
INALCO

Master II Ingénierie Multilingue Année 2014–2015

# MÉMOIRE DE STAGE

# Exploitation de reformulations pour l'acquisition d'un vocabulaire expert/non expert

Edwige Antoine

1er mai 2015 – 31 octobre 2015

Responsable de stage : Natalia Grabar
Responsable de formation : Jean-Michel Daube

# Table des matières

## Liste des tableaux



## Table des figures

Je tiens à offrir mes plus sincères remerciements à ceux qui m'ont aidée tout au long de cette formation.

Aux professeurs de Plurital et de l'Inalco qui m'ont beaucoup marquée, pour leurs compétences, leur disponibilité et leur passion de l'enseignement dans une ambiance toujours détendue.

A Natalia Grabar, pour son extrême gentillesse, sa confiance, sa disponibilité, la richesse de ses remarques et ses innombrables relectures.

A mon conjoint, qui a su me supporter durant toutes ces années.

A mes très chers collègues de master, qui eux seuls savent mes angoisses et à qui je dois ces fous rires quotidiens inoubliables.

A ma sœur et mes amis proches, qui ont toujours été là pour moi, et m'ont soutenue à leur façon, avec leurs mots et leur présence.

A ceux qui ne sont plus là pour partager ces moments avec moi, dont j'ai hérité la persevérance et la détermination qui m'ont permise de poursuivre ces années d'études malgré tout, et qui m'ont toujours soutenue malgré mon absence. A mes parents.

# 1 Introduction

Quel que soit le domaine de spécialité, la communication entre une personne profane et les experts peut s'avérer difficile, du fait que le langage spécialisé n'est pas toujours partagé par ces deux types d'interlocuteurs. Le domaine médical n'échappe pas à cette règle : l'incompréhension du discours médical par les patients (ou les personnes proches de patients) n'est pas rare et peut perdre le patient dans un flux d'informations opaques et inaccessibles. Ceci l'amène souvent à l'incapacité de prendre une décision face à un traitement, à comprendre les conséquences de la maladie et du traitement sur son quotidien, ou tout simplement à comprendre sa maladie (Deléger & Zweigenbaum, 2008). Aujourd'hui, cette constatation devient encore plus importante du fait que les patients ont un accès accru aux informations en ligne, et peuvent les consulter par leurs propres moyens (Zeng & Tse, 2006). De manière paradoxale, la disponibilité des informations agrandit et accentue la barrière de l'incompréhension entre le langage des patients, qui est généralement le langage utilisé au quotidien, et le jargon du milieu médical. Cette situation est due à la différence dans les connaissances du domaine que présentent ces deux catégories de personnes (patients et médecins). Ainsi, un patient, étant non expert du milieu médical, ne peut pas avoir les connaissances nécessaires pour comprendre parfaitement les informations données par le corps médical, ni même pour rechercher les informations dont il a besoin. En effet, le processus de recherche est complexe, qu'il se fasse sur internet, dans la presse ou ailleurs : trouver une information implique la rencontre de nouveaux concepts, qui eux-mêmes peuvent nécessiter une nouvelle recherche, et ainsi de suite (Boubé & Tricot, 2010) ; le patient n'est pas face à un médecin pour interagir et obtenir des réponses comme cela se passe dans une situation de dialogue (Vergely *et al.*, 2009). En effet, s'il y a bien deux types de discours (celui des patients et celui des médecins), au cours d'une conversation, nous pouvons observer une "mise en commun" lors d'une demande de reformulation, de répétition, de clarification de la part du patient en cas d'incompréhension (Zeng & Tse, 2006). Notons cependant que, face à un médecin, le patient ne demande pas toujours les informations dont il a besoin. Souvent, pour le patient, il s'avère plus acceptable et facile de consulter les sources disponibles sur internet, ou plus précisément sur les forums, pour récupérer ces informations. En effet, d'après une enquête Solucient [1], 45% des patients tendent à utiliser internet, et seulement 16% se réfèrent à leur médecin (Zielstorff, 2003). La communication entre ces deux catégories de personnes peut en effet être compliquée.

## 1.1 Alphabétisation médicale

La compréhension des informations médicales s'effectue grâce à deux facteurs : l'environnement et les connaissances du patient, partagées ou non avec les professionnels de santé (Zeng *et al.*, 2005a). De ce fait, on parle d'*alphabétisation médicale* (ou *health literacy*) définie comme *"the degree to which individuals have the capacity to obtain, process, and understand basic health information and services needed to make appropriate*

1. http ://www.bloomberg.com/Research/stocks/private/snapshot.asp ?privcapId=1318959

*health decisions"* (Ratzan & Parker, 2000). Il s'agit en particulier de la compréhension d'une information médicale, de sa lecture et interprétation, ou bien de la capacité de pouvoir rechercher les informations d'un domaine. Il peut s'agir par exemple de savoir choisir les bons mots-clés en interrogeant un moteur de recherche pour optimiser le résultat. L'alphabétisation médicale est alors différente d'un individu à l'autre, et dépend de l'éducation et de la catégorie socio-professionnelle d'une personne donnée, mais aussi de son histoire médicale (Zeng *et al.*, 2005a). Un patient possède son vocabulaire qui lui est propre, répondant à un choix lexical, une structure logique, à des concepts, tout comme le discours médical a aussi sa propre structure. En effet, un patient atteint d'une maladie précise a souvent une meilleure compréhension des informations médicales relatives à cette maladie : grâce à la maîtrise de certaines notions propres à cette atteinte, du fait de les entendre systématiquement lors des entretiens avec le corps médical, ou de les rencontrer régulièrement au cours de diverses recherches.

Notons aussi que Zeng & Tse (2006) précisent que le vocabulaire profane médical est *dynamique*, c'est-à-dire qu'il est aussi instable qu'un langage à proprement parler et varie dans le temps et en fonction des personnes. Cette instabilité pose un problème fondamental dans la conception d'un vocabulaire profane médical, car les connaissances des patients évoluent. Cette problématique est aussi valable du côté du patient : ses connaissances doivent évoluer aussi, car il doit et peut assimiler des informations nouvelles entre deux consultations.

## 1.2 Mise à jour des connaissances des patients

A travers ses recherches, le patient tente de combler son manque de connaissances : il le fait, partiellement ou complètement, grâce aux notions qu'il ne perçoit sans doute pas de la même manière que les médecins. Le terme de *dépression*, par exemple, considéré comme une véritable maladie chez les médecins, peut être utilisé par un non expert pour signifier un état de tristesse plus ou moins important, avec une notion de gravité bien moindre (Zeng & Tse, 2006). En effet, si nous regardons une terminologie médicale (UMLS, voir section 4.1.1, le terme *dépression* y apparaît deux fois : *C0344315.T033.DISO dépression* et *C0344315.T033.DISO dépression*. Dans cette notation, le *C* introduit un identifiant du concept, le *T* introduit un identifiant du type sémantique, *DISO* marque le groupe sémantique et correspond ici à *disorder* (maladie). Par ailleurs, il apparaît de nombreuses fois au sein d'autres termes plus spécifiques, tels que : *C1269683.T048.DISO dépression majeure*, ou *C0027804.T048.DISO dépression névrotique*. Ainsi, ce terme, bien familier dans le langage quotidien, est utilisé par les profanes avec une sémantique moins subtile que celle des experts. Nous pouvons aussi prendre comme exemple le terme *paranoïa* (que l'on trouve dans l'UMLS sous différents concepts : *C1456784.T048.DISO paranoïa*, *C0011251.T048.DISO paranoïa* ou *C0152119.T048.DISO paranoïa alcoolique*), qui est une véritable maladie mentale connue des psychiatres, tandis que ce terme montre souvent une utilisation plus ou moins péjorative par un non expert. Dans ce sens, on parle souvent du processus de *traduction* (Zeng & Tse, 2006) qu'un non expert effectue cognitivement. Autrement dit, il traduit un terme technique vers une expression plus facilement compréhensible pour lui selon

son degré de connaissances : *taux de sucre* pour le terme expert *glycémie*, etc. Le patient peut aussi garder le même terme avec une signification au niveau de ses connaissances.

Lors de cette *traduction*, le patient risque de perdre alors une partie des informations que le médecin essaie de lui transmettre. De ce fait, le non expert utilise généralement des termes plus génériques, moins précis que ceux utilisés par un expert. Il peut aussi utiliser un terme sans connaître sa signification précise ou bien en lui associant une signification spécifique, comme avec l'exemple de *depression*. Dans ce même article, Zeng & Tse (2006) précise également que les dictionnaires généraux proposent des significations "standards" des termes médicaux, c'est-à-dire génériques et peu précises, proches des définitions que s'en font les patients, contrairement aux définitions techniques sous lesquelles les médecins entendent ces mêmes termes. L'auteure prend l'exemple du terme *surpoids*, défini par un dictionnaire de la langue anglaise comme *plus de poids que la normale, nécessaire ou autorisé*. Pour ce même terme, l'UMLS propose *index de masse corporelle entre 25 et 30*, alors que *obésité* est définie comme *index de masse corporelle supérieure à 30*.

Notre travail consistera à analyser les reformulations des termes médicaux dans le discours médical, afin d'obtenir leurs paraphrases mieux accessibles et compréhensibles pour les non experts.

# 2 État de l'art

Nous nous intéressons aux notions de paraphrases (section 2.1) et de reformulations (section 2.2). Nous présentons ensuite des études effectuées sur l'extraction de paraphrases (section 2.3), ainsi qu'à quelques types de corpus utilisés dans ce type de travaux (section 2.4).

## 2.1 Notion de paraphrase

La paraphrase est la notion qui va nous suivre tout au long de ce travail. C'est une notion très utilisée dans plusieurs domaines de recherche en TAL (la recherche d'information, la génération de texte, la synthèse de texte, la didactique des langues), mais aussi tout simplement dans le langage quotidien, comme les écrits journalistiques ou la langue orale. L'utilisation de la paraphrase peut alors être consciente ou non. Cependant, malgré son omniprésence, c'est une notion complexe et difficile à définir (Bhagat & Hovy, 2013) ; il faut expliciter les frontières de la paraphrase, préciser de quel type de paraphrase il s'agit, différencier les notions de synonymie, d'équivalence, ou encore des relations pragmatiques. En effet, il existe divers types de paraphrases, et c'est ce point qui la rend difficile à définir et à délimiter.

Globalement, la paraphrase permet d'exprimer un même concept par des moyens linguistiques différents. Autrement dit, les paraphrases ont une similarité sémantique, mais une différence de forme :

- syntaxique (catégorie syntaxique, agencement des mots) : *endomètre* → *muqueuse de l'utérus* ; nous avons une transformation syntaxique et morphologique (base

supplétive/mot français) ;
- sémantique avec la généralisation du sens : *benzodiazépine → un tranquillisant* ;
- différence de registre expert/non expert : *réactions immunitaires → réactions de défense.*

La difficulté pour définir la paraphrase en est une aussi, par conséquent, pour le TAL ; cette difficulté réside notamment dans les frontières de segments à relever. Cependant, il a été démontré que ces frontières ont tout de même une structure relativement fixe, et de ce fait, des classifications ont été proposées dans la littérature (Bhagat & Hovy, 2013). Il est évident que lors d'une étude, il est important de s'en tenir à quelques types de paraphrase seulement, car il est compliqué de proposer la même méthode pour traiter tous les types de paraphrases à la fois. En effet, certaines paraphrases ont un rôle ou une structure spécifiques. Voici les types de paraphrases proposés (Bhagat & Hovy, 2013) :

- substitution par synonymie : *Google a acheté Youtube → Google a acquis Youtube*
- substitution par antonymie : *Pat a mangé → Pat n'a pas faim*
- substitution conversationnelle (qui utilise l'inversion des éléments et le contexte) : *Google a acheté Youtube → Youtube a été vendu à Google*
- changement de voie : *Christie aime Pat → Pat est aimé par Christie*
- changement de personne : *Pat a dit j'aime le foot → Pat a dit qu'il aimait le foot*
- substitution par co-référent ou par un pronom : *Pat aime Christie car elle est intelligente → Pat aime Christie car Christie est intelligente*
- répétition, ellipse : *Pat peut courir vite et Christie peut courir vite aussi → Pat peut courir vite et Christie aussi*
- variation de fonction des mots : *Pat a montré une belle démonstration → la démonstration de Pat était belle*
- substitution acteur/action : *je n'aime pas les conducteurs imprudents → je n'aime pas la conduite imprudente*
- substitution verbe/nom : *Pat enseigne à Christie → Pat est l'enseignant de Christie*
- substitution manipulateur/technique : *le pilote a décollé malgré la tempête → l'avion a décollé malgré la tempête*
- substitution générale ou spécifique : généralisation ou précision d'un concept : *je n'aime pas les conducteurs imprudents → je n'aime pas les automobilistes imprudents*
- substitution par métaphore : *j'ai du conduire à travers le brouillard aujourd'hui → j'ai du conduire à travers un mur de brouillard aujourd'hui*
- substitution partie/tout : *les avions américains ont livré une défense → l'armée de l'air américaine a livré une défense*
- changement de catégorie (verbe/nom, adj, adv ; nom/adj ; : *Pat aime Christie → Pat est amoureux de Christie*
- substitution verbe-préposition/nom : *les finalistes joueront dans le Giants stade → le Giants stade sera le terrain pour les finalistes*
- changement de temps : *Pat aime Christie → Pat aimait Christie*

– changement d'aspect : *Pat arrive → Pat arrive pour aujourd'hui*
– changement de modalité : *Google doit acheter Youtube → Google a acheté Youtube*
– implication sémantique : *Google a négocié l'achat de Youtube → Google à acheté Youtube*
– équivalence numérique approximative : *Disneyland est à 40 km d'ici → Disneyland est à 30 minutes d'ici*
– substitution par connaissances du monde : *le gouvernement a déclaré la victoire en Irak → Bush a déclaré la victoire en Irak*

Comme nous le présentons dans le section 2.3, la paraphrase est l'objet de plusieurs travaux en TAL, où l'objectif est de proposer des méthodes automatiques de reconnaissance de paraphrases.

## 2.2 Notion de reformulation

La reformulation est l'action de redire à nouveau quelque chose qui a déjà été dit (Bot *et al.*, 2008), mais d'une manière différente, augmentant ainsi les chances d'être compris. La reformulation peut être effectuée à la demande d'un interlocuteur, ou sous décision du locuteur lui-même, conscient qu'un terme puisse être difficile à comprendre par son interlocuteur (ou lecteur). La reformulation peut se manifester par des schémas, parfois liées par un marqueur tel que *c'est-à-dire*, *autrement dit*, l'utilisation de parenthèses, etc. De ce fait, ces marqueurs peuvent indiquer la présence d'une reformulation. La reformulation n'est pas toujours paraphrastique car les segments liés par les marqueurs peuvent aussi marquer une incise, des disfluences, des relation de causalité, etc. Au cas où la reformulation est paraphrastique, le marqueur apparaît au sein d'une paire de paraphrases et permet ainsi de détecter cette paraphrase, qu'il serait difficile de déceler autrement du fait de sa différence structurelle (morpho-syntaxe, fonctions, ponctuation, etc) et de registre langagier (Gulich & Kotschi, 1983).

## 2.3 Travaux existants en détection automatique de paraphrases

Diverses méthodes automatiques de reconnaissance de paraphrases ont été proposées. Elles peuvent être appliquées aux données de la langue générale ou spécialisée. Nous proposons ici quelques travaux.

McCray *et al.* (1999) ont proposé un traducteur automatique de termes médicaux vers des expressions profanes et inversement : le MEDLINE*plus*, se trouvant sur le site de la NLM (National Library of Medecine [2]). Présenté sous forme de moteur de recherche, il est organisé selon divers sujets médicaux (cancers, obésité, etc), et fait le lien avec différentes ressources médicales telles que le MeSH (Medical Subject Headings (NLM, 2001)), le AIRS (AIRS Taxonomy of Human Services [3]) ou des sites médicaux, pour assister les personnes désirant trouver une information médicale sur le net et proposer une traduction du discours source au discours cible. Le moteur de recherche se veut "souple", c'est-à-dire qu'il est capable de pardonner une erreur de la part de l'utilisateur.

2. http ://www.nlm.nih.gov/
3. https ://211taxonomy.org/

Par exemple, une personne tapant le terme *nurologue* se verra proposer des termes tels que *neurologue*, ou *urologue*.

