ces reformulations ont été les plus annotées dans le corpus **ClassYN GP**, à **13,77%**, **ClassYN SP (9,02%)** et **GrandMed-Ro2 (7,17%)**, par rapport au corpus **CLEAR SP (3,18%)** et très peu présentes dans le corpus **CLEAR GP (0,67%)**. Il y a une différence importante entre les textes du projet **CLEAR** et **CLASSYN** : les articles scientifiques de ce dernier s'adressent partiellement aux étudiants de licence et de master, alors que dans **CLEAR**, les textes scientifiques s'adressent aux spécialistes confirmés.

Nous envisageons de continuer les annotations sur le corpus roumain afin d'identifier davantage de reformulations médicales correctes en roumain. Notre hypothèse actuelle est que la précision risque de se diminuer lorsque le nombre de phrases annotées augmentera. L'appel à un autre annotateur est aussi nécessaire pour avoir une double annotation et il est possible d'avoir des désaccords entre les annotateurs. Néanmoins, notre **liste élargie de marqueurs de reformulation en roumain** nous permettra d'identifier un nombre plus grand de reformulations correctes par rapport à notre liste initiale sur le français (et réduire le nombre de données bruitées) et de construire un corpus entièrement avec double annotation.

***

Nous présentons par la suite les expériences avec des outils d'apprentissage automatique tels que les **réseaux de neurones** et les **modèles de langues** de type ***Transformers*** afin d'identifier automatiquement des reformulations médicales en français et en roumain.

3. Expériences d'apprentissage automatique neuronal

Dans cette section nous présentons nos expériences avec des outils d'apprentissage automatique par réseaux de neurones et l'analyse des résultats obtenus. Notre objectif est, d'un côté, de **générer d'autres reformulations médicales à partir de nos données annotées** précédemment sur les deux corpus, français et roumain, et ainsi **agrandir automatiquement nos corpus de reformulations**. Nous expliquons ci-dessous notre méthode étape par étape. Pour nos expériences de **génération de reformulations à partir du terme**, nous utilisons **l'architecture neuronale APT** (*Adversarial Paraphrasing Task*) (Nighojkar et Licato, 2021), et un modèle de langue de type **Transformer T5** (présenté dans la **Partie III**, **sous-chapitre 2.6.1.2**). D'un autre côté, nous avons mené des expériences de **classification automatique des reformulations** à l'aide d'une **architecture LSTM**, utilisant les mêmes corpus annotés présentés dans les chapitres précédents. Ainsi, nous identifions les phrases qui peuvent contenir des reformulations.

3.1.1 Nos expériences et les données d'entraînement

Nous avons mené nos expériences de **génération de la reformulation** avec l'architecture **APT**. L'objectif est de présenter un terme en entrée et d'obtenir des nouvelles reformulations en sortie. **APT** a été testé par ses développeurs (Nighojkar et Licato, 2021) utilisant le Transformer **T5** *(Text-to-Text Transformer)* de Google (Raffel *et al*., 2020). Nous avons décidé de le tester également, car il contient quatre langues, dont nos langues d'étude : l'allemand, l'anglais, **le français et le roumain**. Nous adaptons le modèle **T5** à l'aide de nos corpus de reformulations, pour en générer des nouvelles reformulations.

Nous avons lancé nos expériences sur le cluster du **Centre de Calcul de l'Université de Strasbourg (CCUS)**[81]. Nous avons adapté et entraîné deux modèles, le Transformer **T5-small** (préentraîné sur 60 millions de paramètres) et sa version complète,

[81] Le Centre de Calcul de l'Université de Strasbourg (CCUS) fait partie de centres de calcul régionaux les plus puissants en France. Il est géré par les pôles CESAR (Calcul Et Services Avancés à la Recherche) et ICS (Infrastructures Cloud et Services) de la Direction du numérique de l'Université de Strasbourg (informations extraites du site https://hpc.pages.unistra.fr/, accédé à 18h57, le 17/11/2022).

T5-base (220 millions de paramètres) pour les deux langues, et **mT5-small** et **mT5-base** pour le roumain :

- avec les **8 626 paires de *terme – reformulation*** issues de nos expériences et annotations sur les corpus français CLEAR Cochrane et ClassYN ;
- avec les **3 027 paires de *terme – reformulation*** du corpus roumain GrandMed-Ro2.

Pour cela, nous avons découpé les données en trois parties (pour chaque langue) :

1) **Un corpus d'entraînement** : **8 146** paires de *terme – reformulation* correctes issues de nos annotations automatiques et évaluations humaines sur les corpus CLEAR Cochrane et ClassYN et **2 727** paires de *terme – reformulation* du corpus roumain GrandMed-Ro2 ;
2) **Un corpus de validation** : l'évaluation est réalisée pendant le processus d'apprentissage sur des blocs de vingt exemples corrects de paires *termes – reformulation* ;
3) **Un corpus de test** : **480** paires *terme – reformulation* extraites de la liste de reformulations du corpus français et **300** paires *terme – reformulation* du corpus roumain (car nombre plus réduit de phrases annotées). Ces exemples n'ont pas été utilisés pour l'entraînement pour éviter les biais.

JEU DE DONNÉES		
Corpus	**Phrases d'entraînement**	**Phrases de test**
CLEAR Cochrane SP	2 528	**480**
CLEAR Cochrane GP	2 668	
ClassYN SP	1 197	
ClassYN GP	2 233	
Total corpus français	**8 626**	
GrandMed-Ro2	**3 027**	**300**

Tableau 57. Jeu de données pour l'entraînement du Transformer T5 issues de nos données annotées en français et en roumain

3.2 Résultats de la génération en français

Nous avons réalisé notre première expérience avec le modèle **T5-small**. L'entraînement sur nos données a duré **8 heures**, mais les reformulations générées restaient très proches du terme médical original. Nous avons lancé une deuxième

expérience avec le modèle plus grand, **T5-base**, qui a **220 millions de paramètres** (3,66 fois plus grand que le modèle T5-small). Nous avons modifié les paramètres pour avoir entre 1 et 5 prédictions de reformulations médicales pour chacun de 480, respectivement 300, termes médicaux de la liste de test. L'entraînement avec **T5-base** a duré **24 heures** et les reformulations générées semblent plus proches de la reformulation originale, présente dans le corpus annoté manuellement. Nous avons utilisé une taille maximale de 256 mots, respectivement 128 mots (pour les reformulations), le taux d'apprentissage (3e-4), 4 epochs et des batchs (échantillons d'entraînement et validation) de taille 20 paires, le paramètre concernant la réduction des poids (0,01), un optimiseur AdamW (l'epsilon 1e-8). Nous évaluons manuellement chaque prédiction pour cette deuxième expérience et nous présentons les résultats de notre analyse.

Dans le prochain point, nous présentons en détail les résultats de notre évaluation et nous analysons les prédictions de reformulations générées automatiquement avec l'architecture proposée pour les deux langues.

3.2.1 Génération de reformulations avec T5-base

Nous avons obtenu **2 268 prédictions de reformulations** médicales générées automatiquement pour les **480 termes médicaux en français** de la liste de test. Les résultats se présentent sous la forme suivante :

```
bronchodilatateurs courants
Truth:
(médicaments utilisés pour élargir les voies respiratoires)
Prediction:
(par exemple PDA)
(bronchodilatateurs courants)
tel que les bronchodilatateurs courants
(par exemple air d'admission dans les airs de l'hôpital)
(par exemple la thoracocib)
```

Nous observons que le terme médical apparait en premier « bronchodilatateurs courants », suivi par la reformulation médicale issue de nos annotations sur le corpus précédée de la mention en anglais *Truth:* « (médicaments utilisés pour élargir les voies respiratoires) ». La reformulation attendue (celle du *Truth*) est suivie par les prédictions automatiques du modèle de langue **T5-base** lancé avec l'architecture **APT**, après la mention en anglais *Prediction:*.

Afin d'évaluer la qualité de la génération automatique, nous créons **une échelle d'annotation et d'évaluation des prédictions**, valable pour les deux langues, que nous présentons en détail ci-dessous.

3.2.2 Échelle d'évaluation de prédictions de reformulations

L'échelle d'annotation conçue pour évaluer les prédictions est construite en trois parties : *une première partie* qui évalue chaque prédiction de reformulation générée, *une deuxième partie* qui calcule la moyenne de toutes les prédictions d'un terme, et *une troisième* qui évalue si au moins une des prédictions générées est correcte parmi toutes les reformulations générées pour un terme médical à la fois.

La première partie contient les valeurs numériques suivantes :

- La valeur **2** : pour les prédictions identiques à la reformulation médicale initiale (*Truth*) ;
- La valeur **1** : pour les prédictions correctes, mais différentes de la reformulation médicale initiale ;
- La valeur **0** : pour les prédictions incorrectes. Même s'il y a des parties de reformulations correctement construites, nous avons attribué la valeur 0 en cas de mots inventés ou des mots inadaptés dans le contexte (par exemple le terme désigne une inflammation de la peau et la reformulation parle des symptômes des infections urinaires) ;
- La valeur **-1** : pour les répétitions du terme médical.

La deuxième partie calcule la moyenne de toutes les prédictions d'un terme ;

La troisième partie contient les valeurs numériques et les interprétations suivantes :

- La valeur **1** : le terme médical a au moins une prédiction de reformulation médicale correcte ;
- La valeur **0** : le terme médical n'a reçu que des prédictions de reformulations médicales qui ne sont pas correctes.

Notre **échelle d'évaluation** prend la forme suivante dans un tableur Excel :

Terme	Truth	Prédictions t5-base ; Légende : Prédiction=Truth : 2 ; Prédiction=OK(autre-ref) : 1 ; Prédiction=NOK : 0 ; Prédiction=Identique-au-terme : -1	Score par prédiction	Moyenne du score par prédiction	Score si 1 préd. correcte : 1 ; All-NOK = 0
maladie d'Alzheimer	Truth: ou démence	Prediction: / maladie Alzheimer ou maladies inflammatoires	0	0.8	1
		, ou d'autres troubles cognitifs	1		
		/ altération d'au moins 12 ml de vie	0		
		et les autres troubles cognitifs	1		
		ou démence	2		

Figure 28. Échelle d'évaluation des prédictions automatiques de reformulations médicales générées avec APT

Nous présentons notre analyse de prédictions automatiques ci-dessous.

3.2.3 Évaluation et analyse des prédictions automatiques

Nous montrons dans le **Tableau 58** les résultats de notre évaluation manuelle de la qualité des prédictions de reformulations générées automatiquement. Parmi les 480 termes médicaux du corpus de test, **180 termes ont été attribués des prédictions de reformulations correctes**, représentant **37,50%** des termes.

Prédiction avec T5-base	N° termes	%
au moins une prédiction correcte	180	**37,50%**
aucune prédiction correcte	300	**62,50%**
Total	**480**	**100%**

Tableau 58. Statistiques sur les résultats des prédictions de reformulations avec T5-base en français

Nous analysons en détail les scores donnés à chacune de **2 268 prédictions** générées automatiquement par l'outil (voir **Tableau 59**). Nous observons que pour **73 prédictions**, le modèle **T5-base** a appris et a généré correctement les reformulations correspondantes pour seulement 27 termes médicaux, c'est-à-dire qu'il a généré la reformulation de la colonne *Truth.* Ces prédictions sont de plusieurs types :

- **Abréviations** :

  Terme : troubles associés à l'entorse cervicale ; **Truth:** (TAEC) ; **Prediction:** (TAEC)
- **Hyperonymes** :

Terme : maladies cardiovasculaires, l'ostéoporose et la démence ; **Truth:** telles que maladies chroniques ; **Prediction:** maladies chroniques

- **Paraphrases** :

  Terme : syndrome confusionnel ; **Truth:** (délirium) ; **Prediction:** (délirium)

- **Dénominations** :

  Terme : sclérose latérale amyotrophique ; **Truth:** / maladie du motoneurone ; **Prediction:** / maladie du motoneurone

Nous avons observé que dans les cas des **dénominations**, l'apparition fréquente de la paire respective de *terme – reformulation* dans le corpus d'apprentissage joue un rôle important dans la précision de la prédiction générée. Par exemple, nous comptons 16 prédictions générées automatiquement de type « maladie du motoneurone » pour le terme « sclérose latérale amyotrophique », qui apparait 11 fois dans la liste de termes du corpus de test. Nous remarquons également que, même si le modèle **T5-base** a été paramétré pour générer **maximum 5 prédictions** pour chaque terme médical, il peut trouver la reformulation correcte exacte (selon *Truth*) avec **un seul essai,** comme pour les termes médicaux polylexicaux « essais contrôlés randomisés » ; Truth: (ECR) ; Prediction: (ECR) ou « Organisation mondiale de la Santé » ; Truth: (OMS) ; Prediction: (OMS).

Prédiction avec T5-base	N° prédictions	%
Score 2	73	3,21%
Score 1	244	10,75%
Scores positifs	***317***	***13,97%***
Score 0	1848	81,48%
Score -1	94	4,14%
Scores négatifs	***1942***	***85,62%***
Total	**2268**	**100%**

Tableau 59. Différents scores de l'échelle d'évaluation de prédictions avec T5-base pour le français

Parmi toutes les prédictions générées, **317** (**13,97%**) des prédictions générées sont des **reformulations correctes** pour les **180 termes** (**37,50%**). Le Transformer a généré également ***d'autres* reformulations correctes que celle de la colonne *Truth*** parmi ses prédictions, **244** (**10,75%**) plus précisément. Cela veut dire que le Transformer a *généré* de **nouvelles reformulations**, en dehors de celles qui ont été données comme exemples lors de l'entraînement. Nous illustrons les exemples suivants de ***reformulations nouvelles*** :

- **66 reformulations médicales** (**2,91%**) ont été générées automatiquement pour le terme « **schizophrénie** » qui apparait 19 fois dans la liste de termes de test, occurrences issues aléatoirement du corpus annoté CLEAR, par exemple :

  schizophrénie
  Truth: est un/e maladie mentale débilitante qui touche environ un pour-cent de la population mondiale
  Prediction:
  / troubles mentaux chroniques
  troubles schizophréniformes
  ou troubles mentaux chroniques
  maladies et troubles mentaux graves

- **T5-base** a généré des reformulations nouvelles qui sont même **plus faciles à comprendre** que la reformulation originale, comme dans l'exemple :

  programme de promotion du port du casque
  Truth: (par exemple incluant des éducateurs, ceux développant les aptitudes telles que la prise de décision et la résistance à la pression des pairs ou améliorant le respect de soi ou la connaissance de ses propres capacités)
  Prediction: c'est-à-dire de mesures destinées à renforcer l'image du port de casque

- Dans le cas des abréviations en anglais, T5 propose **l'abréviation en français** pour le terme médical correspondant, comme dans l'exemple ci-dessous :

  facteur neurotrophique ciliaire
  Truth: (CNTF pour Ciliary Neurotrophic Factor)
  Prediction: (FNC)

T5-base a généré **1 942** reformulations incorrectes, qui représentent **85,62**% de toutes les reformulations. Celles-ci ont été générées pour **393** termes médicaux (**81,87%**). Parmi ces reformulations incorrectes, **94** (**4,14%**) sont des répétitions de **16** termes médicaux (**3,33%**). Nous avons considéré dans cette catégorie également les répétitions du terme médical avec un marqueur de reformulation, de type :

hospitalisation pour une opération chirurgicale
Truth: ou traumatisme
Prediction: (par exemple l'hospitalisation pour une opération chirurgicale)

dissection de l'artère carotide
Truth: désigne déchirure dans la paroi de l'un des principaux vaisseaux sanguins transportant le sang vers le cerveau
Prediction: (définie comme désection de l'artère carotide)

Pour éliminer les répétitions du même mot dans la prédiction générée, nous avons mené une deuxième expérience (**Exp 2**) avec **T5-base** de l'architecture **APT** (**sans répétition du même mot**) (la technique d'apprentissage profond s'appelle *fine-tuning*, présentée dans le **sous-chapitre 3.1.1, Partie III**). Nous avons évalué les prédictions générées lors de cette deuxième expérience pour le corpus français et nous comparons les deux versions (**Exp 1** et **Exp 2**, sans répétitions) ci-dessous. La précision des termes avec au moins une prédiction correcte a augmenté à **44,79%** (**7,29%** de plus que dans l'**Exp 1**).

	CLEAR et ClassYN			
Modèle	**T5-base Exp 1**		**T5-base Exp 2**	
Données statistiques	**N° trm**	**%**	**N° trm**	**%**
au moins une prédiction correcte	180	37,50 %	**215**	**44,79%**
aucune prédiction correcte	300	62,50 %	265	55,20%
Total	480	100%	480	100%

Tableau 60. Statistiques sur les résultats de prédictions de l'expérience 1 (répétitions possibles) et 2 (sans répétitions) pour le français

Les prédictions correctes (*score 1* et *score 2*) restent quand même limitées : **381 (16,80%)**. La contrainte sans répétitions imposée à la deuxième expérience augmente le nombre de prédictions correctes (*scores positifs*) de **2,83%** pour arriver à un pourcentage de **16,80%**.

	CLEAR et ClassYN			
Échelle d'évaluation	**T5-base Exp 1**		**T5-base Exp 2**	
	N°	**%**	**N°**	**%**
Score 2	73	3,21%	50	2,20%
Score 1	244	10,75%	331	14,60%
Scores positifs	***317***	***13,97%***	***381***	***16,80%***
Score 0	1848	81,48%	1796	79,22%
Score -1	94	4,14%	89	3,92%
Scores négatifs	***1942***	***85,62%***	***1885***	***83,14%***
Total	**2268**	**100%**	**2267**	**100%**

Tableau 61. Statistiques sur les scores de l'échelle d'évaluation de prédictions en français : expériences 1 et 2

Nous avons analysé les prédictions annotées avec le *score 1* afin d'identifier celles qui sont des **nouvelles reformulations** par rapport au jeu de données d'entraînement (les 8 146 paires terme-reformulation). Nous analysons les prédictions uniques, sans doublons, et nous observons que lors de l'**Exp 2**, **T5-base** a généré **81,55%** de **nouvelles**

reformulations. Les nouvelles reformulations obtenues sont en général des reformulations assez simples, des variantes du nom de la maladie sous forme d'adjectif (*schizophrénie : une maladie schizophrénique*). Les scores négatifs sont dus également aux mots inventés ou mal orthographiés, générés par le Transformer. Nous avons identifié seulement **8** mots inventés **(0,35%)**, de type « maladie *nichéolaire* », « une maladie caractérisée par une *ote* de cœur ».

	CLEAR et ClassYN			
Transformer	**T5-base Exp 1**		**T5-base Exp 2**	
	N° ref	**%**	**N° ref**	**%**
nouvelle prédiction	135	**68,52%**	84	**81,55%**
prédiction entraînement	62	31,47%	19	18,44%
Total sans doublons	*197*	*100%*	*103*	*100%*

Tableau 62. Statistiques sur les prédictions automatiques qui sont des nouvelles reformulations en français

D'autres erreurs de génération concernent *l'insertion des mots ou séquences de mots* dans la reformulation, qui n'ont pas de lien avec le contexte. Le terme « cholangite sclérosante primitive » (« maladie intestinale inflammatoire », reformulation de référence) est reformulé comme « une maladie chronique qui peut être caractérisée par des troubles cholestatiques, des antécédents et des taches causés par une inflammation des voies et des structures de la peau ». « Les structures de la peau » n'ont pas de lien direct avec les maladies intestinales. Ce type d'erreur est difficilement identifiable, mais apparait fréquemment dans les paraphrases annotées avec le score 0.

***

Nous avons mené une autre expérience avec **T5-base** sur le français, cette fois en modifiant **le paramètre de la longueur** de la reformulation. Si nos expériences précédentes ont été paramétrées pour générer une reformulation de maximum **256 mots**, nous avons lancé une expérience avec la longueur maximale de **128 mots**. Les résultats montrent que parmi les **2 180 prédictions uniques** générées, **2 047 (93,89%)** sont des **prédictions nouvelles** par rapport à **l'expérience 2** (*sans répétitions*) dont le paramètre de la longueur était de 256 mots. Le modèle de langue a généré uniquement **133 prédictions identiques** entre les deux expériences. Cette expérience montre que la modification d'un seul paramètre de l'architecture peut radicalement changer les prédictions générées.

***

Nous analysons les prédictions incorrectes et nous proposons une **classification des prédictions les plus surprenantes**, en mettent en valeur la « créativité » des Transformers.

3.2.4 Classification de prédictions de reformulations en français

Lors de notre évaluation manuelle de prédictions, nous avons identifié des prédictions très diverses. Parmi les **1 848** exemples avec un score négatif (*zéro*) (**81,48%**), nous avons identifié **58** (**3,13%**) prédictions de reformulation qui se démarquent par plusieurs caractéristiques uniques. Pour illustrer ces particularités, nous avons créé une **classification manuelle des prédictions de reformulations générées automatiquement**. Notre classification met en avant la composante sémantique et l'impact pragmatique de ces reformulations automatiques auprès d'un évaluateur humain. Nous justifions nos choix et nous illustrons cette liste avec des **exemples surprenants** qui peuvent être considérés comme :

- **Humoristiques** : certaines prédictions contiennent des associations des mots qui peuvent faire sourire les lecteurs humains, comme « une stagiaire ayant tendance à se dissoudre rapidement », « trouble de la fonction fonction publique », « les troubles du génie génétique », et encore d'autres. Le mot est proposé dans un contexte inapproprié.

strabisme
Truth: est un/e affection dans laquelle les yeux ne sont pas alignés normalement
Prediction: est un/e maladie associée à une stagiaire ayant tendance à se dissoudre rapidement

acouphène
Truth: en tant que trouble principal
Prediction: un/e trouble de la fonction fonction publique

autisme, asthme, leucémie, rhume des foins, diabète de type 1, troubles de la marche, maladie de Crohn, maladies démyélinisantes, infections bactériennes ou virales
Truth: : affections
Prediction: maladies génétiques et les troubles du génie génétique

recherches supplémentaires
Truth: par exemple les agents anticancéreux modernes (y compris lesdits nootropes) et l'intégration à la thérapie néoadjuvante (telle que la radio-chimiothérapie préopératoire)

Prediction: telles que l'addition de requins à un traitement primaire ou général

- **Répétitives** : pour le premier exemple ci-dessous, la répétition du mot « sanguin » peut être justifiée par le fait qu'il apparait trois fois également dans le terme, donc le Transformer essaie de reproduire cette structure. Pourtant, cette justification ne fonctionne pas pour le deuxième exemple, dont le terme contient un seul mot, « asthme ».

débit sanguin veineux lent, une augmentation de la coagulation sanguine et des lésions des parois des vaisseaux sanguins
Truth: font partie des facteurs favorisant l'apparition de ces affections
Prediction: définie comme addition de coagulants sanguins dans les vaisseaux sanguins dans les autres sangs sanguins
(par exemple, un revêtement sanguin lent, un saignement sanguin ou sanguin réactif (SVR), un renfermement sanguin sanguin de l'air et une épaisse coagulation sanguine)
(un traitement important pour réduire les taux sanguins sanguins sanguins de plus en plus courant)
défini/e ; définie comme une augmentation de la coagulation sanguine sanguine en aval et des lésions paroissiales de l'appareil sanguin sanguin

asthme
Truth: est un/e maladie courante provoquant respiration sifflante, toux et difficultés à respirer chez les adultes et les enfants
Prediction: : maladie asthmatique qui se caractérise par l'apparition d'une sensation asthmatique d'asthme.

- **Basées sur un morphème** : le Transformer considère de manière erronée que les termes médicaux qui commencent par le morphème « myo » sont synonymes, ce qui expliquerait les prédictions suivantes, où « un sombre myocarde » est la reformulation de « myopie » et « maladies rares du myocarde » du terme médical « troubles myotoniques ».

myopie
Truth: est un/e défaut de la vision qui se trouble lorsque des objets sont observés à distance
Prediction: , tel que un sombre myocarde

troubles myotoniques
Truth: connue sous le nom de maladies musculaires
Prediction: des difficultés à améliorer la fonction motrice des myotines
maladie maladies rares du myocarde

- **Absurdes :** nous avons identifié certaines prédictions dont la dimension sémantique n'est pas assurée correctement par le Transformer, ce qui résulte dans des associations absurdes ou artificielles des mots, comme « col de l'audace » ou « œil résiduel ».

cancer buccal
Truth: maladie
Prediction: / tel que la maladie du col de l'audace

démences communes
Truth: telles que maladie d'Alzheimer
Prediction: définie comme une démences avec une ou plusieurs des composantes de l'oeil résiduel.

- **« Néologismes »** : mots inventés par le Transformer qui n'existent pas dans le lexique du français, comme « maladie *nichéolaire* »

bronchiolite
Truth: est un/e maladie respiratoire grave qui affecte les bébés
Prediction: maladie nichéolaire

- **Malaisantes :** les prédictions générées automatiquement ne suivent pas une éthique humaine, ce qui peut mettre mal à l'aise les lecteurs humains. Dans l'exemple ci-dessous, la phrase « les enfants sont remis à l'abattoir » est générée pour reformuler une procédure médicale lors d'une opération.

intestin est remis en place dans l'abdomen sous anesthésie en salle d'opération
Truth: , cette procédure
Prediction: (par exemple des espaces de congés formé pendant la première période d'opération, les enfants sont remis à l'abattoir), l'intestin est étanchéité c'est-à-dire remise dans l'enceinte de l'abdomen en présence de l'étoile

L'exemple ci-dessous considère qu'une maladie liée au cycle menstruel féminin « dont la cause est inconnue » est « une affection très fréquente chez les personnes souffrant d'un trouble maniaque ». Cet exemple peut être issu des textes qui décrivent les pratiques des médecins d'antan qui ne prêtait pas beaucoup d'attention aux douleurs que les femmes disaient ressentir, particulièrement quand aucune cause physique n'était facile à identifier. La femme en question était alors considérée comme ayant des troubles psychiatriques (dont « trouble maniaque »).

dysménorrhée primaire
Truth: correspondait à douleurs dont la cause est inconnue
Prediction: est un/e une affection très fréquente chez les personnes souffrant d'un trouble maniaque

Le choix d'un Transformer de générer une prédiction en particulier n'est pas traçable, car les réseaux de neurones fonctionnent comme des boîtes noires. Cela veut dire que nous ne pouvons pas identifier les textes ou les phrases exactes qui lui ont servi comme inspiration pour l'association entre « dysménorrhée primaire » et « trouble maniaque ». Cependant, nous justifions cette prédiction comme issue du corpus web Common Crawl, un corpus qui contient des textes d'une grande variété de sujets provenant de la toile.

***

Les expériences de génération automatique de reformulations en français nous ont permis d'obtenir des **reformulations correctes** pour **44,79%** (**215**) termes lors de la deuxième expérience. Même si ce pourcentage reste modéré, les reformulations correctes générées peuvent être ajoutées à notre corpus de reformulations après une double annotation. L'analyse des prédictions incorrectes nous servent à ajuster les paramètres de l'outil. L'expérience avec le paramètre de la longueur du contexte réduit de 256 à 128 mots montre que **93,89%** de prédictions générées sont des **reformulations *nouvelles*** par rapport aux expériences antérieures. Ceci prouve que l'outil APT *apprend* à générer des nouvelles reformulations médicales.

***

Nous présentons par la suite les expériences de **génération automatique de reformulations** réalisées pour la langue roumaine.

3.3 Résultats de la génération en roumain

Dans la même lignée, nous avons lancé plusieurs expériences pour le roumain. Nous avons d'abord utilisé la version réduite du **Transformer multilingue de T5** (Xue *et al.*, 2020) (qui couvre 101 langues, dont le roumain), intitulée **mT5**. Ce modèle, à l'instar du **T5**, n'a pas été entraîné pour une tâche en particulier (comme la tâche *paraphrase et similarité de phrases*), ce qui peut donner des résultats moins précis pour une tâche telle que la génération de la paraphrase. Nous analysons en détail les prédictions obtenues successivement avec le modèle **mT5-small** (300 millions paramètres), **mT5-base** (580 millions paramètres) et **T5-base** (modèle utilisé également pour le français, avec 220 millions de paramètres).

3.3.1 Génération de reformulations avec mT5-small, mT5-base, T5-base

Le Transformer **mT5-small** a généré **1 499 prédictions de reformulations** médicales générées automatiquement pour les **300 termes médicaux** de la liste de test. Le transformer **mT5-base** a généré **1 491 prédictions et T5-base, 1 490 prédictions**, une moyenne de **4,97%** reformulations par terme (nous rappelons que l'architecture APT a été paramétrée pour générer entre 1 et 5 reformulations par terme). Les résultats se présentent sous la même forme que pour le français (voir **sous-chapitre 3.2.1**).

Transformer	N° prédictions pour 300 termes
mT5-small	1 499
mT5-base	1 491
T5-base	1 490

Tableau 63. Prédictions de reformulations en roumain générées automatiquement avec les Transformers

Nous avons analysé les prédictions automatiques selon la même **échelle d'annotation** utilisée pour le français, dont la conception et les valeurs sont décrites dans le **sous-chapitre 3.2.2**. Nous présentons les résultats de cette analyse ci-dessous et nous comparons la qualité des prédictions des trois Transformers sur la même liste de termes.

3.3.2 Évaluation et analyse des prédictions automatiques

Nous observons que parmi les 300 termes, uniquement **85 termes (28,33%)** ont reçu des **prédictions de reformulations correctes** avec le Transformer **mT5-base** et **80 (26,66%) termes** avec **T5-base**. Le modèle **mT5-small** a donné très peu de prédictions

correctes, seulement pour **21 termes médicaux (7%)**. Les résultats **(26,66%)** avec **T5-base** sont plus bas que pour le français avec le même Transformer (**T5-base**), dont **37,5%** de termes ont reçu une reformulation correcte. Ceci s'explique par le nombre réduit de données disponibles pour l'entraînement pour le roumain (3 027 reformulations annotées par rapport à 8 626 pour le français) et également par la syntaxe plus complexe du roumain (déclinaisons, cas, article enclitique, signes diacritiques présents ou absents).

Transformer	**mT5-small**		**mT5-base**		**T5-base**	
Données statistiques	**N° termes**	**Précision %**	**N° termes**	**Précision %**	**N° termes**	**Précision %**
Au moins une prédiction correcte	21	**7%**	85	**28,33%**	80	**26,66%**
Aucune prédiction correcte	279	**93%**	215	**71,66%**	220	**73,33%**
Total	**300**	**100%**	**300**	**100%**	**300**	**100%**

Tableau 64. Statistiques sur les résultats de prédictions de reformulations avec mT5-small, mT5-base et T5-base pour le roumain

Concernant notre échelle d'évaluation de prédictions, nous remarquons que le Transformer **mT5-base** donne un plus grand nombre de prédictions correctes, avec **143 (9,59%)**, suivi de près par **T5-base** avec **120 prédictions correctes (8,05%)**. Le Transformer mt5-small prédit uniquement **24 (1,60%)** reformulations correctes, ce qui est justifié par la taille réduite du modèle, par rapport aux autres (voir **Tableau 65**).

	mT5-small		**mT5-base**		**T5-base**	
Échelle d'évaluation	**N° préd**	**%**	**N° préd**	**%**	**N° préd**	**%**
Score 2	0	0%	2	0,13%	3	0,2%
Score 1	24	1,60%	141	9,45%	117	7,85%
Scores positifs	***24***	***1,60%***	***143***	***9,59%***	***120***	***8,05%***
Score 0	1 474	98,33%	1 310	87,86%	1 320	88,59%
Score -1	1	0,06%	38	2,54%	50	3,35%
Scores négatifs	***1 475***	***98,39%***	***1 348***	***90,40%***	***1 370***	***91,94%***
Total	**1 499**	**100%**	**1 491**	**100%**	**1 490**	**100%**

Tableau 65. Différents scores de l'échelle d'évaluation de prédictions pour le roumain

Concernant les prédictions correctes, nous observons que **la majorité (98,25%)** des prédictions générées sont des **reformulations nouvelles** (*score 1*) par rapport aux données d'apprentissage, les reformulations annotées (*score 2*). Cela signifie que les Transformers ont *construit* de nouvelles reformulations pour les 300 termes médicaux de la liste de test.

Par exemple, pour le terme « **Infectiile herpetice »** (*les infections par l'herpès*), dont la reformulation initiale était **«** sunt afectiuni eruptive de natura inflamatorie » (*sont des affections éruptives de nature inflammatoire),* **mT5-base** a généré la reformulation « cum ar fi o infectie virala contagioasa » (*telle qu'une infection virale contagieuse*). Nous illustrons ces nouvelles reformulations par plus d'exemples ci-dessous :

> **Dermatita *(Dermatite)***
> **Truth:** este o afectiune frecvent intalnita, caracterizata prin inflamarea pielii si prurit *(est une affection courante caractérisée par une inflammation de la peau et des démangeaisons)*
> **Prediction:** este o afectiune inflamatorie a pielii *(est une affection inflammatoire de la peau)* **(mT5-base)**
>
> **Infarct *(Infarctus)***
> **Truth:** sau alte leziuni la nivelul inimii *(ou d'autres lésions au cœur)*
> **Prediction:** sau alte traumatisme cardiace *(ou autres traumatismes cardiaques*) **(T5-base)**
>
> **Cistita acuta simpla *(Cystite aiguë simple)***
> **Truth:** reprezinta inflamatia mucoasei vezicii urinare produsa de agenti fizici, chimici sau infectiosi *(est une inflammation de la paroi de la vessie causée par des agents physiques, chimiques ou infectieux)*
> **Prediction:** cum ar fi boli ale organelor genitale *(comme des maladies des organes génitaux)* **(mT5-small)**

Nous avons identifié un grand nombre de prédictions qui contiennent des **répétitions** (**mT5-small** : 105 prédictions (7%), **mT5-base** : 107 (7,17%), **T5-base :** 112 (7,51%)) de type « boala boala » *(maladie maladie)*, « un tip tip » *(un type type)*, « cum ar fi un virus viral viral » *(comme un virus viral viral)*, « fiind afectiune cronica cronica » *(étant une affection cronique cronique)*, « afectiune afectiune » *(affection, affection)*. D'autres reformulations contiennent également des répétitions dont les mots sont déclinés ou conjugués, comme le montrent les deux exemples ci-dessous :

> **Depresia *(Dépression)***
> **Truth:** cunoscuta intre specialisti sub numele de tulburare depresiva majora *(connue par les spécialistes sous le nom de trouble dépressif majeur)*
> **Prediction:** este o boala boala boala boala bipolara care se manifesta la nivelul depresiei - adica sub nivelul depresiului *(est une maladie maladie maladie maladie bipolaire qui se manifeste au niveau de la dépression – c'est-à-dire en-dessous du niveau de la dépression)* **(T5-base)**
>
> **Tripanosomiaza umana africana *(Trypanosomiase humaine africaine)***

Truth: este o boala parazitara transmisa de mustele tete din genul Glossina si cauzata de un grup de paraziti cunoscuti sub numele Trypanosoma brucei *(est une maladie parasitaire transmise par le genre de mouches Glossina et causée par un groupe de parasites connus sous le nom de Trypanosoma brucei)*
Prediction: este o afectiune afectiune care afecteaza pentru o afectiune afectiune *(est une affection affection qui affecte pour une affection affection)* **(mT5-small)**

Pour éliminer les répétitions du même mot dans la prédiction générée, nous avons mené une deuxième expérience (**Exp 2**) avec **T5-base** en changeant un paramètre du script de l'architecture **APT**. Nous avons évalué les **1 485 prédictions générées** lors de cette deuxième expérience et nous comparons les deux versions (**Exp 1** et **Exp 2, sans répétitions**) ci-dessous. La précision des résultats s'est améliorée avec **19%** et nous avons **137 termes** qui ont reçu *au moins une prédiction* de reformulation correcte **(45,66%)**, par rapport à seulement **80 termes** lors de la première expérience **(26,66%) (Tableau 66)**.