Schwartz & Hearst (2003) ont décrit un algorithme simple de détection d'abréviations et de leur définition, travail dont nous nous sommes inspirée pour nos extractions (section 5.2.1). L'algorithme repère les formes situées entre parenthèses, car elles indiquent la présence d'une abréviation ou de sa définition. Les auteurs traitent ainsi deux cas de figure :

- abréviation (sa définition)
- définition d'une abréviation (abréviation correspondante)

L'algorithme reconnaît le type de patron selon que les parenthèses contiennent un mot isolé, auquel cas c'est le patron *définition (abréviation)*, ou plusieurs mots, qui correspond au patron *abréviation (définition)*. Si l'abréviation est entre les parenthèses, chaque lettre de celle-ci est enregistrée, puis les termes situés avant les parenthèses sont parcourus à l'envers : si la lettre des termes correspond à une lettre de l'abréviation, le terme est enregistré. Le principe est identique si c'est la définition qui se situe entre les parenthèses : chaque première lettre des termes est enregistrée, et le terme situé avant les parenthèses, considéré comme l'abréviation, est parcouru lettre par lettre. Ainsi, chaque lettre est mise en correspondance avec le terme de la définition.

Zeng *et al.* (2005a) utilisent la mesure de familiarité entre des termes liés à la médecine, dans un jargon médical et dans un langage des patients, à l'aide de questionnaires à choix multiples soumis à des volontaires non experts. Dans les données étudiées, chaque concept possède deux synonymes : un provenant du discours patient et un autre provenant du jargon médical. Sur cette base, deux versions du questionnaire sont générées : une utilisant les synonymes du discours non expert, et l'autre utilisant les synonymes du discours technique. Les participants non experts doivent alors indiquer quelle est la sémantique des termes. À la figure 1, nous présentons un exemple d'une question dans les deux discours : la version A contient les synonymes experts, tandis que la version B contient les synonymes non experts. Les auteurs utilisent comme *baseline* les termes qui ont plus de 3 syllabes ou qui ne sont pas présents dans la Dale-Chall List (liste d'environ 3000 mots fréquents (Dale & Chall, 1948)). En effet, selon les auteurs, ces termes sont potentiellement difficiles à comprendre : comme ils ont une longueur assez importante et ne font pas partie des mots les plus fréquents de la langue, cela peut opacifier leur sémantique.

Zeng *et al.* (2006) ont développé un lexique patient (le Consumer Health Vocabulary ou CHV) afin de favoriser la recherche d'information pour les patients et de diminuer la différence entre les discours des patients et des médecins. À terme, l'initiative a pour but d'identifier les expressions profanes évoquant un terme médical, de lier ces expressions aux termes qu'elles sous-entendent, et d'intégrer les connaissances à partir de diverses disciplines médicales. Les premières expériences ont commencé par identifier les expressions profanes en se fondant sur les formes de surface (syntagmes, mots). Les auteurs ont observé trois types d'expressions :

- les CFD (Consumers-Friendly Display) : les termes compris par le grand public (Zeng *et al.*, 2005b),

```
Version A:
1. A geriatric person is one who is ________.
A. Very old
B. lanky and good looking
C. well groomed
D. aggressive and loud

Version B:
1. An elderly person is one who is ________.
A. Very old
B. lanky and good looking
C. well groomed
D. aggressive and loud
```

Fig. 1 – Exemple de questionnaires présentés aux participants (Zeng *et al.*, 2005a).

- les mots communs aux profanes et médecins, mais ayant une distance sémantique significative à cause d'une compréhension partielle, comme l'illustrent les exemples avec les termes *dépression* et *paranoïa* (comme présenté dans la section 1.2),
- les expressions profanes inexistantes dans le vocabulaire médical.

Concernant les travaux effectués sur le français, Deléger & Zweigenbaum (2008) ont travaillé avec un corpus monolingue comparable pour étudier des types de syntagmes utilisés par les experts et les non experts. Après avoir effectué un alignement en phrases des corpus, les auteurs ont utilisé des indicateurs statistiques pour observer les fréquences des mots et des catégories syntaxiques. Les auteurs ont utilisé des patrons morphosyntaxiques afin de retrouver les phrases nominales dans le corpus des médecins, puis les phrases verbales dans le corpus des patients. Ensuite, la mesure de similarité cosinus est calculée afin de détecter les paires de paraphrases sémantiquement proches du point de vue distributionnel, c'est-à-dire, les paires de syntagmes contenant les patrons syntaxiques recherchés : $N\ ADJ$ et $V\ ADV$. Les auteurs ont montré que les médecins ont plutôt recours à des syntagmes nominaux (comme *traitement*) là où, pour des termes sémantiquement proches, les non experts utilisent des syntagmes verbaux (comme *traiter*).

Cartoni & Deléger (2011) exploitent des n-grammes lexicaux extraits à partir de corpus médicaux comparables expert/non expert en français. La longueur des n-grammes varie entre 2 et 6 mots, dont les auteurs ne gardent que ceux ayant une cohérence syntaxique, c'est-à-dire que les n-grammes doivent comporter au moins 2 mots pleins, ne doivent pas comporter de ponctuation, ou se terminer par un déterminants ou une préposition. Les n-grammes sont racinisés dans chaque corpus (expert et non expert) et mis en correspondance entre ces deux corpus avec la méthode de sacs de mots. Les auteurs ne gardent que les n-grammes qui comportent des racines correspondantes

pour en faire leurs paraphrases candidates, comme dans cet exemple *consommation régulière*/*consommer de façon régulière*. Par la suite, les auteurs ont généralisé la méthode grâce à l'exploitation de parties du discours et de patrons spécifiques, tout en gardant les liens entre les correspondances. Voici un exemple de schéma lexical avec sa généralisation en patron syntaxique :

expert → non expert
*Traitement du patient → traiter un patient*
*N1 Prep N2 → V1 Det N2*

Les patrons sont filtrés à travers leur fréquence (le nombre de paraphrases qu'ils représentent) et leur pertinence (définie à l'aide d'annotateurs) dans les deux types de discours. Ils ont ensuite été classés selon le type de transformation : morphosémantique (dérivation), inversion simple, flexion verbale et variation zéro.

Bouamor *et al.* (2012) utilise différentes techniques de reconnaissances de paraphrases, afin de profiter de leur complémentarité, sur des corpus monolingues parallèles :

- L'apprentissage statistique d'alignement de mots, en proposant à Giza++ (Och & Ney, 2003) des paires de paraphrases possibles ;
- La description de la variation lexicale, grâce au logiciel Fastr (Jacquemin, 1994), qui, à l'aide de patrons de réécriture morphosyntaxiques et de variations lexicales, repère des variantes lexicales telles que *protéger de façon permanente = protection constante*, où *protéger* et *protection* sont formés sur la même racine ; *de façon permanente* et *constante* ont une sémantique identique mais forment une variation lexicale. Bouamor *et al.* (2012) utilisent le système dans les deux sens : recherche de segments paraphrastiques de la première phrase dans une seconde phrase, et inversement. Sont retenus comme résultats l'intersection des deux recherches ;
- La structure syntaxique des énoncés, à partir de l'algorithme de (Pang *et al.*, 2003), qui permet de donner les arbres syntaxiques d'une paire de paraphrases et les fusionne lorsque leurs catégories filles sont identiques. L'algorithme a été amélioré grâce à l'analyseur probabiliste de Berkeley (Petrov & Klein, 2006), en utilisant les $k$ meilleurs résultats ;
- Le calcul de la transformation des mots, utilisant la mesure TER p (Translation Edit Rate plus) pour calculer la distance entre une hypothèse de traduction et une traduction de référence. Les opérations acceptées sont l'insertion, la suppression, la substitution de mots, le déplacement et la substitution de segments. Dans l'exemple qui suit, nous observons une substitution de segments : *Ce dégrèvement → Cet allongement*, une substitution lexicale : *équivaut → revient*, et donc une paire de paraphrases : *Ce dégrèvement fiscal équivaut → Cet allongement fiscal revient* ;
- L'exploitation des équivalences de traduction par langue pivot, se servant de la probabilité de paraphrasage entre deux segments (Bannard & Callison-Burch, 2005). Un premier segment est traduit dans une langue pivot, puis retraduit vers la langue d'origine. Bouamor *et al.* (2012) ont utilisé cette technique sur le corpus des débats parlementaires Europarl en anglais et en français, préalablement traduit par MOSES (Koehn *et al.*, 2007) et aligné en mots par Giza++ .

Les paraphrases récupérées à l'aide de ces cinq techniques ont été soumises à une classification par apprentissage automatique à partir de traits linguistiques et statistiques.

Grabar & Hamon (2015) ont proposé une méthode pour la détection de paraphrases pour les composés médicaux néoclassiques, c'est-à-dire formés sur des racines grecques et/ou latines. La méthode exploite l'analyse morphologique de Derif (Namer, 2009), logiciel qui segment les mots en morphèmes et pose des étiquettes sur ces morphèmes :

*myocardique/A : [[[myo N*] [carde N*] NOM] ique ADJ]*

Une fois l'analyse morphologique effectuée, une traduction des racines néoclassiques est proposée, comme par exemple :

*myocardiaque : myo=muscle, cardia=cœur*

L'utilisation de corpus français monolingues, étiquetés morphologiquement et syntaxiquement permet d'extraire et d'aligner les syntagmes, qui contiennent les mots de la décomposition des termes néoclassiques, avec ces termes. Par exemple, pour *myocarde* ou *myocardique*, le syntagme *muscle du cœur* est proposé. Cette méthode vise à créer un vocabulaire en français pour faciliter la compréhension de termes médicaux complexes, ces derniers étant essentiellement fondés sur des racines grecques et latines.

## 2.4 Types de corpus

Il existe plusieurs types de corpus sur lesquels les études concernant l'extraction de la paraphrase se sont appuyées (Madnani & Dorr, 2010) :

- Les corpus monolingues simples, souvent utilisés avec des méthodes de similarité distributionnelle pour l'extraction de paraphrases ;
- Les corpus monolingues comparables, c'est-à-dire deux corpus évoquant le même sujet mais proposant une structure ou un discours différent, dans lesquels nous pouvons trouver des formes sémantiquement proches au niveau du document, non des phrases elles-mêmes. Par exemple, un corpus non expert et un corpus expert traitant du même sujet. La difficulté de détection de paraphrases est cependant accrue ;
- Les corpus monolingues parallèles, c'est-à-dire deux corpus très semblables sur le même sujet, contenant des paires ou listes de phrases sémantiquement proches ou équivalentes. Ces corpus peuvent être issus de traduction ;
- Les corpus bilingues parallèles, c'est-à-dire un corpus et une traduction dans une langue cible, généralement générée par la traduction automatique statistique à l'aide d'alignement (mots, phrases). C'est cet alignement qui se révèle précieux pour la détection de paraphrases, puisque les deux corpus contiennent des mots et expressions sémantiquement équivalents. Notons que la traduction est aussi une méthode de détection de paraphrase à elle-même, car lister les différentes traductions possibles d'une phrase d'une langue source à une langue cible permet d'obtenir une liste de phrases sémantiquement équivalentes.

La plupart des études existantes sont des études contrastives de la langue des patients et des médecins. Elles œuvrent sur des corpus comparables experts/non expert, du fait d'une plus grande disponibilité de ceux-ci, et afin d'avoir des données comparables entre

un discours non expert traitant de la médecine et un discours médical. Cependant, la difficulté à obtenir de tels corpus peut freiner les travaux de recherche sur cette thématique.

# 3 Objectifs

Notre objectif est d'acquérir un vocabulaire pour expliquer et faciliter la compréhension de termes médicaux. De ce fait, nous allons travailler sur la reformulation dans le discours médical et exploiter un corpus médical rédigé par des spécialistes, afin de garantir une plus grande fiabilité des extractions. Cela nous permet d'obtenir des couples de syntagmes appartenant aux langages expert et non expert, et ayant une équivalence sémantique. Notre travail est basé sur différentes hypothèses :

- Lorsqu'un professionnel de santé reformule un concept (mot ou syntagme), cela indique qu'il s'agit d'une expression potentiellement inconnue du public non expert ;
- L'acte de reformulation d'une notion par le professionnel médical nous permet d'extraire le terme et sa reformulation.

Les paraphrases extraites sont, pour la plupart, formellement très éloignées. Nous classifions ensuite ces paraphrases en nous inspirant d'une classification existante (Bhagat & Hovy, 2013), et en faisons une discussion. L'objectif principal est donc de mettre en place un extracteur de reformulations, en utilisant l'information syntaxique, afin d'extraire des couples de paraphrases et expliciter ainsi certains concepts techniques.

# 4 Matériel

Nous utilisons trois types de matériel : les terminologies médicales (section 4.1), les corpus (section 4.2) et les ressources linguistiques (section 4.3).

## 4.1 Les terminologies médicales

Il existe plusieurs terminologies, ontologies ou nomenclatures du domaine médical, de même que des outils qui permettent de manipuler ces ressources terminologiques. Les terminologies ont pour objectif de décrire le domaine médical et sont généralement créées pour des professionnels de santé. Relativement à notre objet d'étude, ces terminologies sont donc peu compréhensibles pour un public profane. En revanche, ces terminologies représentent une bonne base pour effectuer des travaux de recherche pour l'acquisition du vocabulaire expert/non expert.

### 4.1.1 UMLS, Unified Medical Language System

L'UMLS (Unified Medical Language System) (Lindberg *et al.*, 1993) est un ensemble de ressources terminologiques et de programmes permettant l'interopérabilité entre les terminologies. Il est multilingue, mais repose principalement sur l'anglais. Seuls 2% des

termes sont issus du français (Delbecque *et al.*, 2005). L'UMLS s'appuie sur trois sources de connaissances :

- le Metathesaurus, qui propose des termes et une codification de ceux-ci provenant de plusieurs vocabulaires médicaux (ICD-10-CM, LOINC®, MeSH®, RxNorm, and SNOMED CT®) et juridique (CPT®) ;
- un réseau sémantique, qui permet de montrer les catégories sémantiques et leurs relations ;
- des outils de TAL autour de ce lexique spécialisé.

Les termes sont groupés au sein de plus de 1 million concepts. Chaque concept a un identifiant unique appelé CUI (Concept Unique Indentifier), et les CUI sont assignés aux types et groupes sémantiques. Il existe 134 types sémantiques et 54 relations sémantiques (Delbecque *et al.*, 2005). Cette structure forme une hiérarchie : par exemple, le type sémantique *Signs or Symptoms* permet de relier tous les symptômes d'UMLS. De façon plus précise, cela peut conduire à la relation sémantique *diagnostiquer*. L'UMLS peut être utilisé pour le développement d'applications, pour la *traduction* entre différents types de terminologies. Il est librement disponible sous licence individuelle. Il contient un vocabulaire appelé le MedlinePlus, disponible en anglais uniquement (McCray, 1989). La NLM (National Library of Medecine) assure le maintien de cette ressource.

### 4.1.2 SNOMED Int (Systematized Nomenclature of Medicine)

La terminologie SNOMED International (Systematized Nomenclature of Medicine) est une nomenclature du domaine médical créée aux États-Unis par le College of American Pathologists (Côté *et al.*, 1993). Elle existe en plusieurs langues, notamment en français. Elle permet d'offrir un standard terminologique pour l'échange électronique d'informations médicales. La Snomed est structurée en plusieurs axes sémantiques représentant les catégories (métiers, morphologie, chimie biomédicale, etc), avec des liens entre les concepts. La Snomed Int en français est disponible sous licence. Elle est diffusée par ASIP santé [4].

## 4.2 Corpus

Nous exploitons deux corpus : un corpus pour le développement sur lequel la méthode est définie et réglée, et un corpus de test, qui permet de tester la méthode sur un grand volume de données. Les deux corpus sont disponibles au format texte brut et avec l'étiquetage morpho-syntaxique par Cordial (Laurent *et al.*, 2009).

### 4.2.1 Le corpus de développement

Le corpus de développement est issu d'un forum de santé *masante.net* [5], créé et modéré par des médecins. Ce site permet aux utilisateurs de poser des questions médicales, auxquelles deux professionnels de santé répondent systématiquement. Le site propose

4. http ://esante.gouv.fr/services/referentiels/referentiels-d-interoperabilite/snomed-35vf
5. masantenet.com

une catégorisation par spécialité médicale, mais il est possible d'accéder à l'ensemble du forum. Au-delà de cette activité de questions-réponses, le site propose aussi des articles variés autour de la médecine (obésité, diabète, etc). Dans notre travail, nous exploitons la partie dédiée aux questions et réponses (le forum). Ce corpus a été constitué dans le cadre du projet Patient's mind [6] porté par l'équipe LIRMM à Montpellier (Abdaoui *et al.*, 2014). C'est un corpus monolingue en français et comparable car il met en parallèle un discours non expert (questions des patients) avec le discours des professionnels de médecine (réponses).

Pour les objectifs de notre travail, nous nous intéressons uniquement à la partie des réponses (texte produit par les professionnels de médecine), afin de travailler sur la reformulation dans le discours médical. Nous supposons en effet que les professionnels de santé essaient d'expliquer les termes médicaux potentiellement compliqués lorsqu'ils les utilisent dans leurs réponses.

Le corpus contient 6 139 lignes, totalisant 315 362 mots. Une réponse type répond au schéma suivant :

> *cher(e) prénom,*
> *texte de la réponse*
> *Bien cordialement.*
> *Ceci n'est pas une consultation médicale et n'a pas pour objet de la remplacer.*

Ce corpus, disponible en ligne, n'est pas anonymisé et peut contenir les pseudonymes et les prénoms des utilisateurs.

#### 4.2.2 Le corpus de test

Le corpus de test est constitué d'articles de la Wikipédia liés au Portail de la Médecine (version de janvier 2015). Ce corpus contient 18 434 articles (15 235 219 occurrences). Il a été créé dans le cadre d'une étude précédente (Grabar & Hamon, 2015).

Le corpus contient des informations encyclopédiques sur plusieurs notions médicales. Les contributeurs ont en général une bonne connaissance des sujets abordés. L'objectif est entre autre de présenter les notions techniques et de les rendre accessibles au grand public. Nous supposons alors que le corpus contient également des reformulations.

### 4.3 Ressources linguistiques

#### 4.3.1 Mots vides

Les mots vides correspondent aux mots qui ne sont pas significatifs dans l'alignement des termes, autrement dit, les mots grammaticaux tels que *de, et, à, ou, etc*, les auxiliaire *est, a* quelque soit leur forme temporelle, et certains adverbes *tout, plusieurs*. Cette liste contient 111 mots vides, et a été adaptée à notre corpus pour mieux traiter les données.

6. https ://www.lirmm.fr/patient-mind/pmwiki/pmwiki.php ?n=Site.Accueil

### 4.3.2 Ressources morphologiques

Les ressources morphologiques comportent 163 823 paires de mots couvrant les dérivations {*aorte*; *aortique*} et les flexions {*aortique*; *aortiques*}. Elles sont issues des travaux précédents (Grabar & Zweigenbaum, 2000 ; Zweigenbaum & Grabar, 2003) et ont été complétées à partir des corpus de notre travail. Ces ressources permettent d'assurer une normalisation lors de l'appariement des expressions extraites du corpus et des termes des terminologies. Leur utilisation est décrite dans la section 5.3.

# 5 Méthodes

L'objectif de notre travail est de détecter un concept médical et sa reformulation, afin de construire un vocabulaire associant les termes techniques avec leurs équivalents grand public. Nous proposons de traiter trois cas de figures :

1. la détection des siglaisons et de leurs formes étendues (section 5.2.1). L'hypothèse est que les sigles présentent des difficultés de compréhension et que leurs formes étendues peuvent faciliter cette compréhension ;
2. la détection des reformulations avec trois marqueurs de reformulation *c'est-à-dire, autrement dit, encore appelé(e)(s)* (section 5.2.2). L'hypothèse est que l'emploi de ces marqueurs par le personnel médical signifie que les notions médicales correspondantes nécessitent une explication. Ces marqueurs sont aussi les déclencheurs dans la recherche de la relation de paraphrase ;
3. la détection des reformulations (hors siglaisons) grâce aux parenthèses (section 5.2.3). Les parenthèses, bien que d'une sémantique plus lâche, permettent de détecter d'autres reformulations.

Ces trois méthodes ont été mises en place grâce à une étude manuelle du corpus de développement. Ainsi, nous avons commencé par relever manuellement une partie des reformulations et les avons analysées afin de définir les schémas récurrents qui se trouvent à la base de trois méthodes :

1. *abréviation (forme étendue)*, *forme étendue (abréviation)*
2. *concept marqueur reformulation*
3. *concept (reformulation)*

Le schéma 2 présente le schéma général de notre approche. Nous décrivons d'abord le pré-traitement de corpus (section 5.1). Nous présentons ensuite l'étape principale qui concerne les trois méthodes d'extration de reformulations (section 5.2). Nous décrivons également la méthode d'alignement des reformulations avec les termes (section 5.3) et l'évaluation des résultats à différentes étapes (section 5.4).

## 5.1 Pré-traitement de corpus

Le corpus de développement a d'abord subi un léger nettoyage :

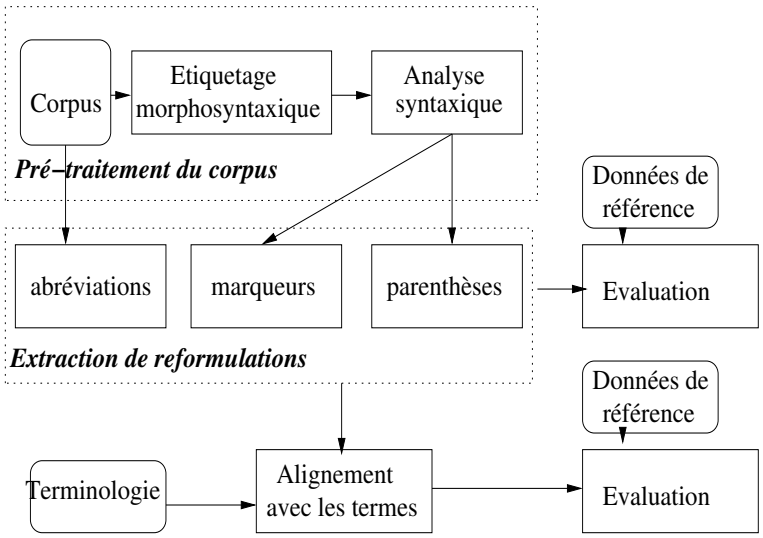



Fig. 2 – Schéma général des trois méthodes

- la restitution du point de fin de phrase manquant avant *Bien cordialement*, car autrement cette expression est collée à la phrase précédente, ce qui peut nuire à l'extraction de reformulations ;
- la normalisation des orthographes du marqueur *c'est-à-dire*, qui peut en effet présenter en corpus plusieurs variations (*c'est -à-dire, c'està-dire*, etc.). Comme nous nous servons de ce marqueur pour repérer les reformulations, il est important de le normaliser.

La ponctuation d'origine est sauvegardée car :

- le point offre un point de repère pour délimiter les reformulations : Cordial travaille en effet sur les phrases et se sert du point comme délimiteur ;
- la virgule permet de restreindre l'extraction de reformulations.

La méthode pour l'extraction de siglaisons fonctionne sur des données texte brut et il n'est pas nécessaire de prétraiter le corpus. En revanche, les corpus doivent être pré-traités pour les deux autres méthodes. Les corpus sont étiquetés et analysés syntaxiquement avec Cordial (Laurent *et al.*, 2009). L'analyse syntaxique permet alors de délimiter les syntagmes syntaxiques et sert également de base pour délimiter les reformulations.

Dans le tableau 1, nous présentons un extrait de l'étiquetage et analyse syntaxe de Cordial pour la phrase *Vous devez les faire brûler par un gastroentérologue spécialisé, c'est-à-dire un proctologue.* Les champs exploités dans notre travail sont les formes, les lemmes correspondant aux formes, et les informations sur les groupes syntaxiques : à quel type de syntagme appartient la forme et dans quel syntagme de la phrase elle s'inscrit. Cordial propose également deux autres types d'information : la position de chaque mot (numéro d'ordre depuis le début de la phrase en nombre de mots et de caractères) et la sémantique associée à différents mots. Par ailleurs, il redonne la phrase complète et

l'encadre des informations sur le *début de phrase* et *fin de phrase*.

| *forme* | *lemme* | *POS* | *POSMT* | *GS* | *type GS* | *Prop* |
|---|---|---|---|---|---|---|
| Vous | vous | PPER2P | Pp2.pn | 1 | S | 1 |
| devez | devoir | VINDP2P | Vmip2p | 2 | V | 1 |
| les | le | PPER3P | Pp3.pa | 3 | C | 2 |
| faire | faire | VINF | Vmn– | 4 | D | 2 |
| brûler | brûler | VINF | Vmn– | 5 | V | 3 |
| par | par | PREP | Sp | 8 | F | 3 |
| un | un | DETIMS | Da-ms-i | 8 | F | 3 |
| gastroentérologue | gastroentérologue | NCMS | Ncms | 8 | F | 3 |
| spécialisé | spécialisé | ADJMS | Afpms | 8 | F | 3 |
| , | , | PCTFAIB | Ypw | - | - | 3 |
| c' | ce | PDS | Pd-..- | 11 | N | 3 |
| est | est | ADV | Rgp | - | p | 3 |
| -à | à | PREP | Sp | 14 | I | 3 |
| -dire | dire | VINF | Vmn– | 14 | I | 3 |
| un | un | DETIMS | Da-ms-i | 16 | D | 3 |
| proctologue | proctologue | NCMS | Ncms | 16 | D | 3 |
| . | . | PCTFORTE | Yps | - | - | - |

TABLE 1 – Un extrait de texte étiqueté et analysé syntaxiquement.

## 5.2 Trois méthodes pour l'extraction de reformulations

### 5.2.1 Extraction des siglaisons

Nous avons traité les siglaisons en nous appuyant sur l'algorithme proposé dans l'état de l'art (Schwartz & Hearst, 2003). L'entrée de cet algorithme est le corpus brut au format texte. Dans un premier temps, l'algorithme détecte les formes entre parenthèses. Une fois extraites, ces formes sont traitées uniquement si elles apparaissent dans les contextes suivants :

patron 1 : *forme étendue/concept (abréviation)*
patron 2 : *abréviation (forme étendue/concept)*

Contrairement au travail d'origine (Schwartz & Hearst, 2003), nous avons pris en compte les majuscules car elles marquent mieux les abréviations. Elles permettent également de différencier les reformulations utilisant les siglaisons et les reformulations utilisant les parenthèses sans les siglaisons (traitées dans la section 5.2.3). La distinction entre les patrons 1 et 2 est effectuée selon le nombre de mots entre parenthèses :

- si un seul mot se trouve entre parenthèses, alors il s'agit du patron 1 ;
- si plusieurs mots se trouvent entre parenthèses, alors il s'agit du patron 2.

Ensuite, il faut parcourir les mots proches de l'abréviation détectée : avant les parenthèses dans le cas du patron 1, entre les parenthèses dans le cas du patron 2. Le traitement de ces deux patrons est différent :

**Patron 1.** Les lettres de l'abréviation sont parcourues en sens inverse. Pour chacune d'elle, nous regardons la première lettre des mots avant les parenthèses, en sens inverse. Si la première lettre du mot correspond à celle de l'abréviation (avec la neutralisation de la casse), alors ce mot est retenu et nous passons au mot qui se trouve devant, etc. Lorsque toutes les lettres de l'abréviation sont traitées, l'algorithme s'arrête ;

**Patron 2.** Le traitement du patron 2 suit le même principe :
- si le mot avant les parenthèses est en majuscules, alors il s'agit d'une abréviation ;
- la première lettre de chaque mot de la forme étendue entre les parenthèses est retenue et comparée avec les lettres de l'abréviation se trouvant juste avant les parenthèses. Si les lettres correspondent (le contraire est rare dans ce cas, car les parenthèses délimitent correctement la forme étendue), nous apparions l'abréviation et la forme étendue.

Cette méthode fournit un ensemble d'abréviations avec leurs formes étendues. Trois cas de figures peuvent se présenter :
- toutes les lettres de l'abréviation sont appariées ;
- une partie des lettres est appariée : nous marquons alors qu'il s'agit d'une *formule incomplète* ;
- aucune lettre de l'abréviation n'est appariée : nous marquons alors que la forme étendue est *inconnue*.

Dans tous les cas, il est nécessaire de gérer les doublons : lorsqu'une abréviation contient plus d'une occurrence d'une lettre donnée, comme dans *LAL*, il faut éviter d'extraire *lymphoblastique aiguë lymphoblastique* au lieu de *leucémie aiguë lymphoblastique*. Dans ce cas, nous vérifions si le mot courant se trouve déjà parmi les mots retenus de la forme étendue. Si c'est le cas, ce mot est alors ignoré. Les abréviations avec des lettres répétées sont très courantes dans le corpus de développement et ce traitement a donc permis de maximiser considérablement nos résultats.

### 5.2.2 Extraction des reformulations avec marqueurs

Pour cette méthode, ainsi que pour la méthode qui traite les reformulations parenthésées (section 5.2.3), nous avons tout d'abord relevé manuellement quelques reformulations afin d'observer l'étiquetage et l'analyse syntaxique proposés par Cordial et pour voir si et comment il est possible d'exploiter ces informations syntaxiques pour effectuer une extraction pertinente de reformulations.

Dans nos analyses, nous employons trois notions :
- les marqueurs : il s'agit d'un ou plusieurs terme(s) qui introduisent la reformulation ; *les réactions immunitaires* ***c'est-à-dire*** *les réactions de défense*

– les concepts ou segments reformulés : il s'agit de la notion reformulée dans notre corpus, celui qui se situe à gauche du marqueur ; ***les réactions immunitaires*** *c'est-à-dire les réactions de défense*
– les reformulations : il s'agit de la partie qui reformule le concept, qui se situe à droite du marqueur ; *les réactions immunitaires c'est-à-dire* ***les réactions de défense***

**5.2.2.1 Choix des informations syntaxiques.** Nous avons considéré trois niveaux d'informations syntaxiques de Cordial :

– *Catégories syntaxiques.* D'après nos premières observations, les catégories syntaxiques (champs *POS* et *POSMT* du tableau 1) ne semblent pas exploitables car il est difficile de dégager des régularités à ce niveau de syntaxe pour différentes reformulations. Proposer une méthode basée sur des patrons morphosyntaxiques rendrait donc l'extraction trop générale et amènerait beaucoup de bruit si un patron commun est reconnu très souvent, ou au contraire amènerait des silences si certaines structures syntaxiques ne sont pas décrites ;
– *Groupes syntaxiques.* Avec les reformulations introduites par les marqueurs, les groupes de mots qui précèdent et suivent le marqueur appartiennent généralement au même groupe syntaxique. Par exemple, dans le tableau 1, comme indiqué dans les colonnes *GS* et *type GS*, il s'agit de groupes 8/F et 16/D. Il s'agit donc d'une observation exploitable par notre méthode. Cette observation s'applique aussi au concept car, dans les reformulations, il n'est pas rare que les premiers termes soient de type relatif, par exemple, et appartiennent donc à un autre syntagme. De plus, nous avons remarqué que certaines reformulations correspondent à deux syntagmes, ce qui peut également rendre l'extraction complète difficile. Dans la phrase suivante *une sécrétion de colostrum, c'est-à-dire une gouttelette venue des canaux galactophores*, la reformulation *une gouttelette venue des canaux galactophores* se situe sur deux syntagmes : *une[17] gouttelette[17] venue[17] des[20] canaux[20] galactophores[20]* (les numéros à côté de chaque forme sont les numéros de syntagmes), comme nous le voyons dans le tableau 2, colonne *GS*.

| *forme* | *lemme* | *POS* | *POSMT* | *GS* | *type GS* | *Prop* |
|---|---|---|---|---|---|---|
| une | un | DETIFS | Da-fs-i | 17 | D | 1 |
| gouttelette | gouttelette | NCFS | Ncfs | 17 | D | 1 |
| venue | venu | ADJFS | Afpfs | 17 | D | 1 |
| des | de le | DETDPIG | Da-.p-i | 20 | B | 1 |
| canaux | canal | NCMP | Ncmp | 20 | B | 1 |
| galactophores | galactophore | ADJPIG | Afpmp | 20 | B | 1 |

TABLE 2 – Extrait de la phrase étiquetée *une sécrétion de colostrum, c'est-à-dire une gouttelette venue des canaux galactophores.*

– *Propositions.* Les codes propositionnels (colonne *Prop* du tableau 1) sont aussi intéressants à exploiter, notamment pour les reformulations, permettant d'aller

au-delà des groupes syntaxiques et donc d'obtenir des extractions probablement plus complètes.