Transformer	**T5-base Exp 1**		**T5-base Exp 2**	
Données statistiques	**N° termes**	**Précision %**	**N° termes**	**Précision %**
au moins une prédiction correcte	80	**26,66%**	137	**45,66%**
aucune prédiction correcte	220	**73,33%**	163	**54,33%**
Total	**300**	**100%**	**300**	**100%**

Tableau 66. Statistiques sur les résultats des prédictions de l'expérience1 et 2 (sans répétition) avec T5-base pour le roumain

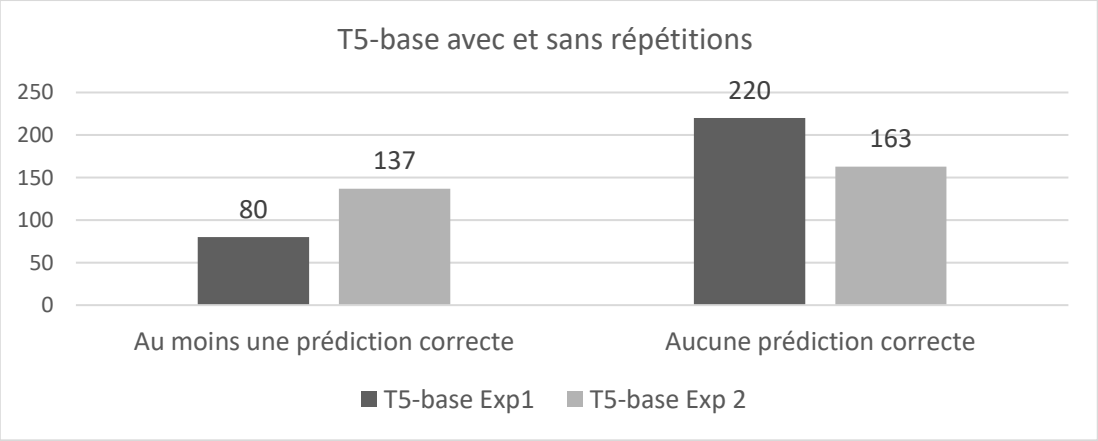

Figure 29. Amélioration des résultats de prédiction de reformulations avec T5-base pour le roumain

Les prédictions correctes sont en nombre de **222** (**14,94%,** *score 1* et *score 2*), dont le score pour les prédictions exactes aux reformulations données en entrée (colonne *Truth*)

est deux fois plus élevé, mais tout en restant très faibles (*score 2* à **0,47%**). Le score de la deuxième expérience augmente le nombre de prédictions correctes (*scores positifs*) avec **6,89%**, ce qui nous permet d'avoir **54,54% de reformulations de plus** par rapport à la première expérience.

	T5-base Exp 1		**T5-base Exp 2**	
Échelle d'évaluation	**N° préd**	**%**	**N° préd**	**%**
Score 2	3	0,2%	7	0,47%
Score 1	117	7,85%	215	14,47%
Scores positifs	***120***	***8,05%***	***222***	***14,94%***
Score 0	1 320	88,59%	1 216	81,88%
Score -1	50	3,35%	45	3,03%
Scores négatifs	***1 370***	***91,94%***	***1 261***	***84,91%***
Total	**1 490**	**100%**	**1 485**	**100%**

Tableau 67. Statistiques sur les différents scores de l'échelle d'évaluation de prédictions pour T5-base lors des expériences 1 et 2 pour le roumain

Nous avons analysé les prédictions annotées avec le *score 1* afin d'identifier celles qui sont des **nouvelles reformulations** par rapport au jeu de données d'entraînement (les **2 727** paires terme-reformulation). Nous analysons les prédictions uniques, sans doublons, et nous observons que lors de **l'Exp 2**, **T5-base** a généré **86%** de **nouvelles reformulations** en roumain. Les nouvelles reformulations obtenues sont en général des reformulations assez simples, des hyperonymes de type « tulburare depresiva majora : cum ar fi tulburari neurologice » (*trouble dépressif majeur : tels que les troubles neurologiques*).

	GrandMed-Ro2			
Transformer	**T5-base Exp 1**		**T5-base Exp 2**	
	N° ref	**%**	**N° ref**	**%**
nouvelle prédiction	94	**84,68%**	172	**86%**
prédiction entraînement	17	15,31%	28	14%
Total sans doublons	*111*	*100%*	*200*	*100%*

Tableau 68. Statistiques sur les prédictions automatiques qui sont des nouvelles reformulations sur le corpus roumain

Nous observons un plus grand nombre de mots inventés qu'en français (**Exp 1** : 71 mots (4,76%) ; **Exp 2** : 129 mots (8,68%)), tels que les mots inventés dans les prédictions suivantes : « reprezinta o afectiune in care se dezvolta un *tumefl rea* in *cortija* » (*représente*

une affection dont se développe [mot inventé] mauvaise [mot inventé]), « numiti *articulatari* » (*nommés [mot inventé]),* « este o boala a virusesului » (*est une maladie de [mot inventé]*).

Nous remarquons que ces mots inventés ont un lemme correct, comme dans les deux derniers exemples, « articulație » (*articulation*) et « virus » (*virus*). La difficulté pour le roumain vient de la déclinaison avec article enclitique (attaché à la fin du mot) et par la présence des cas en roumain. Dans l'exemple « este o boala a *virusesului* », la forme correcte du dernier mot serait la forme articulée, cas génitif du mot virus, c'est-à-dire « a virusului » (*du virus*). La forme incorrecte « a virusesului » montre que le Transformer n'a pas trouvé la bonne particule à ajouter à la fin du mot pour illustrer le cas génitif (qui exprime la possession en roumain) ou qu'il a ajouté la particule au mot anglais *viruses* (forme au pluriel de *virus*). Nous observons ce problème également avec le mot « articulatari », dont la forme correcte est « articulații » (*articulations*). L'absence de signes diacritiques du roumain (ă, î, â, ț, ș), problème récurrent dans les textes de vulgarisation sur la toile, complexifie encore la tâche du Transformer.

***

Nous présentons par la suite la **classification des prédictions de reformulations incorrectes**, mais qui mettent en valeur la « créativité » de génération de textes en roumain des Transformers.

3.3.3 Classification de prédictions de reformulations en roumain

Nous avons classé les reformulations surprenantes en roumain selon les mêmes critères que pour le français, introduites dans le **sous-chapitre 3.2.4**. Nous présentons d'abord les **catégories** qui ont été identifiées aussi pour le français (*humoristiques, néologismes, absurdes, malaisantes*) et ensuite nous introduisons et nous définissons les nouvelles catégories identifiées (*sarcastiques, philosophiques, poétiques*).

Nous illustrons toutes les catégories avec des exemples des prédictions en roumain avec la mention du Transformer d'origine et traduites par nous en français :

- **Humoristiques** :

  psihoza (psychose)
  Truth: inseamna o pierdere de contact cu realitatea si reprezinta mai degraba un simptom al mai multor boli mintale, decat o afectiune medicala de sine statatoare *(signifie une perte de contact avec la réalité et représente un symptôme de*

plusieurs maladies mentales plutôt qu'une condition médicale à part entière)
Prediction: inseamna ca ficatul nu are un sentiment de pierdere a increderii si a increderii *(signifie que le foie n'a pas le sentiment d'avoir perdu confiance en lui et de la confiance)* **(mT5-base)**

disfunctii erectile *(dysfonctionnements érectiles)*
Truth: reprezinta imposibilitatea obtinerii si mentinerii erectiei si se datoreaza, in majoritatea cazurilor, unui flux sangvin deficitar catre penis *(représente l'impossibilité d'obtenir et de maintenir une érection et est due, dans la plupart des cas, à un débit sanguin déficient dans le pénis)*
Prediction: cum ar fi infarctul erectilor *(comme l'infarctus des érections)* **(T5-base)**

- **« Néologismes »** : « reprezinta o afectiune in care se dezvolta un *tumefl rea in cortija* » *(représente une affection dont se développe [mot inventé] mauvaise [mot inventé])* **(T5-base)** ; « (toetare articulara) » *([mot inventé] articulaire)* ***(mT5-base)***

- **Absurdes :**

Nistagmus *(Nystagmus)*
Truth: este o afectiune caracterizata de miscari necontrolate ale ochilor (adesea, din cauze neurologice)*(est une affection caractérisée par des mouvements oculaires incontrôlés (souvent dus à des causes neurologiques))*
Prediction: inseamna ca persoana nu are suficienti de multi ani in timpul unei saptamani ale timpului liber *(signifie que la personne n'a pas assez d'années pendant la semaine du temps libre)* **(mT5-base)**

Dizartria *(Dysarthrie)*
Truth: reprezinta o afectiune care presupune o tulburare a vorbirii, aparuta ca rezultat al slabirii muschilor implicati in vorbire *(est une affection impliquant un trouble de la parole qui se produit à la suite d'un affaiblissement des muscles impliqués dans la parole)*
Prediction : este o boala in care necontroleaza oamenii de la un nivel inferior *(est une maladie dans laquelle les personnes d'un niveau inférieur sont hors de contrôle)* **(T5-base)**

- **Malaisantes :**

Spalarea pe maini si dezinfectarea suprafetelor *(Le lavage des mains et la désinfection des surfaces)*
Truth: reprezinta metode foarte bune de protectie, chiar si impotriva multor alte boli *(sont de très bonnes méthodes de protection, même contre de nombreuses autres maladies)*
Prediction: cum ar fi spânzurarea suprafetelor *(comme la pendaison des surfaces)* **(T5-base)**

LADA *(LADA)*
Truth: inseamna Diabet latent autoimun al adultului *(signifie Diabète adulte auto-immun latent)*
Prediction: numita infectie a sacrificiului ocular *(appelé infection du sacrifice de l'œil)***(T5-base)**

- **Sarcastiques** : nous considérons le paroxysme « trouble dépressif majeur » reformulé avec « comportement anormalement de confortable » comme une marque de sarcasme :

tulburare depresiva majora *(trouble dépressif majeur)*
Truth:(sau tulburare unipolara)*(ou un trouble unipolaire)*
Prediction: , un comportament anormal de confortabil *(un comportement anormalement de confortable)* **(T5-base)**

- **Philosophiques :** certaines reformulations contiennent des expressions qui expriment une *réflexion philosophique sur le monde*, de type « situations au-delà de la réalité », « trouble de l'ordre », « symptômes anormaux de la vie », voire une des plus grandes peurs concernant les intelligences artificielles, comme celle d'un « ordinateur qui aide à décomposer le système immunitaire » :

Tiroglobulina *(Thyroglobuline)*
Truth: cunoscuta si dupa prescurtarea TG *(également connu sous l'abréviation TG)*
Prediction: este o boala neurologica rara care apare dupa instalarea unui computer care ajuta la descompunerea sistemului imunitar *(est une maladie neurologique rare qui survient après l'installation d'un ordinateur qui aide à décomposer le système immunitaire)* **(mt5-base)**

Dizartria *(Dysarthrie)*
Truth: reprezinta o tulburare a vorbirii care poate afecta viata multor personae *(est un trouble de la parole qui peut affecter la vie de nombreuses personnes)*
Prediction: este o tulburare tulburare a ordinii in, care persoana se dezvolta si se împotmoleste, se extinde in directii ce duc la apatie *(est un trouble trouble de l'ordre dans lequel la personne se développe et se bloque, se développe dans des directions qui mènent à l'apathie)* **(T5-base)**

Fisura anala *(Fissure anale)*
Truth: reprezinta o ruptura la nivelul mucoasei anusului *(est une déchirure de la muqueuse de l'anus)*
Prediction: este o afectiune afectiune in, care celulele pot duce la extinderea densitatii pielii si dificultatile pentru a trata situatiile dincolo de realitate *(est une affection affection dans laquelle les cellules peuvent conduire à l'extension de la densité de la peau et la difficulté de traiter des situations au-delà de la réalité)* **(mT5-base)**

Fibromialgia *(Fibromyalgie)*
Truth: este o afectiune caracterizata prin dureri musculoscheletale cronice larg raspandite pe suprafata corporala *(est une affection caractérisée par des douleurs musculo-squelettiques chroniques généralisées sur le corps)*
Prediction: numite simptome anormale ale vietii *(appelés symptômes anormaux de la vie*) **(mT5-base)**

- **Poétiques** : nous notons cet exemple d'utilisation d'une figure de style littéraire (*hyperbole*) pour reformuler la hernie :

Hernia *(Hernie)*
Truth: reprezinta protruzia unui organ printr-o zona de slaba rezistenta in urma ruperii sau slabirii tesutului muscular *(représente la protrusion d'un organe à travers une zone de faible résistance suite à la rupture ou à l'affaiblissement du tissu musculaire)*
Prediction: hiperbolii la nivelul pielii) *(hyperboles au niveau de la peau))* **(mT5-base)**

Ces exemples montrent la **grande diversité** de prédictions générées par les **Transformers T5-base, mT5-small** et **mT5-base**. Même si ces prédictions ne sont pas toutes des reformulations correctes pour les termes médicaux, nous considérons qu'il est important d'analyser toutes les prédictions afin de mieux ajuster les paramètres. Notre adaptation d'un seul paramètre pour éviter les répétitions prouve que l'architecture **APT** permet **d'améliorer drastiquement la qualité des reformulations générées** (par **19%** sans les répétitions). Nous exploiterons ces pistes de recherche dans des expériences futures.

3.4 Score inter-annotateur Kappa : prédictions

Nous avons calculé le **score inter-annotateur Kappa** (Cohen, 1960) pour **1 196 prédictions** en français et **1 234 prédictions** en roumain (générées pour **250** termes de chaque langue), annotées par deux annotateurs francophones, non-spécialistes du domaine de la médecine. Nous avons obtenu un score inter-annotateur Kappa de **0,44** pour le français et de **0,48** pour le roumain. Ces accords sont *modérés,* car ils concernent quatre valeurs d'annotation différentes selon le guide d'annotation de prédictions automatiques (*2, 1, 0* et *-1*). Par conséquent, ces scores sont très précis pour chaque valeur.

Nous avons calculé également le score Kappa pour les **250 termes annotés par langue** afin d'identifier si ces termes ont reçu **au moins une prédiction de reformulation correcte**. Pour les 250 termes annotés en français, le score est de **0,42**, et, en roumain, de

0,55, également des scores inter-annotateur Kappa *modérés*. Les scores modérés s'expliquent par la difficulté de la tâche : il est difficile d'identifier la reformulation correcte (surtout quand les mots inventés utilisent des préfixes ou suffixes utilisés couramment pour créer des termes, et les annotateurs ne sont pas experts du domaine pour les identifier).

3.5 Bilan des prédictions de reformulations avec APT

Nos expériences montrent que l'architecture **APT** peut être utilisée également pour générer des **reformulations sous-phrastiques médicales** avec une précision de **44,79%** pour le français et **45,66%** pour le roumain (si l'on considère les termes qui ont au moins une bonne prédiction). Si Nighojkar et Licato (2021) ont généré un grand nombre de paraphrases de la langue générale en anglais (car ils utilisent des données de Twitter ou du PPDB), nos expériences sont menées sur des données spécifiques du domaine médical en français et en roumain, ce qui rend la tâche plus difficile. Les résultats sont comparables entre les deux langues traitées. Il n'y a pas beaucoup de prédictions correctes, car nous avons utilisé les modèles de langue générale et une quantité assez réduite de données médicales.

	CLEAR et ClassYN				**GrandMed-Ro2**			
Transformer	**T5-base Exp 1**		**T5-base Exp 2**		**T5-base Exp 1**		**T5-base Exp 2**	
Données statistiques	**N° trm**	**%**	**N° trm**	**%**	**N° trm**	**%**	**N° trm**	**%**
au moins une prédiction correcte	180	37,50%	**215**	**44,79%**	80	26,66%	**137**	**45,66%**
aucune prédiction correcte	300	62,50%	265	55,20%	220	73,33%	163	54,33%
Total	480	100%	480	100%	300	100%	300	100%

Tableau 69. Analyse quantitative des résultats de prédictions de l'expérience 2 (sans répétition) sur les corpus français et roumain

Notre adaptation d'un seul paramètre pour éviter les répétitions montre que l'architecture **APT** permet **d'augmenter** la précision (concernant le nombre de termes ayant au moins une reformulation correcte générée automatiquement) à **45%**, valeur qui reste modérée. Nous exploiterons les erreurs observées afin de les éviter dans des expériences futures pour améliorer nos résultats de prédictions sur les deux langues d'étude, en particulier pour éviter les mots inventés. Nous observons qu'en moyenne **84%** des prédictions de reformulations générées lors de la deuxième expérience (sans doublons)

sont des **nouvelles reformulations** par rapport à la première expérience, ce qui prouve la capacité de génération de nouveau contenu textuel du modèle de langue **T5**.

	CLEAR et ClassYN				**GrandMed-Ro2**			
Transformer	**T5-base Exp 1**		**T5-base Exp 2**		**T5-base Exp 1**		**T5-base Exp 2**	
	N° ref	**%**	**N° ref**	**%**	**N° ref**	**%**	**N° ref**	**%**
nouvelle prédiction	135	**68,52%**	84	**81,55%**	94	**84,68%**	172	**86%**
prédiction entraînement	62	31,47%	19	18,44%	17	15,31%	28	14%
Total sans doublons	*197*	*100%*	*103*	*100%*	*111*	*100%*	*200*	*100%*

Tableau 70. Analyse quantitative des prédictions automatiques qui sont de nouvelles reformulations sur le corpus français et roumain

Les reformulations nouvelles correctes peuvent contribuer à élargir le corpus de reformulation dans les deux langues, si les annotateurs humains les valident.

***

Nous continuons nos expériences automatiques avec la **classification automatique** des reformulations correctes et incorrectes. Nous présentons quelques expériences de classification automatique réalisées sur nos données médicales annotées en termes médicaux et en reformulations avec **l'architecture neuronale LSTM**. Cette architecture permet de sélectionner les phrases qui contiennent potentiellement des reformulations. Nous analysons les résultats obtenus dans le chapitre suivant.

3.6 Expériences de classification avec LSTM

Nous utilisons **l'architecture neuronale LSTM** (*Long Short Term Memory*) (Hochreiter et Schmidhuber, 1997), permettant l'application des techniques **d'apprentissage profond** par **réseaux de neurones** (concepts présentés dans la **Partie II**, **sous-chapitre 1.3**) pour une tâche de classification : il s'agit d'identifier les contextes qui contiennent des reformulations et ceux qui n'en contiennent pas. En premier temps, nous présentons les résultats des expériences menées sur nos données en français annotées par deux étudiants[82] lors d'un projet transdisciplinaire en Sciences des données. En

[82] Les étudiants sont : Laura Halimi et Quentin Bacquelé.

deuxième temps, nous montrons les résultats des expériences réalisées par nous, en modifiant des paramètres de l'architecture pour améliorer les résultats.

3.6.1 Classification avec *stemming* (racinisation)

Les expériences ont été menées sur les **2 689** phrases annotées du corpus français **ClassYN SP**, dont **776** contiennent des reformulations correctes et **1 913** ne contiennent pas de reformulation. Le jeu de données de test est de **433** phrases, sélectionnées aléatoirement, et celui d'entraînement est de **2 295** phrases.

Afin de préparer les données, plusieurs opérations ont été mises en place :

- **Nettoyage des données** : suppression de la ponctuation, des majuscules et des minuscules et des mots vides ;
- **Encodage des étiquettes** : les statuts *oui* et *non* (qui indiquent si une phrase contient une reformulation correcte ou ne contient aucune reformulation correcte) sont remplacés avec les valeurs binaires *1* et *0* ;
- **Stemming (racinisation)** : les mots sont remplacés par des racines de mots traités, par exemple le terme médical *insuffisance cardiaque* devient *insuffis cardiaqu*. Cette étape vise à garder uniquement le morphème qui contient le plus de sens.
- **Tokenization et étiquetage** : le tokenizeur et l'étiqueteur NLTK (Bird *et al.*, 2009) attribue une catégorie grammaticale à chaque partie du discours ;
- **Plongements des mots** : avec le modèle de langue **GloVe** (Pennington *et al.*, 2014. Ce modèle donne un contexte plus large pour chaque mot présent dans la phrase. Pourtant, ce modèle est constitué sur des données de langue générale, il n'est pas spécifique au domaine médical. Des modèles entraînés pour le domaine médical pour le français ne sont pas libres d'accès, nous avons utilisé un des modèles disponibles.

L'architecture **LSTM** a été paramétrée de la manière suivante :

- **epochs** : 7 (cycles d'apprentissage avec l'utilisation de toutes les données) ;
- **contexte du plongement des mots** : 300 caractères ;
- **la fonction de perte (val_loss)** : binary_crossentropy (calcule la perte d'entropie croisée entre les vraies reformulations et les reformulations prédites) ;
- **mesure d'évaluation** : précision.

Les résultats de l'apprentissage prouvent que la précision de classification sur les données de test est relativement bonne (**0,7905**). Pourtant, la précision est très haute **(0,9489**) sur les données d'entraînement par rapport aux données de test. Cela indique une situation de *surapprentissage (overfitting),* c'est-à-dire que l'algorithme a trop appris les particularités de chacun des exemples donnés pour l'entraînement. La fonction de perte est de **0,5190**, une valeur trop grande, ce qui prouve que le modèle n'apprend pas correctement à partir des données. La meilleure valeur de la fonction de perte se retrouve au cycle 6 (epoch) d'apprentissage, avec une valeur de **0,4603** et une précision de **0,8024**. Les données sont déséquilibrées, avec plus d'exemples négatifs que positifs.

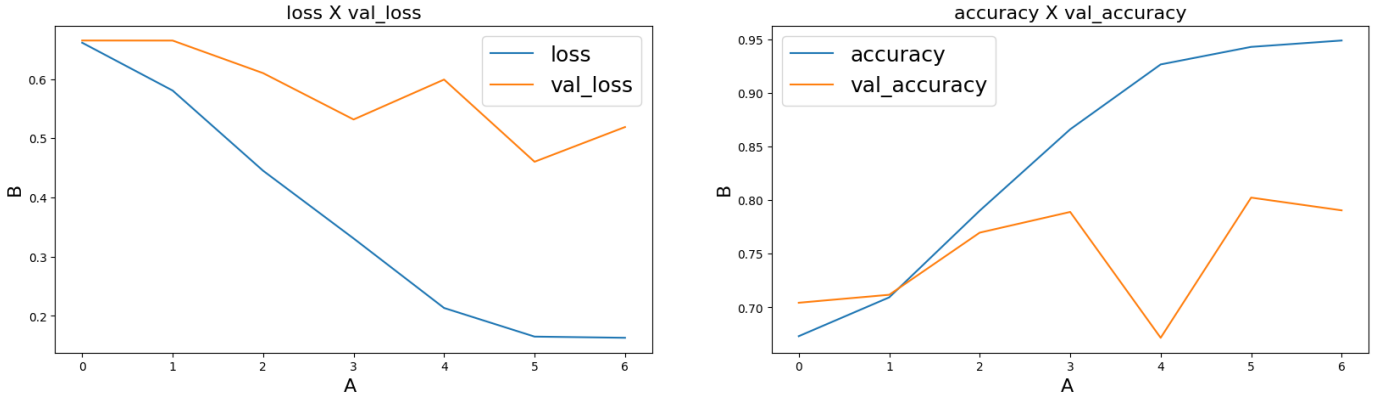

Figure 30. Résultats de classification automatique des phrases avec ou sans reformulations avec l'architecture neuronale LTSM - stemming

Afin de réduire le taux de surapprentissage, nous menons une deuxième expérience avec LTSM en changeant l'étape de *racinisation* avec une étape de *lemmatisation*, qui nous permet de garder la forme du dictionnaire du mot. Nous augmentons également le jeu de données. Nous expliquons notre méthode dans le sous-chapitre suivant.

3.6.2 Classification avec lemmatisation

Le lemme est l'entrée du dictionnaire d'un mot, plus précisément l'absence de particule qui indique le nombre, le genre, l'accord et absence de tout déterminant. La lemmatisation est la procédure automatique de transformation d'un mot dans le discours dans sa forme du dictionnaire (le terme *les maladies myocardiques* devient *maladie myocardique*). Nous menons cette expérience sur la totalité de nos phrases annotées en français, en nombre de **16 072 phrases** des corpus **CLEAR** et **ClassYN**. Nous avons gardé le même paramétrage pour **LSTM** que l'expérience précédente.

Les résultats montrent que la précision de classification sur les données d'entraînement arrive à **0,8143.** Nous notons une légère augmentation par rapport à la

première expérience. La précision sur les données de test baisse très légèrement à **0,9417**, comme illustrée dans la **Figure 31**.

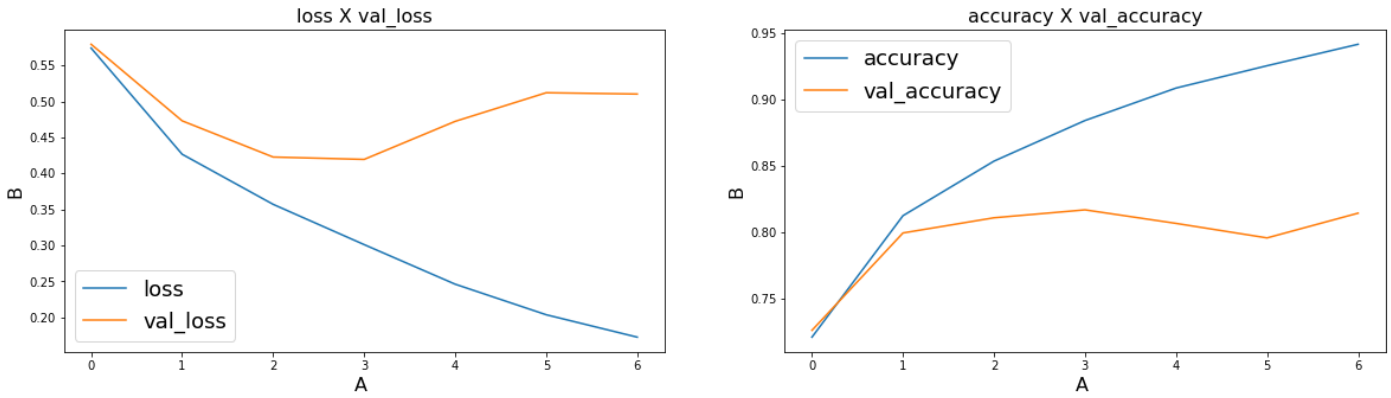

Figure 31. Résultats de classification automatique des phrases avec ou sans reformulations avec l'architecture neuronale LTSM, sur l'intégralité du jeu de données en français, avec lemmatisation

Si la fonction de perte à la fin de l'apprentissage (epoch 7) est très légèrement plus basse que pour l'expérience précédente, **0,5102**, elle baisse à **0,4194** au cycle 4 (avec une précision augmentée légèrement à **0,8168**), par rapport à la première expérience, dont la valeur la plus basse était de **0,4603**. La courbe pour la fonction de perte **val_loss** (qui est calculée sur l'ensemble de données de test) montre que l'erreur diminue, mais ensuite elle augmente à nouveau (0,6). La baisse de la valeur prouve que LSTM apprend mieux à partir du jeu de données lors de l'epoch 4. Pourtant, cette valeur augmente à la fin de l'apprentissage, ce qui signifie que la qualité de l'apprentissage décroît. Le minimum pour la fonction de perte n'a pas été retrouvé.

Afin d'améliorer les résultats et réduire les pertes en fin d'apprentissage (epoch 7) nous menons d'autres expériences en modifiant les *paramètres* de l'algorithme LTSM. Nous avons changé la fonction de perte *binary_crossentropy* avec **mean_squared_error**. Cette fonction calcule la moyenne des carrés de différences entre chaque valeur de référence (phrases avec les étiquettes *oui* et *non*) et chaque valeur estimée (les prédictions de classification des reformulations).

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Figure 32. Formule de calcul de la fonction mean_squared_error

Lors de cette expérience, nous obtenons une précision plus basse, de **0,7626** lors du dernier cycle d'apprentissage (epoch 7) avec une perte diminuée à **0,2028**. Pourtant, lors de l'epoch 4, la précision augmente à **0,823** d'exactitude, ce qui représente une légère amélioration par rapport à l'expérience précédente (0,81). En revanche, l'erreur de la fonction de perte (*mean_squared_error*) diminue davantage à **0,1363** au lieu de 0,51 ou 0,41 (pour l'expérience précédente), ce qui signifie une amélioration importante dans l'apprentissage lors de l'epoch 5.

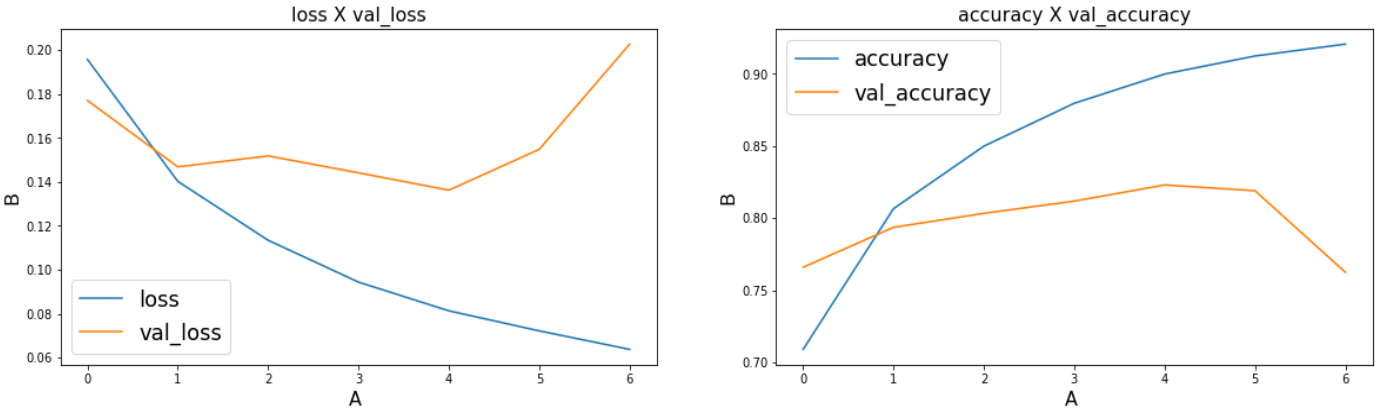

Figure 33. Résultats de classification automatique des phrases avec ou sans reformulations avec l'architecture neuronale LTSM, sur l'intégralité du jeu de données en français, avec lemmatisation et la fonction de perte mean_squared_error

D'autres modifications de paramètres sont possibles pour améliorer davantage les résultats de classification des reformulations avec LSTM, comme :

- l'augmentation du **batch** (le nombre d'échantillons de texte traités avant la mise à jour du modèle), de 32 à 64. Pourtant, l'augmentation du nombre d'échantillons demande une plus grande puissance de calcul et de mémoire de la part de l'ordinateur ;
- le changement du nombre de **cycles d'apprentissages** de 7 à 5 ;
- utilisation d'autres **fonctions de perte**, de type categorical_crossentropy, cosine_similarity, mean_squared_logarithmic_error, etc. ;
- changement de **l'optimiseur** (un algorithme qui identifie des paramètres qui peuvent améliorer la prédiction).

Nous explorons l'architecture LSTM avec trois nouvelles expériences :

1. Nous testons d'abord l'augmentation du **batch** à 64 échantillons, en gardant la fonction de perte **mean_squared_error**. Nous remarquons peu de différences par rapport à l'expérience antérieure (avec batch 32) :

- lors de l'epoch 7 (dernier cycle) : la précision est de **0,8089** et la perte de **0,1583** ;
- lors de l'epoch 5 : la perte diminue très peu à **0,1540**, pourtant la précision diminue également à **0,7897** (au lieu d'augmenter).

2. Nous menons une autre expérience en modifiant uniquement l'optimiseur **rmsprop** *(Root Mean Square Propagation),* qui permet de pondérer le facteur d'amorti de l'itération précédente avec le carré du gradient courant, avec **Adam** *(Adaptive Moment Estimation)*, qui est une méthode de descente de gradient basée sur l'estimation adaptative des moments de l'ordre 1 et 2 des paramètres. Nous gardons la fonction **mean_squared_error** et le **batch** à 32 (pour la rapidité du calcul). Nous observons :
   - lors de l'epoch 7 : la précision est de **0,7952** et la perte de **0,1771** ;
   - lors de l'epoch 5 : la perte est la plus basse **(0,1484)** et la précision plus grande **(0,8101)**.
3. Nous réalisons une dernière expérience en modifiant l'optimiseur **rmsprop** avec **SGD** *(Stochastic Gradient Descent)*, algorithme qui sélectionne aléatoirement un sous-ensemble de données en entrée pour estimer l'erreur et le gradient. Nous testons une autre fonction de perte, **mean_squared_logarithmic_error**, qui calcule la moyenne des carrés des erreurs entre les étiquettes initiales et les prédictions. Nous gardons le batch avec 32 échantillons. La courbe d'apprentissage augmente après chaque epoch. Lors de la dernière epoch (7), la précision est très basse par rapport aux autres expériences, de **0,6941**, pourtant la perte dans l'apprentissage est beaucoup améliorée : **0,0970**.

3.6.3 Bilan des classifications automatiques avec LSTM

Les expériences réalisées avec l'architecture de réseaux de neurones LSTM montrent la complexité de la tâche de classification automatique des reformulations médicales. Nous avons utilisé toutes nos données annotées pour la langue française, au nombre de **16 072** phrases du corpus **CLEAR** et **ClassYN**. Nous avons fait varier les paramètres du modèle afin d'améliorer les résultats d'apprentissage, éviter le surapprentissage et réduire le taux de perte entre l'étape d'apprentissage et l'étape de prédiction.

Nos meilleurs scores (perte : **0,1484** ; précision : **0,8101**) ont été obtenus lors de l'epoch 5 avec les paramètres suivants :

- fonction de perte : **mean_squared_error** ;
- batch : **32** ;

- optimiseur : **Adam**.

Nous observons que le problème de surapprentissage par mémorisation des reformulations persiste, à cause des jeux de données peu équilibrées. Nous envisageons de mener ce type d'expériences automatiques sur les données en roumain dans des travaux futurs.

***

Ces expériences montrent que les reformulations **générées automatiquement** par **APT** ou détectées par la **classification automatique** avec **LSTM** doivent être validées manuellement pour les inclure dans le corpus final de reformulation. Les reformulations générées automatiquement peuvent contenir des erreurs (termes inventés, mots utilisés dans des contextes inappropriés), une validation manuelle reste nécessaire. Il est de même pour les résultats de la classification automatique, les performances du classifieur restent perfectibles (par exemple en faisant appel à un jeu de données plus équilibré en termes de vraies et fausses reformulations).

Dans le prochain chapitre, nous nous intéressons à la *réception par le grand public* des reformulations médicales identifiées lors de nos annotations. Nous évaluons **le niveau de lisibilité et de compréhension des reformulations** et nous analysons les annotations réalisées par des annotateurs non-spécialistes du domaine de la médecine.

4. Évaluation du niveau de lisibilité des reformulations

Les annotateurs humains francophones ont évalué des **reformulations en français** de différents types (*définition, paraphrase, explication, abréviation, etc.*) issues de nos corpus médicaux d'étude, validées manuellement[83], en fonction de **trois niveaux de lisibilité** :

- **Niveau 1** - facile à comprendre : la reformulation médicale est plus facile à comprendre que le terme médical (il y a des mots de la langue commune dans la reformulation) ;
- **Niveau 2** - même complexité : même niveau de complexité ou de technicité entre le terme médical et sa reformulation, c'est-à-dire que le sens de deux parties est difficile à comprendre par l'annotateur ;
- **Niveau 3** - difficile à comprendre : la reformulation médicale est plus complexe ou plus technique que le terme, et, par conséquent, plus difficile à comprendre.

Nous développons ces niveaux dans une **échelle de lisibilité** que nous introduisons ci-dessous.

4.1 Critères d'évaluation de la lisibilité

Nous présentons les **critères d'évaluation de la lisibilité** qui ont été complétés au fur et à mesure de l'annotation avec les observations des annotateurs. Cette échelle se construit sur plusieurs critères : lexical, sémantique, syntaxique, longueur de la reformulation, connaissances universelles partagées. Dans les exemples ci-dessous issus de nos annotations, les termes médicaux sont soulignés et la reformulation (sans marqueur de reformulation) apparait à droite des doubles points.

1. **Lexical et sémantique** :

   Niveau 1 : la reformulation est considérée facile à comprendre si elle contient plusieurs mots de la langue générale, dont le sens est plus accessible pour le grand public.

[83] L'évaluation de la lisibilité n'a pas été réalisée pour les reformulations générées automatiquement.