En conséquence, nous testons deux versions de la méthode pour l'extraction de reformulations introduites par des marqueurs : la *méthode par groupes syntaxiques* et la *méthode par groupes propositionnels*. Notons que quelle que soit la méthode utilisée, nous utilisons toujours les syntagmes pour le repérage des concepts, car les codes propositionnels risquent de "noyer" le concept. Le principe est exactement le même dans les deux méthodes ; seule l'utilisation des codes diffère. Si nous regardons l'exemple 1, les deux types de méthodes fonctionnent aussi bien, car les formes de la reformulation *un proctologue* partagent le même code syntaxique et le même code propositionnel. Ce n'est pas toujours le cas : plusieurs formes ayant un code syntaxique différent peuvent avoir le même code propositionnel, comme nous l'avons vu dans le tableau 2.

Cette méthode se décompose en plusieurs étapes : (1) détection des phrases avec les trois marqueurs de reformulation (*c'est-à-dire, autrement dit, encore appelé(e)(s)*) ; (2) détection de segments reformulés ; (3) premiers tests et ajustement de la méthode.

**5.2.2.2 Détection de phrases avec les marqueurs de reformulation.** La première étape consiste à détecter les phrases contenant les marqueurs de reformulation. Ceci est effectué sur la phrase d'origine réimprimée par Cordial. Lorsque l'un des marqueurs de reformulation est détecté, la phrase est retenue pour l'analyse.

**5.2.2.3 Détection de segments reformulés.** Pour récupérer les segments reformulés, nous nous repérons par rapport au marqueur, aux positions des mots et aux informations syntaxiques. Ainsi, en partant du marqueur, nous parcourons les mots situés avant en sens inverse : nous les considérons comme le concept. Si ces mots appartiennent au même syntagme, nous conservons ces mots. Nous parcourons ensuite les mots situés après le marqueur : nous les considérons comme la reformulation. Selon la méthode utilisée, ces mots sont conservés s'ils appartiennent au même syntagme ou à la même proposition. Notons que c'est toujours le premier syntagme du parcours ou la première proposition qui est gardé(e), c'est-à-dire, celui/celle situé(e) immédiatement après le marqueur.

**5.2.2.4 Test et ajustement de la méthode.** Les premiers tests montrent que la *méthode par groupes syntaxiques* atteint vite ses limites, comme nous nous y attendions. Nous obtenons seulement quelques résultats corrects (53 %). Pour le reste, les reformulations sont souvent incomplètes car elles s'étendent sur plusieurs syntagmes ; notamment lorsque le syntagme commence par une relative, comme dans l'exemple *Vous ne devez pas employer de savons ou des laits sophistiqués, c'est-à-dire contenant plusieurs composants.* du tableau 3.

Dans l'exemple qui suit, nous voyons que la reformulation tient sur deux syntagmes, ayant pour codes 17 et 19. Dans ce cas, retenir juste le premier syntagme n'est pas suffisant pour l'extraction de la reformulation :

| *forme* | *lemme* | *POS* | *POSMT* | *GS* | *type GS* | *Prop* |
|---|---|---|---|---|---|---|
| Vous | vous | PPER2P | Pp2.pn | 1 | S | 1 |
| ne | ne | ADV | Rpn | 3\|1 | S | 1 |
| devez | devoir | VINDP2P | Vmip2p | 3 | V | 1 |
| pas | pas | ADV | Rgn | 3 | Q | 1 |
| employer | employer | VINF Vmn | – | 5 | D | 2 |
| de | de | PREP | Sp | 7 | D | 2 |
| savons | savon | NCMP | Ncmp | 7 | D | 2 |
| ou | ou | COO | Cc | 7 | F | 2 |
| des | de le | DETDPIG | Da-.p-i | 10\|7 | F | 2 |
| laits | lait | NCMP | Ncmp | 10\|7 | F | 2 |
| sophistiqués | sophistiqué | ADJMP | Afpmp | 10\|7 | F | 2 |
| , | , | PCTFAIB | Ypw | - | - | 2 |
| c’ | ce | PDS | Pd-..- | 13 | N | 2 |
| est | est | ADV | Rgp | - | p | 2 |
| -à | à | PREP | Sp | 16 | F | 2 |
| -dire | dire | VINF | Vmn– | 16 | F | 2 |
| contenant | contenant | NCMS | Ncms | 17 | D | 2 |
| plusieurs | plusieurs | ADJIND | Dt-.p- | 19 | D | 2 |
| composants | composant | NCMP | Ncmp | 19 | D | 2 |

TABLE 3 – Analyse syntaxique et étiquetage de la phrase *Vous ne devez pas employer de savons ou des laits sophistiqués, c’est-à-dire contenant plusieurs composants.*

CONCEPT : *savons ou des laits sophistiqués*
MARQUEUR : *c’est-à-dire*
REFORMULATION : *contenant*

Avec la *méthode par groupes propositionnels*, nous obtenons de bien meilleurs résultats (72%) car les reformulations sont plus complètes. L’inconvénient, à l’inverse, est que certaines frontières sont trop larges et mériteraient d’être restreintes. Voici ce que donne l’exemple du tableau 2 avec la méthode par groupes propositionnels :

CONCEPT : *savons ou des laits sophistiqués*
MARQUEUR : *c’est-à-dire*
REFORMULATION : *contenant plusieurs composants*

Nous voyons donc que le résultat correspond à ce qui est attendu, car nous avons pu prendre en compte l’ensemble de la proposition, qui tient sur les deux syntagmes évoqués plus haut. C’est donc cette méthode que nous retenons pour la suite des travaux, car nous la considérons plus efficace.

### 5.2.3 Extraction des reformulations parenthésées

En ce qui concerne la troisième méthode, effectuant l’extraction des reformulations marquées au moyen de parenthèses, nous avons aussi travaillé sur le corpus analysé

syntaxiquement par Cordial. Nous présentons ici la délimitation des segments de la reformulation, analysons la sémantique des parenthèses, proposons des filtres pour éliminer les emplois non reformulatifs, et indiquons le déroulement de cette méthode.

**5.2.3.1 Détection et délimitation des segments.** Nous considérons que le concept se trouve avant les parenthèses et la reformulation entre les parenthèses.

Les segments avec des reformulations parenthésées sont assez aisés à repérer :

- Le concept est récupéré selon le même principe qu'avec la méthode de marqueurs : en utilisant les codes syntaxiques fournis par Cordial.
- La reformulation est délimitée par des parenthèses : les frontières sont ainsi toutes tracées. Par exemple, dans *myopie (difficulté à voir de loin)*, le concept est *myopie* et la reformulation est *difficulté à voir de loin*. Nous pouvons tout de même rencontrer des termes non pertinents au sein des reformulations, comme par exemple d'autres marqueurs de reformulation : *boutons (on parle d'éruption cutanée)*. En effet, *on parle* introduit la reformulation et marque lui aussi sa présence, malgré l'utilisation des parenthèses, mais ne fait pas partie intégrante de la reformulation.

**5.2.3.2 Sémantique des parenthèses.** La plus grande difficulté dans l'extraction de reformulations parenthésées ne réside donc pas dans la détection des frontières, tel que cela a été le cas avec les reformulations avec marqueur, mais dans la sémantique et donc la pertinence de l'extraction. En effet, les parenthèses peuvent être utilisées dans plusieurs contextes, comme par exemple :

- une incise : *de fièvre (en avez-vous)*
- un exemple : *des infections sexuellement transmissibles (papillomavirus)*
- une précision : *d'une maladie génétique (pas d'une dégénérescence)*
- une énumération : *un pansement intestinal (type smecta)* ; *prescrire (primpéran, vogalène, llioresal par exemple)*

**5.2.3.3 Création de filtres.** Nous avons analysé les extractions non pertinentes afin de repérer si nous pouvions tirer des éléments en commun. Il s'avère que la présence de certains termes ou marques typographiques peuvent être utilisés pour limiter de telles extractions et réduire ainsi le bruit :

- Lorsque les reformulations contiennent *type*, *comme* ou *exemple*, car ces termes introduisent un exemple et non pas une reformulation ;
- Lorsque les reformulations contiennent des mots comme *pharmacien* ou *médecin*, car ces termes font en général partie de formules typiques comme *demandez l'avis à votre pharmacien* ou *votre pharmacien vous conseillera*. Ainsi, l'occurrence de parenthèses dans ce contexte :

  *...du paracétamol (votre pharmacien vous conseillera)*

  peut être éliminée des candidats à reformulation paraphrastique ;
- Lorsque les reformulations contiennent des énumérations, souvent marquées par des suites de virgules et par la présence du marqueur *etc.* Le patron exploité pour les énumérations est alors *NOM + PUNC + NOM + PUNC* au moins deux fois.

Dans le cas d'application de ce filtre sur le corpus de développement, une seule bonne reformulation est éliminée :

*un bilan (thyroïde, syndrome inflammatoire ...*

- Lorsque les reformulations débutent par *et* ou *surtout*, car ces termes indiquent un ajout d'information, non pas une reformulation :

  *allergisant (surtout dans l'allergie à l'œuf)*

- Lorsque les reformulations contiennent la ponctuation forte comme *?* ou *!*, car il s'agit dans ces cas de situations spécifiques : recherche d'information ou remarque de la part du médecin, respectivement. Les reformulations ne sont pas paraphrastiques dans de tels contextes ;
- Lorsque les reformulations débutent par *en*, *si* ou *qui*, car ces termes introduisent une précision d'information, au moyen d'une subordonnée le plus souvent, et non pas une reformulation :

  *il s'agit du Prucalopride (qui n'est pas remboursé)*

  ici, le médecin apporte une précision à propos du Prucalopride ;
- Lorsque les reformulations débutent par *mais* : c'est une précision par opposition

  *S'il existe un risque augmenté, il est vraiment très faible (mais donc pas nul)*

- Lorsque les reformulations commencent par *car, puisque, parce que...* car ces marqueurs indiquent une causalité ;
- Lorsque les reformulations portent une notion temporelle (*ans, régulièrement, rarement*, etc). Il s'agit alors généralement de conseil de posologie, de durée d'un traitement, d'un inconfort dû à une maladie, ou d'une fréquence précisée par le médecin :

  *régulièrement (tous les ans)*

- Lorsque les reformulations contiennent le signe % et des chiffres. Les chiffres par eux-mêmes correspondent généralement aux fréquences de prise de médicaments et ne sont presque jamais pertinents surtout accompagnés du signe %. Nous choisissons donc d'ignorer de tels contextes. Mais cela nous permet de garder des contextes pertinents comme :

  *à jeun (10 heures sans manger et sans boire)*

Grâce à ce filtre, nous limitons les extractions non pertinentes et obtenons des résultats plus satisfaisants. Cependant, certaines extractions non pertinentes restent, car ne répondent pas aux structures prévues par les filtres : *est trop sucrée (vous avez un poids un peu élevé par rapport à votre taille).*

**5.2.3.4 Déroulement de la méthode.** Le déroulement global de la méthode pour l'extraction de formulations avec des parenthèses est le suivant :

- les mots entre parenthèses sont détectés au sein d'une phrase,
- le segment entre les parenthèses et le segment qui précède les parenthèses sont enregistrés ;

– l'application des filtres permet de vérifier si certains motifs, considérés comme faisant partie d'une extraction non pertinente, y apparaissent. Si ces motifs sont trouvés, alors les candidats à reformulation sont ignorés ;
– de même, nous ignorons les reformulations parenthésées portant sur les siglaisons car leur traitement est pris en charge par une méthode dédiée (section 5.2.1). Ici, nous ignorons donc les phrases contenant des mots en majuscules, que nous considérons comme des abréviations.

## 5.3 Alignement de segments extraits avec une terminologie médicale

L'objectif de l'alignement des segments extraits avec une terminologie médicale est double :

– nous voulons vérifier la pertinence des extractions. En effet, si au moins un des segments extraits se trouve dans la terminologie médicale, alors il s'agit certainement d'une reformulation concernant une notion médicale et donc pertinente pour l'étude. Cela permet d'éviter des extractions non pertinentes comme :

  *en fibres (pas trop vite sinon vous serez ballonnée)*

– cela nous permet également d'associer les segments extraits aux termes médicaux avérés et rend possible l'exploitation de la ressource constituée dans des contextes de recherche d'information, d'indexation, etc.

### 5.3.1 Méthode d'alignement

L'alignement est effectué avec une méthode exploitée dans un travail précédent (Grabar & Hamon, 2015). Cette méthode permet de comparer les segments extraits avec les termes de la terminologie médicale, en effectuant la désaccentuation, une normalisation morphologique avec des ressources décrites dans la section 4.3.2 et en supprimant les mots dits vides (section 4.3.1). Cette méthode permet aussi de contrôler le taux d'alignement entre les deux séquences comparées (segment extrait et terme) :

– le taux d'alignement des segments extraits, ce qui permet de contrôler si tous les mots de ces segments sont alignés,
– le taux d'alignement des termes techniques de la terminologie, ce qui permet de contrôler si tous les mots de ces termes sont alignés.

Nous pouvons donc exploiter ces deux paramètres pour optimiser les alignements. Ainsi, sur le corpus de développement, nous testons différents seuils pour ces deux taux d'alignement, en les variant par 10 entre 40 et 100 %. Nous considérons que les taux d'alignement inférieurs (30 % ou moins) sont trop faibles pour fournir des résultats acceptables. Lorsque des alignements complets (100 % pour le segment et 100 % pour le terme) sont possibles, il sont proposés en priorité. Les codes indiqués après le terme correspondent aux codes UMLS et/ou Snomed International :

*AINS : ains.C0003211/C-60300*
*anti inflammatoires non stéroïdiens : anti inflammatoires steroidiens.C0003211*

En revanche, lorsque les alignements détectés sont partiels, ils sont tous proposés comme candidats :

*les réactions de défense : defense/T-540A0, defense hote/F-C0480*

Si cette méthode ne trouve aucun candidat, aucune proposition n'est faite :

*est trop sucrée : –*
*(vous avez un poids un peu élevé par rapport à votre taille) : –*

## 5.4 Évaluation

Nous présentons ici différents aspects de l'évaluation : création de jeux de référence (section 5.4.1) et les principes de l'évaluation (section 5.4.2). Ces deux aspects concernent les deux principales étapes de la méthode : l'extraction de reformulations et l'alignement avec la terminologie.

### 5.4.1 Création de jeux de référence

Nous créons deux jeux de référence : extraction de segments en relation de reformulation (section 5.4.1.1) et alignement (section 5.4.1.2). Deux annotateurs sont impliqués dans le processus et une séance de consensus est effectuée à la fin. Nous utilisons le Kappa de Cohen (Cohen, 1960) pour calculer l'accord entre deux annotateurs. Le résultat est compris entre 0 et 1, et indique les niveaux d'accord, où 1 correspond à l'accord parfait entre les annotateurs (Landis & Koch, 1977).

**5.4.1.1 Jeux de référence pour l'évaluation des extracteurs.** Afin d'évaluer la performance de nos méthodes, nous avons élaboré des données de référence. Pour ce faire, nous avons extrait les phrases entières contenant les reformulations, à partir des mêmes données que celles traitées par nos méthodes : les données brutes pour les abréviations, et les données analysées syntaxiquement par Cordial pour les reformulations avec marqueurs et avec parenthèses. Nous avons créé un guide d'annotation (voir l'annexe A page 57) pour que les annotateurs respectent les mêmes spécifications lors des annotations.

Le processus d'annotation est effectué manuellement : les annotateurs détectent les reformulations et les balisent. Les concepts sont balisés *<C>concept</C>*, quel que soit le type de reformulation. Les marqueurs sont balisés *<M>marqueur</M>*. Quant aux reformulations, nous distinguons plusieurs situations :

- les reformulations génériques *<Rgen>* avec des informations globales et complètes sur un concept : *<C>mycose</C>, <M>c'est à dire</M> des <Rgen>champignons</Rgen>*
- les reformulations spécifiques *<Rspe>*, qui sont dépendantes du contexte et d'une situation donnée :
  *<C>un bruxisme</C>, <M>c'est-à-dire</M>, <Rspe> des mouvement automatiques des mâchoires pendant le sommeil</Rspe>*, où *pendant le sommeil* est trop précis pour une reformulation globale. Pour qu'elle soit globale, il aurait fallu supprimer *pendant le sommeil*.

Pour les reformulations avec parenthèses, nous ajoutons deux autres types de balises :

– pour les exemples, les énumérations et en cas de présence de marqueurs *par exemple, etc, du type, comme* : *<C>des médicaments qui inhibent la sécrétion de l'acidité de l'estomac</C> <Ex>(de type oméprazole, en vente libre ou remboursés sur prescription)</Ex>*
– pour les précisions, qui ne sont pas des reformulations à proprement parler mais un affinage de l'information à propos du concept : *<C>une cause liée au foie</C> <Pre>(prise de sang complémentaire pour explorer le foie+échographie)</Pre>*

Les exemples peuvent ne pas être pris en compte, selon le souhait. En revanche, les précisions font partie des parenthésages que nous ignorons dans l'étude actuelle. Ces distinctions peuvent être intéressantes pour les travaux et les analyses futurs.

Les reformulations avec parenthèses ont été les plus compliquées à annoter, du fait de la complexité des différents rôles des parenthèses.

**5.4.1.2 Jeux de référence pour l'évaluation de l'alignement.** À partir des propositions d'alignement (section 5.3) obtenues avec les taux de recouvrement 40 % pour le segment et 40 % pour le terme, nous constituons les données de référence grâce à une validation manuelle par deux évaluateurs. Les sorties des trois méthodes de détection de reformulations (abréviations, avec marqueurs, avec parenthèses) sont traitées. L'objectif est d'invalider les propositions qui ne semblent pas être pertinentes et de valider les propositions qui semblent être pertinentes. Dans ce dernier cas, les alignements complets et partiels sont considérés. Les propositions considérées comme non pertinentes sont la plupart du temps des concepts plus précis ou partiels par rapport à ce qui est attendu. Dans l'exemple :

*les réactions immunitaires :*
*- competence immunitaire/C0020987.T043.PHYS*
*- domaine immunitaire/C0312545.T039.PHYS*
*- elimination immunitaire/C0312550.T039.PHYS*
*- reactions nevrotiques/C0235172.T048.DISO*

nous présentons un extrait des propositions concernant le concept *réactions immunitaires*. Il s'agit de propositions d'alignement partiel. Ces propositions n'ont pas été retenues car elles ne fournissent pas les informations nécessaires : les termes contenant *immunitaire* ou *reaction* sont tous partiels et hors sujet. L'évaluation indépendante par chaque évaluateur est suivie par une séance de consensus afin d'avoir un seul jeu d'annotation de référence. Ce jeu de référence permet d'évaluer les différents seuils d'alignement testés afin de détecter les seuils optimaux.

### 5.4.2 Principes d'évaluation

**5.4.2.1 Évaluation des extractions.** Deux aspects sont évalués : l'accord inter-annotateurs et la performance des méthodes d'extraction de reformulations. Pour l'accord inter-annotateur, le calcul est effectué à deux niveaux :

– *phrases* : il s'agit d'évaluer si les deux annotateurs prennent les mêmes décisions sur la présence et absence de reformulations dans une phrase donnée ;

– *tokens* : il s'agit d'évaluer l'accord sur les frontières des segments en reformulation, telles qu'annotées par les deux annotateurs.

Cet accord consiste à comparer les choix de balisage de chaque annotateur de façon binaire : d'accord (O) ou pas d'accord (N).

Le calcul de la performance des méthodes d'extraction de reformulations est effectué au niveau des deux segments : concept et reformulation. L'évaluation prend en compte les frontières de ces segments. Cette évaluation est effectuée avec le script d'évaluation de la tâche 3 de la campagne DEFT 2015 [7]. L'évaluation est effectuée sur les segments extraits exacts (où les frontières sont respectées) et inexacts (où des recouvrements partiels sont acceptés). Le même type de représentation est donc généré pour les données de référence et les segments extraits, où chaque *token* reçoit un numéro depuis le début de la phrase. Dans les données de référence, l'annotation des reformulations est normalisée vers *<R>reformulation</R>* sans distinguer les cas spécifiques, génériques, exemples, etc., car la méthode automatique ne l'effectue pas.

**5.4.2.2 Évaluation des alignements.** La qualité de l'alignement est évaluée par rapport aux données de référence, en prenant en compte les propositions correctes et incorrectes. Cette évaluation est effectuée sur le corpus de développement, pour lequel les données de référence sont créées. Cela permet de définir les seuils optimaux d'alignement de segments extraits et de termes, qui seront appliqués au corpus de test.

**5.4.2.3 Mesures d'évaluation.** Nous utilisons les mesures d'évaluation classiques dans les travaux de TAL :

– *La précision* correspond au nombre de résultats pertinents obtenus par rapport au résultat total obtenu. Une bonne précision (proche de 1) indique que les résultats obtenus sont tous, ou presque pertinents. Une mauvaise précision (proche de 0) souligne du bruit, c'est-à-dire un grand nombre de résultats non pertinents ;
– *Le rappel* correspond au nombre de résultats pertinents obtenus par rapport au nombre de résultats attendus. Un mauvais rappel (proche de 0) signifie qu'il y a beaucoup de silence, c'est-à-dire que de nombreux résultats n'ont pas été récupérés ;
– *La F-mesure* permet de pondérer le rappel et la précision. Ainsi, une bonne précision et un mauvais rappel peuvent amener à une bonne F-mesure ; idem pour une mauvaise précision et un bon rappel. Les résultats s'interprètent entre 0 et 1, où 1 correspond à une F-mesure parfaite.

Dans certains cas, pour avoir une vue d'ensemble sur les résultats, les moyennes de précision ou de rappel sont utilisées.

7. https ://deft.limsi.fr/2015/evaluation.fr.php ?lang=fr

# 6 Résultats

## 6.1 Résultats d'extraction des siglaisons

De manière générale, lorsque les formes étendues correspondant aux abréviations sont présentes dans le corpus, elles peuvent être récupérées. En revanche, nous pouvons aussi avoir des cas où ces formes sont absentes ou bien partielles :

- *Pas d'extractions.* Dans la phrase *comment sont les ALAT(ou SGPT) et les ASAT (ou SGOT)*, nous rencontrons deux abréviations, *ALAT* et *ASAT*, situées devant des parenthèses, dans lesquelles leurs formes étendues pourraient se trouver. Comme ces formes ne sont pas présentes, notre méthode ne reconnaît pas les termes correspondants aux lettres des abréviations et aucune extraction n'est effectuée ;
- *Résultat partiellement correct et exploitable.* Pour la phrase *Si vous n'habitez pas à Paris, vous devriez consulter un spécialiste dans un hôpital universitaire de votre vile (CHU)*, notre méthode propose l'extraction *CHU : hôpital universitaire.* Cette extraction est compréhensible et exploitable mais la lettre *C* n'y est pas représentée. Idéalement, nous aimerions avoir *Centre Hospitalier Universitaire*, mais la phrase contient seulement les termes *hôpital* et *universitaire* repérés par le programme ;
- *Résultat partiel inexploitable.* Pour la phrase *Vous devez faire une prise de sang (NFS) mais celle-ci peut ne pas être concluante.*, la méthode propose l'extraction *NFS : faire sang.* Nous récupérons ici les termes dont les lettres initiales se trouvent a proximité de l'abréviation et sont reconnues au sein de l'abréviation. En revanche, il ne s'agit pas de la forme étendue correcte *numération formule sanguine*, car elle est absence dans cette phrase. Ce résultat est donc inutilisable.

Les reformulations partielles peuvent en effet être utilisées pour donner une idée générale de la sémantique de l'abréviation.

Sur le corpus de développement, nous obtenons 75 résultats dont 35 sont corrects, 33 partiels et exploitables, et 7 non exploitables. Nous avons 29 types d'abréviations différents, ce qui signifie qu'une bonne partie des abréviations est au moins en double, et 42 types différents et leur forme étendue, ce qui signifie que certaines abréviations ont plusieurs formes étendues. Cela peut correspondre à une faute de frappe, mais aussi à la disponibilité de plusieurs formes étendues dans le corpus. Dans l'exemple qui suit, nous voyons que l'abréviation *AINS* comprend 3 formes étendues différentes, respectivement, une au singulier, une au pluriel, et une contenant une faute de frappe (*stroïdiens*) :

*AINS : anti inflammatoire non stéroïdien*
*AINS : anti inflammatoires non stéroïdiens*
*AINS : anti inflammatoires non stroïdiens*

Notons que certains résultats, jugés partiels par la méthode, peuvent correspondre aux formes étendues correctes. Il s'agit souvent de formes composées :

- *CIV : communication interventriculaire*, où *I* et *V* représentent le terme *Inter-Ventriculaire*, alors que la méthode détecte seulement la lettre *I* ;

– *TG : thyroglobuline*, où *T* et *G* représentent *ThyroGlobuline*, alors que la méthode détecte une seule lettre *T*.

Dans cette situation, le risque est de repérer un autre terme commençant par la seconde lettre du composé : par exemple, la séquence *thyroglobuline globale* est considérée comme correcte pour l'abréviation *TG*.

Concernant les extractions sur le corpus de test, les résultats sont comparables à ceux du corpus de développement, et nous obtenons les mêmes cas de figure :

– Extractions correctes : *ESF : Editions Sociales Françaises*
– Extractions partielles et correctes : *CHUM : Université Montréal*
– Extractions partielles incorrectes : *SEPP : plus*
– Absence d'extractions

Avec le corpus de test, nous obtenons 88 762 occurrences, dont 52 165 sont partielles.

Au niveau des types, cela correspond à 5 566 abréviations, et à 8 106 formes étendues, ce qui signifie qu'une même abréviation peut avoir plusieurs formes étendues différentes. Il peut s'agir de variations, de formes partielles, complètes, ou incorrectes. Nous obtenons, par exemple, 338 occurrences de l'abréviation *VIH*, mais seulement 7 occurrences de la formule complète *VIH : virus immunodéficience humaine*. Dans l'exemple qui suit, l'abréviation *TOC* contient 3 formes étendues différentes (une partielle et inexploitable, une au singulier et une au pluriel, respectivement) :

*TOC : tout*
*TOC : trouble obsessionnel compulsif*
*TOC : troubles obsessionnels compulsifs*

Nos résultats montrent aussi que la méthode peut être appliquée à d'autres langues, en extrayant par exemple *PYLL : potential years life lost*.

En revanche, la méthode récupère également des expressions de type *CC1c3nc2ccccc2 n3CCOCC : O=COCCCc4cccCCN1CCC*, car elles contiennent des suites de majuscules. Une amélioration à prévoir est donc de filtrer ce type d'expressions. Nous avons tenté de ne pas prendre en compte les suites de majuscules contenant des nombres, mais cela enlève de bonnes extractions comme *CD26 : cluster de différenciation 26*. Un filtre plus sophistiqué doit donc être proposé.

De manière générale, cette méthode est assez aisée à implémenter, sauf la gestion de mots extraits en double, comme indiqué à la section 5.2.1.

## 6.2 Résultats d'extraction des reformulations avec marqueurs

### 6.2.1 Difficultés rencontrées lors du règlement de la méthode

Nous avons rencontré plusieurs difficultés lors du règlement de la méthode avec les marqueurs. Nous en présentons ici quelques unes. Le principal défaut de la *méthode par groupes propositionnels* est le bruit qu'elle peut générer.

**6.2.1.1 Reformulations avec virgules.** Les frontières des reformulations varient : certaines sont longues, et nous avons alors limité la longueur à 15 mots ; alors que d'autres reformulations ne tiennent que sur un seul mot. Dans le cas de reformulations longues,

nous nous sommes rendus compte que, généralement, lorsqu'une virgule y apparaît, la frontière du segment doit s'arrêter avant la fin du code propositionnel. Nous limitons alors la frontière du segment à la première virgule rencontrée. Ainsi, dans l'exemple suivant, la non prise en compte de la virgule donne :

- CONCEPT : la chirurgie bariatrique
- MARQUEUR : c'est-à-dire
- REFORMULATION : de l'obésité, s'impose

alors que la prise en compte de la virgule permet d'obtenir une meilleure extraction :

- CONCEPT : la chirurgie bariatrique
- MARQUEUR : c'est-à-dire
- REFORMULATION : de l'obésité

**6.2.1.2 Chevauchement de syntagmes syntaxiques.** De manière plus rare, nous pouvons aussi rencontrer du bruit lors de l'extraction des concepts lorsque certains syntagmes se chevauchent, comme dans l'exemple du tableau 4. Ainsi, nous repérons le code *7* jusqu'au début de phrase. Nous voyons que ce code se chevauche avec le syntagme *10* (colonne *GS*). Comme le code *7* est commun à toute la séquence, nous considérons que les syntagmes *7* et *10|7* doivent être extraits. Mais il existe des cas où le syntagme *10|7* correspond au concept souhaité. Nous avons donc essayé de limiter les extractions de manière à ne garder que le code syntaxique minimal le plus proche, comme *10|7* dans l'exemple du tableau 4. Cependant, avec cette limitation, nous obtenons de moins bons résultats car de nombreuses extractions deviennent tronquées. Nous avons donc abandonné cette variante de la méthode.

| *forme* | *lemme* | *POS* | *POSMT* | *GS* | *type GS* | *Prop* |
|---|---|---|---|---|---|---|
| de | de | PREP | Sp | 7 | D | 2 |
| savons | savon | NCMP | Ncmp | 7 | D | 2 |
| ou | ou | COO | Cc | 7 | F | 2 |
| des | de le | DETDPIG | Da-.p-i | 10\|7 | F | 2 |
| laits | lait | NCMP N | cmp | 10\|7 | F | 2 |
| sophistiqués | sophistiqué | ADJMP | Afpmp | 10\|7 | F | 2 |

TABLE 4 – Extrait étiqueté et analysé syntaxiquement de la phrase *Vous ne devez pas employer de savons ou des laits sophistiqués, c'est-à-dire contenant plusieurs composants.*

**6.2.1.3 Traitement de mots vides.** Avec la *méthode par groupes syntaxiques*, nous nous sommes rendu compte que la suppression de certains mots vides permettrait d'extraire un syntagme complet et pertinent. Par exemple, pour une reformulation commençant par *qui*, le syntagme s'arrête là car *qui* est un syntagme à lui tout seul. Nous avons donc établi une liste de mots dits *vides*, que la méthode ignore et passe au mot

ou syntagme suivant. Les résultats n'étaient finalement pas plus probants car les extractions étaient souvent tronquées. En l'occurrence, cela supprime également les mots grammaticaux qui peuvent avoir un impact sémantique, comme par exemple *après*, qui apparaît à plusieurs endroits avec des rôles différents et plus ou moins indispensables. Une meilleure adaptation de la liste de mots vides est compliquée car trop dépendante des extractions. Nous avons donc décidé de ne pas ajouter cette variante à la méthode.

**6.2.1.4 Traitement de marqueur en début de phrase.** Le marqueur *autrement dit* peut se situer en tout début de phrase : *Malgré le fait que vous ne donnez pas les normes du laboratoire, les chiffres que vous indiquez sont négatifs. Autrement dit, vous n'avez rien.* Le traitement de ce cas a demandé des adaptations spécifiques. Ainsi, la notion de phrase est dépassée, et les deux phrases autour du marqueur sont considérées.

**6.2.1.5 Amélioration de la détection des frontières des reformulations.** Avec la *méthode par groupes propositionnels* subsistent encore quelques rares cas où la reformulation est incomplète, ce qui est la difficulté principale de la *méthode par groupes syntaxiques*. Nous avons observé que ces reformulations débutent par un pronom relatif *qu', que, qui*. Nous considérons donc que, lorsque le premier mot de la reformulation contient *qu*, la méthode doit récupérer tous les mots qui suivent et s'arrêter à la frontière (un point, une parenthèse fermante). Le résultat s'est montré probant : à part une reformulation devenue trop large, cela ne détériore pas les bons résultats que nous obtenions jusque là. A l'inverse, d'autres extractions sont trop larges. Nous nous sommes rendu compte que les parties pertinentes de ces extractions ont pour fonction *D, d, H* ou *N*. Nous avons tenté d'exploiter ces codes. Cela n'était pas intéressant car détériorait d'autres résultats. Cette variante de la méthode a aussi été abandonnée.

**6.2.1.6 Amélioration de la détection des frontières des concepts.** L'extraction des concepts est le plus souvent correcte, et nous avons essayé d'améliorer quelques imperfections, comme nous l'avons vu dans les cas où les syntagmes se chevauchent. Nous avons observé quatre types de patrons morphosyntaxiques fréquents des concepts : N + ADJ (*alopécie androgénique*), D + N (*une lithiase*), N + P + N (*sécrétion de colostrum*) et V + ADJ (*sont erratiques*). Nous avons alors tenté de limiter les concepts à ces patrons. Les résultats se sont montrés bons globalement, mais moins bons qu'avec l'utilisation des syntagmes.