- *L'ostéogenèse imparfaite : maladie des os de verre ;*

Niveau 2 : la reformulation contient un nom de médicament / trouble / affection / maladie peu connu et un cooccurrent utilisé dans le langage courant ; reformulation qui est une paraphrase proche du terme médical, sans apporter d'informations supplémentaires.

- *Les opiacés forts : l'oxycodone la morphine et le fentanyl ;*
- *Les méthylxanthines : la caféine, la théophylline et l'aminophylline ;*
- *La neuropathie paraprotéinémique : les neuropathies associées à une paraprotéine.*

Niveau 3 : la reformulation contient des termes médicaux techniques, des acronymes, des molécules, des procédures ou noms de médicaments très peu connus par le grand public ;

- *maladie cardio-vasculaire établie : maladie coronarienne, maladie cérébro-vasculaire ischémique ou maladie artérielle périphérique connue ;*
- *procédures endovasculaires : artériographie diagnostique, angioplastie, cathétérisation cardiaque ;*
- *maladies chroniques du foie : la cirrhose biliaire primitive et la cholangite sclérosante primitive ;*
- *plantes médicinales chinoises traditionnelles : TCMH.*

2. Connaissances générales partagées par le grand public :

Niveau 1 : reformulation qui contient des termes ou des noms de médicaments qui sont devenus des mots courants (*des médicaments contre la douleur => à base de paracétamol*) ou des acronymes généralement connus, comme AVC, VIH, MST, etc. ;

Niveau 2 : reformulation qui contient une majorité de termes complexes et le terme médical est trop technique ;

- *L'amyotrophie spinale proximale liée au gène SMN1 : maladie de la corne antérieure de la moelle épinière, maladie du motoneurone, amyotrophie spinale antérieure (ASA), amyotrophie spinale infantile (ASI) ou, plus fréquemment, SMA (pour spinal muscular atrophy).*

Niveau 3 : reformulation médicale plus complexe ou plus technique que le terme. La reformulation contient des acronymes, des noms ou des marques de médicaments peu ou pas du tout connus par les non-spécialistes.

- *dystrophie des ceintures de type C : LGMD1C ;*
- *acétaminophène : Tylenol*.

3. Syntaxe et longueur de la reformulation :

Niveau 1 : reformulation courte, de moins de 10 mots, syntaxe simple, présence de conjonction de coordination uniquement (*mais, ou, et, donc, or, ni, car*), peu ou pas de ponctuation ;

- *la fosse iliaque droite : la zone en bas et à droite de l'ombilic ;*
- *troubles digestifs : nausées, vomissements ;*
- *La fibrose pulmonaire idiopathique (FPI) : une maladie associée à une morbidité et une mortalité élevées.*

Niveau 2 : reformulation de la même longueur et de la même complexité syntaxique que la phrase originale ;

- *l'hépatite C : maladie progressive ;*
- *symptômes moteurs parkinsoniens : le tremblement de repos ;*
- *diurétiques : thiazidique.*

Niveau 3 : reformulation longue, de plus de 10 mots, avec une syntaxe complexe (hors énumérations), une grande diversité de connecteurs logiques : conjonction de subordination, adverbe, locution, ponctuation. La complexité de niveau 3 est également soulignée par la nécessité de relire *plusieurs fois* la reformulation afin de la comprendre. La difficulté réside dans un grand nombre d'éléments de ponctuation, de connecteurs logiques, la technicité des termes, la syntaxe très complexe.

- *La dépendance à la cocaïne : un trouble pour lequel il n'existe aucun traitement pharmacologique dont l'efficacité a été éprouvée, mais des avancées dans le domaine de la neurobiologie pourraient orienter le développement de médicaments futurs ;*
- *La fibromyalgie : un syndrome douloureux chronique de l'appareil locomoteur associé à des manifestations fonctionnelles très variées : troubles du sommeil, céphalée, points douloureux, seuil de douleur abaissé, fatigue générale, troubles digestifs, anxiété générale, dépression.*

Nous présentons par la suite l'analyse des résultats d'annotation du niveau de lisibilité des reformulations.

4.2 Résultats d'annotation manuelle de la lisibilité des reformulations

Afin d'évaluer le **niveau de compréhension** de reformulations identifiées lors de nos expériences d'annotation semi-automatique, quatre annotateurs non-spécialistes du domaine de la médecine ont évalué un échantillon de reformulations. Trois annotateurs sont étudiants en Sciences du Langage et un annotateur est étudiant en master de Technologies des Langues[84]. Deux annotateurs sont francophones, deux ont un niveau avancé en français.

Nous avons d'abord évalué le niveau de lisibilité des reformulations médicales issues de nos corpus d'étude français, **CLEAR** et **ClassYN.** Nous avons sélectionné un échantillon de **2 000 reformulations** qui ont été étiquetées comme correctes selon le score inter-annotateur Kappa : 500 reformulations pour chaque type de sous-corpus (**CLEAR SP**, **CLEAR GP**, **ClassYN SP**, **ClassYN GP**).

4.3 Analyse des résultats de l'annotation manuelle de la lisibilité

Chaque sous-corpus a été annoté par deux annotateurs, de la façon suivante :

- Les sous-corpus CLEAR SP et CLEAR GP par Annotateur 1 et Annotateur 2 ;
- Les sous-corpus ClassYN SP et ClassYN GP par Annotateur 3 et Annotateur 4.

Nous analysons deux types d'annotations :

1. L'annotation du niveau de la lisibilité réalisée par chaque annotateur, sans consultation ou échange entre les annotateurs (**Tableau 71**) ;
2. L'adjudication (la mise en accord) entre les paires d'annotateurs par type de sous-corpus. Les annotateurs ont examiné ensemble leurs annotations et se sont mis d'accord pour choisir un niveau de lisibilité pour les reformulations dont le niveau était différent (**Tableau 72**).

Les annotations individuelles sur les deux sous-corpus **CLEAR** montrent une très petite différence entre le corpus grand public (**67%**) et le corpus expert (**66,6%**), concernant le pourcentage de reformulations de niveau 1 (faciles à comprendre). Pour le corpus **ClassYN**, les discrépances sont plus grandes, avec **57,4%** de reformulations plus faciles à

[84] Les annotateurs sont : Laurianne Gully, Lisa Matias-Gaspar, Anna Kalinina, Panagiotis Tsolakis.

comprendre que le terme (*niveau 1*) pour le corpus grand public, par rapport aux **46,3%** pour le corpus expert.

Le nombre de reformulations plus difficiles à comprendre que le terme médical (*niveau 3*) est **deux fois plus grand** dans le corpus ClassYN que le corpus CLEAR dans leur intégralité, avec une moyenne de **23,15%** par rapport à une moyenne de **10,55%**. Nous observons une faible différence de pourcentage entre les corpus CLEAR expert et CLEAR grand public (**11%**, respectivement **10,1%**), tandis que dans le corpus ClassYN, **presque trois fois plus grand** dans le ClassYN EX (**33,8%**) que ClassYN GP (**12,5%**).

Niveau de lisibilité	**CLEAR SP**		**CLEAR GP**		**ClassYN SP**		**ClassYN GP**	
	Annot1	**Annot2**	**Annot1**	**Annot2**	**Annot3**	**Annot4**	**Annot3**	**Annot4**
	N° (%)	**N° (%)**	**N° (%)**	**N° (%)**	**N° (%)**	**N° (%)**	**N° (%)**	**N° (%)**
Niv 1	328 (65,6%)	338 (67,6%)	336 (67,2%)	334 (66,8%)	247 (49,4%)	216 (43,2%)	283 (56,6%)	291 (58,2%)
Niv 2	91 (18,2%)	133 (26,6%)	115 (23%)	114 (22,8%)	103 (20,6%)	96 (19,2%)	155 (31%)	146 (29,2%)
Niv 3	81 (16,2%)	29 (5,8%)	49 (9,8%)	52 (10,4%)	150 (30%)	188 (37,6%)	62 (12,4%)	63 (12,6%)
Total	**500 (100%)**		**500 (100%)**		**500 (100%)**		**500 (100%)**	

Tableau 71. Résultats d'annotation de la lisibilité de reformulations par quatre annotateurs

Lors de leur deuxième annotation, les stagiaires se sont mis d'accord pour uniformiser leurs annotations. Nous observons que le nombre de reformulations faciles (niveau 1) a augmenté à une moyenne de **70%** dans le corpus CLEAR et à **62%** pour le corpus ClassYN. Si pour le corpus CLEAR, la moyenne de reformulations difficiles (niveau 3) a baissé à **7,1%** (par rapport au 10,55%), pour le corpus ClassYN, la moyenne a légèrement augmenté à **24,6%** (par rapport au 23,15%).

Niveau de lisibilité après accord	**CLEAR SP**		**CLEAR GP**		**ClassYN SP**		**ClassYN GP**	
	A1	**A2**	**A1**	**A2**	**A3**	**A4**	**A3**	**A4**
	N° (%)		**N° (%)**		**N° (%)**		**N° (%)**	
Niv 1	359 (71,8%)		341 (68,2%)		235 (47%)		385 (77%)	
Niv 2	109 (21,8%)		120 (24%)		82 (16,4%)		52 (10,4%)	
Niv 3	32 (6,4%)		39 (7,8%)		183 (36,6%)		63 (12,6%)	
Total	**500 (100%)**		**500 (100%)**		**500 (100%)**		**500 (100%)**	

Tableau 72. Résultats de l'annotation de la lisibilité des reformulations après adjudication entre les annotateurs

Lors de leur adjudication, les stagiaires ont discuté les raisons qui les ont aidés à prendre la décision de choisir un certain niveau de lisibilité de la reformulation. Ces raisons ont été synthétisées dans l'échelle de critères d'évaluation présentée dans la **section 4.1**.

4.4 Bilan de l'analyse sur la lisibilité

En guise de conclusion, nous notons que la tâche d'évaluation du niveau de lisibilité et compréhension d'une reformulation médicale dépend largement de la longueur de celle-ci et de la complexité de la syntaxe. L'utilisation des hyperonymes du domaine médical facilite la compréhension d'un terme médical très technique. Nous observons que les connaissances universelles partagées jouent également un rôle important dans la compréhension des reformulations médicales et influencent le choix du niveau de lisibilité. Dans des expériences futures, nous souhaitons tester l'hypothèse sur la variation du niveau de lisibilité des reformulations médicales par rapport aux relations lexicales et aux fonctions sémantico-pragmatiques et évaluer les reformulations en roumain également.

***

Nous concluons notre thèse de doctorat avec la **Partie V** qui présente le bilan général, la conclusion et les perspectives de recherches envisagées lors de futures recherches.

V. BILAN GÉNÉRAL, CONCLUSIONS ET PERSPECTIVES DE RECHERCHE

1. Bilan général de la thèse

Notre travail de thèse a comme objectif la mise en place d'une **méthode semi-automatique de construction de corpus de reformulations médicales**. Notre avons exploré plusieurs travaux, théories et pistes de recherche pour emprunter la méthode la plus adaptée pour notre tâche.

Nous avons exploré **l'état de l'art** sur la question de la *reformulation* et la *paraphrase* en contexte francophone et international et nous avons analysé les théories sur la reformulation dans le discours général et spécialisé (domaine de la médecine). Dans cette thèse, nous considérons **la reformulation sous-phrastique médicale** comme l'équivalence au sens large, basée sur un noyau sémantique commun, qui contribue à la vulgarisation de termes médicaux et qui ne dépasse pas le cadre d'une phrase. Nous avons analysé les travaux sur les *marqueurs lexicaux, grammaticaux et orthographiques* qui indiquent la présence d'une reformulation dans la phrase et nous les avons appliqués dans deux langues, le français et le roumain. Nous avons consulté les travaux sur la *simplification de textes*, sur le *traitement automatique* de la reformulation ou la paraphrase (en contexte anglophone) et sur *l'apprentissage automatique statistique* et par *réseaux de neurones*. Nous avons traité la reformulation comme procédé de la *vulgarisation scientifique* adaptée à un *public cible* (adulte expert ou grand public).

Nous avons travaillé sur deux langues, le *français* et le *roumain*, langue romane moins dotée en ressources textuelles adaptées pour le TAL. Les corpus exploités sont des *corpus comparables, écrits*, qui traitent des sujets médicaux, mais qui se différencient par leurs lecteurs cibles : les spécialistes et le grand public qui veulent s'informer sur les notions médicales. Nous avons construit *le premier corpus de textes de vulgarisation du domaine médical en roumain* (**GrandMed-Ro2**), d'une taille importante (**6 440 951 tokens**), par exploitation des sites de vulgarisation de la toile à l'aide de la plateforme en ligne Sketch Engine.

La méthode automatique de constitution de corpus de reformulation que nous avons développée est constituée des plusieurs étapes de traitement et annotation. Les phrases contenant des termes médicaux ont été extraites automatiquement en partant des

terminologies médicales pour le français et des *listes de termes médicaux* pour le roumain. Nous avons constitué des *listes de marqueurs de reformulation* à partir de l'état de l'art et de nos propres observations sur nos corpus de textes médicaux. Ces marqueurs ont été recherchés dans les phrases qui contiennent des termes médicaux, afin de tester leur utilité dans le marquage d'une reformulation médicale.

Nous avons réalisé un *travail fastidieux d'annotation manuelle* des phrases afin d'identifier davantage de reformulations en dehors des éléments marqués automatiquement (termes médicaux et marqueurs de reformulations issus de recherches automatiques). La *liste de marqueurs* a été élargie sur la base d'observations dans un corpus et la méthodologie appliquée sur un deuxième corpus (avec une liste élargie de marqueurs), dans les deux langues.

Nous avons annoté manuellement un total de **19 890 phrases** :

- **16 227 phrases** du corpus français (**8 667** du corpus **CLEAR Cochrane** et **7 560** du corpus **ClassYN**) ;
- **3 663 phrases** du corpus roumain, **GrandMed-Ro2**.

Pour obtenir des données valides, nous avons mené des campagnes d'annotation sur la base d'un **guide** (qui identifie des critères pour délimiter une reformulation, les relations lexicales entre termes et reformulations, ainsi que leur fonction sémantico-pragmatique). Une proportion de **65%** de ces phrases ont une *double annotation* et *adjudication* entre annotateurs pour obtenir un accord sur une annotation unique.

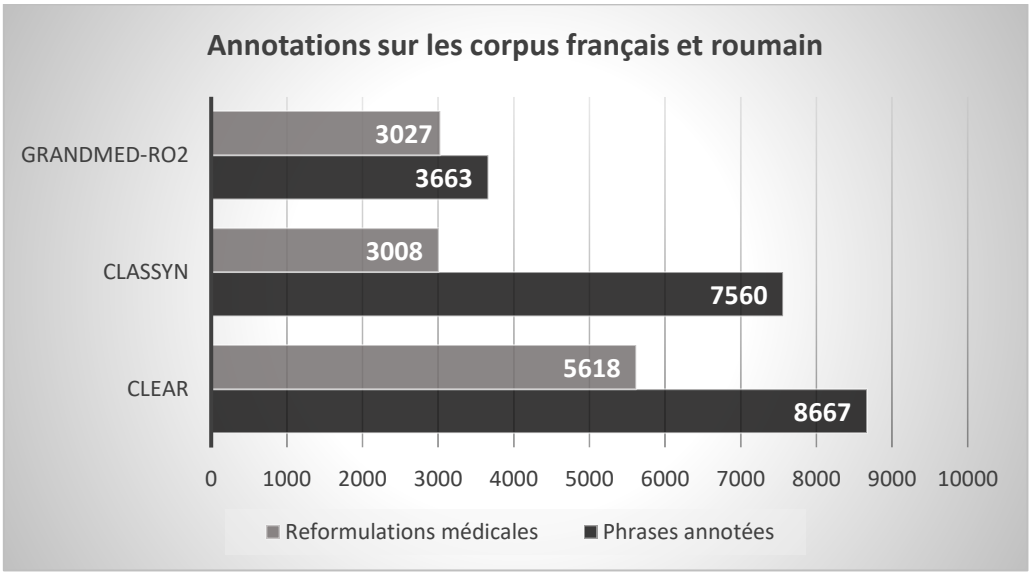

Figure 34. Bilan de résultats d'annotation des reformulations sur les corpus français et roumain

À partir de ces annotations, nous avons identifié **5 618** *paires de termes médicaux avec leurs reformulations médicales correctes* du corpus **CLEAR (100% avec double annotation)**, **3 008** du corpus **ClassYN (35% avec double annotation)** et **3 027** du corpus **GrandMed-Ro2 (33% avec double annotation)**. Avec ces données annotées, nous avons constitué un corpus de reformulations médicales validées par des annotateurs non-spécialistes. Nous appelons notre corpus **RefoMed** (*Reformulations Médicales*). Le corpus **RefoMed** est construit de **11 653** *paires de termes médicaux – reformulations médicales* et il est divisé en deux parties :

- le corpus **RefoMed-Fr**, qui contient **8 626** paires de termes médicaux – reformulations médicales en français ;
- le corpus **RefoMed-Ro**, qui contient **3 027** paires de termes médicaux – reformulations médicales en roumain.

Notre corpus est disponible sur la plateforme en ligne **github** en libre accès pour les chercheurs (sur demande), avec une licence **Creative Commons NonCommercial 4.0 International (CC BY-NC 4.0)**[85] => https://github.com/ibuhnila/refomed

Même si notre objectif est de proposer *une méthode automatique de création de corpus de reformulation*, nous nous sommes intéressés également à la qualité et la diversité des reformulations identifiées et leur utilisation des textes naturels. Nos **11 653 reformulations médicales validées** sont annotées en **relations lexicales** (*hyperonymie, hyponymie, synonymie, méronymie*) et en **fonctions sémantico-pragmatiques** (*définition, exemplification, paraphrase, dénomination, explication*).

Nous avons analysé et évalué un total de **11 314 prédictions de reformulations** générées automatiquement par **l'architecture neuronale APT** avec le modèle de langue de type **Transformer T5** :

- **6 834** prédictions de reformulation en français ;
- **4 480** prédictions de reformulation en roumain.

Nous avons créé des **ressources annotées** et vérifiées manuellement, libres de droits, et un **guide d'annotation et d'évaluation des reformulations** issues des textes naturels et des générations automatiques. Les guides peuvent s'appliquer pour la constitution et l'évaluation d'autres jeux de données (termes et de leur reformulation) pour d'autres domaines. Nous avons analysé les nouvelles reformulations générées

[85] https://creativecommons.org/licenses/by-nc/4.0/

automatiquement et nous avons interprété les résultats afin de les intégrer dans le corpus de reformulation. Pourtant, la tâche d'annotation reste difficile pour des annotateurs non-spécialistes du domaine de la médecine, aussi bien pour l'identification de la reformulation, des relations lexicales et des fonctions sémantico-pragmatiques associées aux reformulations ou à la lisibilité de la reformulation.

Des expériences de classification automatique par apprentissage statistiques avec **l'architecture LSTM** nous montrent que nous avons besoin d'un grand nombre de données annotées équitablement distribuées pour l'identification des reformulations, éviter le surapprentissage et diminuer la perte dans l'apprentissage.

Afin d'analyser **la lisibilité des reformulations** annotées manuellement, **2 000** phrases avec des reformulations correctes ont été annotées par des annotateurs francophones et il y a eu adjudication pour créer une ressource annotée de qualité. Ainsi, nous avons défini un **guide d'annotation de la lisibilité**, qui peut s'appliquer dans une campagne d'annotation similaire.

2. Conclusion générale

L'originalité de notre travail consiste dans le travail sur la reformulation sous-phrastique dans des corpus de grande taille, écrits et comparables du domaine médical. Concernant la langue française, notre apport est une méthode de construction d'un jeu de données spécifiques, des termes médicaux polylexicaux et de leurs reformulations. Ainsi, nous avons exploité et élargi une liste exhaustive de marqueurs de reformulations (à partir de la littérature et des observations en corpus), et nous l'avons combiné avec l'identification de termes à base de terminologies. De plus, la validation manuelle consiste à analyser et annoter des relations lexicales et des fonctions des reformulations issues des textes médicaux écrits dans un but de vulgarisation. Pour le français, nous constituons un **corpus de reformulations médicales** annotées et validées manuellement **(RefoMed-Fr)**. Notre travail d'analyse et d'annotation est **complètement novateur** pour la langue roumaine, ainsi que la constitution de deux ressources textuelles en roumain : un corpus de texte de vulgarisation médicale **(GrandMed-Ro2)** et un corpus de reformulations médicales en roumain **(RefoMed-Ro)**.

Notre étude a validé **l'hypothèse** que les reformulations sous-phrastiques médicales sont présentes en plus grand nombre dans les corpus de vulgarisation. Celles-ci sont également plus faciles à comprendre, à la fois pour les tâches d'annotation et pour la compréhension des lecteurs non spécialistes. Nos **contributions** concernent l'analyse linguistique et l'évaluation des relations lexicales et des fonctions sémantico-pragmatiques qui peuvent être identifiées entre le terme médical et sa reformulation. Nos analyses ont montré que les relations lexicales *d'hyperonymie, l'hyponymie et synonymie* aident à identifier des reformulations médicales valides. Le même résultat est observé pour les fonctions sémantico-pragmatiques de *définition, exemplification et paraphrase*. Cette analyse a permis la création de **plusieurs corpus annotés** en termes de reformulation, de relations lexicales, fonctions sémantico-pragmatiques, et en lisibilité. Les guides d'annotation ainsi que les corpus sont disponibles pour la communauté scientifique.

Notre travail de thèse apporte une **méthodologie d'identification de reformulations médicales** dans des corpus médicaux en français et en roumain. 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 de la langue générale et aident à l'identification des relations entre des termes (contexte définitoire, hyperonymie, méronymie ou synonymie). La méthode est applicable à d'autres langues romanes proches du français.

Le corpus **RefoMed** peut être utile pour des expériences et des recherches dans des tâches de traitement du langage naturel telles que la simplification de textes et des termes médicaux, la réponse à des questions médicales, la génération de textes. La version annotée de notre travail peut être utilisée pour développer des outils destinés aux patients et au grand public afin de mieux comprendre les concepts médicaux (par le biais de chatbots pour les patients ou de documentation de vulgarisation médicale).

3. Perspectives de recherche

Certaines difficultés rencontrées pendant notre travail nous ouvrent plusieurs perspectives de recherche futures. À court terme et dans une perspective d'améliorer les performances des outils et d'agrandir le corpus de reformulations, notre thèse pourra se diriger vers plusieurs axes :

- Travailler sur les autres sous-corpus de CLEAR : les encyclopédies Wikipédia et Vikidia ;
- Exploiter les adaptations des traductions français-roumain et roumain-français des marqueurs de reformulation ;
- Utiliser le module RNER (Mitrofan et Păiș, 2022) pour l'identification des entités nommées médicales sur des textes bruts médicaux en roumain et pour élargir la liste des reformulations ;
- Exploiter le marqueur de synonymie « ou » pour chercher automatiquement des paraphrases synonymiques ;
- Chercher également des lemmes des termes polylexicaux (surtout pour le roumain qui est une langue flexionnelle) et les faire correspondre à leur équivalent dans le discours ;
- Évaluer le niveau de lisibilité des reformulations médicales par rapport aux relations lexicales et aux fonctions sémantico-pragmatiques. Évaluer la lisibilité des reformulations en roumain ;
- Explorer d'autres paramètres pour les architectures neuronales afin d'augmenter la précision des prédictions correctes générées pour les termes médicaux.
- Compléter les expériences de classification avec une architecture LSTM pour le roumain, avec un jeu de données plus conséquent.

Vu le contexte numérique de la société actuelle et l'utilisation croissante des outils d'intelligence artificielle par le grand public, nous envisageons également des perspectives de recherches transdisciplinaires, de type :

- Évaluer l'impact carbone de l'utilisation des réseaux de neurones en TAL afin d'affiner les expériences et réduire le temps de calcul, en faisant appel à des jeux de

données annotées pour une tâche précise (comme notre corpus annoté en reformulations sous-phrastiques médicales, RefoMed) ;

- Analyser l'enjeu éthique de l'utilisation des données issues du web, pour éviter la génération automatique de reformulations discriminatoires ou malaisantes (comme nous avons pu le constater lors de l'analyse des prédictions automatiques générées avec le modèle de langue T5) ;
- Trouver les moyens d'implémenter les reformulations médicales identifiées dans plusieurs supports : des sites de *traduction* du langage scientifique vers un langage simplifié, des chatbots pour des patients, des notices médicales simplifiées, des applications de simplification de textes techniques.

4. Références bibliographiques

A

Abadi Martın, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, …, Xiaoqiang Zheng. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *ArXiv:1603.04467v2*., pp. 19.

Adam Jean-Michel. (1990). *Éléments de linguistique textuelle : théorie et pratique de l'analyse textuelle*. Éditions Mardaga. 272 pages.

Adam Jean-Michel & Revaz Françoise. (1989). Aspects de la structuration du texte descriptif : les marqueurs d'énumération et de reformulation. *Langue française* 81, pp. 59-98.

Agirre Eneko, Banea Carmen, Cer Daniel, Diab Mona, Gonzalez-Agirre Aitor, Mihalcea Rada, Rigau German & Wiebe Janyce (2016). Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. pp. 497-511.* ACL (Association for Computational Linguistics).

Andreopoulos Bill, Alexopoulou Dimitra & Schroeder Michael. (2008). Word Sense Disambiguation in biomedical ontologies with term co-occurrence analysis and document clustering. *Int. J. Data Min. Bioinformatics*, vol. 2, no. 3, pp. 193-215.

Antoine Edwide & Grabar Natalia. (2016). Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert. In *TALN 2016 : Traitement Automatique des Langues Naturelles*. Paris, France. https://hal.archives-ouvertes.fr/hal-01426816.

Authier-Revuz Jacqueline. (1995). *Ces mots qui ne vont pas de soi : boucles réflexives et non-coïncidences du dire*. t. 1-2, Paris, Larousse.

B

Bannard Colin & Callison-Burch Chris. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05*, pp. 597–604. Ann Arbor, Michigan: Association for Computational Linguistics.

Barbu Mititelu Verginica. (2011). Hyponymy Patterns in Romanian. *Memoirs of the Scientific Sections of the Romanian Academy.* vol. XXXIV. pp. 1-13.

Barbu Mititelu Verginica & Mitrofan Maria. (2020). The Romanian medical treebank-SiMoNERo. In *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2020*, pp. 7–16.

Barzilay Regina & McKeown Kathleen R. (2001). Extracting Paraphrases from a Parallel Corpus. In *Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pp. 50–57. Toulouse, France : Association for Computational Linguistics. https://doi.org/10.3115/1073012.1073020.

Beeching Kate. (2007). La co-variation des marqueurs discursifs "bon", "c'est-à-dire", "enfin", "hein", "quand même", "quoi" et "si vous voulez" : une question d'identité ? *Langue française*, 154(2), pp. 78–93.

Benveniste Claire-Blanche. (2010). Lexique et grammaire dans les reformulations. *La rectification à l'oral et à l'écrit,* M. Candea & Reza Mir-Samii (Dir.), Paris, Ophrys, pp. 77-89.

Benveniste Claire-Blanche, Bilger Mireille, Rouget Chirstine & Van Den Eynde Karel. (1990). Le français parlé. Études grammaticales. Paris: *CNRS Éditions*.

Bhagat Rahul & Hovy Eduard. (2013). Squibs: What Is a Paraphrase?. *Computational Linguistics 39* (3), pp. 463–472.

Bidu-Vrânceanu Angela. (2007). Lexicul specializat în mișcare. De la dicționare la texte. *București. Editura Universității din București*. 266 pages.

Bird Steven, Klein Ewan & Loper Edward (2009). Natural Language Processing with Python. *O'Reilly Media Inc*. https://www.nltk.org/book/

Bodenreider Olivier. (2004). The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research 32* (90001): 267D - 270. https://doi.org/10.1093/nar/gkh061.

Bohnet Bernd. (2009). Synchronous Parsing of Syntactic and Semantic Structures. In *Proceedings of the Fourth International Conference on Meaning-Text Theory*, Montreal, pp. 77-86.

Boroș Tiberiu, Dumitrescu Ștefan Daniel & Burtică Ruxandra. (2018). NLP-Cube: End-to-End Raw Text Processing With Neural Networks. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, pp. 171-179.

Bouamor Houda. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues*. Orsay : Université Paris Sud - Paris XI. https://tel.archives-ouvertes.fr/tel-00717702.

Bouamor Houda, Illouz Gabriel, Max Aurélien & Vilnat Anne. (2012). Validation sur le Web de reformulations locales : application à la Wikipédia. Dans *Actes de la conférence conjointe JEP-TALN-RECITAL 2012, 2 : TALN*, pp. 197-210. Grenoble. http://www.aclweb.org/anthology/F12-2015.

Bowker Lynne & Pearson Jennifer. (2002). Working with Specialized Language. *A Practical Guide to Using Corpora*. Londres, New York: Routledge.

Bowman Samuel R., Gauthier Jon, Rastogi Abhinav, Gupta Raghav, Manning Christopher D. & Potts Christopher. (2016). A Fast Unified Model for Parsing and Sentence Understanding. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pp. 1466-1477.

Brants Thorsten & Franz Alex. (2006). Web 1T 5-gram version 1. *Philadelpha: Linguistic Data Consortium*, 2006.

Brixhe Daniel & Specogna Antoinette. (1999). Actes de reformulation et progression du savoir. *Pratiques 103 (1)*, pp. 9-27. https://doi.org/10.3406/prati.1999.1858.

Brockett Chris & Dolan William B. (2005). Support Vector Machines for Paraphrase Identification and Corpus Construction. Dans *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pp. 1-8. http://aclweb.org/anthology/I/I05/I05-5001.

Brown Tom B., Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel M., Wu Jeffrey, Winter Clemens, …, Amodei Dario. (2020). Language Models are Few-Shot Learners. *OpenAI*, arXiv:2005.14165. https://doi.org/10.48550/arXiv.2005.14165.

Brown, Elliot G., Wood Louise & Wood Sue. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Saf., 20(2)*, pp.109–117.

Brouwers Laetitia, Bernhard Delphine, Ligozat Anne-Laure & François Thomas. (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*, pp. 211–224. Grenoble, France : ATALA/AFCP.

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

Buhnila Ioana. (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*, Iasi, Roumanie (online), pp. 139-152.

Buhnila Ioana. (2022a). Le Rôle Des Marqueurs et Indicateurs Dans l'analyse Lexicale et Sémantico-Pragmatique de Reformulations Médicales. *8e Congrès Mondial de Linguistique Française (CMLF)*, 4-8 juillet 2022, Orléans, France, SHS Web of Conferences 138: 10005. https://doi.org/10.1051/shsconf/202213810005.

Buhnila Ioana. (2022b). Identifying Medical Paraphrases in Scientific versus Popularization Texts in French for Laypeople Understanding. In *Proceedings of the Third Workshop on Scholarly Document Processing*. *COLING 22'*, Gyeongju, Republic of Korea, pp. 69-79. Association for Computational Linguistics.

C

Cai Shu & Knight Kevin. (2013). Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2*: Short Papers, pp. 748–752. Sofia, Bulgaria: Association for Computational Linguistics.

Callison-Burch Chris, Koehn Philipp, Monz Christof & Schroeder Josh. (2009). Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pp. 1-28. Athens, Greece : Association for Computational Linguistics.

Cardon Rémi. (2018). Approche lexicale de la simplification automatique de textes médicaux. In *Actes de la conférence Traitement Automatique de la Langue Naturelle, TALN 2018*, pp. 159-73. Rennes, France.

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

Cardon Rémi & Grabar Natalia. (2019). Automatic detection of parallel sentences in comparable biomedical corpora. In *TALN 2019*. Toulouse, France. https://hal.archives-ouvertes.fr/hal-02430446.

Cardon Rémi & Grabar Natalia. (2021). Simplification automatique de textes biomédicaux en français : lorsque des données précises de petite taille aident (French Biomedical Text Simplification : When Small and Precise Helps). In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pp. 275–277, Lille, France. ATALA.

Carreras Xavier & Marquez Lluıs. (2005). Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*. (CoNLL-2005), pp. 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Chandrasekar Raman & Srinivas Bangalore. (1997). Automatic induction of rules for text simplification. *Knowledge-Based Systems*, vol. 10, no 3, pp. 183-190.

Chang Chih-Chung & Lin Chih-Jen. (2011). LIBSVM: A library for support vector machines. In *ACM Transactions on Intelligent Systems and Technology*, 2 : 27:1– 27:27.

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

Chaumartin François-Régis. (2007). Extraction de paraphrases désambiguïsées à partir d'un corpus d'articles encyclopédiques alignés automatiquement. In *RECITAL*, pp. 457-466. France. https://hal.archives-ouvertes.fr/hal-00611241.

Chen MeiHua, Chen YiChun, Huang ShihTing & Chang Jason S. (2013). Augmentable Paraphrase Extraction Framework. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 706–711. Nagoya, Japan: Asian Federation of Natural Language Processing. http://www.aclweb.org/anthology/I13-1083.

Chen David L. & Dolan Bill. (2011). Building a persistent workforce on mechanical turk for multilingual data collection. In *Building a Persistent Workforce on Mechanical Turk for Multilingual Data Collection*.

Chéria Najah. (2010). Reformulation paraphrastique et non paraphrastique dans La Jalousie de Robbe-Grillet : l'exemple de c'est-à-dire vs en fait et en réalité. *L'Information grammaticale 127*, pp. 43-47.

Cislaru Georgeta & Thierry Olive. (2018). Les jets textuels de révision : un point de vue dynamique sur la « reformulation ». *Langages*, vol. 212, no. 4, 2018, pp. 69-86.

Coates-Stephens Sam. (1991). Coping with Lexical Inadequacy - the Automatic Acquisition of Proper Nouns from New Text. In *Proceedings of the 7th Annual Conference of the UW Centre for the New OED and Text Research: Using Corpora*, Oxford, pp. 154-169.

Coavoux Maximin & Crabbé Benoît. (2017). Incremental discontinuous phrase structure parsing with the gap transition. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics : Volume 1, Long Papers*, pp. 1259–1270, Valencia, Spain : Association for Computational Linguistics.

Cohen Jacob. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas., 20*, pp. 27-46.

Collins Michael & Duffy Nigel. (2001). Convolution kernels for natural language. *Advances in Neural Information Processing Systems*. *MIT Press 14*, pp. 625–632.

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

Conneau Alexis, Khandelwal, Kartikay, Goyal Naman, Chaudhary Vishrav, Wenzek Guillaume, Guzmán, Francisco, Grave Edouard, Ott Myle, Zettlemoyer Luke & Stoyanov Veselin. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Constant Matthieu & Sigogne Anthony. (2011). MWU-Aware Part-of-Speech Tagging with a CRF Model and Lexical Resources. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pp. 49–56, Portland, Oregon, USA. Association for Computational Linguistics.

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

Coseriu Eugen. (1967). Structures lexicales et enseignement du vocabulaire. *Les Théories linguistiques et leurs applications*, Strasbourg, Conseil de l'Europe, pp. 9-51.

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

Côté Roger A., Rothwell David J., Palotay James L., Beckett Ronald S. & Brochu Louise. (1993). The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International. *Northfield, IL: College of American Pathologists*.

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

Côté Roger A. (1998). Systematized Nomenclature of Human and Veterinary Medicine: SNOMED International. Version 3.5. *Northfield, IL: College of American Pathologists*.

Creutz Mathias. (2018). Open Subtitles Paraphrase Corpus for Six Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 13, pp. 64-69. Miyazaki, Japon. http://arxiv.org/abs/1809.06142.

Culioli Antoine. (1983). Notes du séminaire de D.E.A. 1983-1984. Paris. Université de Paris 7. [disponible en ligne : https://www.scribd.com/doc/15735/Culioli-Notes-de-seminaire-1983].

D

Daille Béatrice, Jacquin Christine, Monceaux Laura, Morin Emmanuel & Rocheteau Jérome. (2011). TTC TermSuite : une chaîne de traitement pour la fouille terminologique multilingue (TTC TermSuite: a processing chain for multilingual terminology mining). In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles. Démonstrations*, pp. 6–6, Montpellier, France. ATALA.

Deléger Louise & Zweigenbaum Pierre. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora*, pp. 2–10. BUCC '09. Suntec, Singapore: Association for Computational Linguistics.

Devlin Jacob, Chang Ming-Wei, Lee Kenton & Toutanova Kristina. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*. http://arxiv.org/abs/1810.04805.