**6.2.1.7 Un système hybride.** Après avoir amélioré les deux méthodes (par groupe syntaxique et par proposition), nous avons voulu les hybrider afin de créer un système *intelligent* qui pourrait choisir la méthode la plus appropriée à la reformulation rencontrée. Pour cela, nous avons isolé les reformulations devant être extraites par l'une ou l'autre méthode, et avons observé manuellement leurs différences et leurs points communs. Nous avons constaté que les reformulations extraites par la *méthode par groupes propositionnels* possèdent moins de 3 codes syntaxiques identiques, alors que les reformulations extraites par la *méthode par groupes syntaxiques* possèdent au moins 3 codes

syntaxiques identiques. Nous avons alors modifié notre programme pour qu'il compte le nombre de codes syntaxiques identiques après les marqueurs de reformulation : s'ils sont inférieurs à 3, nous utilisons la *méthode par groupes propositionnels* ; sinon, nous utilisons la *méthode par groupes syntaxiques*. Contre toute attente, les résultats ont été décevants, car la *méthode par groupes syntaxiques* n'est pas assez utilisée. Nous avons alors changé le seuil : le nombre de codes syntaxiques identiques doit être inférieur ou égal à 3, pour l'utilisation de la *méthode par groupes propositionnels*. Nous obtenons de meilleurs résultats sur certaines reformulations. Par exemple, pour la même reformulation :

- $< 3$ : *d'une persistance de l'image* (incorrect car incomplet)
- $\leq 3$ : *d'une persistance de l'image sur la rétine* (correct car la méthode syntaxique a été utilisée)

Malheureusement, ces bons résultats se font au détriment d'autres :

- $< 3$ : *du saignement intermittent* (correct)
- $\leq 3$ : *du saignement intermittent aux hémorragies* (incorrect car trop large, car c'est la *méthode par groupes syntaxiques* qui doit être utilisée)

Une deuxième difficulté, qui, à l'inverse, donnait à la *méthode par groupes syntaxiques* une utilisation de trop importante, venait de la gestion des codes syntaxiques dans les reformulations. Jusque là, nous prenions en compte les chevauchements des syntagmes identiques : si un syntagme a pour code *10*, et que le suivant a pour code *10/12*, tous les deux étaient pris en compte. Cette méthode fournissait souvent des reformulations de taille importante. Nous l'avons modifiée pour prendre en compte les codes syntaxiques identiques sans chevauchements. Nous avons gagné 4 résultats corrects mais en avons perdu 2. Nous avons donc gardé l'ancien réglage, car les résultats incorrectement obtenus nous donnent des reformulations trop larges : nous avons privilégié les reformulations trop larges plutôt qu'incomplètes.

### 6.2.2 Résultats obtenus

Suite aux différents tests, nous privilégions la *méthode par groupes propositionnels*. Sur le corpus de test, nous obtenons 96 résultats, dont 91 types de concepts différents, et 96 types de concepts différents et leur reformulation. Cela signifie que certains concepts sont reformulés plusieurs fois, mais qu'aucune reformulation n'est identique. Avec l'exemple suivant, nous voyons que le concept *des canaux galactophores* a deux reformulations différentes :

*des canaux galactophores c'est-à-dire qui fabriquent le lait de la femme*
*des canaux galactophores c'est-à-dire qui sécrètent le lait*

La répartition de l'utilisation des marqueurs est la suivante : 80 avec le marqueur *c'est-à-dire*, 8 avec *encore appelé* et ses variantes en genre et en nombre, et 8 avec *autrement dit*, qu'il soit en milieu ou début de phrase. La présence du marqueur *c'est-à-dire* est nettement supérieure aux deux autres marqueurs. Les marqueurs *c'est-à-dire* et *autrement dit* peuvent se substituer mutuellement, car ils introduisent une explication et une autre façon d'exprimer un concept. En revanche, le marqueur *encore appelé* introduit une reformulation ayant pour fonction la synonymie :

*CONCEPT : celle de troubles fonctionnels intestinaux*
*MARQUEUR : encore appelés*
*REFORMULATION : colopathie fonctionnelle*

Nous voyons dans cet exemple que *colopathie fonctionnelle* est synonyme de *troubles fonctionnels intestinaux*. Dans notre corpus, ce marqueur marque bien une différence de niveau du langage : il reprend généralement un concept exprimé en langage non expert pour donner le synonyme dans le langage expert :

*CONCEPT : syndrome algo-dystrophique de la'appareil mandibulaire*
*MARQUEUR : encore appelé*
*REFORMULATION : de COSTEN*

Les reformulations ne sont pas nécessairement des définitions universelles, mais peuvent être particulières au contexte, comme dans :

*CONCEPT : des effets antagonistes*
*MARQUEUR : c'est-à-dire*
*REFORMULATION : que le Lopressor inhibe l'effet du Bricanyl*

où la reformulation explique le concept *des effets antagonistes*, par rapport à la situation du patient qui pose une question sur son traitement avec deux médicaments précis.

Les reformulations avec une relative amènent des explications correspondant à des phrases complètes :

*CONCEPT : du faisceau nerveux pyramidal*
*MARQUEUR : c'est-à-dire*
*REFORMULATION : qui commande les mouvements des membres*

Certaines reformulations ont une forme définitoire *lexème-définition*, avec un patron morpho-syntaxique typique *DET N -marqueur- DET N ADJ DET N* :

*CONCEPT : un bruxisme*
*MARQUEUR : c'est-à-dire*
*REFORMULATION : des mouvement automatiques des mâchoires*

Concernant le corpus de test, nous obtenons 2 757 résultats, dont 2 485 types de concepts différents, et 2 710 types de concepts différents et leur reformulation. Cela signifie que certains concepts apparaissent à plusieurs reprises (comme les deux occurrences identiques de la forme complète *les vaisseaux lymphatiques encore appelés vaisseaux de lait*), et que peu d'entre eux possèdent plusieurs formes étendues.

La répartition des marqueurs est la suivante : 1 929 avec le marqueur *c'est-à-dire*, 86 avec *encore appelé* et ses variantes, et 145 avec *autrement dit*. Ici aussi, le marqueur *c'est-à-dire* est nettement plus présent que les autres.

Les trois marqueurs exploités ont une fonction reformulative. Notons que nous avons envisagé d'autres marqueurs comme *l'équivalent de*, que nous avons abandonnés, car ils peuvent jouer d'autres rôles. Les trois marqueurs traités s'avèrent tous utiles, même si *c'est-à-dire* est de loin le plus utilisé. Ce dernier nécessite par ailleurs une attention particulière lors du pré-traitement, car étant donné son orthographe (caractère accentué, apostrophe, des tirets), il est sujet à de nombreuses fautes de frappes.

## 6.3 Résultats d'extraction des reformulations avec parenthèses

Nous obtenons 312 extractions sur le corpus de développement, dont 297 types de concepts différents, et 305 types de concepts différents et leur reformulation. Cela signifie que quelques concepts apparaissent plusieurs fois dans les résultats, et que seulement quelques uns apparaissent plusieurs fois avec une même reformulation. Dans l'exemple qui suit, le concept *un proctologue* possède deux reformulations différentes :

*un proctologue (c'est souvent un gastroentérologue spécialisé dans les lésions anales)*
*un proctologue (gastroentérologue)*

Concernant le corpus de test, une première analyse des extractions montre que :

- beaucoup d'extractions concernent des entités nommées, telles que : *Hôpital des Peupliers (Paris)*. Il est difficile de les filtrer, à moins de créer un dictionnaire supplémentaire de mots à ignorer ;
- des propositions non pertinentes et difficiles à filtrer peuvent être extraites, comme *énergétique (carence plutôt liée au marasme)*, où il ne s'agit pas d'une reformulation mais d'un ajout d'information, introduit par l'adverbe *plutôt* ;
- certains concepts peuvent être trop larges, comme *une greffe de valve prothétique (valve mécanique artificielle)*, où le concept correct correspond à *valve prothétique* ;
- plusieurs extractions sont correctes : *Cette hyper contraction (ou spasmes)*.

Dans les extractions de reformulations du corpus de test, nous avons rencontré le marqueur *également nommé*, qui joue le même rôle que *encore appelé*. Même s'il est rarement utilisé (10 fois sur 100 103 résultats), il serait intéressant de l'exploiter dans l'avenir.

Avec le corpus de test, nous obtenons 100 103 extractions, dont 82 537 types de concepts différents, et 92 971 types de concepts différents et leur reformulation. Autrement dit, peu de concepts sont associés avec plusieurs reformulations. Par exemple, la reformulation *des phases fondamentales du Shen (l'esprit)* apparaît 2 fois, tandis que le concept médical *la tomodensitométrie* possède également deux reformulations différentes :

*la tomodensitométrie (ou scanner abdominal)*
*la tomodensitométrie (ou scanner X)*

## 6.4 Résultats d'alignement avec la terminologie

Nous obtenons plusieurs types d'alignements :

- Proposition pertinente (alignement complet) :

  *syndrome polyalgique idiopathique diffus : syndrome polyalgique idiopathique diffus.C0016053.T047.DISO/C0027073.T047.DISO/C0751152.T047.DISO*

- Proposition avec la variation morpho-syntaxique de *troubles fonctionnels intestinaux* (alignement partiel) :

  *troubles gastrointestinaux fonctionnels/C0559031.T047.DISO ;*
  *troubles gastro intestinaux fonctionnels/C0559031.T047.DISO*

- Proposition partielle (alignement partiel) :

*semaines amenorrhee : amenorrhee/C0002453.T047.DISO*

– Proposition compositionnelle (alignements partiels qui peuvent composer la sémantique complète du segment extrait), sur l'exemple de *cause de pus* :

*cause/C0085978.T078.CONC/C0678226.T169.CONC/C1314792.T169.CONC*
*pus/C0034161.T031.ANAT/C0333369.T169.CONC/C0854358.T201.PHYS*

– Proposition non pertinente :

*LCR : ph lcr/C0853364* (trop précis)
*liquide cerebro : regime liquide/C-F2300*

– Aucune proposition :

*NFS : —*

Pour une extraction complète (son concept et sa reformulation), l'alignement peut être fait sur une seule partie, sur les deux ou sur aucune :

– Alignement total sur une extraction (dans cet exemple, *fibromyalgie* et *SPID* sont équivalents) :

*d'une fibromyalgie : fibromyalgie.C0016053.T047.DISO*
*SPID (syndrome polyalgique idiopathique diffus) : syndrome polyalgique idiopathique diffus/C0016053.T047.DISO*

– Alignement sur une des deux parties d'une extraction :

*TSH : –*
*thyroïde : thyroide.C0040132.T023.ANAT*

– Aucun alignement dans une même extraction :

*HAS : –*
*Haute Autorité Santé : –*

L'utilisation de mots vides peut poser des difficultés. Dans l'exemple *AINS*, nous obtenons une seule proposition pour le sigle et pour sa forme étendue : *anti inflammatoire non stéroidiens*. Lors de l'alignement de la forme étendue, *non*, qui fait partie des mots vides, est éliminé. Nous obtenons ainsi l'alignement avec *anti inflammatoire stéroidiens*. Il s'agit cependant d'un alignement correct, car l'abréviation (*AINS*) et la forme étendue (*anti inflammatoire non stéroidiens*) font partie du même concept UMLS C0003211.

Pour effectuer les alignements sur le corpus de test, nous avons appliqué les seuils optimaux (voir section 6.5.3) : 50/100 (abréviations), 80/100 (marqueurs et parenthèses). Par rapport au corpus de développement, les alignements obtenus sont plus rares. Souvent, un seul élément est aligné, comme dans :

*TPD : –*
*Thérapie photodynamique : therapie photodynamique.C0031740.T061.PROC.100.100*

Mais nous avons aussi quelques alignements complets :

*SARM : sarm.C0343401.T047.DISO/C1265292.T007.LIVB.100.100*
*Staphylococcus aureus résistant méticilline : staphylococcus aureus resistant meticilline.C1265292.T007.LIVB.100.100*

Nous pouvons aussi rencontrer des alignements inintéressants, comme dans l'exemple suivant où *ville* n'est pas un terme médical, même s'il se trouve dans la terminologie de

référence. Ce cas est fréquent, et pourrait être amélioré en évitant le groupe sémantique *GEOG* :

*à Skaryszew : –*
*ville : villes.C0008848.T083.GEOG.100.100*

Nous remarquons aussi que la désaccentuation des termes peut induire des erreurs et nous donner des alignements incorrects. Dans l'exemple qui suit, le terme *aîné* sans accent devient un terme anatomique présent dans la terminologie médicale :

*Jean : –*
*(l'aîné) : aine.C0018246.T029.ANAT/C0337546.T033.DISO*

Certaines extractions peuvent être médicales sans être des reformulations. Cela se retrouve particulièrement dans les extractions avec parenthèses :

*tabagisme : tabagisme.C0040332.T048.DISO/C0037369.T055.ACTI/C1306274.T048.DISO...*
*tabac.C0740009.T002.LIVB/C0040329.T109.CHEM/C0086707.T002.LIVB/...*
*(aggrave les problèmes chroniques de santé, et aggrave fortement le risque de cancer de poumon) : –*

A l'inverse, certaines extractions qui forment une reformulation sur le thème médical peuvent ne pas être alignées avec la terminologie, soit parce que le terme n'y est pas présent, soit parce que les seuils sont trop élevés :

*entre le volume télédiastolique : –*
*(ventricule plein) : –*

Notre hypothèse que l'alignement avec la terminologie médicale permet d'éliminer les extractions qui ne concernent pas le domaine médical semble être assez correcte. Par exemple, cette séquence n'est pas alignée :

*la Trotula major : –*
*(Loi de Franck-Starling) : –*

En ce qui concerne les marqueurs, les non reformulations non alignées sont rares, mais pas impossibles. Il peut s'agir de cas où le concept n'est pas situé immédiatement avant le marqueur, comme dans l'exemple complexe qui suit :

*en permanence : –*
*qu'une personne de vue parfaite - le Dr Bates appelle vue parfaite ou vue normale toute vue possédant une capacité visuelle de 10 / 10 ou plus : –*

Nous indiquons, dans le tableau 5, le nombre d'alignements par méthode et par corpus. Nous remarquons que les paires de segments non alignées sont nombreuses, proportionnellement aux nombre d'extractions. Cela s'explique par le fait que les seuils d'alignement sont élevés, afin de maximiser le nombre d'alignements corrects. Les alignements totaux (sur les deux segments) sont moins fréquents que les alignements partiels (sur un seul segment).

## 6.5 Évaluation

Nous présentons et discutons trois types d'évaluation : l'accord inter-annotateur lors de la création des données de référence (section 6.5.1), les résultats d'extraction des segments en relation de reformulation (section 6.5.2), et les résultats d'alignement des segments extraits avec la terminologie (section 6.5.3).

| | *Développement* | | | *Test* | | |
|---|---|---|---|---|---|---|
| | *Abrév* | *Marq* | *Par* | *Abrév* | *Marq* | *Par* |
| *Nb occurrences* | 75 | 96 | 312 | 88 762 | 2 757 | 100 103 |
| *Total* | 11 | 5 | 38 | 154 | 42 | 3 738 |
| *Partiel* | 44 | 37 | 123 | 1 634 | 557 | 25 708 |
| *non alignés* | 20 | 54 | 150 | 6 318 | 1 937 | 60 928 |

TABLE 5 – Nombre d'alignements totaux et partiels sur les deux corpus (développement et test).

### 6.5.1 Accords inter-annotateur

L'accord inter-annotateur se trouve dans le tableau 6 : pour les extractions et les alignements. Pour les extractions, l'accord sur les *tokens* est nettement supérieur à celui sur les phrases : nous pensons que cela est dû à la taille de la population, largement plus grande avec les *tokens*. L'accord concernant les reformulations avec parenthèses est modéré, tandis que les deux autres accords sont presque parfaits. Le taux modéré des reformulations par parenthèses peut être expliqué par le nombre de catégories plus important que les autres types de reformulations, et par la complexité plus grande du jugement de la pertinence. Les résultats de l'accord inter-annotateur des alignements se trouvent dans la deuxième partie du tableau 6. Nous pouvons voir que l'accord est difficile à obtenir sur les abréviations (seulement 0,208), alors qu'il est proche du parfait avec les marqueurs et les parenthèses.

| | *Extraction* | | *Alignement* |
|---|---|---|---|
| | *Phrase* | *Token* | |
| *Abréviations* | 0,661 | 0,967 | 0,208 |
| *Marqueurs* | 0,24 | 0,816 | 0,714 |
| *Parenthèses* | 0,651 | 0,575 | 0,817 |

TABLE 6 – Résultat de l'accord inter-annotateur des extractions et des alignements pour chaque méthode (au niveau des phrases et des *tokens* pour les extractions).

### 6.5.2 Évaluation des données extraites

L'évaluation des extractions est faite à partir des données de référence annotées du corpus de développement (section 5.4.1.1). Les résultats de l'évaluation sont présentés dans le tableau 6. Comme l'unité d'évaluation est une phrase et comme l'ensemble d'évaluation est fermé, les trois mesures d'évaluation sont identiques. Nous pouvons voir qu'avec les abréviations, les extractions exactes et inexactes montrent des performances assez élevées et proches. En revanche, avec les marqueurs et parenthèses, il existe une grande différence entre les évaluations exactes et inexactes. Ces résultats peuvent

| | *Abréviations* | | | *Marqueurs* | | | *Parenthèses* | | |
|---|---|---|---|---|---|---|---|---|---|
| | *P* | *R* | *F* | *P* | *R* | *F* | *P* | *R* | *F* |
| *exact* | 0.74 | 0.74 | 0.74 | 0.24 | 0.24 | 0.24 | 0.23 | 0.23 | 0.23 |
| *inexact* | 0.94 | 0.94 | 0.94 | 0.98 | 0.98 | 0.98 | 0.68 | 0.68 | 0.68 |

TABLE 7 – Précision, rappel et F-mesure des extractions pour chaque méthode.

être expliqués par les difficultés de régler les frontières des segments en se basant sur les informations syntaxiques : par exemple, des mots grammaticaux ou des extensions des syntagmes diminuent la performance de l'évaluation exacte. Par contre, l'évaluation inexacte montre que les extractions proposées contiennent souvent les propositions attendues ou chevauchent avec elles. Dans le cas des extractions inexactes, soit le segment extrait est acceptable et utilisable, soit il doit être soumis à des post-traitements spécifiques. Voila quelques exemples qui montrent les différences de frontières entre la référence et l'extraction automatique :

Référence : *Votre rapport HDL, c'est-à-dire, le bon cholestérol*
Extraction trop large : *Votre rapport HDL, c'est-à-dire, le bon cholestérol est excellent*

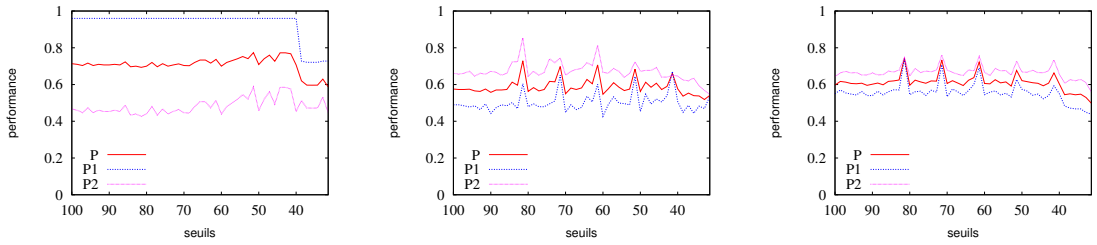
Référence : *la tache sanguine, c'est-à-dire, l'endroit où se rencontrent les vaisseaux sanguins*
Extraction insuffisante : *la tache sanguine, c'est-à-dire, l'endroit*

### 6.5.3 Évaluation des alignements

Nous effectuons une évaluation des performances de l'alignement avec :
- precision moyenne des deux segments,
- precision moyenne du segment 1 (le concept ou segment reformulé),
- precision moyenne du segment 2 (la reformulation).

Cette évaluation est effectuée sur le corpus de développement. Les résultats sont présentés dans la figure 3 pour les trois jeux de données. L'objectif de cette évaluation est de définir les seuils d'alignement optimaux afin de les appliquer ensuite sur le corpus de test.

Sur les figures 3, en abscisse, les performances indiquent la précision. Les seuils des alignements sont exprimés en ordonnée. Entre les deux seuils d'alignement de segments, se trouvent les seuils de termes. Si nous prenons le seuil de segment 100, nous avons 100(segment)-100(terme), puis 100(segment)-90(terme), 100(segment)-80(terme)... 90(segment)-100(terme), 90(segment)-90(terme), et ainsi de suite. Nous voyons que pour les reformulations avec marqueurs ou parenthèses, la meilleure précision se situe aux seuils 80-100, 70-100, 60-100, où l'alignement du segment peut être partiel mais l'alignement du terme est complet. Cela permet d'obtenir les alignements complets des deux ou bien les alignements compositionnels, comme présenté dans la section 6.4. Nous remarquons que le segment 2 est souvent plus facile à aligner que le segment 1. Pour les abréviations, les meilleures précisions sont entre les seuils 50-100 et 40-100. La précision pour l'alignement des abréviations est très élevée, car celles-ci sont plus aisées à extraire et à




(a) Abréviations (b) Marqueurs (c) Parenthèses

Fig. 3 – Précision et rappel de l'alignement des segments extraits avec la terminologie, sur le corpus de développement.

aligner, comparé aux segments extraits avec les marqueurs et les parenthèses. Les seuils optimaux sont donc appliqués lors des alignements des extractions du corpus de test.

## 6.6 Comparaison entre les méthodes

### 6.6.1 Comparaison générale des méthodes

Nous avons proposé trois méthodes d'extraction de reformulations. La première concerne les abréviations. Du point de vue de la forme, les sigles ainsi que leurs formes étendues sont bien reconnus par la méthode. La principale difficulté est due au fait que certains sigles sont utilisés sans leur définition (complètes), auquel cas les résultats restent manquants ou partiels. Pour la méthode d'extraction des reformulations avec les marqueurs, nous avons exploité la *méthode par groupes syntaxiques* et la *méthode par groupes propositionnels*. Dans la première, les reformulations extraites étaient trop souvent incomplètes. Nous avons donc privilégié la méthode par groupes propositionnels, qui fournit des résultats globalement satisfaisants et exploitables. Certaines imperfections peuvent subsister avec une mauvaise détection de frontières (trop larges le plus souvent). Pour les reformulations avec des parenthèses, la reconnaissance est correcte pour la partie reformulation, car celle-ci se situe entre parenthèses. En revanche, la détection des concepts présente les mêmes difficultés qu'avec les marqueurs : parfois incomplète ou parfois trop large. Mais la difficulté principale est que le parenthésage ne concerne pas que les reformulations mais aussi d'autres type d'informations. Nous avons donc créé des filtres appropriés pour limiter les extractions non pertinentes.

Quelle que soit la méthode d'extraction utilisée, il est aisé de distinguer les reformulations qui simplifient (simplification) ou complexifient (complexification) un concept :

- avec une simplification, la reformulation a tendance à avoir une expression complexe, composée de termes compréhensibles par les non experts, d'une fréquence plus ou moins importante dans le corpus ;
- avec une complexification, la reformulation contient très peu de mots, parfois un seul, généralement ayant une fréquence faible et exprimant un terme technique, un nom de maladie ou de médicaments.

Pour résumer, l'extraction d'abréviation et de leur définition est la méthode la plus fiable. A l'inverse, l'extraction de reformulations par parenthèses est la plus complexe car il est difficile de pallier le bruit, du fait de la polysémie des parenthèses.

| | *Développement* | | | *Test* | | |
|---|---|---|---|---|---|---|
| | *Abrév* | *Marq* | *Par* | *Abrév* | *Marq* | *Par* |
| *Nb occurrences* | 75 | 96 | 312 | 88 762 | 2 757 | 100 103 |
| *Nb types* | 29 | 91 | 297 | 5 566 | 2 485 | 88 537 |
| *Nb types + reformulation* | 42 | 96 | 305 | 8 106 | 2 710 | 92 971 |

TABLE 8 – Récapitulatif des extractions avec les trois méthodes, en occurrences et types, sur les deux corpus (développement et test).

Dans le tableau 8, nous proposons un bilan comparatif entre les trois méthodes. Nous indiquons les nombres de types et d'occurrences extraits dans chaque corpus. Nous pouvons observer que les types de concepts et de leur reformulation sont plus nombreux avec les marqueurs et les parenthèses, certainement parce qu'il s'agit de constructions linguistiques plus libres. Un concept donné peut être reformulé plusieurs fois et sa reformulation peut être exprimée de différentes manières, alors qu'avec les abréviations la structure et le contenu sont plus fixes. Notons que les volumes d'extractions avec les abréviations et les parenthèses sont proportionnels entre le corpus de développement et le corpus de test, alors qu'avec les marqueurs, le corpus de test est moins abondant en ce type de constructions.

### 6.6.2 Typologie des reformulations extraites

Nous nous appuyons sur une classification existante (Bhagat & Hovy, 2013) pour décrire et classifier les reformulations extraites. Nous avons analysé les extractions avant et après leur alignement avec la terminologie. Nous avons ainsi constaté qu'il est très difficile de classifier les reformulations avant l'alignement, car les segments extraits peuvent souvent être très complexes linguistiquement. Nous présentons ici quelques observations générales sur les données avant l'alignement :

- nous trouvons essentiellement *synonyme*, *général/spécifique* et *partie/tout* sur le corpus de test ;
- les reformulations avec marqueurs sont souvent définitoires (*avec l'éducation c'est-à-dire les règles sociales formulées ou induites par l'entourage de l'enfant*), apportent une précision par substitution spécifique (*Les extractions instrumentales c'est-à-dire par ventouses ou forceps*) ou montrent une substitution par synonymie (*par le suivi de la santé animale c'est-à-dire la santé du bétail*) ;
- avec les parenthèses, nous trouvons souvent des substitutions par synonymie avec une complexification (*à la douleur (hyperalgie)*, *ou nerveux (hystérie)*, *une petite mâchoire (micrognathie)*) ou des précisions (*viande (boeuf, porc et volaille)*, *l'huile (huile d'olive, huile de pépins de raisin)*).

Après l'alignement, nous avons relevé un échantillon d'extractions dont au moins un des deux segments est aligné. La classification est alors plus aisée. Il s'agit souvent de *synonymes*. Nous avons aussi introduit deux nouvelles classes dans la classification, *définitoire* et *cause à effet*, très fréquentes dans nos données.

**6.6.2.1 Reformulations avec les parenthèses.** Avec les parenthèses, nous observons trois classes dont nous présentons quelques exemples ($a$ en exposé signifie que le segment a été aligné avec la terminologie) :

- synonymes : *nerveux$_a$ (hystérie$_a$)* ; *ou une cyanose$_a$ (bleuissement)* ; *trouble affectif$_a$ (de l'humeur$_a$)* ; *une petite mâchoire (micrognathie$_a$)* ; *Pyrexie$_a$ (fièvre$_a$)* ; *pour la transplantation de moelle (par cytaphérèse$_a$)* ; *Col de l utérus$_a$ (cervix)* ; *Cette hyper contraction (ou spasmes$_a$)* ; *Maladie du charbon$_a$ (Anthrax$_a$)* ; *vasodilatation$_a$ (dilatation des veines)* ;
- définitions : *une scoliose$_a$ (courbure de la colonne vertébrale)* ; *la mégalomanie$_a$ (dans laquelle l'individu pense posséder des pouvoirs hors du commun)* ; *pyurie$_a$ (présence de pus dans l'urine$_a$)* ;
- la relation cause à effet : *d'ulcère tropical$_a$ (moisissures de la jungle)* ; *d'une borréliose$_a$ (due à la bactérie spirochète Borrelia vincentii).*

**6.6.2.2 Reformulations avec les marqueurs.** Avec les marqueurs, nous trouvons essentiellement deux types de reformulation (synonymes et définitions) :

- synonymes : *l'interruption naturelle ou accidentelle de la grossesse, c'est-à-dire, un avortement spontané$_a$* ; *une mort cellulaire$_a$, c'est-à-dire, une nécrose$_a$*,
- définitions : *la contractilité myocardique$_a$, c'est-à-dire, la capacité des cellules musculaires myocardiques à se contracter en réponse à un potentiel d'action* ; *anabolisants$_a$, c'est-à-dire, facilitent l'anabolisme (la croissance des cellules* ; *est invaginé$_a$, c'est-à-dire, qu'il est en forme de creux$_a$* ; *la respiration$_a$, c'est-à-dire, les échanges gaz eux entre l'organisme et l'environnement* ; *l'érythème$_a$, c'est-à-dire, que la peau exposée est rouge* ; *de glaucome congénital$_a$, c'est-à-dire, présent dès la naissance et donc responsable d une buphtalmie* ; *une xérostomie$_a$, c'est-à-dire, une sècheresse buccale$_a$* ; *L'hypertrichose$_a$, c'est-à-dire, la croissance d'une pilosité à un endroit non désiré* ; *macrostomie$_a$, c'est-à-dire, malformation de la bouche$_a$*

Notons que cet échantillon contient essentiellement le marqueur *c'est-à-dire*, qui est le plus productif dans nos corpus.

### 6.6.3 Complémentarité des méthodes

Nous avons comparé les résultats des trois méthodes afin d'analyser leur complémentarité. Nous n'avons effectué cette tâche uniquement sur les résultats du corpus de test, car ceux du corpus de développement ne sont pas suffisamment volumineux. La comparaison entre les extractions avec marqueurs et avec parenthèses nous a apporté 7 lignes identiques, dont 3 reformulations :

*en ambulatoire → sans hospitalisation*
*microcytaires → de petite taille*
*une jaunisse → ictère*

soit 0,007% du nombre total de 102 860 reformulations comparées, le reste s'agissant d'extractions propres à chaque méthode. Cela signifie que quelques reformulations qui utilisent les parenthèses peuvent aussi apparaître avec des marqueurs :

*en ambulatoire (sans hospitalisation)*
*en ambulatoire, c'est-à-dire, sans hospitalisation*

Par ailleurs, les comparaisons entre les abréviations et les reformulations avec marqueurs ou avec parenthèses n'ont montré aucune extraction commune. Cela s'explique par le fait que le traitement des abréviations est différent des deux autres méthodes, car il s'agit exclusivement de lettres d'un concept (un sigle), contrairement aux autres méthodes qui utilisent les groupes syntaxiques et propositionnels.

Ces résultats indiquent que les méthodes sont bien complémentaires et fournissent des extractions très différentes.

### 6.6.4 Comparaison avec les travaux existants

Nous avons expliqué à la section 2.1 que les frontières de la paraphrase sont difficiles à définir. Leur délimitation a été le problème fondamental de notre travail, qui a illustré cette difficulté.

Dans les études précédentes, certaines avaient pour but de simplifier les termes potentiellement difficiles à comprendre pour un profane (Grabar & Hamon (2015)), ou, à l'inverse, de récupérer un vocabulaire profane (Zeng *et al.* (2006)).

D'autres études dont nous nous rapprochons plus, avaient pour but d'extraire les paires de paraphrases contenues dans un texte. Pour cela, plusieurs méthodes ont été proposées :

- les n-grammes de mots (citeCartoni-TALN2011),
- la mesure de familiarité (Zeng *et al.* (2005a)),
- la mesure de similarité cosinus Deléger & Zweigenbaum (2008).
- un mélange de méthodes (Bouamor *et al.* (2012)) entre l'apprentissage statistique, la description de la variation lexicale, la structure syntaxique, le calcul de la transformation des mot, et les équivalence de traduction.

Nous nous éloignons de ces recherches car nous avons travaillé plus précisément sur la reformulation et non sur la détection de paraphrases. De ce fait, nous nous sommes appuyée sur certains marqueurs qui introduisent une reformulation : *c'est-à-dire*, *autrement dit*, et *encore appelé*, ainsi que les parenthèses, pour les reconnaître, car les reformulations n'ont pas de structures définies. Nous avons proposé une méthode de reconnaissance à partir des groupes syntaxiques et propositionnels, obtenus par le biais d'un étiquetage morpho-syntaxique. Il est alors difficile de comparer les résultats obtenus par ses études aux nôtres.

En revanche, pour la détection d'abréviation et de leur forme étendue, nous avons repris l'algorithme proposé par Schwartz & Hearst (2003), à la différence que nous nous

intéressons aux majuscules pour détecter les abréviations. C'est un algorithme connu pour être fiable et efficace (96% de précision), et c'est cette méthode qui montre de meilleurs résultats dans notre travail. Notre précision est plus faible que celle obtenue par Schwartz (74%), car notre corpus contient un certain nombre de formes étendues incomplètes (ECBU = Examen Urines).