Dinh Duy & Tamine Lynda. (2010). Recherche d'information sémantique dans les documents biomédicaux : approche basée sur le sens précis des concepts. *INFormatique des Organisations et Systèmes d'Information et de Décision*, INFORSID 2010, pp. 261-74.

Dolan Bill, Quirk Chris & Brockett Chris. (2004). Unsupervised construction of large paraphrase corpora: exploiting massively parallel news sources. In *COLING '04 : Proceedings of the 20th international conference on Computational Linguistics*, pp. 350–356. Geneva, Switzerland: International Conference on Computational Linguistics.

Donnelly Kevin. (2006). SNOMED-CT: The Advanced Terminology and Coding System for EHealth. *Studies in Health Technology and Informatics*, 121, pp. 279-90.

Dras Mark. (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text*. Ph.D. Thesis, Sydney: Macquarie University.

Drescher Martina. (2008). La reformulation dans la prévention contre le HIV/SIDA. *Pragmatique de la reformulation*. *Types de discours - Interactions didactiques*. Sous la direction de Martine SCHUWER. Marie-Claude LE BOT, Elisabeth RICHARD, pp. 39-54. Rennes : Presses Universitaires de Rennes.

Drouin Patrick (2003). Term extraction using non-technical corpora as a point of leverage. In *Terminology*, vol. 9, no 1, pp. 99-117.

Dufour Françoise. (2005). Reformulation métalinguistique et recatégorisation du référent : civilisation et développement dans la formation discursive du progrès. *Matérialités de l'activité de nomination : Formes, discours, représentations.* Presses Sorbonne-Nouvelle, pp. 165-76.

Dumitrescu Stefan, Avram Andrei-Marius & Pyysalo Sampo. (2020). The birth of Romanian BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4324–4328, Online. Association for Computational Linguistics.

Dupuch Marie, Hamon Thierry & Grabar Natalia. (2013). Grouping of terms based on linguistic and semantic regularities in a cross-lingual context (Groupement de termes basé sur des régularités linguistiques et sémantiques dans un contexte cross-langue) [in French]. *In Proceedings of TALN 2013 (Volume 1: Long Papers)*, pp. 62–75. Les Sables d'Olonne, France : ATALA. https://www.aclweb.org/anthology/F13-1005.

Dutrey Camille, Bouamor Houda, Bernhard Delphine & Max Aurélien. (2011). Typologie des modifications dans les révisions de Wikipédia. *LIMSI Technical Report 2011-01 N° : 2011-01*, pp. 1-24.

E

Elhadad Noemie & Sutaria Komal. (2007). Mining a Lexicon of Technical Terms and Lay Equivalents. In *Biological, translational, and clinical language processing*, pp. 49–56. Prague, Czech Republic: Association for Computational Linguistics. https://www.aclweb.org/anthology/W07-1007.

Erjavec Tomaž. (2010). MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Proceedings of the Seventh International Conference on Language*

Resources and Evaluation (LREC'10), pp. 2544-2547. Valletta, Malta. European Language Resources Association (ELRA).

Eshkol-Taravella Iris, Baude Olivier, Maurel Denis, Hriba Linda, Dugua Céline & Tellier Isabelle. (2011). Un grand corpus oral « disponible » : le corpus d'Orléans 1 1968-2012. *Traitement Automatique des Langues 53 (2):* pp. 17-46.

Eshkol-Taravella Iris & Grabar Natalia. (2014). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. In *21 ème Traitement Automatique des Langues Naturelles*, pp. 304-15. Marseille. http://www.aclweb.org/anthology/F14-1027.

Eshkol-Taravella Iris & Grabar Natalia. (2017). Taxinomie dans les reformulations du point de vue de la linguistique de corpus. *Syntaxe et Sémantique*, vol. 18, no. 1, pp. 149-184.

Eshkol-Taravella Iris & Grabar Natalia. (2018). Reformulations : de l'étude outillée dans les corpus disponibles vers leur détection automatique*. Langages N° 212 (4)*, pp. 5-16.

F

Fikri Aji Alham, Eko Prasojo Radityo, Noor Fatyanosa Tirana, Arthur Philip, Fitriany Suci, Qonitah Salma, Zulfa Nadhifa, Santoso Tomi & Data Mahendra. (2021). ParaCotta: Synthetic Multilingual Paraphrase Corpora from the Most Diverse Translation Sample Pair. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pp. 533-542, Shanghai, China. Association for Computational Lingustics.

Filice Simone & Moschitti Alessandro. (2016). Learning to Recognize Ancillary Information for Automatic Paraphrase Identification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1109–1114. San Diego, California: Association for Computational Linguistics. http://www.aclweb.org/anthology/N16-1129.

Flanigan Jeffrey, Thomson Sam, Carbonell Jaime, Dyer Chris, Smith Noah A. (2014). A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pp. 1426-1436.

Fløttum Kjersti. (1995). *Dire et redire. La reformulation introduite par « c'est-à-dire ».* Thèse de l'Université de Stavanger. Norvège.

Franckel Jean-Jacques. (2005). De l'interprétation à la glose: vers une méthodologie de la reformulation. *D'une langue à l'autre*, pp. 51-78.

François Thomas. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère*. Thèse de Doctorat. Université Catholique de Louvain. Louvain, France.

Freelon Deen. (2013). ReCal OIR: Ordinal, Interval, and Ratio Intercoder Reliability as a Web Service. *International Journal of Internet Science 8 (1)*, pp. 10-16.

Fuchs Catherine. (1980). Quelques réflexions sur la paraphrase dans les théories du langage. *L'information grammaticale 6 (1)*, pp. 37 44.

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

Fuchs Catherine. (1994). *Paraphrase et énonciation*. Éditions OPHRYS, 185 pages.

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

G

Gala Nuria, Thomas François, Bernhard Delphine & Fairon Cédrick. (2014). A model to predict lexical complexity and to grade words. In *Actes de la 21e conférence sur le Traitement Automatique des Langues Naturelles, TALN'2014*, pp. 91-102. Marseille, France. https://hal.archives-ouvertes.fr/hal-01001916.

Gala Núria, Billami Mokhtar B., François Thomas, Bernhard Delphine. (2015). Graded lexicons: new resources for educational purposes and much more. In *22nd Computer-assisted language learning conference (EUROCALL-2015)*, pp. 204-209.

Ganitkevitch Juri & Callison-Burch Chris. (2014). The Multilingual Paraphrase Database. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 4276–4283. Reykjavik, Iceland: European Language Resources Association (ELRA).

Ganitkevitch Juri, Van Durme Benjamin & Callison-Burch Chris. (2013). PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 758–764. Atlanta, Georgia: Association for Computational Linguistics.

Garcia-Debanc Claudine. (2015). La reformulation : usages et contextes. *Corela [En ligne]*, HS-18 | 2015, mis en ligne le 15 novembre 2015, consulté le 11 septembre 2018. 1-7. https://doi.org/10.4000/corela.4032.

Gardin Bernard. (1987). Les enjeux sociaux des reformulations, *Études de linguistique appliquée* 68, pp. 95-110.

Gaudan, Sylvain, Kirsch Harald, Rebholz-Schuhmann Dietrich. (2005). Resolving abbreviations to their senses in Medline. *Bioinformatics, vol. 21*, no. 18, pp. 3658-3664.

Gorrell Genevieve, Song Xingyi & Roberts Angus. (2018). Bio-YODIE: A Named Entity Linking System for Biomedical Text. *arXiv:1811.04860* [cs], http://arxiv.org/abs/1811.04860, pp. 1-5.

Grabar Natalia & Zweigenbaum Pierre. (2000). A General Method for Sifting Linguistic Knowledge from Structured Terminologies. *JAMIASUP*, pp. 310-314.

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

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

Grabar Natalia & Eshkol-Taravella Iris. (2016b). Prédiction automatique de fonctions pragmatiques dans les reformulations. In *TALN 2016: Traitement Automatique des Langues Naturelles*, pp. 1-15. Paris, France: hal-01426814.

Grabar Natalia & Hamon Thierry. (2014). Automatic extraction of layman names for technical medical terms. In *2014 IEEE International Conference on Healthcare Informatics. IEEE*, pp. 310-319.

Grabar Natalia & Hamon Thierry. (2016). Exploitation de la morphologie pour l'extraction automatique de paraphrases grand public des termes médicaux. *Traitement Automatique des Langues, Varia*, *57 (1)*, pp. 85-109.

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

Grau Brigitte, Ligozat Anne-Laure & Gleize Martin. (2015). Recherche d'information précise dans des sources d'information structurées et non structurées : défis, approches et hybridation. *Traitement Automatique des Langues, LIMSI, CNRS, ENSIIE*, Université Paris-Saclay 56; 3.

Green Georgia M. & Olsen Margaret S. (1986). Preferences for and comprehension of original and readability-adapted materials. *Technical report, Champaign, Ill.:* University of Illinois at Urbana-Champaign, Center for the Study of Reading.

Gülich Elisabeth & Kotschi Thomas. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de Linguistique française* 5, pp. 305-351.

Gülich, Elisabeth & Thomas Kotschi. (1987). Les actes de reformulation dans la consultation: La Dame, de Caluire." L'analyse des interactions verbales, la dame de Caluire-une consultation: *actes du colloque tenu à l'Univ. de Lyon 2* du 13-15 décembre 1985. Vol. 18.

Gupta Ankush, Agarwal Arvind, Singh Prawaan & Rai Piyush. (2018). A deep generative framework for paraphrase generation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

H

Hahn Udo, Honeck Martin, Piotrowsky Michael, Schulz Stefan. (2001). Subword segmentation - leveling out morphological variations for medical document retrieval. *AMIA*, pp. 229-233.

Harris Zellig S. (1954). Distributional Structure. *WORD 10:2-3*, pp. 146-62. https://doi.org/10.1080/00437956.1954.11659520.

Harris Zellig S. (1976). *Notes du cours de syntaxe*. Le Seuil. Paris.

Hazem Al Saied, Matthieu Constant & Marie Candito. (2017). The ATILF-LLF System for Parseme Shared Task: a Transition-based Verbal Multiword Expression Tagger. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, pp. 127–132, Valencia, Spain. Association for Computational Linguistics.

He Hua & Lin Jimmy. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement. (2016). In *Proceedings of the 2016 conference of the north American chapter of the Association for Computational Linguistics: human language technologies*. pp. 937-948.

Hearst Marti A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2, 539–545. COLING '92*. Nantes, France: Association for Computational Linguistics. https://doi.org/10.3115/992133.992154.

Heiden Serge, Magué Jean-Philippe, Pincemin Bénédicte. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In *I. C. Sergio Bolasco (Ed.), Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010)* (Vol. 2, pp. 1021-1032). Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. Online.

Hochreiter Sepp & Jurgen Schmidhuber. (1997). Long short-term memory. *Neural computation, 9(8)*, pp. 1735–1780.

Hölker Klaus. (1988). *Zur Analyse von Markern*. Stuttgart : Franz Steiner.

Honeyfield James E. (1997). Simplification. *TESOL Quarterly*, pp. 431-40.

Howard Jeremy & Ruder Sebastian. (2018). Universal Language Model Fine-tuning for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328-339.

Hwang William, Hajishirzi Hannaneh, Ostendorf Mari & Wu Wei. (2015). Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL&HLT*, pp. 211–217.

I

Inkova Olga. (dir.). (2020). Autour de la reformulation. *Droz, collection Recherches et rencontres*, Publication de la Faculté des Lettres de l'Université de Genève, no. 36, 216 pages.

Inkova Olga & Guryev Alexander. (2018). K voprosu o tak nazyvaemyh pojasnitelnyh otnošenijah v russkom jazyke [À propos des relations dites « explicatives » dans la langue russe]. *Russkij jazyk v naučnom ošveščenii* 35, pp. 46-73.

Ion Radu, Irimia Elena, Ștefănescu Dan & Tufiș Dan. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*. Istanbul, Turkey: European Language Resources Association (ELRA).

Ion Radu. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian (in Romanian).* Ph.D. thesis, Romanian Academy.

Ion Radu. (2018). TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian. In *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language (ConsILR 2018)*, November 22-23, 2018, Iași, Romania.

Ion Radu, Elena Irimia & Verginica Barbu Mititelu. (2018). Ensemble Romanian Dependency Parsing with Neural Networks. In *Proceedings of the International Conference on Language Resources and Evaluation*, pp. 1574-1579.

Issa Fuad, Damonte Marco, Cohen Shay B., Yan Xiaohui & Chang Yi. (2018). Abstract Meaning Representation for Paraphrase Detection. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 442–452. New Orleans, Louisiana: Association for Computational Linguistics. http://www.aclweb.org/anthology/N18-1041.

J

Jacquemin Christian. (1999). Syntagmatic and paradigmatic representations of term variation. In *Proceedings of ACL*, College Park, États-Unis.

Jakobson Roman. (1959). On linguistic aspects of translation. Reuben A. Brower (ed.), *On translation*, Cambridge, Mass.: Harvard University, Press., pp. 232-39.

Jakobson Roman. (1963). *Essais de linguistique générale*. Traduit de l'anglais et préfacé par Nicolas Ruwet. Minuit. 260 pages.

Joshi Mahesh, Pedersen Ted, Maclin Richard. (2005). A Comparative Study of Support Vector Machines Applied to the Word Sense Disambiguation Problem for the Medical Domain. *IICAI'05*, pp. 3449- 3468.

K

Kanaan Layal. (2011). *Reformulations, contacts de langues et compétence de communication : analyse linguistique et interactionnelle dans des discussions entre jeunes Libanais francophones*. Thèse de doctorat, Université d'Orléans, Orléans.

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

Kara Mohamed & Wiederspiel Brigitte. (2007). Anaphores résomptives et reformulations. M. Kara (éd.), *Usages et analyses des reformulations*, Metz, Université de Metz, 97-121.

Kilgarriff Adam, Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel & Suchomel Vít. (2014). The Sketch Engine: ten years on. *Lexicography 1*, pp. 7-36.

Kim Jin-Dong, Ohta Tomoko, Tateisi Yuka & Tsujii Junichi. (2003). Genia corpus a semantically annotated corpus for bio-text mining. *Bioinformatics 19(suppl 1)* : i180–i182.

Kim Donghyeon, Lee Jinhyuk, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung & Jaewoo Kang. (2019). A Neural Named Entity Recognition and Multi-Type Normalization Tool for Biomedical Text Mining. *IEEE Access 7:* 73729-40. https://doi.org/10.1109/ACCESS.2019.2920708.

Kingsbury Paul & Palmer Martha. (2002). From TreeBank to PropBank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, pp. 1989-93. Las Palmas, Canary Islands, Spain: European Language Resources Association (ELRA).

Klein Guillaume, Hernandez François, Nguyen Vincent & Senellart Jean. (2020). The OpenNMT Neural Machine Translation Toolkit: 2020 Edition. *AMTA* 2020.

Koehn Philipp. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the Tenth Machine Translation Summit (MT Summit X)*, pp. 79-86. Phuket, Thailand.

Koehn Philipp, Och Franz J. & Marcu Daniel. (2003). Statistical Phrase-Based Translation. In *Proceedings of NAACL-HLT*, pp. 48–54, Edmonton, Canada.

Koehn Philipp, Hoang Hieu, Birch Alexandra, Callison-Burch Chris, Federico Marcello, Bertoldi Nicola, Cowan Brooke, Shen Wade, Moran Christine, Zens Richard, Dyer Chris, Bojar Ondřej, Constantin Alexandra & Herbst Evan. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the association for computational linguistics companion volume proceedings of the demo and poster sessions*, pp. 177-180.

Koptient Anaïs, Cardon Rémi & Grabar Natalia. (2019). Simplification-induced transformations: typology and some characteristics. In *BioNLP 2019*. Florence, Italy. https://doi.org/10.18653/v1/W19-5033.

Koptient Anaïs & Grabar Natalia. (2020). Rated Lexicon for the Simplification of Medical Texts. *The Fifth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing HEALTHINFO 2020*, Oct 2020, Porto, Portugal. ⟨hal-03095275⟩.

L

Labrak Yanis, Bazoge Adrien, Dufour Richard, Rouvier Mickael, Morin Emmanuel, Daille Béatrice & Gourraud Pierre-Antoine. (2023). DrBERT: A Robust Pre-trained Model in French for Biomedical and Clinical domains. arXiv:2304.00958

Laframboise Yvon. (1978). La lisibilité : Qu'est-ce que la lisibilité ? Quels éléments rendent un texte lisible et un autre pas ? *Québec français 32*, pp. 27-29.

Lamy Jean-Baptiste, Venot Alain & Duclos Catherine. (2015). PyMedTermino: an open-source generic API for advanced terminology services. *Studies in Health Technology and Informatics*, IOS Press, 2015, 210, pp. 924-8. ⟨hal-03650024⟩.

Lan Wuwei & Xu Wei. (2018). Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3890–3902.

Laurent Dominique, Nègre Sophie & Séguéla Patrick. (2009). Apport des co-occurrences à la correction et à l'analyse syntaxique. *TALN*. Session posters, Senlis, 24-26 juin 2009.

Le Bot Michel, Schuwer Martine & Richard Élisabeth. (éds). (2008). La reformulation, Marqueurs linguistiques et stratégies énonciatives. Rennes, *Presses Universitaires de Rennes*, Collection « Rivages linguistiques ».

Le Hang, Vial Loïc, Frej Jibril, Segonne Vincent, Coavoux Maximin, Lecouteux Benjamin, Allauzen Alexandre, Crabbé Benoît, Besacier Laurent & Schwab Didier. (2020). FlauBERT: Unsupervised Language Model Pre-training for French. *arXiv:1912.05372 [cs]*. http://arxiv.org/abs/1912.05372.

Le Minh, Postma Marten, Urbani Jacopo & Vossen Piek. (2018). A Deep Dive into Word Sense Disambiguation with LSTM. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 354–365. Santa Fe, New Mexico, USA: Association for Computational Linguistics. http://www.aclweb.org/anthology/C18-1030.

Lebaud Daniel & Ploog Katja. (2013). Paraphrases, reformulations et gloses : points de vues linguistiques. <halshs-00821809>, pp. 1-18.

Lee Jinhyuk, Yoon Wonjin, Kim Sungdong, Kim Donghyeon, Kim Sunkyu, Ho So Chan & Jaewoo Kang. (2020). BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics 36 (4)*, pp.1234-40. https://doi.org/10.1093/bioinformatics/btz682.

Leroy Gondy & Rindflesch Thomas C. (2005). Effects of Information and Machine Learning Algorithms on Word Sense Disambiguation with Small Datasets. *International Journal of Medical Informatics, MedInfo 2004, 74 (7)*, pp. 573-85. https://doi.org/10.1016/j.ijmedinf.2005.03.013

Leroy Gondy, Helmreich Stephen, Cowie James R., Miller Trudi & Zheng Wei. (2008). Evaluating Online Health Information: Beyond Readability Formulas. *AMIA Annual Symposium Proceedings 2008*, pp. 394-98.

Leroy Gondy, Kauchak D. & Mouradi O. (2013). A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty. *Int J Med Inform, 82(8)*, pp. 717–730.

Lété Bernard, Sprenger-Charolles Liliane & Colé Pascale. (2004). MANULEX: A Grade-Level Lexical Database from French Elementary School Readers. *Behavior Research Methods, Instruments & Computers 36 (1)*, pp. 156-66. https://doi.org/10.3758/BF03195560.

Levenshtein Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physice-Doklady*, 10:707-710.

Li Linlin, Roth Benjamin & Sporleder Caroline. (2010). Topic Models forWord Sense Disambiguation and Token-based Idiom Detection. In *ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, p. 1138-47. Uppsala, Sweden.

Ligozat Anne-Laure, Grouin Cyril, Garcia-Fernandez Anne & Bernhard Delphine. (2013). Studying frequency-based approaches to process lexical simplification (Approches à base de fréquences pour la simplification lexicale) [in French]. In *Proceedings of TALN 2013 (Volume 1: Long Papers)*, pp. 493–506. Les Sables d'Olonne, France: ATALA. https://www.aclweb.org/anthology/F13-1036.

Lin Dekang, Church Kenneth, Ji Heng, Sekine Satoshi, Yarowsky David, Bergsma Shane, Patil Kailash, Pitler Emily, Lathbury Rachel, Rao Vikram, Dalwani Kapil & Narsale Sushant. (2010). New tools for web-scale n-grams. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). Valletta, Malta. European Language Resources Association (ELRA), pp. 2221-2227.*

Lindberg Donald, Humphreys Betsy & Mccray Alexa. (1993). The Unified Medical Language System. *Methods Inf Med, vol. 32,* no 4, pp. 281-291.

Lipscomb Carolyn E. (2000). Medical Subject Headings (MeSH). *Bulletin of the Medical Library Association 88 (3)*, pp. 265-66.

Lison Pierre & Tiedemann Jorg. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pp. 923-29.

Liu Hongfang, Teller Virginia & Friedman Carol. (2004). A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. *J Am Med Inform Assoc, vol. 11*, no 4, pp. 320-31.

Liu Yinhan, Ott Myle, Goyal Naman, Du Jingfei, Joshi Mandar, Chen Danqi, Levy Omer, Lewis Mike, Zettlemoyer Luke & Stoyanov Veselin. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*. https://doi.org/10.48550/arXiv.1907.11692.

Loffler-Laurian Anne-Marie. (1984). Vulgarisation scientifique : formulation, reformulation, traduction. *Langue française, Français technique et scientifique : reformulation, enseignement*, no 64, pp. 109-125.

M

Magri Véronique. (2018). Marqueurs de reformulation : exploration outillée et contrastive dans deux corpus narratifs. *Langages No 212 (4)*, pp. 35-50.

Magri-Mourgues Véronique. (2012). Reformulation et textualité dans les contes de La Maison Tellier de Maupassant. *SHS Web of Conferences 1*, pp. 1143-59. https://doi.org/10.1051/shsconf/20120100024.

Magri-Mourgues Véronique. (2013a). Reformulation et dialogisme dans le récit de voyage. In *Echos des voix, échos des textes*, édité par Odile Gannier, pp. 467-82. Classiques Garnier. https://hal.archives-ouvertes.fr/hal-01226868.

Magri-Mourgues Véronique. (2013b). Reformulation et récit de voyage. *Travaux de littérature, no XXVI*, pp. 221-30.

Manning Christopher D., Surdeanu Mihai, Bauer John, Finkel Jenny, Bethard Steven J. & Mc-Closky David. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-5010.

Manzotti Emilio. (1999). Spiegazione, riformulazione, correzione, alternativa : sulla semantica di alcuni tipi e segnali di parafrasi. *Parafrasi. Dalla ricerca linguistica alla ricerca psicopedagogica*, L. Lumbelli & B. Mortara Garavelli (éds), Alessandria, *Edizioni dell'Orso*, pp. 169-206.

Martin Robert. (1976). *Inférence, antonymie et paraphrase : éléments pour une théorie sémantique*. Paris: C. Klincksieck.

Martin Louis, Muller Benjamin, Suárez Pedro Javier O., Dupont Yoan, Romary Laurent, Villemonte de La Clergerie Eric, Sagot Benoît & Seddah Djamé. (2020). Les modèles de langue contextuels Camembert pour le français: impact de la taille et de l'hétérogénéité des données d'entraînement. In *JEP-TALN-RECITAL 2020-33ème Journées d'Études sur la Parole, 27ème Conférence sur le Traitement Automatique des Langues Naturelles, 22ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues. ATALA*, pp. 54-65.

Martin Louis, Fan Angela, De la Clergerie Éric, Bordes Antoine & Sagot Benoît. (2022). MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 1651–1664, Marseille, France. European Language Resources Association.

Martinot Claire. (2003). Pour une linguistique de l'acquisition La reformulation : du concept descriptif au concept explicatif. *Langage et société, no 104*, pp. 147-51. https://doi.org/10.3917/ls.104.0147.

Martinot Claire. (2010). Reformulation et acquisition de la complexité linguistique ? *Travaux de linguistique* 2/ n°61.

Martinot Claire, Bošnjak Tomislava Bošnjak, Gerolimich Sonia & Paprocka-Piotrowska Urszula. (2018). *Reformulation et acquisition de la complexité linguistique*. Londres, ISTE Editions.

Mel'čuk, Igor. (1988). Paraphrase et lexique dans la théorie linguistique Sens-Texte. *Lexique et paraphrase*, pp. 13-54.

Meyer Ingrid. (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.

Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg S. & Dean Jeff. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editeurs, *Advances in Neural Information Processing Systems* 26, pp. 3111–3119. Curran Associates, Inc.

Miller George A. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Miller George A., Leacock Claudia, Tengi Randee, Bunker Ross T. (1993). A semantic concordance. In *Human Language Technology: Proceedings of a Workshop* Held at Plainsboro, New Jersey, pp. 21-24.

Minard Anne-Lyse, Ligozat Anne-Laure & Grau Brigitte. (2012). Simplification de phrases pour l'extraction de relations (Sentence Simplification for Relation Extraction) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pp. 1-14. Grenoble, France: ATALA/AFCP. https://www.aclweb.org/anthology/F12-2001.

Mititelu Barbu Verginica, Tufiș Dan & Irimia Elena. (2018). The Reference Corpus of the Contemporary Romanian Language (CoRoLa). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1178-85. Miyazaki, Japon: European Language Resources Association (ELRA).

Mitrofan Maria & Ion Radu. (2017). Adapting the TTL Romanian POS Tagger to the Biomedical Domain. In *Proceedings of the Biomedical NLP Workshop*, pp. 8-14. Varna, Bulgaria.

Mitrofan Maria, Barbu Mititelu Verginica & Mitrofan Grigorina. (2019). MoNERo: A Biomedical Gold Standard Corpus for the Romanian Language. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 71-79. Florence, Italy: Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5008.

Mitrofan Maria & Tufiș Dan. (2018). BioRo: The Biomedical Corpus for the Romanian Language. In *Proceedings of the 11th Edition of the Language Ressources and Evaluation Conference*, pp. 1192-96. Japan.

Mitrofan Maria, Barbu Mititelu Verginica & Mitrofan Grigorin. (2018). Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language. *Data 3(4)*, 53. 12 pages. https://doi.org/10.3390/data3040053.

Mitrofan Maria & Păiș Vasile. (2022). Improving Romanian BioNER Using a Biologically Inspired System. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pp. 316-322, Dublin, Ireland. Association for Computational Linguistics.

McHugh Mary L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3), pp. 276-282.

Mohammad Saif & Pedersen Ted. (2004). Combining Lexical and Syntactic Features for Supervised Word Sense Disambiguation, *CoNLL'04*, pp. 25-32.

Mohan Sunil & Li Donghui. (2019). MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. *ArXiv:1902.09476 [Cs]*, pp. 1-13.

Moore Robert C. (2001). Towards a Simple and Accurate Statistical Approach to Learning Translation Relationships among Words. In *Proceedings of the Workshop on Data-Driven Machine Translation. ACL*.

Moore Robert C. & William Lewis. (2010). Intelligent selection of language model training data. *ACL*, pp. 220–224.

Moschitti Alessandro. (2006). Efficient convolution kernels for dependency and constituent syntactic trees. In *Proceedings of ECML'06*, pp. 318–329.

N

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

Napoles Courtney, Gormley Matt & Van Durme Benjamin. (2012). Annotated gigaword. In *Proceedings of AKBC-WEKEX 2012*.

Narayan Shashi, Gardent Claire, Cohen Shay & Shimorina Anastasia. (2017). Split and rephrase. In *Proceedings of the 22nd Conference on Empirical Methods in Natural Language Processing*, pp. 617-627.

Névéol Aurélie, Grouin Cyril, Leixa Jeremy, Rosset Sophie & Zweigenbaum Pierre. (2014). The Quaero French medical corpus: A ressource for medical entity recognition and normalization. In *Proceedings BioTextM.* Reykjavik, Iceland.

Nguyen-Son Hoang-Quoc, Miyao Yusuke & Echizen Isao. (2015). Paraphrase Detection Based on Identical Phrase and Similar Word Matching. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pp. 504–512. Shanghai, China: Association for Computational Linguistics. http://aclweb.org/anthology/Y/Y15/Y15-1058.

Nie Yixin & Bansal Mohit. (2017). Shortcut-stacked sentence encoders for multi-domain inference. *arXiv preprint arXiv:1708.02312*.

Nisioi Sergiu, Štajner Sanja, Ponzetto Simone Paolo & Dinu Liviu P. (2017). Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (Volume 2: Short Papers), pp. 85–91, Vancouver, Canada. Association for Computational Linguistics.

Nighojkar Animesh & Licato John. (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)*, pp. 7106–7116, Online. Association for Computational Linguistics.

O

Och Franz Josef & Ney Hermann. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics 29 (1)*, pp. 19–51. https://doi.org/10.1162/089120103321337421.

Ohtake Kiyonori & Yamamoto Kazuhide. (2003). Applicability analysis of corpus-derived paraphrases toward example-based paraphrasing. In *Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pp. 380-391.

P

Page Lawrence, Brin Sergey, Motwani Rajeev & Winograd Terry. (1999). The pagerank citation ranking: Bringing order to the web. In *Technical report*, Stanford InfoLab.

Parker Robert, Graff David, Kong Junbo, Chen Ke & Maeda Kazuaki. (2011). English Gigaword Fifth Edition. *Philadelphia: Linguistic Data Consortium*. Web Download.

Pavlick Ellie, Ganitkevitch Juri, Chan Tsz Ping, Yao Xuchen, Van Durme Benjamin & Callison-Burch Chris. (2015). Domain-Specific Paraphrase Extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 57-62. Beijing, China: Association for Computational Linguistics. http://www.aclweb.org/anthology/P15-2010.

Păiș Vasile, Tufiș Dan & Ion Radu. (2019). Integration of Romanian NLP tools into the RELATE platform. In *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR 2019*, pp. 181-192.

Pecout Anaïs, Tran Thi Mai & Grabar Natalia. (2019). Améliorer la diffusion de l'information sur la maladie d'Alzheimer : étude pilote sur la simplification de textes médicaux. Ela. *Études de linguistique appliquée No 195 (3)*, pp. 325-41.

Pennec Blandine. (2006). *La reformulation en anglais contemporain : indices linguistiques et constructions discursives*. Thèse en linguistique, Université Rennes 2. <tel-00199413>, 337 pages.

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

Pennington Jeffrey, Socher Richard & Manning Christopher D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-1543.

Péry-Woodley Marie-Paule & Rebeyrolle Josette. (1998). Domain and genre in sublanguage text: definitional microtexts in three corpora, *LREC*, pp. 987-992.

R

Raffel Colin, Shazeer Noam, Roberts Adam, Lee Katherine, Narang Sharan, Matena Michael, Zhou Yanqi, Li Wei & Liu Peter J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research (JMLR)*, 21(140).

Rajkumar Anupriya & Chitra Rajan T. (2010). Paraphrase Recognition using Neural Network Classification. *International Journal of Computer Applications (0975 - 8887), Volume 1 – No 29*, pp. 43-48.

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

Reiss Katharina & Vermeer Hans J. (2013). Towards a General Theory of Translational Action: Skopos Theory Explained (*C. Nord, Trans.; 1st ed.). Routledge*. https://doi.org/10.4324/9781315759715.

Rey-Debove Josette. (1978). *Le métalangage : Étude linguistique du discours sur le langage*. Le Robert. Collection L'ordre des mots. Paris, 318 pages.

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

Rossari Corinne. (1994). *Les opérations de reformulation : analyse du processus et des marques dans une perspective contrastive français-italien*. 1e édition. Bern/Berlin/Bruxelles/New York/Oxford/Wien, Peter Lang.

Rossari Corinne. (1997). *Les opérations de reformulation : analyse du processus et des marques dans une perspective contrastive français-italien*. 2e édition. Bern/Berlin/Bruxelles/New York/Oxford/Wien, Peter Lang.

Roulet Eddy. (1987). Complétude interactive et connecteurs reformulatifs. *Cahiers de linguistique française* 8, pp. 111-140.

Roulet Eddy, Filliettaz Laurent, Grobet Anne & Burger Marcel. (2001). *Un modèle et un instrument d'analyse de l'organisation du discours*. (Vol. 62). Peter Lang Gmbh, Internationaler Verlag Der Wissenschaften.

S

Saggion Horacio. (2017). Automatic Text Simplification. *Synthesis Lectures on Human Language Technologies 10 (1)*, pp. 1-137. https://doi.org/10.2200/S00700ED1V01Y201602HLT032.

Săpoiu Camelia. (2010). Tipuri de hiperonime în definiţiile lexicografice ale termenilor medicali. Edited by Rodica Zafiu, Adina Dragomirescu and Alexandru Nicolae. *Limba română: controverse, delimitări, noi ipoteze. Actele celui de-al 9-lea Colocviu al Catedrei de Limba Română, I, Section Lexic, semantică, terminologii*. Editura Universității din București, 297 pages.

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

Schmid Helmut. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*, pp. 44-49. Manchester, UK.

Schnedecker Catherine & Landragin Frédéric. (2014). Les chaînes de référence : présentation, *Langages 195, Armand Colin*, Paris, France, pp. 3-22.

Schiffrin Deborah. (1982). Cohesion in Everyday Discourse: The Role of Paraphrase. *Sociolinguistic Working Paper 97*, pp. 3-17.

Schmidhuber Jürgen. (2015). Deep learning in neural networks: An overview. *Neural networks*, vol. 61, pp. 85-117.

Schuwer Martine, Le Bot Michel & Richard Élisabeth (2009). Pragmatique de la reformulation : Types de discours, interactions didactiques. Rennes, *Presses Universitaires de Rennes*, Collection « Rivages linguistiques ».

Sellam Thibault, Das Dipanjan & Parikh Ankur. (2020). BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online. Association for Computational Linguistics.

Severyn Aliaksei & Moschitti Alessandro. (2012). Structural relationships for large-scale learning of answer re-ranking. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pp. 741–750.

Sekizawa Yuuki, Kajiwara Tomoyuki & Komachi Mamoru. (2017). Improving japanese-to-english neural machine translation by paraphrasing the target language. In *Proceedings of the 4th Workshop on Asian Translation*, pp. 64-69.

Shardlow Matthew. (2014). A Survey of Automated Text Simplification. *International Journal of Advanced Computer Science and Applications 4 (1)*. https://doi.org/10.14569/SpecialIssue.2014.040109.

Shirai Satoshi, Yamamoto Kazuhide & Bond Francis. (2001). Japanese-English Paraphrase Corpus. In *Proceedings of the Workshop on Language Resources in Asia, NLPRS* (2001).

Siddharthan Advaith. (2014). A Survey of Research on Text Simplification. *Recent Advances in Automatic Readability Assessment and Text Simplification, ITL - International Journal of Applied Linguistics*, pp. 259–298.

Snover Matthew G., Madnani Nitin, Dorr Bonnie & Schwartz Richard. (2009). TER-Plus: Paraphrase, Semantic, and Alignment Enhancements to Translation Edit Rate. Machine Translation 23 (2), pp. 117-27. https://doi.org/10.1007/s10590-009-9062-9.

Socher Richard, Huang Eric H., Pennington Jeffrey, Ng Andrew Y. & Manning Christopher D. (2011). Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection. In *NIPS'11 Proceedings of the 24th International Conference on Neural Information Processing Systems*, pp. 801-809. Granada, Spain: Association for Computing Machinery.

Spackman Kent A., Campbell Keith E. & Côté Roger. A. (1997). SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA Annual Fall Symposium*, pp. 640-44.

Specia Lucia, Kumar Jauhar Sujay & Mihalcea Rada. (2012). SemEval-2012 task 1: English Lexical Simplification. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics - Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pp. 347–355. SemEval '12. Montréal, Canada: Association for Computational Linguistics.

Steinberger Ralf, Pouliquen Bruno, Widiger Anna, Ignat Camelia, Erjavec Tomaž, Tufiș Dan & Varga Daniel. (2006). The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, pp. 2142-47. Genoa, Italie.

Steuckardt Agnès & Niklas-Salminen Aïno (éds). (2003). Le mot et sa glose. *Langue et langage no 9* : Publications de l'Université de Provence.

Steuckardt Agnès. (2005). Les marqueurs de reformulation formés sur dire. *Dans A. Steuckardt & A. Niklas-Salminen (éds). Les marqueurs de glose*. Aix-en-Provence. Publications de l'Université de Provence, pp. 51-65.

Steuckardt Agnès. (2018). Les marqueurs de reformulation formés sur dire : exploration outillée. *Langages No 212 (4)*, pp. 17-34.

Stevenson Mark, Guo Yinkun, Gaizauskas Robert & Martinez David. (2008). Knowledge sources for word sense disambiguation of biomedical text. In *Proceeding of BioNLP'08*, pp. 80-87.

Straka Milan, Hajič Jan & Straková Jana. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4290-4297. Portorož, Slovenia: European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1680.

T

Taghipour Kaveh & Ng Hwee Tou. (2015). One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, pp. 338-344.

Tchechmedjiev Andon. (2012). État de l'art : mesures de similarité sémantique locales et algorithmes globaux pour la désambiguïsation lexicale à base de connaissances (State of the art : Local Semantic Similarity Measures and Global Algorithmes for Knowledge-based Word Sense Disambiguation) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012*, volume 3: RECITAL, pp. 295–308. Grenoble, France: ATALA/AFCP. https://www.aclweb.org/anthology/F12-3023.

Tchechmedjiev Andon, Abdaoui Amine, Emonet Vincent, Zevio Stella & Jonquet Clement. (2018). SIFR annotator: ontology-based semantic annotation of French biomedical text and clinical notes. *BMC bioinformatics, 19(1)*, 405.

Teston-Bonnard Sandra. (2008). Je veux dire est-il toujours une marque de reformulation ? La Reformulation. Marqueurs linguistiques. *Stratégies énonciatives*, M.-Cl. Le Bot, M. Schuwer & E. Richard (éds), Rennes, Presses Universitaires de Rennes, Collection « Rivages linguistiques », pp. 51-69.

Thierry Grass. (2022). L'erreur n'est pas humaine. *Traduire*, 246 | 2022, pp. 10-23.

Todirascu Amalia, Pado Sebastian, Krisch Jennifer, Kisselew Max & Heid Ulrich. (2012). French and German Corpora for Audience-Based Text Type Classification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pp. 1591–1597. Istanbul, Turkey: European Language Resources Association (ELRA).

Tufiș Dan & Ceaușu Alexandru. (2008). DIAC+: A Professional Diacritics Recovering System. In *Proceedings of the 6th Language Resources and Evaluation Conference*, pp. 1-8. Marrakech, Morocco: ELRA - European Language Resources Association.

Tufiș Dan, Barbu Mititelu Verginica, Irimia Elena, Păiș Vasile, Ion Radu, Diewald Nils, Mitrofan Maria & Onofrei Mihaela. (2019). Little strokes fell great oaks. Creating CoRoLa, the reference corpus of contemporary Romanian. *RRL, LXIV, 3.*, pp. 227-240.

U

Urieli Assaf & Ludovic Tanguy. (2013.) L'apport du faisceau dans l'analyse syntaxique en dépendances par transitions : études de cas avec l'analyseur Talismane. In *(publication en ligne). https://halshs.archives-ouvertes.fr/halshs-00953754.*

V

Vanrullen Tristan. (2003). Vers une analyse syntaxique à granularité variable. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA) 22*, pp. 186-21.

Vapnik Vladimir N. (1995). *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.

Váradi Tamás, Nyéki Bence, Koeva Svetla, Tadić Marko, Štefanec Vanja, Ogrodniczuk Maciej, Nitoń Bartłomiej, Pęzik Piotr, Barbu Mititelu Verginica, Irimia Elena, Mitrofan Maria, Tufiș Dan, Garabík Radovan, Krek Simon & Repar Andraž. (2022). Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 100–108, Marseille, France. European Language Resources Association.

Vargas Élodie. (2008). Un comportement de type céramique, c'est-à-dire cassant : les reformulations intratextuelles dans les émissions de vulgarisation télévisées allemandes. In *Pragmatique de la reformulation. Types de discours - Interactions didactiques*. Sous la direction de Martine SCHUWER. Marie-Claude LE BOT, Elisabeth RICHARD, pp. 21-38. Rennes: Presses Universitaires de Rennes.

Vassiliadou Hélène. (2004). *Les connecteurs « c'est-à-dire (que) » en français et « ðilaði » en grec. Analyse syntaxique et sémantico-pragmatique*. Thèse de doctorat, Université de Strasbourg.

Vassiliadou Hélène. (2008). Quand les voies de la reformulation se croisent pour mieux se séparer : à savoir, autrement dit, c'est-à-dire, en d'autres termes. *La reformulation : marqueurs linguistiques, stratégies énonciatives*, M.-Cl. Le Bot, M. Schuwer & E. Richard (éds), Rennes, Presses Universitaires de Rennes, Collection « Rivages linguistiques », pp. 35-50.

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

Vassiliadou Hélène. (2013b). La formation de c'est-à-dire (que) et de ses correspondants dans les langues romanes : quelques remarques. In *Actes del 26é Congrés de Lingüística i Filologia Romàniques*, E. Casanova Herrero & C. Calvo Rigual (éds), Berlin, Walter de Gruyter, pp. 453-464.

Vassiliadou Hélène & Steuckardt Agnès. (2017). La glose, ses gloses et leurs instruments : que nous apprend le français préclassique ? *Le Français préclassique 19*, pp. 117-141.

Vassiliadou Hélène. (2014). Je veux épouser une Tahitienne, à savoir Maeva vs L'auteur du Lac, c'est-à-dire Lamartine : histoires kleiberiennes de référence. *Res-per-nomen IV : Théories du sens et de la référence*. Hommage à Georges Kleiber, E. Hilgert, S. Palma, R. Daval & P. Frath (éds), Reims, EPURE, pp. 253-268.

Vassiliadou Hélène. (2016). Mouvements de réflexion sur le dire et le dit : c'est-à-dire, autrement dit, ça veut dire. *Histoires de dire. Petit glossaire des marqueurs formés sur le verbe dire*, L. Rouanne & J.-C. Anscombre (éds), Bern/Berlin/Bruxelles/New York/Oxford/Wien, Peter Lang, pp. 339-364.

Vassiliadou Hélène. (2017). C'est-à-dire (que), des emplois propositionnels aux emplois discursifs : poursuite du débat sur les connecteurs et les marqueurs discursifs. *À l'articulation du lexique, de la grammaire et du discours : marqueurs grammaticaux et marqueurs discursifs*, G. Dostie & F. Lefeuvre (éds), Paris, Honoré Champion, pp. 399-414.

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

Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Łukasz & Polosukhin Illia (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Vezin Liliane. (1976). Les paraphrases : étude sémantique, leur rôle dans l'apprentissage. *L'année psychologique 76 (1)*, pp. 177-197.

Vila Marta, Martí Antònia M. & Rodríguez Horacio. (2011). Paraphrase concept and typology. A linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural 46*, pp. 83-90.

Vion Robert. (2006). Reprise et mode d'implication énonciative. *La linguistique 42*, pp. 11-28.

W

Wahle Jan Philip, Ruas Terry, Meuschke Norman & Gipp Bella. (2021). Are neural language models good plagiarists? A benchmark for neural paraphrase detection. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL). IEEE*, pp. 226-229.

Williams Adina, Nangia Nikita & Bowman Samuel R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of NAACL-HLT 2018*, pp. 1112–1122, New Orleans, Louisiana, June 1 - 6, 2018. Association for Computational Linguistics. arXiv preprint arXiv:1704.05426.

X

Xu Wei, Callison-Burch Chris & Dolan William B. (2015). Semeval-2015 task 1: Paraphrase and semantic similarity in twitter (pit). In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 1-11.

Xue Linting, Constant Noah, Roberts Adam, Kale Mihir, Al-Rfou Rami, Siddhant Aditya, Barua Aditya & Raffel Collin. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.

Y

Yang Yinfei, Zhang Yuan, Tar Chris & Baldridge Jason. (2019). PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yuan Dayu, Richardson Julian, Doherty Ryan, Evans Colin & Altendorf Eric. (2016). Semi-supervised Word Sense Disambiguation with Neural Models. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1374–1385. Osaka, Japan: The COLING 2016 Organizing Committee. https://www.aclweb.org/anthology/C16-1130.

Z

Zhang Xingxing & Lapata Mirella. (2017). Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 584–594, Copenhagen, Denmark: Association for Computational Linguistics. doi: 10.18653/v1/D17-1062.

Zhang Yuan, Baldridge Jason & He Luheng. (2019). PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* Volume 1 (Long and Short Papers), pp. 1298-1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhou Zhong, Sperber Matthias & Waibel Alex. (2021). Paraphrases as Foreign Languages in Multilingual Neural Machine Translation. In *Proceedings of 57th Annual Meeting of the Association for Computational Linguistics Student Research Workshop, 2019*, dernière revision en 2021. arXiv:1808.08438.

Zweigenbaum Pierre. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, pp. 1-23.

5. Index des notions

R

S

T

V

6. Annexes

6.1 Guide d'annotation

Nous présentons en détail le guide d'annotation. La **Figure 35** montre un aperçu du document de travail pour l'annotation : un tableur avec différentes colonnes pour annoter le terme reformulé, le marqueur de reformulation, l'indicateur, la reformulation identifiée et les relations lexicales, sémantico-pragmatiques des reformulations médicales identifiées.

Statut	Lignes avec marqueurs + termes médicaux (texte original CLEAR GP)	Terme médical	Marqueur	Indicateur	Reformulation	Relations lexicales	Relations sémantico-pragmatiques
non	Les personnes atteintes de troubles bipolaires et les professionnels de la santé qui les suivent.			trouble/s			
oui	Les troubles de l'humeur dits bipolaires sont un problème de santé mentale fréquent.	Les troubles de l'humeur	dits	trouble/s	bipolaires	synonymie	paraphrase
oui	Le topiramate est un médicament utilisé dans le traitement de l'épilepsie, qui pourrait jouer un rôle dans celui des troubles bipolaires.	Le topiramate	est un/e	trouble/s non<h>; médicament	un médicament utilisé dans le traitement de l'épilepsie, qui pourrait jouer un rôle dans celui des troubles bipolaires	hyperonymie	définition
oui+2		Les troubles de l'humeur	sont un		un problème de santé mentale fréquent	hyperonymie	définition

Figure 35. Exemples d'annotations dans le document de travail

6.1.1 Les phrases qui contiennent des termes médicaux et marqueurs

Les phrases sur lesquelles nous travaillons ont été extraites si elles respectent deux conditions :

- des termes médicaux y sont identifiés avec l'annotateur biomédical SIFR-BioPortal (Tchechmedjiev *et al.*, 2018), présenté dans la **section 2.2 (Partie III)** ;
- des marqueurs de reformulation qui font partie de la liste de marqueurs de la littérature (Péry-Woodley et Rebeyrolle, 1998 ; Charolles et Coltier, 1986 ; Vassiliadou, 2013a ; Grabar et Eshkol-Taravella, 2016a ; Steuckardt, 2018) et

des observations issues de l'analyse des corpus (liste de marqueurs dans la **Partie III**, **sous-chapitre 2.3**).

Ces phrases sont ensuite validées lors de l'annotation manuelle pour vérifier si la présence des termes et des marqueurs dans une phrase implique également une reformulation médicale et si c'est bien le terme annoté qui a été reformulé.

6.1.2 Le terme médical reformulé dans la phrase

Dans la terminologie sélectionnée, nous travaillons aussi bien avec **des termes médicaux simples** qu'avec **des termes polylexicaux** (composés de plusieurs unités lexicales). Plusieurs patrons morphosyntaxiques correspondent aux termes polylexicaux :

- **Le groupe nominal simple** (avec déterminant défini / indéfini / pas de déterminant) : *la bronchectasie ; placebo ;*
- **Le groupe nominal avec expansion** : *un dysfonctionnement systolique du ventricule ;*
- Si le groupe nominal a une phrase relative en expansion, **la relative n'est pas prise en compte dans la structure du terme médical** *: <terme>dialyse péritonéale</terme> ~~<relative>où un cathéter est introduit de manière permanente dans le péritoine</relative>~~ ;*
- **Le groupe verbal** : *fermer le péritoine*.

A	B	C	D	E
Statut	**Lignes avec marqueurs + termes médicaux SIFR**	**Terme médical**	**Marqueur**	**Reformulation**
oui	La grippe est une maladie respiratoire aiguë provoquée par des virus grippaux A et B.	La grippe	est une <maladie>	maladie respiratoire aiguë provoquée par des virus grippaux A et B

Figure 36. Exemple de terme médical

Parfois, le terme annoté automatiquement par l'outil n'est pas celui qui a été reformulé dans la phrase. Dans l'exemple suivant, le terme reformulé « mélasma » n'est pas annoté par l'outil (les termes annotés automatiquement sont balisés avec **<t>** ; **</t>**, les marqueurs sont soulignés par nous), pourtant, il y a une reformulation :

- *<terme>**Le mélasma**</terme> <u>est une</u> <ref>**<t>maladie</t> cutanée psychologiquement stressante**</ref> <u>connue également sous le nom de</u> <ref>**"<t>chloasma<t>" ou "masque de <t>grossesse</t>"**</ref>)*

6.1.3 Le marqueur de reformulation présent dans la phrase

Nous avons annoté automatiquement les marqueurs et les indicateurs de reformulations dans leur variante orthographiée avec et sans accents afin de couvrir également les cas dont les accents manquent dans les textes d'origine.

Nous recherchons les marqueurs de reformulations et les indicateurs de reformulations suivants :

- **marqueurs lexicaux** : *c'est-à-dire, c'est à dire, c'est a dire, ça veut dire, veut dire pour dire autrement, autrement dit, signifie, désigne, ce qu'on appelle, ce que l'on appelle, est aussi appelé, aussi appelé, doit être compris comme, au sens de ;*
- **indicateurs lexicaux** (mots qui aident à identifier les reformulations médicales) : *affection, affections, maladie, maladies, trouble, troubles, par exemple, tel que*, *défini, définie, defini, definie, défini comme, définition.*

A	B	C	D	E
Statut	**Lignes avec marqueurs + termes médicaux SIFR**	**Terme médical**	**Marqueur**	**Reformulation**
oui	La grippe est une maladie respiratoire aiguë provoquée par des virus grippaux A et B.	La grippe	est une <maladie>	maladie respiratoire aiguë provoquée par des virus grippaux A et B

Figure 37. Exemple de marqueur de reformulation annoté

Nous utilisons des **étiquettes spécifiques (ou des balises)** pour l'annotation des marqueurs et indicateurs de reformulations de la manière suivante[86] :

- **est un/une** : pour faire des recherches automatiques plus précises sur le marqueur de reformulation générique « est un/une » nous avons rajouté des mots qui indiquent la présence d'une reformulation médicale, tels que **<maladie>, <trouble>** et **<affection>**, afin de réduire le bruit et le nombre d'identifications erronées ;

  *<terme>**La démence**</terme> est une <ref>**maladie progressive qui affecte principalement les personnes âgées**</ref>.*

- **()** : parenthèses accompagnées ou non d'un autre marqueur de reformulation ;

[86] Dans les exemples donnés (extraits à partir de nos corpus français), les termes médicaux sont marqués en gras et avec les balises **<terme> </terme>**, les marqueurs de reformulation sont **surlignés** et les reformulations médicales sont annotées en gras et avec les balises **<ref> </ref>**.

*<terme>**La dysménorrhée primaire**</terme> correspond à <terme>**des douleurs dont la cause est inconnue**</terme> (c'est-à-dire <ref>**qu'aucune affection médicale n'est identifiée**</ref>).*

- **, (virgule)** : la reformulation est délimitée par une virgule ;

  *<terme>**Les œstrogènes et la testostérone**</terme>, <ref>**des hormones sexuelles stéroïdes**</ref>, affectent un certain nombre de ces facteurs de risque, en particulier le cholestérol et la coagulation, et peuvent être utiles dans les maladies vasculaires périphériques.*

- **: (double point) :** la reformulation est délimitée par de doubles points accompagnés ou non d'un autre marqueur de reformulation ;

  *Les principaux critères de jugement de la revue étaient <terme>**la fausse couche**</terme> : définie comme <ref>**la perte spontanée d'une grossesse avant la viabilité fœtale**</ref>, le décès du bébé (la mortinaissance ou la mortalité néonatale) et la mortalité maternelle.*

- **; (point-virgule)** : la reformulation est délimitée par un point-virgule.

  *Les individus atteints d'une <terme>**fonction rénale diminuée**</terme> (<terme>**maladie rénale chronique**<terme> ; <ref>**MRC**</ref>) subissent des changements dans les niveaux de calcium et de phosphore dans le sang.*

6.1.4 La reformulation du terme médical

Nous délimitons la reformulation en fonction de sa taille dans la structure de la phrase : on annote des mots simples, des syntagmes mais aussi des subordonnées. Dans ce sens, nous définissons plusieurs **types de reformulations** illustrés par des exemples extraits de nos corpus français :

- **nom simple** : *L'amputation en-dessous du genou peut être nécessaire pour les personnes souffrant de l'ischémie aiguë des membres inférieurs causée par une maladie vasculaire au stade avancé ou de <ref>**l'infection**</ref> du pied diabétique (<terme>**sepsis**</terme>), lorsque aucune autre option de traitement n'est possible.*

- **groupe nominal** : *Orthophonie versus <terme>**placebo**</terme> ou <ref>**absence d'intervention**</ref> pour le traitement des troubles de la parole dans la maladie de Parkinson.*
- **groupe nominal avec des expansions (par exemple, une relative) :** *<terme>**La polyarthrite rhumatoïde**</terme> (PR) est une <ref>**maladie inflammatoire chronique et systémique qui affecte principalement les petites articulations des mains et des pieds** </ref>.*

- **groupe verbal** : *<terme>**La dysménorrhée primaire**</terme> correspond à <terme>**des douleurs dont la cause est inconnue**</terme> (c'est-à-dire <ref>**qu'aucune affection médicale n'est identifiée**</ref>).*

- **une phrase subordonnée** : *En cas d'<terme>**indication**</terme> (c'est-à-dire <ref>**si la pulpe dentaire à l'intérieur de la dent meurt**</ref>), un traitement supplémentaire peut s'avérer nécessaire, par exemple un traitement canalaire.*

A	B	C	D	F
Statut	**Lignes avec marqueurs + termes médicaux SIFR**	**Terme médical**	**Marqueur**	**Reformulation**
oui	La grippe est une maladie respiratoire aiguë provoquée par des virus grippaux A et B.	La grippe	est une <maladie>	maladie respiratoire aiguë provoquée par des virus grippaux A et B

Figure 38. Exemple de reformulation annotée

6.1.5 Statut de la reformulation

Dans notre document de travail, le **statut** indique si la phrase contient ou pas une reformulation médicale. Pour définir les différents types de statuts possibles, nous avons créé la liste suivante de valeurs illustrées par des exemples extraits de nos corpus français :

- **non** : la phrase ne contient pas de reformulation ;

  La cryoplastie pour la maladie artérielle périphérique.

- **oui** : la phrase contient des reformulations ;

  *En cas d'<terme>**indication**</terme> (c'est-à-dire <ref>**si la pulpe dentaire à l'intérieur de la dent meurt**</ref>), un traitement supplémentaire peut s'avérer nécessaire, par exemple un traitement canalaire.*

- **oui <inv>** : la phrase contient une reformulation inversée, avec la reformulation à gauche et le terme médical à droite ;

  *Les acides biliaires sont utilisés pour diverses <ref>**maladies chroniques du foie**</ref>, principalement <terme>**la cirrhose biliaire primitive**</terme> **et** <terme>**la cholangite sclérosante primitive**</terme>.*

- **oui <2+>** : 2 ou plusieurs reformulations dans la même phrase ;

  *<terme>**L'eczéma**</terme> (<ref>**dermatite atopique**</ref>) est une <ref>**maladie cutanée très courante et de longue durée due à des facteurs génétiques et environnementaux**</ref> et qui se déclare souvent pendant la petite enfance et l'enfance.*

- **oui <2+> <inv>** : 2 ou plusieurs reformulations dans la même phrase, dont la phrase contient au moins une reformulation inversée ;

  *<ref>**La mauvaise circulation veineuse des jambes installée sur le long terme**</ref> est une <terme>**maladie chronique**</terme> appelée <ref>**insuffisance veineuse chronique** </ref> (<ref>**IVC**</ref>).*

	A	B	C	D	E
1	**Statut**	**Lignes avec marqueurs + termes médicaux SIFR**	**Terme médical**	**Marqueur**	**Reformulation**
48	oui	L'incontinence urinaire (énurésie) nocturne consiste en la perte involontaire d'urine pendant la nuit, non-causée par une maladie organique sous-jacente.	L'incontinence urinaire	()	(énurésie)
49	oui <2+>		L'incontinence urinaire (énurésie)	consiste en	la perte involontaire d'urine pendant la nuit, non-causée par une maladie organique sous-jacente.

Figure 39. Exemple de statut d'une reformulation

6.1.6 Relations et fonctions entre la reformulation et le terme médical

Nous menons des analyses qualitatives **lexicales et sémantico-pragmatiques** sur les reformulations identifiées afin de réaliser une **catégorisation** et une possible **taxinomie de reformulations** et leur utilisation dans les textes médicaux. Pour cela, nous analysons, d'une part, les relations lexicales entre les termes et la reformulation, et, d'autre part, les

fonctions sémantico-pragmatiques des reformulations. Nous nous inspirons des travaux de Săpoiu (2013) et Barbu Mititelu (2011) pour construire notre processus d'annotation de relations lexicales reliant des reformulations médicales et les termes. Les définitions des fonctions sémantico-pragmatiques sont inspirées de la taxinomie d'Eshkol-Taravella et Grabar (2017) et adaptées par nous aux textes médicaux écrits.

1. **Relations lexicales** : montrent le lien lexical entre les deux segments, le terme médical et la reformulation médicale.

   - **synonymie** : la reformulation et le terme sont dans une relation d'équivalence sémantique ;

     *L'amputation en-dessous du genou peut être nécessaire pour les personnes souffrant de l'ischémie aiguë des membres inférieurs causée par une maladie vasculaire au stade avancé ou de <ref>**l'infection**</ref> du pied diabétique (<terme>**sepsis**</terme>), lorsque aucune autre option de traitement n'est possible.*

   - **hyperonymie** : la reformulation est en relation hiérarchique générique avec le terme, dont le terme est un sous-type de la reformulation (et celle-ci est le type générique) ;

     *<terme>**Le typhus des broussailles**</terme> est une <ref>**maladie bactérienne**</ref> prévalente dans les régions de l'Asie et du Pacifique.*

   - **hyponymie :** la reformulation est en relation hiérarchique spécifique avec le terme, dont le terme est plus générique que la reformulation et celle-ci est un sous-type spécifique ;

     *<terme> **Des antibiotiques** </terme> (<ref>**chloramphénicol, tétracycline et doxycycline**</ref>) sont utilisés dans le traitement de cette maladie.*

   - **méronymie** : la reformulation se construit à travers une partie / composante pour expliciter le terme (partie / tout) ;

     *<terme> **La maladie artérielle périphérique** (MAP) < /terme> implique <ref> **une obstruction des grandes artères**< /ref>.*

2. **Fonctions sémantico-pragmatiques** : représentent les raisons pragmatiques qui poussent le locuteur à utiliser la reformulation dans le discours de type scientifique et, respectivement, grand public :

- **définition** : le terme est défini, car il est considéré comme étant trop technique ou spécialisé et donc, difficile à comprendre ;

  <terme> ***La polyarthrite rhumatoïde****</terme> (PR) est une <ref>* ***maladie inflammatoire chronique et systémique qui affecte principalement les petites articulations des mains et des pieds*** *</ref>.*

- **dénomination** : le terme est reformulé à l'aide d'un autre nom (ou terme) ;

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

- **exemplification** : la reformulation est constituée d'exemples qui aident à illustrer le sens du terme à travers plusieurs entités du même type ;

  <terme> ***Des antibiotiques*** *</terme> (<ref>****chloramphénicol, tétracycline et doxycycline****</ref>) sont utilisés dans le traitement de cette maladie.*

- **explication** : le terme est suivi par une situation ou une procédure en particulier et la reformulation donne une explication en apportant des détails en plus ;

  <terme> ***La mort subite cardiaque****</terme> signifie que <ref>****le coeur, puis la circulation s'arrêtent****</ref>.*

- **paraphrase** : le sens du terme est exprimé dans la reformulation avec d'autres mots dans le cadre d'un syntagme avec le but de simplifier le terme, tout en gardant son sens ;

  *Orthophonie versus <terme>****placebo****</terme> ou <ref>****absence d'intervention****</ref> pour le traitement des troubles de la parole dans la maladie de Parkinson.*

6.2 Paramètres de l’architecture bidirectionnelle LSTM

embedding_24 (Embedding)	(None, 155, 300)	5145000
bidirectional_24 (Bidirectional)	(None, 155, 310)	565440
global_max_pooling1d_24 (GlobalMaxPooling1D)	(None, 310)	0
batch_normalization_24 (BatchNormalization)	(None, 310)	1240
dropout_72 (Dropout)	(None, 310)	0
dense_72 (Dense)	(None, 155)	48205
dropout_73 (Dropout)	(None, 155)	0
dense_73 (Dense)	(None, 155)	24180
dropout_74 (Dropout)	(None, 155)	0
dense_74 (Dense)		

6.3 Liste de mots de la langue courante erronément identifiés comme termes médicaux par SIFR-BioPortal

accéléré
actuel
actuellement
affecté par
aggravé
aggravé par
amélioré par
annuel
antérieur à
après
associé à
atténué
atténué par
aucun antécédent de
aucun signe de
avant-pied
avec une durée
avec une période
besoin pour
besoins du patient
bihebdomadaire
causant
causé par
cinq
cinquième
compatible avec
condition durant
confirmation de
confirmé par
conséquence de
contact avec
contrôlé par
d'accord avec
de la prostate
début de
début de la maladie
demande pour
dépendance à
dépendance à une substance
déterminé par
deux côtés
deuxième
deuxième phase de l'accouchement
deuxième trimestre de la grossesse
deuxième vertèbre cervicale
deuxièmement
disponibilité de
disponible
dix
dixième
dont la qualité est
dont le stade est
douze
dû à
durant
durée
en plus de
entrant dans
épisode de
épisodique
exposé
exposé à
exposition à
faible risque de
fréquemment
fréquent
haut risque de
hebdomadaire
huit
huitième
imminent
indépendant de
indication de

inséré dans
introduit par
la configuration
la consistance
la couleur
la forme
la grosseur
la nuit
laisse entrevoir
le jour
le temps
matin
menaçant
mensuel
métastatique de
midi
minuit
neuf
neuvième
non disponible
occasionnel
occasionnellement
onze
partie de
pas de signe de
patient affecté
période définie
période limitée
périodique
périodiquement
plus
plus grand
plus gros
plus petit
possible
premier
premier trimestre de la grossesse
premièrement
probable
provenant de
quatre
quatrième
quotidien
rare
rarement
récemment
récent
résultant de
retardé
retour à
révision de
risque de
risque élevé de
risque modéré de
saisonnier
sans pouls
sans stress
sans-abris
semble indiquer
sept
septième
signe de
simultané
simultanément avec
six
sixième
soir
soulagé par
souvent
suivant
suivi par
traité avec
traité par
traité pour
transmis par
trois
troisième
troisièmement
trouvé dans
un côté
une distance
une distribution
une durée
une évaluation
une évolution
une intensité
une odeur

Les ressources complètes des extraits sont consultables sur le site github

6.4 Termes médicaux reformulés du corpus CLEAR SP *(extrait)*

asthme tussif
5-aminosalicylates
ablation par radiofréquence
absence d'alimentation
absence de traitement
abus sexuel d'enfants et d'adolescents
accident vasculaire cérébral
accidents vasculaires cérébraux
accouchement prématuré
acémétacine
acétate de desmopressine
acétate de zuclopenthixol
achalasie
acide acétylsalicylique
acides gras d'origine marine
acides gras n-3 polyinsaturés
Acides gras oméga 3
acides gras oméga-3 d'origine marine
Acides gras polyinsaturés
acides gras polyinsaturés à longue chaîne
acidose
acné
activités basiques de la vie quotidienne
activités de la vie quotidienne
activités quotidiennes
activités quotidiennes de base
acupuncture
adénocarcinome de l'endomètre
administration précoce d'une solution d'acides aminés
adultes en soins palliatifs
aérobique
affection
affection respiratoire sous-jacente
affection très fréquente
affections
affections chroniques
affections courantes
affections coûteuses et invalidantes
affections cutanées inflammatoires
affections inflammatoires chroniques
affections inflammatoires des voies respiratoires
affections liées au HPV
affections médicales spécifiques
affections neuro-musculaires
affections respiratoires
âge gestationnel
agent
agent oral
agents anti-angiogéniques supplémentaires
agents chimioprotecteurs
agents cytotoxiques
Agents de Santé Communautaires
agents dopaminergiques
agents pathogènes
agitation au réveil
agonistes

agonistes du GABA
agressivité
akathisie
alcoolisme
alimentation entérale
alimentation entérale précoce
alimentation par sonde
alimentation supplémentaire
alprazolam
altérations métaboliques
amblyopie
amblyopie par privation de stimulus
amélioration de la dépression
amélogenèse imparfaite
amélogenèse imparfaite (AI)
amputation majeure unilatérale ou bilatérale de l'extrémité inférieure
analgésie régionale
analgésiques simples
anémie
anémie aplasique grave acquise
anémie due à une carence en fer
angine de poitrine stable
angioplastie transluminale percutanée
antagoniste de la nicotine
Antagonistes des récepteurs de l'adénosine diphosphate
antagonistes des récepteurs de l'angiotensine
antagonistes des récepteurs de l'angiotensine II
antécédents cardiovasculaires
antiagrégants plaquettaires
antibiotiques
antibiotiques oraux
anticholinergiques
anticoagulants
anticorps monoclonaux anti-CD20
antidépresseur
antidépresseurs
antidépresseurs de seconde génération
antidépresseurs tricycliques
Antidépresseurs tricycliques pour troubles du spectre autistique
antihypertenseurs
anti-inflammatoire non-stéroïdien
anti-inflammatoires non stéroïdiens
antipsychotiques
antipsychotiques de deuxième génération
antirhumatismal modificateur de la maladie
antirhumatismaux modificateurs de la maladie
antithrombine
aphasie
appareil vibratoire du corps entier
appendicite compliquée
application intratympanique de gentamicine
apport calorique et nutritionnel
approches chirurgicales
apraxie
apraxie de la parole
aripiprazole
aromathérapie
artère fémorale superficielle
artériopathie cérébrale autosomique dominante avec infarctus sous-corticaux et leucoencéphalopathie
artérite à cellules géantes
artérite à cellules géantes (ACG)
arthrite inflammatoire
arthrite inflammatoire associée à des maladies des tissu conjonctifs
arthroplastie totale de la hanche
arthroplastie totale de la hanche (ATH)
arthroplastie totale du genou
arthrose
arthrose du genou et de la main
aspirine
asthme
asthme et le reflux gastro-oesophagien
asthme sévère
ataxie de Friedreich
atélectasie postopératoire
attrition
audit
audit avec retour d'information
augmentation de la morbidité
augmentation de l'apport énergétique
autisme
autorité parentale
autre intervention non-chirurgicale
autre traitement actif
autres interventions
autres types de schizophrénie
auxiliaires dentaires
AVC
bactériurie significative
baisse de la tolérance au glucose et une hausse de la résistance à l'insuline
baisse de la vision
benzodiazépines
bétahistine
bien-être psychosocial
biomarqueur
biopsie hépatique
bléomycine, d'étoposide et de cisplatine
blépharite
blessure sportive ou due au surmenage
blessures aux tendons
blocage régional ou paracervical
bonne fonction neurologique
bonne vision fonctionnelle
borderline
botulisme
BPCO
bronchectasie
bronchiolite
bronchite aiguë
bronchoconstriction
bronchopneumopathie chronique obstructive
bronchopneumopathie chronique obstructive (BPCO)
broncho-pneumopathie chronique obstructive (BPCO)
bronchopneumopathie obstructive chronique
bursite sous-acromiale/tendinite calcifiante de courts rotateurs de l'épaule
cancer
cancer de l'endomètre
cancer du sein
cancer du sein métastatique
cancer épithélial de l'ovaire
cancer épithélial de l'ovaire (CEO)
cancer gastrique
cancer précoce
cancer rectal non métastatique localement avancé
cancers gynécologiques
candidose buccale
capacités aérobiques maximums
capacités fonctionnelles globales d'une personne
carcinome in situ
carcinomes épidermoïdes oropharyngés associés au papillomavirus humain
cardiomyopathie du péripartum
cardiomyopathie du péripartum (CMPP ou CPMO)
cardiomyopathie hypertrophique
cardiomyopathie hypertrophique (CMH)
cardiopathie congénitale
carie dentaire
caries
caries dentaire
cataplexie
cataracte
catégorie d'attachement de l'enfant
cathétérisme à long terme
cathéters ayant d'autres modifications antimicrobiennes

cathéters veineux centraux (CVC)
causes de maladies respiratoires
cellulite et l'érysipèle
centres de soins sans rendez-vous
centrolobulaire, péricellulaire et périportale
céphalées
certains critères d'évaluation psychologiques
cervicalgie
charge virale indétectable
Chi
chimiothérapie à haute dose
chimiothérapie adjuvante
chimiothérapie postopératoire
chimiothérapie systémique
chirurgie
chirurgie conservatrice
chirurgie en cas de détresse ou d'urgence
chirurgie étendue
chlorpromazine
cholangite sclérosante primitive
cholélithiase
choléra
cholestérol à lipoprotéines de basse densit
cholestérol à lipoprotéines de faible densité
cholestérol à lipoprotéines de haute densité
cinétose
circoncision
cirrhose
classe d'analgésiques
Classification internationale des maladies
claudication intermittente
'client'
clozapine
coagulopathie
colique infantile
coliques infantiles
colite ulcéreuse
colite ulcéreuse (CU)
collaboration en soin de santé pour les troubles mentaux sévères
combinaison de paracétamol et d'orphénadrine
combinaisons d'affections définies qui coexistent fréquemment
communication à sens unique
communication bidirectionnelle
comorbidité
comorbidités
comparateurs actifs inhalés
complications
complications au niveau de la plaie
complications cardiovasculaires
complications iatrogènes
complications obstétriques les plus graves
complications périopératoires
comportement
comportement de marginalisation de l'enfant
comportement d'introversion de l'enfant
comportement opérant
comportement perturbateur
comportement tabagique
comportements répétitifs récurrents
composés d'or
concentration initiale " faible "
condition de contrôle
conditions
conséquences de l'hypersalivation
consommation excessive d'alcool
consommation excessive de sodium alimentaire
constipation
contraceptifs
contrôle
convulsions tonicocloniques
coqueluche
cornée
corticoïdes inhalés
corticoïdes intranasaux
corticoïdes oraux
corticoïdes topiques
corticostéroïdes
corticostéroïdes par voie orale
co-traitement(s)
court terme
créatine
créatinine sérique
crises
Critères de classification et de diagnostic des maladies mentales
critères de jugement cliniques liés aux soins
Critères diagnostiques de recherche
critères secondaires
cross-linking
cryptosporidiose
curcumine
cyclosporine
cytomégalovirus
cytomégalovirus (CMV)
cytoréduction complète
Dans la mucoviscidose (MV), les maladies pulmonaires
danse-thérapie
danshen
de maladie de Parkinson avec démence
Débit de filtration glomérulaire estimé
début de la grossesse
décès du bébé
décompression orbitaire
défauts graves
déficience intellectuelle
déficit cognitif d'origine vasculaire
déficit cognitif léger
déficits ischémiques ultérieurs
dégénérescence maculaire liée à l'âge
déglutition
degré 0
degré 4
délais de récupération
délirium
démence
démence avec corps de Lewy
démence cérébrovasculaire
démence d'Alzheimer
démence d'Alzheimer (MA)
démence de la maladie de Parkinson
démence fronto-temporale
démence mixte
démence vasculaire
démences
démences rares
densité minérale osseuse
dentistes
dents avant inférieures
dépendance à la cocaïne
dépendance aux drogues
dépistage passif
dépression
dépression et l'anxiété
dépression postnatale
dermite séborrhéique
déshydratation
déshydratation par perte d'eau
détartrage et de polissage de routine
deux pro-nucleate
développement
dexibuprofène
diabète
diabète de type 1
diabète de type 2
diabète gestationnel
diacéréine
diagnostic par laparoscopie
dialyse péritonéale
diazépam
didanosine
différence moyenn
différence moyenne
différence moyenne standardisée
différences moyennes standardisées
difficulté d'apprentissage
difficultés à allaiter
difficultés comportementales
difficultés en termes de parole et de voix
dilatation du col

dilemmes éthiques
dimorphique
dispositif intra-utérin
dispositif intra-utérin hormonal
dispositifs médicaux invasifs
dissection aortique
diurétique agissant sur l'anse de Henle
diverticulite
diverticulose
dosage élevé
dosage faible
douleur
douleur abdominale aiguë
douleur associée à l'endométrios
douleur chronique
douleur latérale du coude
douleur neuropathique diabétique
douleur persistante
douleur résultant d'une cause " médicale "
douleurs d'origine professionnelle au niveau du bras, du cou ou de l'épaule
douleurs neuropathiques
drains transtympaniques
drépanocytose
DSM
dysarthrie
dysenterie bacillaire
dysfonctionnement érectile
dysfonctionnement érectile (DE)
dysfonctionnement systolique du ventricule gauche
dysfonctionnements du système nerveux autonome
dysfonctionnements thyroïdiens
dyskinésie
dyskinésie tardive
dyskinésie tardive (DT)
dyslexie
dyslexie développementale
dysménorrhée
dysménorrhée primaire
dyspepsie fonctionnelle
dyspepsie fonctionnelle (DF)
dysphagie
dysphagie oropharyngée
dysphonie spasmodique
dysplasie broncho-pulmonaire
dysplasie broncho-pulmonaire (DBP)
dysplasie développementale de la hanche (DDH) non corrigée
dyssynergie vésico-sphinctérienne (DVS)
échelle CGI-I
échelle de somnolence de l'université Karolinska
échelle d'évaluation unifiée pour la maladie de Parkinson
école du dos
ECR en grappes
eczéma
éducation sur les médicaments
effet secondaire anticholinergique
effets extrapyramidaux
effets indésirables
effets indésirables chez les femmes
effets indésirables chez lesles nourrissons
effets indésirables de la kétamine
effets indésirables des opiacés
effets indésirables mineurs
effets secondaires
effets secondaires anticholinergiques
[…]

6.5 Termes médicaux reformulés du CLEAR GP *(extrait)*

ablation
ablation par radiofréquence
accident vasculaire cérébral
accident vasculaire cérébral (AVC)
accidents vasculaires cérébraux
accidents vasculaires cérébraux (AVC)
accouchement avant terme
acétate de desmopressine
acétylcholine
achalasie
acide aminosalicylique
acide tranexamique
acides gras
Acides gras oméga 3
acides gras polyinsaturés
acné
acouphènes
activités quotidiennes
adénosine diphosphate
affection
affection médicale
affection psychologique
affections
affections à long terme
affections anales
affections auto-immunes
affections chroniques
affections coûteuses et invalidantes
affections du nez et des sinus
affections dues à un excès d'acide urique
affections invalidantes
aortiques et cardiaques
affections médicales
affections musculosquelettique
affections orthopédiques
affections pulmonaires
agent fibrinolytique
agent permettant de disperser le produit
agents anticancéreux modernes
agents antigrippaux
agents dopaminergiques
agents infectieux
agents pathogènes
agonistes de la dopamine
AGPI alimentaires
agranulocytose
AI
akathisie
albuminurie continue
alcaloïdes de l'ergot
alcoolisme
'alimentation par voie intraveineuse
allergies
allopurinol
alvéolite sèche
amblyopie
amblyopie par privation de stimulus (APS)
amélogenèse imparfaite
amélogenèse imparfaite (AI)
aminoadamantanes
AMLA
amoxicilline et la clindamycine
amputées
AMS de type III
amyotrophie spinale
amyotrophie spinale (AMS)
analgésiques
analgésiques administrés par voie orale ou dans une veine
analogues de la gonadolibérine
anémie
anémie drépanocytaire
anesthésiant
anesthésique local

anévrisme
angine
angine de poitrine
angioplastie par ballonnet
angle de Cobb
anomalies coronariennes
anomalies de la peau
anomalies hémostatiques
anomalies sensorielles
anorexie mentale
anorexie mentale (AM)
antagoniste des récepteurs H2
antagonistes de l'aldostérone
antagonistes des opiacés
antagonistes des récepteurs de l'adénosine diphosphate
antagonistes des récepteurs de type 1 de l'angiotensine II
antagonistes muscariniques à longue durée d'action
antécédents de consommation d'amphétamines
antibiotiques
antibiotiques systémiques
anticholinergiques
anticoagulant
anticoagulants
antidépresseurs
antidépresseurs sérotonergiques
anti-douleur
antidouleurs simples
antigrippale
antihistaminiques
anti-inflammatoires non stéroïdiens traditionnelle
antioxydant
antioxydants
anxiété
anxiété ou l'inquiétude
aortite
apnée du sommeil
appareil orthodontique
appellation cardiopathie ischémique
appellation prévention secondaire
appendicite compliquée
appendicite gangreneuse
apraxie de la parole
APs
artémisinine
artère hépatique
artères du coeur
artères terminales
artériopathie cérébrale autosomique dominante avec infarctus sous-corticaux et leucoencéphalopathie
artériopathie oblitérante de l'artère iliaque
artérite à cellules géantes
artérite à cellules géantes (ACG)
arthrite
arthrite inflammatoire
arthrite psoriasique
arthroplastie totale de la hanche
arthroplastie totale de la hanche (ATH)
arthrose
arthrose (ostéoarthrite ou OA)
articulation
articulations
arythmies
aspect spirituel ou religieux
aspects de la santé physique
aspects de la santé psychosociale
aspects de la vie qui constituent la QV
aspirine
assistance respiratoire
assistance respiratoire invasive
asthme
asthme professionnel
astragale
ataxie
ataxie de Friedreich
atélectasie pulmonaire
athérome
athérosclérose
attachement de l'enfant
attachement insécure " évitants "
attachement insécure "résistants"
attelle de contention idéale doit être passive
aucun traitement
autorité parentale
autre intervention
autres dits " négatifs "
autres spondylarthrites
AVC
axillaires
bactéries présentent une résistance à des antibiotiques
bactéries présentent une résistance aux antibiotiques de routine
barotraumatisme
bébés nés avant le terme
bébés nés avant terme
bébés nés avant-terme
bébés nés trop tôt
bénéfices attribués à l'acupuncture
benzodiazépines
bêta2-agonistes à courte durée d'action
bêta-agonistes à action prolongée
bêta-agonistes à action prolongée (BAAP) inhalés
biais
biopsie du foie
blépharite
blessures pénétrantes
blocages nerveux régionaux
bon cholestérol
botulisme
bouche sèche
bourdonnements dans les oreilles
BPCO
BPCO - bronchite chronique
bronchectasie
bronchiolite
bronchite et l'emphysème liés au tabac
bronchodilatateurs
bronchopneumopathie chronique obstructive
broncho-pneumopathie chronique obstructive
bronchopneumopathie chronique obstructive (BPCO)
broncho-pneumopathie chronique obstructive (BPCO)
BZD
caillot sanguin
calcium et de phosphore
cancer
cancer bronchopulmonaire
cancer colorectal
cancer de la bouche
cancer de la prostate
cancer de l'endomètre
cancer de l'ovaire
cancer de l'utérus (utérin/endométrial)
cancer épithélial de l'ovaire
cancer épithélial de l'ovaire (CEO)
cancer systémiques
cancers gynécologiques
Candida
candidose buccale
capacité émotionnelle
capacité fonctionnelle
capacité immunitaire
carbohydrate
cardiomyopathie hypertrophique
cardiomyopathie hypertrophique (CMH)
cardiopathie congénitale
carie dentaire
cataplexie
cataracte
cathéter
cathéter veineux périphériques
cathéter veineux périphériques/canule (CVP)
cathétérisme à long terme
cathétérisme cardiaque
cathéters périphériques
causes iatrogènes
cavités contenant du liquide (ventricules) du cerveau
cellules malignes
cérébrolysine
cerveau
cervicalgie

ceux qui réduisent l'intérêt sexuel par d'autres mécanismes
changements appliqués au mode de vie
changements dans le système immunitaire
changements de mode de vie
changements du mode de vie
chemo-fog
chimiothérapie
chimiothérapie (traitement médicamenteux) en même temps que la radiothérapie
chimiothérapie préopératoire ou néoadjuvante
chimiothérapie systémique
chirurgie
chirurgie de la cataracte
chirurgie de l'obésité
chirurgie de réduction tumorale
chirurgie en cas de détresse ou d'urgence
chirurgie mineure
chirurgie mini-invasive
chirurgie propre
cholangite sclérosante primitive
cholangite sclérosante primitive (CSP)
cholécystectomie laparoscopique
cholélithiase
cholestérol
cholestérol des lipoprotéines de basse densité
chorioamniotite
chrome
chronique
cirrhose
cirrhose biliaire primitiv
cirrhose biliaire primitive
cirrhose du foie
claudication intermittente
claudication intermittente (CI)
clopidogrel
clotiapine
clozapine
CNTF
coagulopathie
codéine
cognition
col
colite ulcéreuse
colite ulcéreuse (CU)
combinaison de traitements antirétroviraux
comorbidité
co-morbidité
compétences
complications
complications à long terme
complications de la grossesse
complications diabétiques
complications impliquant le coeur, les poumons, ou les voies respiratoires
complications liées au VIH
complications microvasculaires du diabète
complications motrices
complications par des maladies sous-jacentes
complications physiques
complications pulmonaires postopératoires
complications rares de la maladie
complications urologiques majeures
complications urologiques majeures (CUM)
comportement sexuel
composants d'origine non végétale
compréhension du langage
congénitales
conisation
conséquences de l'hypersalivation
conséquences neurologiques à vie
consommation de substances
constipation
convulsif
convulsions récurrentes
coqueluche
Cordyceps sinensis
cornée
correction du trouble de réfraction
corticoïdes inhalés
corticostéroïdes
coumarine
courbure de la colonne égale ou supérieure à 10 deg
court terme
COX-2
crampes douloureuses dans les jambes ou les fesses à la marche
crèmes antibiotiques
crises cardiaques
crises convulsives
crises modérées
critère de jugement principal
critères de jugement
critères de jugement importants
critères de jugement importants pour les patients
critères de jugement importants qui étaient rarement, voire jamais, inclus
critères de jugement liés aux patients
critères de jugement nutritionnels
croissance restreinte
cryptosporidiose
curcumine
cyclo-oxygénase 2
cystite bactérienne
cytomégalovirus
cytoréduction complète
cytoréduction optimale
de façon systémique
de syndromes génétiques
de troubles de vascularisation des jambes
décompensation
déconditionnement
DEF
déficience intellectuelle
défocalisation
hypermétropique périphérique
dégénérescence maculaire liée à l'âge
dégénérescence maculaire liée à l'âge (DMLA)
degré 0
degré 4
délires
délirium
démangeaisons généralisées
démence
démence avec corps de Lewy
démence dans la sclérose en plaques
démence de la maladie de Parkinson
démence et les troubles cognitifs
démence fronto-temporale
démence vasculaire
démence vasculaire (DV
démence vasculaire (DVa)
démences
démences communes
démences rares
dengue
dépendance à la cocaïne
dépendance à l'alcool
dépendance à une substance
dépendance aux opiacés
dépistage
dépression
dépression diagnostiquée cliniquement
dépression majeure
dépression post-partum
dérivation urinaire
dermatomyosite et la polymyosite
derrière le tympan
destruction par la chaleur à l'aide radiofréquences
deuxième trimestre
développement cognitif de l'enfant
développement de complications
développement intellectuel
diabète
diabète de type 1
diabète de type 2

diabète gestationnel
diabète gestationnel (DG)
dialyse
diarrhée
diarrhée aiguë
diarrhée persistante
diazépam
différentes composantes des soins
différents types de chimiothérapie
difficultés neuropsychologiques
difficultés psychologiques
discectomie lombaire
dispositifs de libération prolongée de fluor
dispositifs médicaux invasifs
dissection aortique
dissection de l'artère carotide
diurétiques
divers événements indésirables
diverticules
diverticulite
DMLA
doigts et des orteils
donépézil
douleur chronique
douleur neuropathique
douleur résultant d'une cause " médicale "
douleurs abdominales récurrentes
douleurs neuropathiques
douleurs pelviennes chronique
douleurs pelviennes chroniques
drain transtympanique
drépanocytaires
drépanocytose
du foie, du pancréas, du coeur, du rein et du poumon
durée de l'infectiosité
dysarthrie
dysfonctionnement érectile
dysfonctionnement érectile (DE)
dysfonctionnement vestibulaire aigu idiopathique
dyskinésie tardive
dyslexie
dysménorrhée
dyspepsie
dyspepsie fonctionnelle
dyspepsie fonctionnelle (DF)
dysphagie
dysphagie oropharyngée
dysplasie broncho-pulmonaire
dystrophine
dysurie
échec du traitement
échelle d'anxiété de Hamilton
échelle visuelle analogique
échographie bidimensionnelle en mode B
échographie Doppler couleur
éclampsie
écoulement involontaire de salive
eczéma
eczéma atopique
effet indésirable
effets
effets cardio-vasculaires
effets délétères
effets indésirables
effets indésirables mineurs
effets indésirables possibles du phénobarbital pour les femmes
effets potentiellement indésirables des agonistes de la dopamine chez les femmes
effets secondaires
effets secondaires avec la déféripone
effets secondaires de la dicyclomine
effets secondaires digestifs
[…]

6.6 Termes médicaux reformulés du corpus ClassYN SP *(extrait)*

18F-fluoro-désoxy-glucose
accident vasculaire cérébral
accidents périphériques
accidents vasculaires cérébraux
acide désoxyribonucléique
acroparesthésies et des angiokératomes
activité de base de la vie quotidienne
activité des méthyltransférases de l'ADN
activité histone-acétylase
affaiblissement récent global
affaiblissement récent sélectif
affection extra-articulaire
affection longue durée
affection néoplasique
affection oculaire
affections
affections organiques
affections articulaires intermittentes chez l'enfant
affections auto-immunes
affections dermatologiques
affections immunologiques
Affections pulmonaires (sans infection)
affections respiratoires chroniques
agents fongiques
Aides mécaniques
AJI
ALD
alkylation des bases
american college of rheumatology
amitiptyline
amphotéricine
amphotéricine B
amylose AL
amylose AL ou d'une maladie de Randall
amylose cardiaque
anesthésie générale
angéite granulomateuse
allergique
anomalies biologiques
anomalies électrocardiographiques
anomalies structurales et cytoarchitecturales
anomalies structurales pouvant expliquer la crise
antagonistes des récepteurs de l'angiotensine-2
antibiotique dit temps dépendant
antibiotiques
anticitrulline
anticoagulant circulant
anticorps anticytoplasme des polynucléraires
anticorps antinucléaires
anticorps antiphospholipides
anticorps anti-phospholipides
antiépileptiques
antifongiques
antigènes spécifiques de Mycobacterium tuberculosis
anti-phospholipides
antirétroviraux
antithrombine
anti-TNFa
anxiété
apoptose Fas
appareillage
apporter une nouvelle fonction

AR
arguments histologiques
arthralgies intermittentes
arthrites réactionnelles
Association de langue française pour l'étude du diabète et des maladies métaboliques
asthme
asthme , mucoviscidose , ventilation mécanique
AT directes
ATM
attaques de sommeil
atteinte cardiaque
atteinte des muscles lisses
atteinte du nerf optique
atteinte bronchique
atteinte du système nerveux central
atteinte glomérulaire du myélome
atteinte hépatique sévère
atteinte interstitielle
atteinte médullaire
atteinte muqueuse
atteinte musculeuse
atteinte pleurale
atteinte pulmonaire
atteinte rénale
atteinte sous-séreuse
atteintes digestives
atteintes extra-neurologiques
atteintes non artéritiques du nerf optique
atteintes rénales
autisme de Kanner
autoAc anti-GAD
autoAc non spécifiques
autoanticorps
autonomie
autres affections immunologiques
avoir ou non un aspect granulomateux
bacille de Klebs-Loeffler
bactéries
bactéries à prolifération extracellulaire
bactéries intracellulaires
Bartonellose
bases alkylées
biopsie des glandes salivaires accessoires
Bloom
bouffées de chaleur , érythermalgie
BPCO
bradykinésie
bronchopneumopathie chronique obstructive
burn out syndrome
CAIX
cancers
capacité pulmonaire totale
carcinoma associated retinopathy
cardiovasculaires
cataractes syndermatotiques
cathéter « queue de cochon » rotationne
causes courantes de syndrome confusionnel
causes infectieuses
CD53
CD81 , CD82 , CD9
CE cérébrales des molécules d'adhérence
cellule endothéliale
cellules
cellules anormales
cellules avec des détergents
cellules du sang périphérique
cellules immunitaires humaines
Céphalées avec signes de gravité
céphalées d'allure commune
certaines comorbidités
certaines hémopathies malignes
certains virus
chaise à primate
choc hémorragique chirurgical
chorée de Huntington
chorée gravidique
chromosome 16
CINCA
CINCA (chronic , infantile , neurological , cutaneous and articular)
clone cellulaire humain
CMI
coagulation intravasculaire disséminée
coefficient de diffusion apparent
Cognitive Symptoms Inventory
co-infection sévère
comorbidité
comorbidités
compartiments légèrement acides (pH 6 ,2) riches en tétraspanines
complexes
complications cardiaques
complications de la corticothérapie chez les sujets âgés
complications du syndrome métabolique
complications relèvent en effet d'un mécanisme hormonal
complications sensorielles
complications sévères
comportement moteur
comportements d'auto-mutilation
comportements oniriques
composés toxiques
comprimés à libération prolongée
comprimés gastro-résistants
concentration de l'antibiotique est supérieure à la CMI du germe
condensation alvéolaire
connectivite
connectivites
coping
cortex cérébral
corticoïdes
coxite
crises généralisées
crises non épileptiques psychogènes
crises partielles simples
crises répétées
cryoglobulines
cryopyrine
cryptorchidisme
CSI
cutanées
cytokines inflammatoires
cytomegalovirus
Dc
DDD
de liens
défauts de pigmentation
déficit en récepteur de IFN-gamma comme celui en IL- 12 et en récepteur de IL-12
déficit auditif
déficit en récepteur de l'IFN-gamma comme celui en IL- 12 et en récepteur de l'IL-12
déficit neurosensoriel
déficit visuel
dégradation
DEP
dérivation
désacétylation des substrats chromatiniens par certaines histone-désacétylases
diabète de type 2
diminution rapide de la synthèse protéique et l'inhibition de la synthèse protéique
discriminateur
diurétiques
divers métaux
DM1
DNases Activées par la Caspase 3
domaine 5Y-CAP
domaine à glycine
domaine topologique
domaines topologiques
DOPAC
douleurs notamment faciales
dysautonomiques
dysfonctionnement cognitif
dysfonctionnements de systèmes neuronaux complexes

dyskinésies
dystrophie myotonique
dystrophie myotonique de type 1
dystrophie myotonique de type 2
ECG
échelle d'évaluation motrice
échelle de Montréal Cognitive Assesment
échographie et/ou la scintigraphie
effets indésirables
effets indésirables périphériques muscariniques
efflux anagène
efflux télogène
EFR
EH
élastance artérielle
électrophysiologie
élément appelé « C
EMAD
embolie pulmonaire
emphysème pulmonaire
encéphalomyélite auto-immune expérimental
encéphalopathie postérieure réversible
endoscopie
endothéline
entérocolopathies inflammatoires chroniques
entéropathie évolutive
entités ophtalmologiques
enzyme de conversion de l'angiotensine
enzymes
épaississement de interstitium pulmonaire périlobulaire
éphaptique
épidurite , troubles rachidiens statiques
épilepsie
épisode dépressif majeur
étiologies
étiologies infectieuses
évolution chronique
existence d'activités des fibres nociceptives d'origine ectopique
exposition à divers métaux
facteur soluble TIMP-1
facteur Stuart
facteurs de croissance
facteurs de risque
facteurs de risque vasculaires dits « non conventionnels »
facteurs de risques cardiovasculaires
facteurs de transcription
facteurs de transcription dans les neurones striataux de la voie directe
facteurs hormonaux
facteurs rhumatoïdes
facteurs susceptibles d'interférer avec l'alimentation
famille d'antithrombotiques
FE
fibromyalgie
fibrose médiastinale
fissure
FixM/F
fluides biologiques
fonction physiologique
fonctions autonomes
forme « bulbaire »
forme « raide »
forme d'une uvéite granulomateuse
forme galénique interdisant l'écrasement
formes « périphériques »
formes cliniques et neuropathologiques rares
formes de neurolupus
fraction d'éjection ventriculaire gauche
fusion d'images
fusion rigide
gastroentérite à éosinophiles
gène
gêne
gènes
gènes de facteurs anorexigènes
gènes de facteurs orexigènes
gènes liés à biologie neurale
gènes liés notamment à la synthèse
génome viral
germes
GINA
Glasgow coma scale
gonocytes
GR
graisses
graisses insaturées
grande circulation
granulocyte agglutination test
granulome tuberculoïde
granulomes choroïdiens
grosses articulations
groupements méthyl
guérison
HAD
hémogramme
hémopathies lymphoïdes
hémopathies malignes
hémophilie A acquise
hépatite C
hirudine
histone-désacétylases
HNA-2
HTAP
HTLV-1
HVA
hybridation in situ par fluorescence
hydrocéphalie
hyperexcitabilité
hyperosmolarité diabétique
hyperostose vertébrale
hyperostoses primitives
hyperparathyroïdie primitive
hypertension artérielle
hypocinésie
hypoglycémie
hyponatrémie
hypophonie
hypospadias
hypothyroïdie congénitale
hypothyroïdie est la plus profonde
iatrogène
IC-FEP
IgE
imagerie par résonance magnétique
imagerie corrélative
imagerie corrélative , de fusion ou « multi-modalités »
imagerie nucléaire
imagerie radiographies comparatives des hanches de face
immunodépression
immunoglobulines
immunosuppresseurs
index pronostique international
infarctus du myocarde
infection
infection chronique de l'os et/ou des structures adjacentes
infection du matériel de stimulation
infection extra-articulaire
infection opportuniste
infection rétinienne adjacente
infections
infections aspergillaires invasives
infections bactériennes
infections virales
infiltration éosinophile
infliximab
inhibiteur
inhibiteurs
INNTR
Insuffisance médullaire
insuffisance rénale chronique
interleukines
intervenants paramédicaux
interventions pharmaceutiques
intralobulaire
intrusion dans état d'éveil d'aspects physiologiques du sommeil paradoxal
IRC
itraconazole
kilobases
Laboratoire de Cancérologie Expérimentale
lésion démyélinisante extensive
lésion retardée

lésion unique ou multiple de grande taille
lésions préexistantes pulmonaires
lésions cérébrales
lésions chroniques
leucémie myéloïde chronique
liaison transmembranaire
liens
lignées
lipase hormonosensible
liquide de perfusion
liquide-céphalorachidien
liquides de l'organisme
lithium
localisations
lupus érythémateux systémique
lupus-like syndrome
lymphomes agressifs
lymphomes indolents
lymphomes peu différenciés
MAI
maladie
maladie autosomale récessive
maladie de Parkinson
maladie auto-immune
maladie complexe
maladie d'Alzheimer
Maladie d'Hartnup
maladie de Behçet
maladie de Carré
maladie de Chagas
maladie de Crohn et rectocolite hémorragique
maladie de Fabry
maladie de Horton et artérite de Takayasu
maladie de Kawasaki
Maladie de Scheuermann
maladie des chaînes légères
Maladie des chaînes légères et lourdes
Maladie des chaînes lourdes
maladie hyperostosante
maladie inflammatoire
maladie lupique
maladie non auto-immune
maladies
maladies génétiques rares
maladies auto-immunes
maladies cutanée
maladies fréquentes
maladies générales
maladies génétiques
maladies héréditaires
maladies infectieuses
maladies inflammatoires chroniques de l'intestin
maladies mentales
maladies neurodégénératives
Maladies neurologiques
maladies sexuellement transmissibles
maladies systémiques
malnutrition
manière prolongée
manifestations fonctionnelles
manifestations articulaires des entérocolopathies inflammatoires chroniques
manifestations associées aux atteintes d'organes extrathoraciques
manifestations du lupus
manifestations neuropsychiatriques
marqueurs tumoraux
matériel prothétique
mécanismes homéostatiques
médiateurs inflammatoires
médicaments
médicaments à la fois néphro- et ototoxiques
médicaments administrés per os
médicaments anti-infectieux
médicaments indispensables
MEFV
membres inférieurs
méningite bactérienne lymphocytaire
[…]

6.7 Termes médicaux reformulés du corpus ClassYN GP *(extrait)*

abdominaux
accident cardio-vasculaire
accident sportif
accident vasculaire cérébral
accident vasculaire cérébral (ou AVC)
accidents vasculaires cérébraux
acétaminophène
acétylcholine
acétylcholinestérase
acides gras
Acné
acné conglobata
acnés sévères
acromégalie
actes douloureux de courte durée
activité physique
activités
adénocarcinome
adénomectomie
Adénovirus
affection allergique
Affection de longue durée
affection longue durée
affection ORL
affection stomatologique ou ophtalmologique
affections articulaires chroniques
affections chirurgicales
Affections de Longue Durée
affections des vaisseaux sanguins
affections génitale
affections neurologiques et neurodéveloppementales
affections opportunistes
affections pulmonaires
agénésie isolée du corps calleux
aide technique
ALFEDIAM
alglucosidase alpha
allergies
Allocation adulte handicapé
allocation d' éducation de l' enfant handicapé
alopécie
alpha-bloquants
alpha-dystroglycanopathies
alpha-thalassémie
altération de l' état général
amaurose
amiantose
amniocentèse
AMS
AMS de forme C (cérébelleuse)
AMS de forme P (parkinsonienne)
amyotrophie
amyotrophie scapulo-péronière
amyotrophie spinale
amyotrophie spinale de type I
amyotrophie spinale de type III
Amyotrophie spinale distale
amyotrophie spinale juvénile
amyotrophie spinale proximale liée au gène SMN1
amyotrophies spinales
amyotrophies spinales distales
amyotrophies spinales proximales
analogues de la somatostatine
analyses sanguines
anatomopathologiste

androgène
anémie
anémie hémolytique autoimmune
anémie hémolytique auto-immune
anémies « hémolytiques »
anémies hypochromes microcytaire
Anesthésie
anesthésie générale
anesthésie locale
anesthésies régionales
anévrisme
anévrysme congénital , malformations artério-veineuses
angine
Angine accélérée
angine de poitrine
angine dite "blanche"
angine érythémateuse
angiographie
angiographie par scan
angioplastie
Anomalie de position de l’articulation
anomalie génétique
anomalie génétique ou chromosomique
anomalies
anorexie et la boulimie
anorexie mentale
antalgiques
antécédent médical
Antécédent médical ou chirurgical
antiagrégants plaquettaires
anti-agrégants plaquettaires
anti-arythmiques
antibiogramme
antibiotiques
anticholinergiques
anticorps
anticorps anti-CCP
anticorps monoclonal
antihypertenseurs
anti-inflammatoire anti-stéroïdien
anti-inflammatoires dits « stéroïdiens »
anti-inflammatoires non stéroïdiens
antispasmodiques
anxiété
anxiété généralisée ; le trouble panique , la phobie sociale , l' agoraphobie , les phobies simples , l' état de stress post-traumatique et le trouble obsessionnel - compulsif
anxiolytiques
AOS
apex
aplasie médullaire
apnée
apnée du sommeil
apnée obstructive du sommeil
apnées
appareil CPAP
appareil d' imagerie
appareil de radioscopie
appendice
appendicite de l' adulte
apraxie
apraxie idéatoire
apraxie idéomotrice
arrêt cardiaque
arsenic
artère
artères
artères carotidiennes
artériosclérose
artérite inflammatoire
artérite temporale
arthrite
arthrodèse
arthrose
arthrose et l' arthrite
arythmie
arythmies
assassin silencieux
Association Française contre les Myopathies
Association Française du Syndrome d' Evans
asthénie
asthme
asthme paroxystique
ataxie
athéromatose artérielle périphérique
athéromatose artérielle périphérique / une maladie vasculaire athérosclérotique
athérosclérose
atlas
atrophie
atrophie multisystématisée
atrophie multisytématisée
atrophie musculaire péronière
atteinte des muscles respiratoires
atteinte du cervelet
atteinte du nerf périphérique
atteinte musculaire
atteinte neurogène
atteinte oculaire
atteinte primitive bilatérale des surrénales
atteintes cardiaques
atteintes génétiques motoneuronales
autorisation de mise sur le marché
Autosomique
Autosurveillance glycémique
autre type d' hormonothérapie
AVC
AVC hémorragique
AVC ischémique
avoir plus de muscle en masse
bacille , bactérie , virus
bassin perfusion
bêta2-glycoprotéine ou la prothrombine
bicalutamide
bilan d' extension du cancer
bilan de la maladie
bilan local
binoculaire
biopsie
biopsie chirurgicale diagnostique
biopsie de la prostate
biopsie musculaire
biothérapies
bisphosphonate
BPCO
bradycardie
bronche
bronches
Broncho-Pneumopathie Chronique Obstructive
Broncho-pneumopathie chronique obstructive (BPCO)
buflomédil
cæcum
caméra à scintillation
Cancer
cancer de l' ovaire
Cancer différencié
cancer du poumon
cancer du sein
cancers
Cancers du larynx et du système digestif
cancers urologiques
candidose
cardiomyopathie
Cardiomyopathie Dilatée
Cardiomyopathie ventriculaire droite arythmogène
Cardiomyopathies
carotidienne
cas isolé
cataractes
cathéter à site d' injection implantable
cathéter-ballon
cathétérisme des sinus pétreux
cathéters
cause rare mais grave
causes de myopathie
cautérisation
cavernome cérébral
cécité légale
cellules anormales
cellules immunitaires
cellules inflammatoires
cellules sanguines
cellulite
cellulite des hanches
centre de la cellule
céphalées
Céphalées liées à une anomalie du métabolisme
cerveau

cervicale
ces dents de sagesse restent incluses
Cette cellulite
Cette masse
chancre
Chimiosensible
chimiosensibles
Chimiothérapeute
chimiothérapie
chimiothérapie avant la chirurgie
chimiothérapie complémentaire
chirurgie
chirurgie (ou prostatectomie totale)
chirurgie de réévaluation
chirurgie du cancer du rectum
chirurgie locale
chlamydiae
chlamydiose
choc mineur
cholécystectomie par laparotomie
Cholestérol
chondrocalcinose
chorée
chorée de Sydenhan
choriocentèse
chromosomes sexuels
chronique
chyme
circulation collatérale
cirrhose
claudication
claudication des mâchoires
claudication intermittente
CMT liée à l'X
coagulation
cœlioscopie
cœlioscopie/célioscopie
Cœur pulmonaire
col de l' utérus
col du fémur
coliques néphrétiques
colon
côlon
coloscopie
colostomie
colostomie , stomie de protection ou anus artificiel
colostomie ou anus artificiel
Complexe de Carney
complication
complications cardio-vasculaires
Complications de la cataracte
complications de la tuberculose pulmonaire
complications oculaires
complications rythmiques
composé radioactif
conductrices
condyle occipital
condylomes
congénital
conjonctives
consommation excessive d' alcool
constipation
consultations d' hématologie
contractions utérines
contractures
corticodépendance
corticoïdes
corticothérapie
coxibs
crépitants
crise cardiaque
crise généralisée tonico-clonique
crises cardiaques
crises généralisées
crises partielles
cryothérapie
curage axillaire
curage ganglionnaire/curage des ganglions
curiethérapie
curiethérapie de la prostate
curiethérapie par implants temporaires
curiethérapie utérovaginale
cyanose
cyclophosphamide
cyphose
cystite
cystites
cytologie
cytomégalovirus
cytoponction
Daltoniens
daltonisme
dardarine
De troubles ostéo-articulaires
déficit intellectuel
déficit neurologique
déficits en protéine C ou S , en antithrombine III et l' hyperhomocystéinémie familiale
déficits neurologiques permanents
déformation du squelette
déformations
déformations des pieds
dégénérescence
Dégénérescence maculaire liée à l'âge
déminéralisation
dénomination « idiopathique »
dépistage
dépression
dermatologue
dermite herpétiforme
déséquilibre hydrique
diabète
diabète de type 1
diabète de type 2
diabète insipide central
diabète insipide néphrogénique
diabète insipide neurogène
diabète non insulino-dépendant
Diagnostic
diarrhée liquide en jet
diastole
difficultés à uriner
difficultés scolaires
diplopie
DMLA
douleur
douleurs au niveau de l’anus et du rectum
douloureux brûlement urinaire
doxorubicine
DPI
dysarthrie
dysfonctionnement urinaire ou érectile
dysgénésie du corps calleux ou agénésie calleuse
dysmorphie
dysphagie intermittente et variable
dysplasie
Dysplasie ventriculaire droite arythmogène
dyspnée
dyspnée d'effort
dystrophie des ceintures de type C
dystrophies musculaires congénitales
dysurie
ecchymoses
échelle
échographie endorectale
échographie ou sonographie
éclampsie
édicaments antalgiques
Education thérapeutique
éducation thérapeutique
effet est suspensif
Effets indésirables neurologiques, psychologiques et comportementaux
Effets indésirables urinaires et sexuels
effets secondaires
effets secondaires de l' aspirine au niveau digestif
effets secondaires de la ménopause
effets secondaires importants
effets secondaires sur la sexualité
électrocardiogramme
électromyogramme
électrophorèse des protéines
embolie pulmonaire
embolisation
embryopathie
emphysème
emphysème sévère

endocardite
endomètre
endométriose
endométrite
endoprothèse
endoprothèse (stent)
endoscope
Entente préalable
entérocolopathies inflammatoires
entorse ou une foulure
environnement psychologique
enzyme
enzyme recombinante
enzymes musculaires
épidémiologie
épilepsie
épiploon
épirubicine
épisodes neurologiques déficitaires transitoires
équilibre flexion/extension
éradication
Ergothérapeute
éructations
érythémato-pultacée
érythème
érythème cutané
espace métatarsien
essai thérapeutique
estrogènes
être à jeun
être absorbées dans l'intestin grêle
Etre séropositif pour le VIH
étude épidémiologique
études de la population
études scientifiques
examen
examen anatomopathologique
examen anatomo-pathologique
examen clinique
examen d' imagerie
examen de dépistage
examen de prévention et des soins nécessaires
examen extemporané
examen IRM
examen sanguin simple
examen standard
examens de médecine nucléaire
examens radiologiques
exercices de renforcement musculaire
exérèse
expression « chaude-pisse »
extrasystoles ventriculaires
facteur de risque
facteurs « psychosociaux »
facteurs aggravants de la maladie asthmatique
Facteurs de risque
faiblesse musculaire
faiblesse musculaire distale
fatigue
fausses-routes alimentaires
fécalome
Fédération française de pneumologie (FFP)
fer
fibrodysplasie ossifiante progressive
fibrodysplasie ossifiante progressive (FOP)
fibrome , malformation , synéchie , hypoplasie , béance du col
fibromyalgie
fibrose
fibrose kystique
filament fin
fistule
fistules
fistulisation
fluoro-uracile
fonction de reproduction des ovaires
fonction motrice
fonctions essentielles
fonctions vitales
formation histologique
forme extrêmement rare du SAPL
forme familiale
forme hépato-musculaire
forme non héréditaire ou sporadique
formes autosomiques dominantes
formes d' amyotrophies spinales
[…]

6.8 Termes médicaux reformulés du corpus GrandMed-Ro *(extrait)*

abdomenul de lemn
abilitatile lor de gandire
ablatie prin cateter
acanthosis nigricans
accident vascular cerebral
accidentare majora
accidentari
accidente vasculare cerebrale si ale vaselor periferice
Accidentele rutiere
Accidentul cerebral
Accidentul vascular cerebral
Accuzide Forte
acetaldehida
acetaminofen sau ibuprofen
acetaminofenul sau ibuprofenul
acid alfa-linolenic
aciditatea
acizi
acizi toxici
acizilor grasi cu lant scurt
Acneea
Acneea vulgara
Acromegalia
activitati normale
activitatile fizice
acumulare de trigliceride
Acumularile de drusen
adenoame
Adenom de prostata
Adenomioza
ADHD
ADN-ul
adrenalina
Afazia
Afazia optica
Afazia progresiva primara
afectiune
afectiune a pielii
afectiune ale stomacului
afectiune cronica
afectiune medicala
afectiune medicala pe termen lung
afectiunea raynaud
afectiuni
afectiuni acute
afectiuni ale inimii
afectiuni ale pielii
afectiuni asociate
afectiuni asociate cu afazia
afectiuni asociate cu obezitatea
afectiuni autoimune
afectiuni benigne ale pielii
afectiuni cardiace
afectiuni cardiace congenitale
afectiuni, care ar putea contribui la bradicardie
afectiuni coloanei vertebrale
afectiuni congenitale
afectiuni cronice

artrita reumatoida, lupusul eritematos sistemic sau sclerodermia
artrita reumatoida, lupusul si sclerodermia
artrocenteza
artroplastie
Artroza
ascita
aspartamul
aspartat aminotransferaza
aspirina
aspirina sau codeina
aspirina sau warfarina
aspirina, acetaminofenul sau ibuprofenul
aspirina, ibuprofen
aspirina, ibuprofen, naproxen
aspirina, ibuprofenul si naproxenul sodic
aspirina, naprosinul si ibuprofenul
Astigmatismul
astm
astm bronsic alergic
astmul
astmul alergic
Astmul bronsic
astrocite
atac de cord
Atac de panica
Atacul de cord
Atacul de panica
Atacurile de panica
Atat hipoglicemia, cat si hiperglicemia
Ataxia Friedreich
Ataxie
ateroscleroza
atrii
Atrofie geografica
Autismul
Autismul nonverbal
Autismul sau tulburarea de spectru autistTSA)
autozomal recesive
azotemia
azotul de uree si creatinina.
Bacillus anthracis
bacilul Koch
bacteria Borrelia burgdorferi
bacteria Mycobacterium Tuberculosis
bacteriemie
bacterii „rele”
Balneoterapia
balsamuri pentru constipatie
banda iliotibiala
batai pe minut
benefic circulatiei sanguine
beneficii antioxidante
benigna
beta-blocante
Bifidobacteria si Lactobacillus
bifosfonati
bilateral
biopsie
Biotina
blefarita
Blefarospasmul
Blocantii histaminici
Boala
boala Addison
Boala Alexander
boala ALS
boala Alzheimer
boala arteriala coronariana si artimia cardiaca
boala arteriala periferica
boala artrozica
boala autoimuna
Boala Basedow Graves
Boala Basedow-Graves
Boala cardiaca ischemica
boala celiaca
boala coronariana
Boala Creutzfeldt-Jakob
Boala Creuzefeldt - Jakobs
Boala Crohn
boala cronica
boala cu transmitere sexuala
boala de baza
boala de reflux gastroesofagian
boala de reflux gastro-esofagian
Boala Fabry
boala gastrointestinala
Boala gingivala
boala grava
Boala Graves
Boala Graves Basedow
Boala Graves-Basedow
Boala Hansen
boala hepatica sau renala
boala Hirschsprung
boala inflamatorie
Boala inflamatorie a intestinului
boala inflamatorie intestinala
Boala inflamatorie pelvina
Boala Kawasaki
Boala mana-gura-picior
Boala Meniere
boala multifactoriala
Boala nodului sinusal
Boala Paget
boala Parkinson
boala Parkinson, boala Huntington si boala Alzheimer
boala pe termen lung
Boala Peyronie
Boala polichistica a ficatului
Boala polichistica hepatica
Boala polichistica la ficat
Boala polichistica renala
boala progresiva
boala pulmonara obstructiva cronica
boala pulmonara obstructiva cronica avansata
boala respiratorie
boala reziduala
boala serului
Boala Sever
Boala Still a adultului
boala varicoasa
boala vasculara periferica
boala von Willebrand
boala Willis-Ekbom
Boala Wilson
Boalamana-gura-picior
Boli
boli ale creierului
Boli ale plamanilor
boli ale sistemului imunitar
boli ale tesutului conjunctiv
boli autoimune
boli cardiace
boli cardiovasculare
boli cerebrale degenerative
boli cerebrovasculare
boli cronice
boli cronice renale
boli cu transmitere sexuala
boli de inima
boli de inima, bronsita cronica, emfizem si astm
boli de piele
boli de piele severe
boli de reflux gastro-esofagian
boli degenerative
boli grave
boli inflamatorii intestinale
boli inflamatorii intestinaleBII)
boli metabolice
boli neurodegenerative
boli neurologice
boli psihice
boli pulmonare subiacente
boli rare
boli respiratorii
Boli terminale
Boli tiroidiene
bolii inflamatorii intestinale
Bolile autoimune
bolile cardiace
Bolile cardiovasculare
bolile cardiovasculare, diabetul de tip 2 si cancerul
bolile cronice
bolile cronice la adulti
Bolile cu transmitere sexuala
Bolile de inima
Bolile digestive
Bolile endocrine
bolile inflamatorii intestinale
Bolile inflamatorii intestinale, sindroamele ulceroase
Bolile psihice
bolile respiratorii
Bolile reumatice
bolile reumatismale
bolile vaselor
bolilor cu transmitere sexuala
bolilor diareice
bolilor inflamatorii intestinale
bolilor intestinale inflamatorii

Botulismul
BPOC
Bradicardia
bradicardie
bronhiilor
Bronho-pneumopatia cronica obstructiva
Bronhopneumopatia obstructiva cronica
Bronho-pneumopatia obstructiva cronica
bronhospasm
Bronsita
Bronsita acuta
Bronsita cronica
bufeurile si perioadele neregulate
Bulimia nervoasa
bursei
bursita
Bursita genunchiului
bypass
bypass gastric
Bypass-ul aorto-coronarian
CA125
CA-125
caile biliare
calciferol
Calcifilaxia
Calculii renali
calculii simptomatici
calculilor
calitati chemopreventive
calmante
canal deferent
canale salivare
Cancer
cancer „tacut”
Cancer de colon ereditar nonpolipozic
cancer de testicule
cancer oral
cancer ovarian
cancere avansate
cancerul
Cancerul colorectal
cancerul de san
Cancerul in faza terminala
Cancerul la gura
Cancerul ovarian
Cancerul pulmonar
Cancerul renal
Cancerul testicular
Candidoza bucala
Cannabisul
capilare
capsula
capsulita adeziva
Carbohidratii
cardiomiopatia hipertrofica
cardiomiopatie
cardiomiopatie de stres
Cardiopatia ischemica
cardioversie
carotenoide
Carotenoizi
Cataracta
cataracta, degenerescenta maculara legata de varsta, pterigionul
cataracta, glaucomul sau dezlipire de retina
cateter
cateter Swan-Ganz
cauze mai grave pentru paralizia faciala
cauze stomatologice
cauzele subiacente periculoase
cauzelor insuficientei renale
cavitatea pleurala
cavitatea vitreana
cavitati in plamani
Cearcane
cefaleea
Cefaleea migrenoasa
Cefaleea tensionala
cele mai dificile aspecte ale agenezei vaginale
Cele trei tipuri de glande salivare majore
celule canceroase
celule CD4 +
celule gigante
celule granuloase
celulele albe
Celulele CD4
celulele crestei neurale
celulele maligne
celulele mezoteliale
celulelor
celulelor stem
cerumen
cerumenului
Cervicita cronica
Cetoacidoza
Child-Pugh
Chirurgia de sterilizare pentru femei
Chirurgia deschisa
[…]

6.9 Paires de terme médical – reformulation : CLEAR SP *(extrait)*

Terme médical [TAB] Reformulation médicale

5-aminosalicylates (5-ASA)

a dépression et l'anxiété , affectent jusqu'à 15 pour cent de la population du Royaume-Uni

a réduction de la quantité de liquide, calculée en ml (différence du volume LO du bras affecté avant et après traitement)

accident vasculaire cérébral (AVC)

acides gras d'origine marine (huile de poisson)

Acides gras oméga 3 (huile de poisson)

acides gras oméga-3 d'origine marine (huile de poisson)

Acides gras polyinsaturés (AGPI)

acides gras polyinsaturés à longue chaîne (AGPI-LC)

activités de la vie quotidienne (AVQ)

aérobique (requérant de l'oxygène)

affection céphalées

affections (l'insuffisance veineuse chronique, le lymphoedème et les hémorroïdes)

affections la douleur chronique

affections une anémie hémolytique, une ataxie cérébelleuse et des difficultés cognitives

affections don’t le cancer, les troubles arthritiques et les maladies cardiovasculaires

affections chroniques : l'asthme, le trouble déficitaire de l'attention avec hyperactivité, l'insuffisance veineuse chronique, le diabète, la dysfonction érectile, l'hypertension et l'arthrose

affections chroniques notamment les douleurs neuropathiques

affections cutanées inflammatoires affection/s ; telles que l'atopie et la séborrhée

affections inflammatoires chroniques telles que l'arthrite rhumatoïde, l'oesophagite et l'inflammation post-opératoire

affections liées au HPV les verrues génitales et la néoplasie intra-épithéliale cervicale (CIN)

agent (par exemple par exemple le carboplatine)

agents chimioprotecteurs (tel que acétylcystéine, amifostine, calcium et magnésium, diéthyldithiocarbamate, glutathion, ORG 2766, oxcarbazépine, acide rétinoïque ou vitamine E)

agents cytotoxiques notamment le cisplatine, le méthotrexate, la doxorubicine et la vinblastine

agonistes (par exemple buprénorphine, méthadone, naltréxone)

altérations métaboliques (ostéoporose, induction ou inhibition des enzymes hépatiques)

Antagonistes des récepteurs de l'adénosine diphosphate (ADP)

antagonistes des récepteurs de l'adénosine diphosphate (ADP)

antibiotiques (traitements)

anticoagulants tel que la warfarine ou l'héparine de bas poids moléculaire (HBPM)

anticoagulants la warfarine ou l'héparine de bas poids moléculaire

anticorps monoclonaux anti-CD20 (le rituximab, l'ofatumumab et le GA101)

antidépresseur (notamment comme des agents non conventionnels, comme l'hypéricum)

antidépresseurs de seconde génération (ASG)

antidépresseurs tricycliques (ATC)

Antidépresseurs tricycliques pour troubles du spectre autistique (TSA)

anti-inflammatoire non-stéroïdien (AINS)

antipsychotiques , les benzodiazépines ou anti-histaminiques

antirhumatismal modificateur de la maladie (ARMM)

antirhumatismaux modificateurs de la maladie (ARMM)

appareil vibratoire du corps entier (EAV)

appareil vibratoire du corps entier comme l'utilisation d'une plate-forme oscillante verticale ou rotative produisant un stimulus physiqu

appendicite compliquée définie comme une appendicite gangreneuse ou perforée

apport calorique et nutritionnel (total issu des suppléments caloriques protéiques oraux et de l'alimentation)

approches chirurgicales notamment la néphro-urétérectomie radicale ouverte et les procédures laparoscopiques

arthrite inflammatoire (polyarthrite rhumatoïde, spondylarthrite ankylosante, arthrite psoriasique, autres spondylarthrites ou arthrite inflammatoire associée à des maladies des tissu conjonctifs)

arthrite inflammatoire associée à des maladies des tissu conjonctifs (comme le lupus érythémateux disséminé)

augmentation de l'apport énergétique y compris l'augmentation de la densité énergétique de l'alimentation au lait et/ou du volume des fluides

autres interventions c'est-à-dire des exercices de respiration ou l'exercice physique

autres types de schizophrénie comme la psychose

auxiliaires dentaires soins bucco-dentaires

bléomycine, d'étoposide et de cisplatine (BEP)

blocage régional ou paracervical (BPC)

bonne vision fonctionnelle (l'acuité visuelle de 6/12 ou plus)

bonne vision fonctionnelle (acuité visuelle de 6/12 ou plus)

borderline (trouble de la personnalité limite)

BPCO (les maladies respiratoires)

bronchectasie maladie respiratoire

bronchectasie ou maladie respiratoire

bronchoconstriction est un/e considérée comme un marqueur de l'inflammation des voies respiratoires

bronchopneumopathie chronique obstructive (BPCO)

bronchopneumopathie chronique obstructive (BPCO)

bronchopneumopathie chronique obstructive la maladie

bronchopneumopathie chronique obstructive (BPCO)

bronchopneumopathie obstructive chronique (BPCO)

cancer (le cancer du sein)

cancer tel que le cancer colorectal métastatique

cancer du sein ou d'autre maladie grave

carcinome in situ cancer précoce superficiellement invasif ou cancer superficiel T-1m T1-a

cathéters ayant d'autres modifications antimicrobiennes par exemple les pansements antiseptiques, embases (hub), cathéters tunnélisés, connecteurs intraveineux sans aiguille ou verrous antibiotiques

causes de maladies respiratoires les causes virales et bactériennes, l'aspiration secondaire du contenu gastro-intestinal et les pathologies pulmonaires prédisposantes

centres de soins sans rendez-vous (walk-in clinics)

centrolobulaire, péricellulaire et périportale : trois types de fibrose cicatricielle le plus couramment observés dans la maladie alcoolique du foie

certains critères d'évaluation psychologiques en particulier, l'anxiété, la dépression et les troubles de l'humeur

Chi (le concept traditionnel chinois traduit par la force vitale ou l'énergie)

cholestérol à lipoprotéines de faible densité (LDL)

cirrhose maladie maladie hépatique

claudication intermittente se caractérise par des douleurs dans les jambes ou les fesses pendant l'effort, disparaissant au repos.

communication à sens unique par exemple des messages textuels, des posters ou des programmes radio

communication bidirectionnelle par exemple discussions face à face entre les parents et les professionnels de la santé

comorbidité des combinaisons d'affections courantes prédéfinies

comorbidité par exemple l'hypertension et une maladie cardio-vasculaire

comorbidités tel que des maladies gastro-intestinales ou hépatiques

comparateurs actifs inhalés par exemple solution saline hypertonique ou dornase alfa

complications (risque de nausées et de vomissements postopératoires (NVPO), admission ou réadmission à l'hôpital, troubles comportementaux postopératoires et complications respiratoires et cardiovasculaires périopératoires)

complications au niveau de la plaie (c'est-à-dire une infection, une cicatrice hypertrophique et une sensibilité de la cicatrice)

complications au niveau de la plaie (c'est-à-dire une infection, une cicatrice hypertrophique et une sensibilité de la cicatrice)

complications obstétriques les plus graves , une pré-éclampsie et une restriction de la croissance intra-utérine

conditions tel que la maladie inflammatoire pelvienne et l'endométriose

contrôle (placebo ou absence de traitement)

corticoïdes topiques (intranasaux)

corticostéroïdes par voie orale (CSO)

co-traitement(s) (par exemple la ribavirine)

crises appelées des douleurs musculaires et osseuses

Critères diagnostiques de recherche (RDC)

d'âge gestationnel (AG)

d'agitation au réveil ou, délire d'émergence

d'alimentation par sonde (gastrique, duodénale ou jéjunale)

Dans la mucoviscidose (MV), les maladies pulmonaires sont caractérisées par l'altération de la clairance mucociliaire

de maladie de Parkinson avec démence (DMP)

Débit de filtration glomérulaire estimé (DFGe)

début de la grossesse (c'est-à-dire moins de 24 semaines de grossesse)

défauts graves (par exemple fort taux d'abandon)

déficit cognitif d'origine vasculaire l'expression des troubles cognitifs importants sans perte de mémoire

déficit cognitif d'origine vasculaire (DCOV)

déficit cognitif léger (DCL)

déficit cognitif léger (DCL)

déficit cognitif léger (DCL)

déficit cognitif léger (DCL)

degré 0 (pas de douleurs ni de signes physiques)

degré 4 (fracture ou luxation)

démence défini/e ; défini comme manifestations d'agressivité, d'agitation ou de psychose

démence troubles cognitifs

démence ou d'autres troubles cognitifs

démence la malade

démence avec corps de Lewy (DCL)

démence cérébrovasculaire (DCV)

démence de la maladie de Parkinson (DMP)

démence mixte est qualifiée de la combinaison de la démence vasculaire et de la maladie d'Alzheimer

démence mixte nom de la démence vasculaire et la maladie d'Alzheimer sont couramment combinées

démences comme la démence dans la maladie de Parkinson

dentistes par exemple soins bucco-dentaires

dents avant inférieures (canines inférieur et incisives inférieures)

dépistage passif (la plupart des patients sont identifiés lors de visites dans des établissements de soins)

dépression couvre dépression majeure, de troubles d'adaptation et de dysthymie

des acides gras n-3 polyinsaturés (également nommés acides gras oméga-3)

des activités basiques de la vie quotidienne (ADLVQ)

des affections y compris l'entérocolite nécrosante et de rétinopathie du prématuré

des affections courantes l'otite moyenne

des affections coûteuses et invalidantes (comme les maladies cardiovasculaires, le cancer, le diabète et les maladies respiratoires chroniques)

des affections inflammatoires des voies respiratoires telles que l'asthme

des affections médicales spécifiques (hémiplégie ou fracture du col du fémur)

des affections respiratoires la BPCO

Des agents dopaminergiques tels que la bromocriptine et la lévodopa,

des agents pathogènes Clostridium difficile

des agonistes du GABA (baclofène, GABA gamma-vinyle, GABA gamma-acétylénique, progabide, muscimol, valproate de sodium et tétrahydroisoxazolopyridine (THIP))

des altérations métaboliques (ostéoporose, induction ou inhibition des enzymes hépatiques)

des antécédents cardiovasculaires (infarctus du myocarde, AVC, maladie artérielle périphérique occlusive ou angine de poitrine)

des anticholinergiques (benzhexol, benzatropine, bipéridène ou biperiden ou d'orphénadrine ou procyclidine ou la scopolamine ou trihexylphenidyl)

des antipsychotiques de deuxième génération , maladie la zotépine et l'amisulpride

des combinaisons d'affections définies qui coexistent fréquemment ,commme le diabète et la dépression

des combinaisons d'affections définies qui coexistent fréquemment comme le diabète et la dépression

des complications cardiovasculaires tels que des effets indésirables à plus long terme

des complications iatrogènes (par exemple un cathétérisme central)

des complications iatrogènes (par exemple un cathétérisme central)

des comportements répétitifs récurrents tels qu' une préoccupation démesurée pour l'hygiène des mains, le contrôle et le perfectionnisme

des Critères de classification et de diagnostic des maladies mentales CCMD-3-R

des difficultés comportementales telles que l'agressivité ou des comportements d'automutilation, une instabilité émotionnelle ainsi que des problèmes liés à l'anxiété

des effets extrapyramidaux (des mouvements involontaires, la maladie de Parkinson et l'acathésie)

des effets indésirables (des troubles du mouvement induits par les médicaments)

des effets indésirables (par exemple lésions nerveuses et vasculaires, infection profonde ou superficielle)

des effets indésirables des diarrhées, des vomissements, une diminution de l'appétit, une intolérance au glucose, des ballonnements et des problèmes de comportement alimentair

des effets indésirables principalement des troubles gastro-intestinaux et des bouffées vasomotrices

des effets indésirables notamment des bouffées vasomotrices et des troubles gastro-intestinaux

des effets indésirables ont été des troubles gastro-intestinaux

[…]

6.10 Paires de terme médical – reformulation : CLEAR GP *(extrait)*

Terme médical [TAB] Reformulation médicale

accident vasculaire cérébral (AVC)

accident vasculaire cérébral (maladie AVC)

accident vasculaire cérébral (AVC) et d'autres lésions cérébrales

accidents vasculaires cérébraux (maladie AVC)

acétylcholine connue sous le nom d maladie une substance chimique

acide aminosalicylique (maladie Les préparations à base de 5-ASA)

acides gras (lipides)

Acides gras oméga 3 (maladie huile de poisson)

acides gras polyinsaturés (trouble/s AGPI)

acouphènes (des sifflements dans les oreilles)

acouphènes (des bourdonnements dans les oreille)

acouphènes (des tintements d'oreilles)

activités quotidiennes telles que ; trouble/s s'habiller, se nourrir et se doucher

affections , notamment ; affection/s l'arthrose

affections telles que ; affection/s la pneumonie, la méningite et la diarrhée

affections y compris ; affection/s la douleur chronique

affections à long terme comme affection/s l'asthme ou le diabète

affections anales telles que ; affection/s les hémorroïdes et les fistules

affections auto-immunes telles que la polyarthrite rhumatoïde

affections auto-immunes telles que la polyarthrite rhumatoïde

affections chroniques (trouble/s ; affection/s l'asthme, le trouble déficitaire de l'attention avec hyperactivité, l'insuffisance veineuse chronique, le

diabète, la dysfonction érectile, l'hypertension, l'arthrose)

affections chroniques (l'asthme (deux études), le trouble déficitaire de l'attention avec hyperactivité (une étude), l'insuffisance veineuse chronique (deux études), le diabète (quatre études), la dysfonction érectile (une étude), l'hypertension (deux études) et l'arthrose (trois études))

affections coûteuses et invalidantes comme maladie ; affection/s les maladies cardiovasculaires, le cancer ou le diabète

affections dues à un excès d'acide urique comme la goutte

affections médicales maladie ; telles que ; affection/s le diabéte, la polyarthrite rhumatoîde, les maladies cardiaques ou encore la naissance prématurée de bébés en sous-poids

affections orthopédiques (affection/s musculosquelettiques)

affections pulmonaires comme maladie ; affection/s les maladies pulmonaires obstructives chroniques

agents antigrippaux (maladie amantadine, rimantadine, zanamivir et oseltamivir)

agents dopaminergiques (maladie médicaments imitant l'effet du neurotransmetteur dopamine)

agents infectieux tels que des champignons, des parasites et des virus

agents pathogènes (maladie responsables de maladies)

agonistes de la dopamine (médicaments imitant la dopamine)

agranulocytose (trouble/s une réduction de la production de globules blancs)

akathisieappelé trouble/s un trouble moteur

akathisie(agitation intérieure et incapacité à rester assis)

'alimentation par voie intraveineuse (parentérale)

allergies telles que l'asthme, la dermatite / l'eczéma, le rhume des foins

aminoadamantanes par exemple la rimantadine

amputées (retrait chirurgical d'une partie du membre)

analgésiques ; tels que le fentanyl

anémie (maladie faible taux d'hémoglobine)

anémie , trouble/s un trouble qui se caractérise par une réduction du nombre de globules rouges et de la capacité du sang à transporter l'oxygène

anémie qui représente trouble/s trouble sanguin

anémie maladie ; appelé un problème sanguin

anesthésiant / congelant

anesthésique local tel que l'améthocaïne, la benzocaïne ou la lidocaïne

anomalies coronariennes maladie lésions cardiaques

antagonistes de l'aldostérone (par exemple la spironolactone ou l'éplérénone)

antagonistes des récepteurs de type 1 de l'angiotensine II (ARA-II)

antécédents de consommation d'amphétamines c'est-à-dire la quantité cumulée et à la fréquence d'exposition aux amphétamines

antibiotiques (clarithromycine, azithromycine et levofloxacine)

antidépresseurs tels que ; des agents tricycliques

antigrippale (maladie contre la grippe)

antihistaminiques comme les stéroïdes topiques

antihistaminiques par exemple Clarityne, Piriton, Zyrtec, Benadryl et Phénergan

antioxydant (maladie protégant les cellules des lésions causées par des radicaux libres)

appareil orthodontique (maladie bagues)

appendicite compliquée définie comme une appendicite gangreneuse (nécrose des tissus mous) ou perforée (rompue)

artères du coeur (maladie coronaires)

artériopathie oblitérante de l'artère iliaque affection/s peut entraîner des douleurs dans les jambes à la marche (claudication intermittente), au repos, voire même des ulcères du pied ou de la jambe

articulations (des hanches et des genoux)

articulations comme ou le genou ou la hanche

articulations telles que e genou ou votre hanche ou les articulations de vos mains

aspects de la santé physique : maladie ; par exemple ; ou le poids, la taille, le tour de tête/bras/jambe, la durée de sommeil en 24 heures ; la durée des pleurs ou des agitations ; la bilirubine sérique et le nombre d'épisodes de maladie

assistance respiratoire (maladie ventilation mécanique)

assistance respiratoire invasive (par exemple l'intubation)

asthme trouble/s ; ou trouble respiratoire chronique

ataxie trouble/s ralentissement de la parole, de troubles de l'élocution ou parce que leur voix rend un son dur ou nasal

atélectasie pulmonaire (c'est-à-dire l'incapacité des poumons à se gonfler totalement)

athérosclérose maladie ; ou durcissement ou à l'obstruction des artères

attachement de l'enfant (la capacité de l'enfant à chercher et à maintenir une relation proche avec le principal proche)

aucun traitement (soins habituels)

autres spondylarthrites (SpA)

axillaires(aisselle)

bactéries présentent une résistance à des antibiotiques , par exemple Staphylococcus aureus résistant à la méthicilline

bactéries présentent une résistance aux antibiotiques de routine , par exemple Staphylococcus aureus résistant à la méthicilline

barotraumatisme défini comme la présence d'un pneumothorax sur la radiographie du thorax ou l'insertion d'un drain thoracique pour pneumothorax avéré ou soupçonné

bêta-agonistes à action prolongée (par exemple le Formotérol)

biais (par exemple surestimation des bénéfices ou sous-estimation des inconvénients des traitements de manière systématique)

biais autrement dit une surestimation des effets bénéfiques et une sous-estimation des effets néfastes

biais c'est-à-dire surestiment les bénéfices et sous-estiment les risques

blessures pénétrantes (par exemple les plaies par balle ou par couteau)

bon cholestérol (lipoprotéines de haute densité)

bouche sèche (sécheresse buccale)

BPCO (trouble/s emphysème et bronchite chronique)

BPCO - bronchite chronique (maladie une maladie pulmonaire chronique avancée)

BPCO - bronchite chronique ou emphysème

bronchectasie (maladie une maladie pulmonaire chronique)

bronchiolite (maladie une maladie respiratoire qui affecte généralement les nourrissons et qui typiquement imite un simple rhume)

bronchopneumopathie chronique obstructive (BPCO)

broncho-pneumopathie chronique obstructive c'est-à-dire BPCO, c'est-à-dire bronchite chronique ou emphysème, ou les deux

broncho-pneumopathie chronique obstructive , une combinaison d'emphysème et de bronchite chronique

broncho-pneumopathie chronique obstructive ou BPCO

broncho-pneumopathie chronique obstructive (BPCO)

bronchopneumopathie chronique obstructive (BPCO) maladie ; ou maladie pulmonaire obstructive chronique (MPOC)

calcium et de phosphore <non-marq> maladie ces minéraux

cancer d'autres affection/s affections douloureuses au long cours

cancer systémiques c'est-à-dire non localisés

cathéter (un tube mince)

cathéter veineux périphériques/canule (CVP) aussi appelé/e voie veineuse

cathétérisme cardiaque appelée une procédure invasive

cathéters périphériques (c'est-à-dire des tubes insérés dans les veines des membres et conçus pour une utilisation à court terme)

causes iatrogènes (problèmes résultant d'un traitement médical)

cellules malignes (cancéreuses)

ceux qui réduisent l'intérêt sexuel par d'autres mécanismes (c'est-à-dire les antipsychotiques et les antidépresseurs sérotonergiques (ISRS))

changements dans le système immunitaire y compris ; , maladie ;trouble/s une modification de la réaction immunologique

changements de mode de vie , par exemple être physiquement plus actif, manger moins de calories et rester assis moins longtemps

chemo-fog ou chemobrain

chimiothérapie (maladie traitement médicamenteux)

chimiothérapie (médicaments anticancéreux)

chimiothérapie (traitement qui cible les cellules lymphatiques)

chimiothérapie préopératoire ou néoadjuvante (connue sous le nom de maladie chimiothérapie avant la chirurgie)

chirurgie de la cataracte (affection/s En retirant le cristallin opaque)

chirurgie de l'obésité (trouble/s Les procédures chirurgicales pour traiter l'obésité)

chirurgie de réduction tumorale (maladie enlever autant que possible le cancer visible)

cholécystectomie laparoscopique connue sous le nom de affection/s l'ablation chirurgicale de la vésicule biliaire à travers une chirurgie endoscopique

cholestérol (maladie une substance graisseuse)

cholestérol (appelée lipide)

cholestérol des lipoprotéines de basse densité (LDL, le mauvais cholestérol)

chorioamniotite (par exemple morbidité maternelle)

chronique indique que affection/s l'affection persiste pendant au moins six semaines

chronique (la maladie peut être douloureuse et de longue durée)

chronique (au long cours)

chronique (persistante)

cirrhose (une maladie hépatique avancée)

claudication intermittente (également connue sous le nom de maladie Une personne atteinte d'une maladie des artères des jambes peut ressentir de la douleur en marchant)

claudication intermittente (des douleurs dans les jambes à la marche)

claudication intermittente (CI) on parle également de maladie une maladie artérielle périphérique suite au rétrécissement des artères des jambes peuvent ressentir des crampes douloureuses dans les jambes ou les fesses à la marche

clopidogrel , maladie un nouveau médicament antiplaquettaire oral

CNTF (Ciliary Neurotrophic Factor)

colite ulcéreuse (CU) , maladie une maladie inflammatoire chronique du côlon

combinaison de traitements antirétroviraux (maladie cART)

comorbidité maladie présence d'autres maladies ou pathologies

co-morbidité (maladie présence d'autres maladies)

complications (maladie ; telles que l'infection)

complications (trouble/s IVU, hydronéphrose et troubles de la fonction rénale)

complications à long terme comme maladie des maladies rénales, nerveuses et oculaires

complications liées au VIH maladie ; telles que ; ou les infections opportunistes ou le décès

complications motrices maladie ; telles qu' une dyskinésie, un mouvement du corps saccadé de pseudo-danse

complications motrices telles que une dyskinésie, un mouvement du corps saccadé de pseudo-danse

congénitales (maladie présentes à la naissance)

conséquences neurologiques à vie notamment ; trouble/s déficiences cognitives et motrices, ainsi que des troubles auditifs et visuels

convulsif (épileptique)

convulsions récurrentes (crises)

Cordyceps sinensis (maladie une herbe médicinale chinoise)

cornée (la partie avant transparente de l'oeil)

corticostéroïdes (également appelés par exemple corticoïdes)

COX-2 (par exemple le célécoxib)

crèmes antibiotiques , maladie la mupirocine et l'acide fusidique

crises convulsives (crises)

crises modérées (c'est-à-dire nécessitant des corticoïdes oraux)

critères de jugement importants , maladie ; tels que la qualité de vie liée à la santé, les maladies (la morbidité) et les coûts économiques

critères de jugement importants qui étaient rarement, voire jamais, inclus (par exemple le sentiment de contrôle pendant le travail, l'allaitement, l'interaction entre la mère et son bébé, les coûts et les résultats concernant le nourrisson)

croissance restreinte (défini/e ; définition les bébés petits mais normalement développés et les bébés qui ne concrétisent pas leur plein potentiel de croissance)

d'analgésiques tels que le fentanyl

d'autres dits négatifs (fatigue, apathie, perte des émotions)

de façon systémique (administrés par voie orale ou par injection)

de syndromes génétiques (par exemple : syndrome de Down, syndrome de Rett

[…]

6.11 Paires de terme médical – reformulation : ClassYN SP *(extrait)*

Terme médical [TAB] Reformulation médicale

18F-fluoro-désoxy-glucose (FDG)

18F-fluoro-désoxy-glucose (FDG)

accident vasculaire cérébral (AVC)

accident vasculaire cérébral (AVC)

accident vasculaire cérébral (AVC)

accidents vasculaires cérébraux ((AVC)

acroparesthésies et des angiokératomes appelé/e affection/s douleurs des extrémités

activité de base de la vie quotidienne (trouble/s ADL)

activité des méthyltransférases de l' ADN (DNMT)

activité histone-acétylase (HAT)

affaiblissement récent global (maladie/s asthénie , anorexie , amaigrissement)

affaiblissement récent sélectif (troubles de la marche , troubles intellectuels)

affection longue durée (affection/s ALD)

affection extra-articulaire : affection/s psoriasis , acné , pustulose , entéropathie , infection génito-urinaire ou digestive , atteinte oculaire

affection extra-articulaire : affection/s psoriasis , acné , pustulose , entéropathie , affection génito-urinaire ou digestive , atteinte oculaire

affection néoplasique (affection/s splénique , gastrique , génito-urinaire , endocrinien , vésicale)

affection oculaire (toute autre affection/s dégénérescence maculaire , glaucome)

affections , comme maladie/s le syndrome de Gougerot-Sjögren ou le syndrome des antiphospholipides

affections dermatologiques (maladie/s acnés conglobata et fulminante , hidrosadénite suppurée ou maladie de Verneuil , pustulose palmoplantaire)

affections dermatologiques (psoriasis , pemphigus)

affections immunologiques (lupus , syndrome des antiphospholipides)

Affections pulmonaires (sans infection) : affection/s asthme , mucoviscidose , ventilation mécanique

affections respiratoires chroniques (asthme , bronchopneumopathie chronique obstructive)

agents fongiques (notamment Histoplasma)

Aides mécaniques (trouble/s cannes , béquilles , déambulateur)

AJI (arthrite juvénile idiopathique)

AJI : arthrite juvénile idiopathique

ALD (affection/s affection de longue durée)

ALD (affection/s affection de longue durée)

ALD (affection/s affection de longue durée)

ALD (affection/s affection de longue durée)

ALD (affection/s affection de longue durée)

ALD : affection/s affection de longue durée

ALD : affection/s affection de longue durée

amphotéricine (AMB)

anesthésie générale (trouble/s AG)

anomalies biologiques (trouble/s syndrome inflammatoire , troubles de l' absorption)

antagonistes des récepteurs de l'angiotensine-2 (ARA-2)

antagonistes des récepteurs de l'angiotensine-2 (ARA-2)

antibiotique dit temps dépendant tels que la vancomycine et les bêtalactamines

anticitrulline [affection/s anti-CCP]

anticorps anticytoplasme des polynucléraires (maladie/s ANCA)

anticorps antinucléaires (maladie/s AAN)

anticorps antinucléaires (FAN)

anticorps antinucléaires (FAN)

anticorps antiphospholipides (trouble/s aPL)

anticorps anti-phospholipides (trouble/s aPL)

antiépileptiques notamment la gabapentine et la prégabaline

antifongiques , l' itraconazole (ITZ) et l' amphotéricine B (AMB)

antifongiques l' itraconazole (ITZ) et l' amphotéricine B (AMB)

antigènes spécifiques de Mycobacterium tuberculosis tels que la catalase-peroxydase

anti-phospholipides (trouble/s SAPL)

antithrombine (AT)

apporter une nouvelle fonction (synthèse de facteurs neurotrophiques)

arguments histologiques (maladie/s granulomes épithélioïdes et gigantocellulaires sans nécrose caséeuse)

Association de langue française pour l'étude du diabète et des maladies métaboliques (maladie/s ALFEDIAM)

Association de langue française pour l'étude du diabète et des maladies métaboliques (maladie/s ALFEDIAM)

asthme (affection/s une affection respiratoire chronique)

asthme , mucoviscidose , ventilation mécanique : affection/s Affections pulmonaires (sans infection)

ATM est un senseur de dommage capable d' activer des voies de signalisation agissant notamment sur le cycle cellulaire et sur la recombinaison homologue

atteinte pulmonaire (trouble/s insuffisance respiratoire restrictive , hypoxémie , hypercapnie)

atteinte muqueuse définie par une infiltration de la muqueuse sans atteinte de la musculeuse ou de la sous-séreuse

atteinte musculeuse définie une obstruction intestinale complète ou incomplète avec infiltration de la musculeuse par des PNE en l'absence d'une ascite à éosinophiles

atteinte musculeuse définie par une obstruction intestinale complète ou incomplète avec infiltration de la musculeuse par des PNE en l'absence d'une ascite à éosinophiles

atteinte sous-séreuse définie une infiltration du tractus gastro-intestinal par des PNE associée à une ascite à éosinophiles

atteinte sous-séreuse définie par une infiltration du tractus gastro-intestinal par des PNE associée à une ascite à éosinophiles

atteintes extra-neurologiques tel que maladie/s le livedo du syndrome de Sneddon [101] , les douleurs acrales et les angiokératomes de la maladie de Fabry , les stries angioïdes et les lésions des grands plis du pseudoxanthome élastique [102] , ou encore l'atteinte rétino-cochléaire du syndrome de Susac

autisme de Kanner (trouble/s les troubles développementaux de la personnalité ; trouble de l' unité)

autoanticorps (maladie/s autoAc)

avoir ou non un aspect granulomateux (PRD en graisse de mouton)

bacille de Klebs-Loeffler (maladie/s corynebacterium diphteriae)

bactéries (Yersinia , Salmonella)

biopsie des glandes salivaires accessoires (BGSA)

Bloom (échange d' ADN trop fréquent amenant à un retard de croissance et de nombreux cancers)

bouffées de chaleur , érythermalgie (trouble/s Troubles vasomoteurs ou de la thermorégulation)

BPCO : bronchopneumopathie chronique obstructive

BPCO bronchopneumopathie chronique obstructive

bradykinésie (ralentissement des mouvements)

bronchopneumopathie chronique obstructive (BPCO)

burn out syndrome connu sous le nom de un contexte d' épuisement et de dépression

cancers notamment sein , des poumons , de la rate , et de l' estomac

cancers notamment sein , des poumons , de la rate , et de l' estomac

cancers notamment digestifs

cancers notamment digestifs

carcinoma associated retinopathy [CAR]

cardiovasculaires (bronchospasmes , hypotension)

cardiovasculaires (bronchospasmes , hypotension)

causes infectieuses : mycobactéries atypiques , en particulier chez des immunodéprimés [21] , lèpre [4] and [7] , histoplasmose [22] , candidose [3] and [5] , toxoplasmose [4] , Epstein-Barr virus

CD53 est un/e un marqueur leucocytaire

CD53 est un/e par exemple un marqueur leucocytaire

CD81 , CD82 , CD9 telles que des compartiments légèrement acides (pH 6 ,2) riches en tétraspanines

cellules anormales telles que des blastes (leucémie) , des schizocytes

cellules immunitaires humaines telles que les cellules T

cellules immunitaires humaines telles que les cellules T

Céphalées avec signes de gravité (trouble/s hypertension intracrânienne , troubles de la vigilance)

céphalées d'allure commune (maladie/s migraine , céphalée de tension)

certaines comorbidités tels que l'insuffisance rénale mais aussi les troubles cognitifs et psychiatriques graves , les associations médicamenteuses qui potentialisent l'activité des sulfamides hypoglycémiants , la polymédication ainsi que l'alcool

certaines hémopathies malignes (syndromes lymphoprolifératifs)

certains virus comme le CMV , le HSV , l'EBV en cas d'hypercellularité dans le LCS , mais également le virus JC

chaise à primate appelé/e système de contention utilisé chez le primate vigile encore

chorée gravidiquenotamment maladie/s pendant une grossesse

chromosome 16 (maladie/s 16q12)

coagulation intravasculaire disséminée (CIVD)

coefficient de diffusion apparent (maladie/s CDA)

comorbidités tels que trouble/s l'insuffisance rénale

comorbidités tels que trouble/s l'insuffisance rénale

compartiments légèrement acides (pH 6 ,2) riches en tétraspanines telles que CD81 , CD82 , CD9 et partiellement CD63

complexes (avec trouble de la vigilance)

complications cardiaques (trouble/s arythmie ou troubles de la conduction)

complications de la corticothérapie chez les sujets âgés , notamment l'ostéoporose , les complications cutanées et les complications morphologiques.

complications relèvent en effet d'un mécanisme hormonal (trouble/s acné , hirsutisme , complications psychiatriques , troubles hydro-électrolytiques)

complications sévères (maladie/s neuropathie périphérique , maladie thromboembolique veneuse profonde)

comportement moteur (et notamment dyskinétique)

composés toxiques comme les sels biliaires

comprimés à libération prolongée (par exemple Kaléorid® , Topalgic® LP)

comprimés à libération prolongée (par exemple Kaléorid® , Topalgic® LP)

comprimés gastro-résistants (par exemple Dépakine® , Dépamide® , Inexium®)

comprimés gastro-résistants (par exemple Dépakine® , Dépamide® , Inexium®)

comprimés gastro-résistants par exemple Dépakine® , Dépamide® , Inexium®

connectivites (par exemple lupus systémique , syndrome de Sjögren , polyarthrite rhumatoïde , myopathies inflammatoires)

coping : il s'agit de maladie/s la capacité à surmonter un événement éprouvant et d'en diminuer l'impact sur son bien être physique et psychique

corticoïdes (méthotrexate , mycophénolate mofétil , cyclophosphamide)

coxite (maladie/s Les formes chroniques exceptionnellement destructrices)

crises non épileptiques psychogènes (Les autres trouble/s crises somatoformes , troubles factices , simulations)

crises répétées (notamment plus de deux crises en 24 heures)

cryoglobulines , maladie/s des immunoglobulines (Ig) anormales qui précipitent lors d' une exposition au froid

cryoglobulines , maladie/s des immunoglobulines (Ig) anormales qui précipitent lors d' une exposition au froid

cryopyrine est appelé/e La protéine correspondante au gène CIAS1 a des régions homologues avec celle de la pyrine

cryptorchidisme (affection/s non descente des testicules dans le scrotum)

cryptorchidisme (non descente des testicules dans le scrotum)

CSI : corticostéroïde inhalé

CSI orticostéroïde inhalé

cutanées (hématomes , lésions purpuriques)

cytokines inflammatoires (TNF? , IL-1? , IFN?)

cytomegalovirus (CMV)

cytomegalovirus (CMV)

d' affections organiques d'ordre rhumatologique , hormonale , neurologique voire métabolique

d' autoAc non spécifiques (facteurs rhumatoïdes , antinucléaires , antiphospholipides)

d' autoAc non spécifiques ((facteurs rhumatoïdes , antinucléaires , antiphospholipides)

d' HVA (4-hydroxy-3-méthoxyphenylacetique acide , Sigma)

d' inhibiteurs tels que le F4P ou le transforming growth factor

d' une infection rétinienne adjacente(par exemple ors d' une nécrose rétinienne aiguë)

d'accident vasculaire cérébral (AVC)

d'affections dermatologiques (psoriasis , pemphigus)

d'autres affections immunologiques (lupus , syndrome des antiphospholipides)

d'encéphalopathie postérieure réversible (maladie/s PRES)

Dc (diagnostic)

Dc : diagnostic

DDD (définie dose définie journalière)

de Bloom (échange d' ADN trop fréquent amenant à un retard de croissance et de nombreux cancers)

de liens appelé/e Linking number , Lk

de liens appelé/e Linking number , Lk

de comprimés gastro-résistants (par exemple Dépakine® , Dépamide® , Inexium®)

de DOPAC (3 ,4 dihydroxyphenylacetique acide Sigma)

de manière prolongée (maladie/s pendant une période minimale de 18 mois)

DEP : débit expiratoire de pointe

DEP débit expiratoire de pointe

dérivation ou dérivatisation

dérivation ou dérivatisation

des anomalies structurales et cytoarchitecturales (trouble/s neuropile , interactions neurone- glie)

des arthralgies intermittentes (par exemple un rhumatisme palindromique)

des défauts de pigmentation , certaines zones sont dépigmentées alors que d' autres sont hyperpigmentées

des défauts de pigmentation , certaines zones sont dépigmentées alors que d' autres sont hyperpigmentées

[…]

6.12 Paires de terme médical – reformulation : ClassYN GP *(extrait)*

Terme médical [TAB] Reformulation médicale

abdominaux (muscles de l' abdomen)

abdominaux (muscles de l' abdomen)

accident cardio-vasculaire (maladie/s AVC)

accident vasculaire cérébral aussi appelé trouble/s AVC

accidents vasculaires cérébraux [ou AVC]

acétaminophène par exemple Tylenol

acétaminophène par exemple Tylenol©

acétylcholine (RACh)

acétylcholinestérase (ACh)

acides gras (maladie/s oméga 3)

Acné est un/e maladie/s une maladie de la peau liée à l' inflammation des follicules pilosébacés

acnés sévères telles que l'acné nodulaire, l'acné conglobata (acné grave du dos) ou toute acné pouvant provoquer des cicatrices définitives

Acromégalie : maladie/s Maladie due à une sécrétion excessive d' hormone de croissance qui entraîne un développement important de certaines parties du corps (mains , pieds , visage)

acromégalie (croissance exagérée du visage)

actes douloureux de courte durée : piqûres, ponctions, petite chirurgie, réductions de fracture ou soins douloureux

activité physique (maladie/s monter un escalier, soulever une charge lourde)

activités (telles que grimper une échelle ou faire de longues marches)

adénomectomie : consiste à intervention qui consiste à enlever un adénome de la prostate au cours d' une opération

Adénovirus : Virus qui a une affinité pour les tissus lymphoïdes

Affection de longue durée (affection/s ALD)

affection longue durée (affection/s ALD)

affection stomatologique ou ophtalmologique (glaucome)

affections articulaires chroniques : affection/s polyarthrite chronique évolutive , rhumatisme psoriasique , ostéonécrose , tuberculose osseuse

Affections de Longue Durée (maladie/s ALD)

Affections de Longue Durée (ALD)

affections des vaisseaux sanguins (malformation vasculaire , AVC, hémorragie méningée , maladie de Horton)

affections opportunistes On appelle maladie/s les maladies ou les infections qui accompagnent le sida

affections pulmonaires telles que affection/s la bronchite chronique

affections pulmonaires telles que affection/s la bronchite chronique

aide technique désigne tout élément qui permet de pallier une fonction déficiente

aide technique désigne tout élément qui permet de pallier une fonction déficiente

ALFEDIAM (maladie/s Association de langue française pour l' étude du diabète et des maladies métaboliques)

alglucosidase alpha (maladie/s une enzyme recombinante)

allergies (en particulier asthme , eczéma , rhume des foins)

Allocation adulte handicapé (AAH)

allocation d' éducation de l' enfant handicapé (AEEH)

alpha-bloquants (des médicaments qui agissent sur la force du muscle de la vessie ou des sphincters)

Altération de l' état général (maladie/s fatigue , amaigrissement , perte de l' appétit)

amniocentèse (maladie/s Le diagnostic prénatal, proposé en début de grossesse)

amniocentèse (maladie/s Le diagnostic prénatal, proposé en début de grossesse)

amniocentèse (prélèvement au niveau du liquide amniotique)

AMS de forme C (cérébelleuse) anciennement atrophie olivopontocérébélleuse

AMS de forme P (parkinsonienne) appelée autrefois maladie/s dégénérescence striatonigrique

amyotrophie (affection/s fonte musculaire)

amyotrophie (affection/s fonte musculaire)

amyotrophie (la fonte des muscles)

amyotrophie (une fonte musculaire)

amyotrophie scapulo-péronière (maladie/s aux muscles des épaules et des chevilles)

Amyotrophie spinale distale ou neuropathie motrice héréditaire distale

amyotrophie spinale juvénile ou maladie/s maladie de Kugelberg-Welander

amyotrophies spinales proximales (ou maladie/s amyotrophies spinales antérieures , ASA)

amyotrophies spinales proximales sont un un groupe de maladies héréditaires qui se caractérisent par une faiblesse musculaire liée à une paralysie plus ou moins importante , et par une fonte ou « atrophie » des muscles de la racine des membres , c' est-à-dire des hanches , des épaules (muscles dits « proximaux ») , ainsi que des muscles du tronc

analogues de la somatostatine notamment l' octréotide et le lanréotide

analyses sanguines (maladie/s prise de sang)

analyses sanguines (maladie/s prise de sang)

androgène : hormone sexuelle mâle

anémie (maladie/s une baisse des globules rouges dans le sang)

anémie (maladie/s la baisse de la quantité de globules rouges dans le sang)

anémie (baisse importante du taux de globules rouges)

anémie on parle d' les globules rouges sont déficitaires

anémie (affection dans laquelle le nombre de globules rouges dans le sang est trop faible , entraînant un apport d' oxygène insuffisant aux tissus et aux organes)

anémie (mauvaise absorption du fer)

anémies « hémolytiques » les autres ; c' està-dire maladie/s dues à une destruction importante des globules rouges

anémies « hémolytiques » (notamment bêta-thalassémie et drépanocytose)

Anesthésie : consiste à acte qui consiste à endormir et rendre insensible le patient (anesthésie générale) ou une partie du corps (anesthésie locale ou partielle) pendant une intervention chirurgicale

anesthésie : consiste à acte qui consiste à endormir et rendre insensible un patient (anesthésie générale) ou une partie du corps (anesthésie locale)

anesthésie : consiste à acte qui consiste à endormir et rendre insensible un patient (anesthésie générale) ou une partie du corps (anesthésie locale)

anesthésie : consiste à acte qui consiste à endormir et rendre insensible un patient (anesthésie générale) ou une partie du corps (anesthésie locale)

anesthésie : consiste à acte qui consiste à endormir et rendre insensible un patient (anesthésie générale) ou une partie du corps (anesthésie locale) pendant une intervention chirurgicale

anesthésie : consiste à acte qui consiste à endormir et rendre insensible un patient (anesthésie générale) ou une partie du corps (anesthésie locale)

anesthésie : consiste à acte qui consiste à endormir et rendre insensible un patient (anesthésie générale) ou une partie du corps (anesthésie locale)

Anesthésie : consiste à acte qui consiste à endormir et rendre insensible la patiente (anesthésie générale) ou une partie du corps (anesthésie locale ou partielle) pendant une intervention chirurgicale

Anesthésie : consiste à acte qui consiste à endormir et rendre insensible la patiente (anesthésie générale) ou une partie du corps (anesthésie locale ou partielle) pendant une intervention chirurgicale

anesthésie générale (endormir et rendre insensible un patient)

anesthésie locale (endormir et rendre insensible une partie du corps)

anévrisme notamment une malformation de la paroi de certains vaisseaux sanguins

anévrysme congénital , malformations artério-veineuses (d' autres maladie/s maladies des artères ou des veines)

Angine accélérée : maladie/s douleur angineuse qui ne répond pas au traitement , pouvant annoncer une crise cardiaque

angine de poitrineappelé/e douleur thoracique

angine de poitrineAutres noms : maladie/s maladie coronarienne

angine de poitrine(douleur semblable à une crampe , intermittente , qui empire lors d' un effort physique)

angiographie (radiographie des vaisseaux après injection d' un produit de contraste radio-opaque)

angiographie (consiste en radiographie des vaisseaux après injection d' un produit de contraste radio-opaque dans une artère ou une veine)

angioplastie (rétrécissement d' une artère peut être traité à l' aide d' un ballon)

Anomalie de position de l’articulation (scoliose , malformation de la hanche)

anomalie génétique (maladie/s mutation)

anomalie génétique (maladie/s mutation)

anomalie génétique (maladie/s mutation)

anomalie génétique (maladie/s mutation)

anomalie génétique (mutation ou délétion du gène SMN1 chez le fœtus)

anomalie génétique ou chromosomique (en particulier trisomie 13 et 18)

anomalies (maladie/s mutations)

antalgiques (médicaments contre la douleur)

antalgiques (médicaments contre la douleur)

antécédent médical : maladie/s traitement déjà subi , en cours ou maladie antérieure

antécédent médical : maladie/s traitement déjà subi , en cours ou maladie antérieure (diabète , maladie cardiaque , etc

antécédent médical : maladie/s maladie antérieure ou opération déjà subie ou en cours (diabète , maladie cardiaque , etc

Antécédent médical ou chirurgical : maladie/s maladie ou traitement déjà subi ou en cours (diabète , maladies cardiaques , etc

antiagrégants plaquettaires par exemple Ticlid® , Plavix® , aspirine

anti-agrégants plaquettaires (par exemple trouble/s Ticlid® , plavix , aspirine)

anti-agrégants plaquettaires (par exemple trouble/s Ticlid® , plavix , aspirine)

anti-arythmiques (trouble/s amiodarone)

antibiotiques (pénicilline)

anticholinergiques (des médicaments qui diminuent l' hyperactivité de la vessie)

anticorps en particulier des anticorps anti-p24 et anti-p1

anticorps anti-CCP (maladie/s peptides cycliques citrullinés)

anti-inflammatoires non stéroïdiens par exemple Apranax®, Naprosyne®, Nifluril®, Surgam®, Voltarène®

apex (appelé/e la pointe du cœur)

apnée (interruptions de la respiration)

apnées (arrêts respiratoires)

appareil de radioscopie appelé/e « simulateur »

apraxie signifie sans action

arrêt cardiaque c'est-à-dire quand le cœur cesse de battre

arrêt cardiaque (mort subite)

artère (vaisseau qui va du coeur vers le cerveau)

artères (vaisseaux qui conduisent le sang du cœur vers les organes)

artères carotidiennes (larges artères du cou qui sont les voies principales d' alimentation du cerveau en sang et en oxygène)

arthrodèse consiste à ; Il s' agit d' une intervention qui consiste à fixer deux (ou plus) vertèbres entre elles , de façon à stabiliser l' étage vertébral défaillant

arthrodèse consiste à ; Il s' agit d' une intervention qui consiste à fixer deux (ou plus) vertèbres entre elles , de façon à stabiliser l' étage vertébral défaillant

arthrose est la maladie/s la plus fréquente des maladies articulaires

arthrose notamment ; ou d'autres maladie/s maladies des os

arthrose ou d' autres maladie/s maladies des os

arythmie (maladie/s atteinte cardiaque)

arythmies (trouble/s troubles du rythme cardiaque)

Association Française contre les Myopathies (maladie/s AFM)

asthénie (une fatigue)

asthme ou d' maladie/s maladie pulmonaire chronique

ataxie (déséquilibre , maladresse)

athéromatose artérielle périphérique / une maladie vasculaire athérosclérotique (maladie/s rétrécissement des artères des jambes ou des bras)

athéromatose artérielle périphérique / une maladie vasculaire athérosclérotique (MVAS)

Athérosclérose , maladie/s maladie des artères

Athérosclérose , maladie/s maladie des artères

Athérosclérose : maladie/s Maladie des artères obstruées par des plaques qui contiennent du cholestérol

athérosclérose aussi appelé durcissement des artères

atrophie (rétrécissement)

atrophie multisytématisée est connue sous de nombreuses appellations maladie/s syndrome de Shy-Drager , dégénérescence striatonigrique , atrophie olivopontocérébelleuse

atrophie multisytématisée (MSA)

atrophie musculaire péronière (une atrophie progressive, plus marquée aux jambes)

atrophie musculaire péronière (une atrophie progressive, plus marquée aux jambes)

atrophie musculaire péronière (une atrophie progressive, plus marquée aux jambes)

atteinte du cervelet (trouble/s syndrome cérébelleux)

atteinte musculaire (faiblesse musculaire)

atteinte neurogène (on parle d' maladie/s une atteinte persistante (chronique) des motoneurones)

atteinte primitive bilatérale des surrénales appelé/e PPNAD (Primary Pigmented Nodular Adrenal Disease)

atteintes génétiques motoneuronales notamment maladie/s la maladie de Fazio-Londe , syndrome de Brown-Vialetto-Van Laere ou certaines variantes d'amyotrophie spinale

autorisation de mise sur le marché (affection/s AMM)

autorisation de mise sur le marché (AMM)

Autosomique signifie maladie/s le gène responsable de la maladie n' est pas lié au sexe du porteur

AVC (maladie/s « caillot au cerveau »)

AVC voir Accidents vasculaires cérébraux

AVC hémorragique (un vaisseau sanguin du cerveau se rompt)

AVC ischémique (un arrêt du fonctionnement cérébral qui peut avoir lieu : - si une partie du cerveau n' est plus irriguée)

avoir plus de muscle en masse c'est-à-dire une augmentation de la masse musculaire

bacille , bactérie , virus (un microorganisme invisible à l' œil nu susceptible de provoquer des maladies)

[…]

6.13 Paires de terme médical – reformulation : GrandMed-Ro2 *(extrait)*

Terme médical [TAB] Reformulation médicale

abilitatile lor de gandire (cognitive)
ablatie prin cateter numita procedura
acanthosis nigricans numita afectiune
accident vascular cerebral cum ar fi probleme grave
accident vascular cerebral (AVC)
accidentare majora cum ar fi o fractura
accidentari, cum ar fi ruptura de ligament de cartilaj
accidente vasculare cerebrale si ale vaselor periferice (leziuni ireversibiele la nivelul vaselor sanguine)
Accidentele rutiere sau alte traumatisme
Accidentul cerebral reprezinta un blocaj la nivelul unui vas de sange responsabil de conducerea fluxului sanguin catre creier
Accidentul vascular cerebral este cea mai frecventa afectiune a creierului
Accidentul vascular cerebral reprezinta o conditie medicala acuta fie de natura ischemica, fie hemoragica in urma careia se produce o perturbare in asigurarea fluxului sanguin normal
Accuzide Forte sau alte medicamente
Acest lucru poate rezulta din ritmuri cardiace anormale (aritmii) sau un dezechilibru de electroliti – minerale cum ar fi sodiu, potasiu si calciu, care mentin echilibrul lichidelor din corp
acetaldehida numit compus
acetaldehida , substanta ce modifica antigenele de la nivelul membranei hepatice

acetaminofen sau ibuprofen cum ar fi medicamente anti-durere si anti-febra
acetaminofenul sau ibuprofenul cum ar fi medicamente pentru febra si durere
acid alfa-linolenic numit un tip de acid gras Omega-3
acid alfa-linolenic (ALA)
aciditatea (pH-ul)
acizi numiti cetone
acizi toxici numiti corpi cetonici
acizilor grasi cu lant scurt cum ar fi butiratul
acizilor grasi cu lant scurt (SCFA)
Acneea este o boala a pielii caracterizata prin puncte negre, puncte albe, inflamatie, eruptii cutanate, piele rosie si uneori leziuni profunde
Acneea este o afectiune plurifactoriala si complexa
Acneea este un tip de boala un tip de boala inflamatorie
Acneea este o afectiune inflamatorie plurifactoriala
Acneea vulgara este o boala de piele care apare din cauza hiperactivitatii glandelor sebacee
Acneea vulgara sau acneea
Acneea vulgara este o afectiune a pielii, si anume a stratului superficial, a epidermei
Acneea vulgara caracterizata prin hipersecretia glandelor sebacee
Acneea vulgara este o boala comuna a pielii care apare atunci cand sebumul si celulele moarte ale pielii astupa porii
Acneea vulgara popular denumita acnee
Acromegalia este o afectiune rara care rezulta in urma productiei excesive a hormonului de crestere
activitati normale cum ar fi actul sexual
activitati normale cum ar fi actul sexual
activitatile fizice cum ar fi efortul fizic general, sportul sau orice miscare intensa
acumulare de trigliceride cunoscute fiind in literatura medicala sindroamele dislipidemice in cadrul carora trigliceridele, colesterolul sau alte particule lipidice cu potential patogen au un nivel crescut
Acumularile de drusen reprezinta mici depuneri galbene de proteine grase, adica lipide, care se acumuleaza sub retina
adenoame numite tipuri de polipi
Adenom de prostata inseamna marirea prostatei
Adenomioza este o afectiune in care mucoasa interioara a uterului se extinde in peretele muscular al uterului
Adenomioza boala inrudita cu endometrioza
ADHD este o boala ce prezinta componenta genetica
ADN-ul este format din sase parti: un glucid, un mineral (rest de acid fosforic) si patru substante chimice speciale numite baze
adrenalina cum ar fi produse chimice asociate cu stresul
adrenalina cum ar fi produsi chimici legate de stres
adrenalina (epinefrina)
Afazia este una dintre una dintre cele mai intalnite tulburari de vorbire
Afazia este una dintre cele mai intalnite tulburari de vorbire
Afazia optica este un sindrom sindrom rar in care pacientii nu pot sa numeasca obiecte prezentate vizual, dar nu au dificultati in denumirea acestor obiecte in urma unei prezentari tactile sau verbale
Afazia progresiva primara este termenul utilizat pentru dificultatea de vorbire care se dezvolta treptat
afectiune cum ar fi bolile de inima sau bolile pulmonare
afectiune afectiune numita fibrom uterin
afectiune a pielii cum ar fi acnee, psoriazis sau eczema
afectiune ale stomacului afectiune se numara aciditatea gastrica
afectiune cronica cum ar fi boli cardiace sau diabet
afectiune medicala cum ar fi afectiuni cardiace sau pulmonare
afectiune medicala pe termen lung cum ar fi diabetul, bolile de inima, afectiunile pulmonare, afectiunile renale sau o boala neurologica
afectiunea raynaud afectiune circulatorie
afectiunea raynaud caracterizata prin reducerea cantitatii de sange care ajunge la nivelul mainilor si picioarelor
afectiuni cum ar fi un accident vascular cerebral sau o tumora pe creier
afectiuni cum ar fi : BPOC, simptome gastrointestinale, boala Parkinson, otravire, astm, ameteala, rau de miscare
afectiuni cum ar fi cancerul biliar sau cancerul pancreatic
afectiuni cum ar fi : hipertensiunea arteriala, ateroscleroza, fibrilatia atriala si diabetul
afectiuni cum ar fi un accident vascular cerebral sau o tumora pe creier
afectiuni cum ar fi: BPOC, simptome gastrointestinale, boala Parkinson, otravire, astm, ameteala, rau de miscare
afectiuni cum ar fi TBC
afectiuni cum ar fi o infectie toracica
afectiuni cum ar fi diabetul sau anemia
afectiuni cum ar fi artrita, diabetul si chiar migrenele
afectiuni cum ar fi artrita reumatoida si diabetul
afectiuni cum ar fi hipertensiunea arteriala, ateroscleroza, fibrilatia atriala si diabetul
afectiuni cum ar fi boala arteriala periferica
afectiuni cum ar fi boala Hodgkin, leucemia si bolile insotite de imunosupresie
afectiuni (cum ar fi dializa)
afectiuni acute cum ar fi traumatismele
afectiuni acute cum ar fi traumatismele

afectiuni ale inimii cum ar fi infarctul
afectiuni ale pielii cum ar fi psoriazisul
afectiuni ale pielii cum ar fi eczema si dermatita seboreica
afectiuni ale pielii cum ar fi o eczema, lichenul plan sau psoriazisul
afectiuni asociate cum ar fi diabetul sau o tiroida subactiva
afectiuni asociate cu afazia cum ar fi boala Alzheimer (sau alte forme de dementa) sau accidentul vascular cerebral
afectiuni asociate cu afazia, cum ar fi boala Alzheimer (sau alte forme de dementa) sau accidentul vascular cerebral
afectiuni asociate cu obezitatea ; diabet de tip 2, hipertensiunea arteriala, apneea in somn, boala de reflux gastroesofagian
afectiuni autoimune cum ar fi: diabetul zaharat, artrita reumatoida
afectiuni benigne ale pielii cum ar fi negi
afectiuni cardiace cum ar fi bolile valvelor inimii si ritmurile cardiace neregulate
afectiuni cardiace cum ar fi bolile valvelor inimii si ritmurile cardiace neregulate
afectiuni cardiace congenitale (innascute)
afectiuni care ar putea contribui la bradicardie cum ar fi o infectie, hipotiroidism sau un dezechilibru electrolitic
afectiuni coloanei vertebrale cum ar fi hernia de disc, spondiloza si spurturile osoase
afectiuni congenitale cum ar fi sindromul Down si defecte cardiace
afectiuni cronice cum ar fi bolile reumatice
afectiuni cronice cum ar fi bolile reumatice
afectiuni cronice cum ar fi reumatoida si lupusul
afectiuni cronice cum ar fi artrita, guta si infectiile
afectiuni cutanate precum eczema
afectiuni de natura infectioasa cum ar fi patologiile cauzate de specii de stafilococi, streptococi sau alte bacterii
afectiuni gastro-intestinale cum ar fi ulcerul gastric si cancerul
afectiuni hepatice cum ar fi hepatita si alcoolismul cronic
afectiuni hepatice cum ar fi hepatita si alcoolismul cronic
Afectiuni hepatice: steatoza hepatica
afectiuni mai grave cum ar fi cancer pulmonar, insuficienta cardiaca, embolie pulmonara sau tuberculoza
afectiuni medicale cum ar fi spasmele pleoapelor si durerile de cap severe
afectiuni medicale cum ar fi spasmele pleoapelor si durerile de cap severe
afectiuni non-canceroase cum ar fi chisturile
afectiuni oculare cum ar fi cataracta, degenerarea maculara legata de inaintarea in varsta si glaucomul
afectiuni oculare cum ar fi cataracta, degenerarea maculara legata de inaintarea in varsta si glaucomul
Afectiuni ortopedice : platfus, slabiciune osoasa
afectiuni psihiatrice cum ar fi depresia sau autismul
afectiuni psihiatrice cum ar fi depresia sau autismul
afectiuni pulmonare incluzand emfizemul
afectiuni pulmonare, cum ar fi pneumotoraxul, embolia pulmonara sau cancerul de plamani
afectiuni subiacente cum ar fi guta, artrita sau diabetul
afectiuni vasculare cum ar fi boala cerebrovasculara (accidente vasculare cerebrale) si boala vasculara periferica
afectiunile cardiace, cum ar fi hipertensiunea arteriala, ateroscleroza si angina pectorala
afectiunile de baza cum ar fi artrita reumatoida sau guta
afectiunile extrapericardice si pericardice cum ar fi : infarctul miocardic acut, infarctul pulmonar, pneumotorax stang, pneumonie si pleurezie stanga, confundarea frecaturilor pericardice cu suflurile cardice
Afectiunile inflamatorii cum ar fi arterita Takayasu
Afectiunile inflamatorii, cum ar fi arterita Takayasu
afectiunile oculare cum ar fi ochii uscati, degenerarea nervilor optici, degenerarea maculei, defectele vizuale si cresterea sensibilitatii la infectii
Afectiunile oculare cum ar fi glaucomul
afectiunilor inflamatorii cum ar fi artrita si astmul
afectiunilor inflamatorii cum ar fi artrita si astmul
afectiunilor inflamatorii cum ar fi artrita si astmul
Aftele si alte infectii cu candida
Ageneza vaginala este o malformatie congenitala rara care se manifesta prin nedezvoltarea vaginului si a uterului
Ageneza vaginala cunoscuta si sub denumirea de aplazie mulleriana
agenti patogeni cum ar fi bacteriile responsabile pentru boala Lyme si virusul gripei H1N1
Agentii de maturare cum ar fi carbura de calciu
agentilor patogeni cum ar fi bacteriile
agentilor patogeni cum ar fi bacteriile
Aici sunt inclusi combustibilii, solventii (cum ar fi tetraclorura de carbon) si plumbul (si vopsele pe baza de plumb, tevi si materiale de lipit)
AINS cum ar fi aspirina sau ibuprofenul
AINS (medicamente antiinflamatoare nesteroidiene)
AINS cum ar fi ibuprofenul sau naproxenul sodic
AINS cum ar fi ibuprofen sau naproxen
alanin aminotransferaza (ALT sau TGO)

alanin aminotransferaza sau aspartat aminotransferaza cum ar fi analizelor de sange de rutina
alcalozei un dezechilibru al ph-ului, caracterizat prin cresterea anormala a alcalinitatii plasmei si a lichidelor intestinale
alcoolismul (consumul excesiv de alcool)
Alcoolismul este o boala cronica ce se caracterizeaza prin consumul excesiv si indelungat al alcoolului etilic
alcoolismul este o boala care se poate trata
Alcoolismul este o boala cronica ce se caracterizeaza prin consumul excesiv si indelungat al alcoolului etilic
aldosteron , hormon secretat de glandele suprarenale si care ajuta la echilibrarea nivelului de sodiului si apei din organism
alergarea sau tenisul cum ar fi sport care creeaza miscari extinse si repetitive
alergarea sau tenisul cum ar fi sport care creeaza miscari extinse si repetitive
alergenilor din aer (cum ar fi polenul plantelor sau sporii mucegaiului)
alergic (hipersensibil)
alergie cum ar fi rinita alergica (febra fanului), urticarie sau eczema
alergie cum ar fi - rinita alergica (febra fanului), urticarie sau eczema
alergie la alergenul respectiv inseamna ca pielea reactioneaza cu aparitia unei tumefactii pruriginoase
alergie la polen-alimentare cunoscuta sub numele de Aceste fructe contin o proteina care poate provoca simptome la persoanele sensibile la polen de mesteacan sau mere
alergie sau simptome comune cum ar fi stranutul, congestia nazala sau respiratia suieratoare
alergii cum ar fi polenul, praful, parul animalelor de companie sau […]
alergii sau alte alte boli sistemice
alergii sau astm bronsic alergic diferite tipuri de reactii alergice
Alergiile reprezinta o reactie de aparare exagerata a organismului deoarece factorul extern este perceput mai agresiv decat este de fapt
Alergiile reprezinta o reactie de aparare exagerata a organismului deoarece factorul extern este perceput mai agresiv decat este de fapt, iar manifestarea acestora nu tine cont de varsta
Alergiile alimentare, cele medicamentoase sau alte tipuri de reactii de hipersensibilitate severe
alfa-1-antitripsina numita proteina
alimentatia (dietele pentru slabit)
alimente cum ar fi carbohidratii sau grasimile
alimente bogate in acizi grasi Omega 3 cum ar fi tonul si somonul
alimente bogate in acizi grasi Omega 3 cum ar fi tonul si somonul
alimente bogate in acizi grasi omega-3, cum ar fi peste, nuci si unele uleiuri
alimente bogate in calciu cum ar fi produse lactate, legume bogate in calciu, conserve de sardine / somon, nuci si fructe bogate in calciu
alimente bogate in fier cum ar fi spanac, soia, mere, masline, caise uscate, sfecla rosie […]

6.14 Codes scripts en Perl et Python (APT et LSTM)

À consulter sur le site github

6.15 Interprétation de valeurs du score Kappa (McHugh, 2012)

Valeur de score Kappa	Niveau d'accord	% de données fiables
0 – .20	Pas d'accord	0 – 4%
.21 – .39	Très bas	5 – 15%
.40 – .59	Bas	16 – 35%
.60 – .79	Modéré	36 – 63%
.80 – .90	Fort	64 – 81%
.90 – .99,99	Presque parfait	82 – 100%

Université de Strasbourg

École doctorale des Humanités (ED520)
Université de Strasbourg

Ioana BUHNILA

Une méthode automatique de construction de corpus de reformulation

Résumé

Notre thèse a comme objectif la mise en place d'une méthode semi-automatique de construction des corpus de reformulations sous-phrastiques médicales, en français et en roumain. Nous définissons la reformulation sous-phrastique comme l'équivalence basée sur un noyau sémantique commun, située dans l'empan d'une phrase, qui contribue à la vulgarisation médicale. Notre méthode consiste, d'une part, dans l'exploitation des corpus comparables et des marqueurs pour identifier automatiquement des termes médicaux et leurs reformulations et, d'autre part, dans l'utilisation des architectures à base de réseaux de neurones pour la reconnaissance et la génération automatique de la reformulation. Nous avons construit le premier corpus de textes de vulgarisation médicale en roumain de grande taille, GrandMed-Ro2. Nous avons annoté manuellement et réalisé une analyse linguistique de 19 890 phrases (57% ont une double annotation). Les 11 653 paires de termes médicaux - reformulations validées constituent le corpus RefoMed. Nous évaluons la lisibilité des reformulations pour le grand public et nous analysons 11 314 prédictions de reformulations générées automatiquement.

Mots-clés : corpus de reformulation médicale ; annotation semi-automatique ; guide d'annotation de reformulations ; marqueurs de reformulation ; analyse lexicale et sémantico-pragmatique ; lisibilité des reformulations ; génération et classification automatique de reformulations

Abstract

The objective of our thesis is to set up a semi-automatic method for the construction of medical subphrastic paraphrase corpora in French and Romanian. We define the sub-phrastic reformulation as the equivalence based on a common semantic core, within a sentence, which contributes to the popularization of medical terms for lay people. Our method consists, on the one hand, in the exploitation of comparable corpora and markers to automatically identify medical terms and their paraphrases and, on the other hand, in the use of neural network architectures for the automatic recognition and generation of the paraphrase. We built the first large corpus of Romanian medical popularization texts, GrandMed-Ro2. We manually annotated and performed a linguistic analysis of 19,890 sentences (57% have a double annotation). The 11,653 validated medical term-paraphrase pairs constitute the RefoMed corpus. We evaluate the readability of the paraphrases for the general public and analyse 11,314 automatically generated paraphrase predictions.

Keywords: medical paraphrase corpus; semi-automatic annotation; paraphrase annotation guide; paraphrase markers; lexical and semantico-pragmatic analysis; readability of paraphrases; automatic generation and classification of paraphrases