Nous nous démarquons aussi des autres travaux car nous utilisons des terminologies médicales afin d'exploiter uniquement les reformulations contenant des notions médicales. Seul McCray *et al.* (1999) en utilise pour une finalité différente de la nôtre, à savoir son traducteur de termes expert/non expert.

# 7 Corpus oral du SAMU

## 7.1 Collecte et préparation du corpus

En parallèle, nous avons travaillé sur un corpus de transcriptions orales. Il s'agit de transcriptions d'appels authentiques au SAMU d'Arras, concernant principalement les AVC (Accident Vasculaire Cérébral) mais aussi quelques cas variés. Il s'agit de conversations entre un appelant et un ARM (Assistant Régulation Médicale), entre l'ARM et le médecin régulateur du SAMU, et enfin entre l'appelant et le médecin régulateur. Chaque transcription a été faite à l'aide du logiciel Transcriber [8] (Barras *et al.*, 1998), permettant d'obtenir en sortie un format XML. Les fichiers XML sont ensuite transformés en texte brut, en respectant les chevauchements de la parole. Il s'agit d'un corpus *sensible* dont la collecte et l'utilisation ont nécessité une charte de confidentialité et une déclaration spécifique auprès de la CNIL. Une des clauses de la déclaration CNIL prévoit que plusieurs informations sont anonymisées lors de la transcription (nom, prénom, ville, rue, numéro de rue, numéro de téléphone...). Lors de la transformation en texte, les balises d'anonymisation sont remplacées par des informations fictives et aléatoires, pour assurer une meilleure lecture. Le corpus total est constitué de 31 fichiers d'une longueur variable, aux deux formats, XML et texte. Ce corpus contient 16 641 occurrences et a nécessité une cinquantaine d'heures pour la transcription. Le corpus est disponible pour la recherche auprès des partenaires spécifiés dans la déclaration CNIL (les chercheurs de l'université Lille 3) pendant une période de 10 ans. Les copies du corpus ne peuvent être transmises que par une clé USB. Toute personne qui l'utilise doit signer la charte de confidentialité.

## 7.2 Exploitation actuelle du corpus

Ce corpus a été créé dans le cade du projet *EQU (Éthique, Qualité, Urgences)*, conduit par Natalia Grabar, CR CNRS et Pierre Valette, chef de service du SAMU au centre hospitalier d'Arras. L'objectif du projet est d'améliorer la qualité de la prise en charge au SAMU. Le projet dispose d'une équipe de chercheurs hétérogène : philosophes, linguistes,

8. *http ://trans.sourceforge.net*

chercheurs en sciences de la communication, ainsi que l'équipe du SAMU d'Arras (chef de service, chef des équipes ARM, gérants du service, etc.)

Le travail sur un corpus authentique permet d'analyser comment se déroule une conversation téléphonique entre l'appelant, l'ARM, et le cas échéant le médecin. Nous pouvons ainsi observer les tours de paroles des interlocuteurs, comment les informations de localisation sont demandées par l'ARM puis transmises par l'appelant, la difficulté de les obtenir, ainsi que l'obtention des informations médicales telles que les antécédents, les traitements en cours, etc. Ces informations ne sont pas simples à obtenir auprès des appelants dans un contexte d'urgence. L'émotion de l'appelant peut être embarrassante pour une conversation intelligible, voire un danger pour le malade, car cela peut prendre beaucoup de temps. Souvent, l'obtention des symptômes pour le diagnostic se fait par des comparaisons (*elle tousse comme un petit chien... non comme un phoque*) ou par rapport *à la normale* (*mon père n'est pas comme d'habitude ce matin*). Dans ce contexte, il est souvent important d'avoir des éléments plus précis pour pouvoir obtenir les symptômes nécessaires à la prise en charge correcte. Le corpus transcrit et actuellement disponible n'est pas suffisant pour effectuer des observations, quantitatives et qualitatives, qui seraient généralisables. Il n'est pas non plus suffisant pour étudier et acquérir le lexique parallèle expert/non expert. Nous pouvons cependant faire un certain nombre de propositions méthodologiques pour le travail futur.

## 7.3 Propositions méthodologiques pour l'exploitation future

L'analyse future peut porter sur deux points : l'interaction verbale (section 7.3.1) et le contenu de l'échange (section 7.3.2). Nous prévoyons de traiter le corpus linguistiquement avec TreeTagger (Schmid, 1994), Flemm (Namer, 2000), Cordial (Laurent *et al.*, 2009) et Dérif (Namer, 2009) afin de traiter différents aspects langagiers. Les données de la transcription (durées, changements de locuteurs...) et les différentes annotations (disfluences, hésitations, pauses...) seront également exploitées. Cette proposition méthodologique s'inspire du travail précédent (Boyé *et al.*, 2014). L'objectif de cette proposition est d'analyser les conversations de manière contrastive afin d'étudier le discours des appelants d'un côté et des médecins et ARM de l'autre côté. Cela devrait permettre de mieux prendre en compte les spécificités du discours des appelants.

### 7.3.1 Analyse de l'interaction verbale

Deux critères de l'interaction communicative peuvent être étudiés :

- *Tours de parole.* Les tours de parole permettent d'étudier la dynamique des échanges ;
- *Temps de parole et chevauchements.* Le temps de parole permet d'étudier la participation des sujets à la conversation. Nous pouvons obtenir un temps de parole propre au sujet ainsi qu'un temps de parole totalisant les moments où les deux locuteurs parlent simultanément (chevauchements).

### 7.3.2 Analyse intrasujet

L'analyse intrasujet concerne différents critères qui s'articulent autour des phénomènes oraux (section 7.3.2.1), lexicaux (section 7.3.2.2) et syntaxiques (section 7.3.2.3).

**7.3.2.1 Phénomènes oraux** Parmi les phénomènes propre à la langue orale, nous étudions les groupes de souffles, les disfluences (Pallaud, 2002 ; Henry & Pallaud, 2004 ; Bove, 2008) et le débit de parole. Les critères à étudier sont les suivants :

- *Groupes de souffle.* Chaque coupure dans le flux de parole est considérée comme le groupe de souffle. Les reprises du souffle peuvent rapportées à des contraintes physiologiques mais également être mises en lien avec le processus d'élaboration des contenus verbaux ;
- *Pauses vides.* Les pauses vides sont constitutives d'un échange conversationnel. Elles participent à la fluence du discours et sont nécessaires, à la réflexion des locuteurs pour constituer leurs énoncés et à leur traitement par les auditeurs. Les pauses vides sont des silences d'au moins 1 seconde entre deux actes de parole ;
- *Les pauses remplies.* Les pauses remplies se manifestent par des *euh* ou *hum* seuls et par des allongements de syllabes ;
- *Les amorces.* Les amorces sont des interruptions de mots en cours d'énonciation (Pallaud, 2002). Elles sont très nombreuses dans les productions orales (Henry & Pallaud, 2004) ;
- *Les répétitions.* Nous distinguons deux sortes de répétitions (Henry & Pallaud, 2004) :
  - les répétitions dites *faits de langue*, qui sont volontaires, comme dans *Il pleuvait il pleuvait*, qui équivaut alors à *il pleuvait beaucoup*,
  - les répétitions dites *faits de parole*, qui ne sont pas volontaires et qui sont alors considérées comme des disfluences, comme dans *Pour arriver aux beaux aux beaux jardins* ou *Mais je je fais quand même.*

  Les répétitions peuvent porter sur un seul mot (*La la mentalité*) ou sur un syntagme (*pour les pour les femmes*). Elles peuvent être continues ou discontinues (*c'était oui c'était la Manche*) ;
- *Les autocorrections.* Les autocorrections sont en quelque sorte une sous-catégorie de répétitions. Il s'agit de répétitions avec substitution d'un ou plusieurs mots par d'autres. Le sujet revient par exemple sur un mot pour :
  - modifier le genre ou le nombre (*ben les le port c'était un peu mieux payé*),
  - pour réaliser une permutation au niveau paradigmatique sans changer la structure syntaxique (*on se réunit tous dans un chez une personne*),
  - pour remodeler la phrase d'un point de vue syntagmatique (*les bâtiments sont c'est pas elle la propriétaire*) ;
- *Les inachèvements.* Les inachèvements apparaissent lorsque le locuteur abandonne un énoncé sans le compléter, le répéter ou le corriger (*Enfin il y a une ...* ou *C'est une chan-*) Bove (2008). Dans ce cas, il peut reprendre son énonciation à partir d'une nouvelle structure, avec ou sans pause intermédiaire, ou arrêter sa production orale. De la même façon que les autres disfluences, la présence d'in-

achèvements n'entrave pas le discours du locuteur et ne gêne pas la compréhension de l'auditeur. Ils sont souvent réalisés dans le but de progresser plus rapidement dans la suite du propos ou de donner la parole plus tôt à l'interlocuteur ;
- *Débit de parole.* Le débit de parole est calculé en faisant le rapport entre le nombre de mots total et le temps de parole total du sujet. Les pauses vides sont exclues car on ne peut pas savoir à quel locuteur attribuer la pause. Pour une estimation du débit oral global, le nombre de mots s'entend à partir des données brutes comprenant les disfluences.

**7.3.2.2 Phénomènes lexicaux** Les phénomènes lexicaux sont essentiellement liés à l'importance, la richesse et la complexité du lexique :
- *Quantité de mots produits.* Il est important d'avoir une indication quantitative globale des productions des sujets. En effet, on pourrait ainsi déduire qu'un sujet produisant peu de mots pourrait avoir, par exemple, une plus faible diversité lexicale qu'un sujet parlant beaucoup. Nous proposons ici une comparaison entre le nombre de mots total (nombre brut) et le nombre de mots sans les phénomènes oraux et disfluences (nombre net) ;
- *Informativité.* Sa faiblesse peut être mise en évidence par l'emploi d'éléments non constitutifs de l'énonciation mais toutefois très fréquents dans les productions orales, comme les interjections (*oh*, *ah*) ou les entités lexicales (*bon*, *là*) ;
- *Énoncés OUI/NON.* Dans les échanges conversationnels étudiés, de nombreux énoncés contiennent uniquement des *oui* ou des *non* (pouvant aller jusqu'à 10 occurrences). De tels énoncés sont aussi moins informatifs que des énoncés normaux mais ne sont généralement pas mentionnés dans la littérature. Il nous a paru donc intéressant de recenser de tels énoncés et d'en observer la proportion dans les productions afin de relativiser le nombre d'énoncés et de prises de parole ;
- *Diversité lexicale.* La diversité lexicale est traduite en nombre de lemmes de substantifs, verbes et adjectifs différents utilisés par un sujet ;
- *Rapport lemmes/mots total.* Le rapport lemmes/mots total est calculé à partir du nombre de mots produits net, car il est plus représentatif au niveau du contenu sémantique du discours ;
- *Complexité morphologique.* La complexité morphologique étudie la structure des mots et leur complexité en terme du nombre de bases et d'affixes. Elle est calculée pour les lemmes des mots non grammaticaux grâce à l'analyse produite par l'analyseur morphologique Dérif ;
- *Fréquence lexicale.* La fréquence lexicale concerne la fréquence d'emploi d'un lemme donné par un sujet dans notre corpus.

**7.3.2.3 Phénomènes syntaxiques** Différents critères liés à la syntaxe des énoncés sont a étudier, essentiellement à partir des corpus lemmatisés. Ils sont essentiellement relatifs à la complexité des énoncés et aux catégories syntaxiques des mots :
- *Longueur moyenne des énoncés (LME).* Dans notre étude, l'énoncé est matérialisé comme étant une phrase. Cette convention a été établie lors de la transcription.

La longueur moyenne des énoncés (LME) est calculée sur les corpus avec un énoncé par ligne, sans les annotations et sans les différents phénomènes oraux (*e.g.* répétitions, autocorrections). Le calcul prend en compte le nombre total de mots par rapport au nombre d'énoncés ;
- *Pronoms personnels.* Nous pouvons étudier les pronoms personnels de deux manières : en nombre d'occurrences et en rapport substantifs/pronoms ;
- *Verbes.* Le nombre moyen de verbes par rapport au nombre total de lemmes est calculé. Nous pouvons calculer la proportion de verbes à un temps et un mode donnés par rapport à tous les verbes conjugués ;
- *Proportion au sein des catégories syntaxiques.* Ce critère couvre la répartition des lemmes en fonction de leurs catégories syntaxiques : substantifs, verbes, adjectifs ;
- *Typologie et complexité des syntagmes.* Ce critère concerne les types de syntagmes (nominaux, verbaux, prépositionnels...) et leur complexité (structure, longueur...).

# 8 Conclusion et Perspectives

Dans le but d'aider les non experts en médecine à mieux comprendre les informations médicales, nous proposons de construire un vocabulaire avec des expressions équivalentes et relatives à deux catégories de personnes : patients et médecins. Pour cela, nous exploitons la reformulation dans le discours médical écrit. Notre corpus de développement provient du forum *masante.net*, où toute personne peut demander une information médicale à des médecins qui y répondent systématiquement. Ce corpus contient les réponses des médecins. Le corpus de test contient les articles du *Portail de la médecine* de wikipédia francophone.

Pour l'extraction de reformulations, nous avons proposé trois méthodes :
- Extraction d'abréviations et de leurs formes étendues,
- Extraction de reformulations introduites par les marqueurs *c'est-à-dire*, *autrement dit* et *encore appelé(e)(s)*,
- Extraction de reformulations marquées par des parenthèses.

Les méthodes sont réglées sur le corpus de développement et ensuite appliquées au corpus de test.

La méthode d'extraction d'abréviations et de leur forme étendue fonctionne à partir du texte brut. Les abréviations et leurs formes étendues sont assez aisées à détecter et extraire, mais les résultats dépendent de la disponibilité et de la complétude de ces informations dans le corpus. Nous obtenons la précision exacte de 0.74, et la précision inexacte de 0.94. Pour les deux dernières méthodes, le corpus est étiqueté and analysé syntaxiquement avec Cordial, ce qui permet de travailler sur les groupes syntaxiques (pour les concepts) et les groupes propositionnels (pour les reformulations). Cette méthode est plus difficile à régler, et nous avons constaté plusieurs difficultés :
- les concepts et les reformulations peuvent être extraits partiellement ou avoir des frontières trop larges,

– les reformulations avec les parenthèses apportent beaucoup de bruit, car la sémantique des parenthèses va au-delà de la reformulation. Nous avons proposé des filtres spécifiques pour limiter le bruit, mais certaines extractions non pertinentes subsistent.

Ces deux méthodes (par marqueurs et par parenthèses) montrent une précision exacte de 0.24 et 0.23, et une précision inexacte de 0.98 et 0.68, respectivement.

Les trois méthodes sont complémentaires. Pour le vérifier, nous avons comparé les résultats à partir du corpus de test, où nous disposons d'un volume de données plus important. Nous avons observé seulement 3 reformulations identiques entre les résultats fournis par les reformulations avec marqueurs et parenthèses, et pas de résultats communs entre les abréviations et les autres méthodes.

Actuellement, nous considérons que les concepts se situent immédiatement avant le marqueur ou les parenthèses. Nous ne traitons donc pas les cas où les deux segments en reformulation sont éloignés, comme dans *l'infection virale dont elle peut-être atteinte, c'est-à-dire, surtout la grippe*, où *la grippe*, qui reformule *l'infection virale*, se trouve à distance dans la phrase. Il s'agit d'un cas relativement rare, mais qui pourrait permettre d'obtenir plus de résultats.

Concernant la méthode d'extraction avec marqueurs, nous avons utilisé les marqueurs les plus fréquents et les plus fiables dans le corpus de développement. Cependant, d'autres marqueurs existent, tels que *l'équivalent de*, *ou encore*, *ou*. Nous avons choisi de ne pas les utiliser, car ils peuvent jouer différents rôles dans la langue. Dans l'avenir, il serait intéressant d'observer leur fonctionnement afin de pouvoir les exploiter également.

Les extractions obtenues avec les trois méthodes sont ensuite alignées avec une terminologie. Ceci fournit 3 934 alignements complets (concept et reformulation) et 27 899 alignements partiels, toutes méthodes confondues. Comme il peut exister plusieurs reformulations pour un concept, il serait intéressant de les classer selon leur pertinence et fiabilité. Notons aussi que les résultats obtenus sur le corpus de test doivent aussi être validés avant leur utilisation. Nous avons classifié les relations qui existent entre les deux segments. Les relations les plus fréquentes sont la synonymie, la définition et la relation de cause à effet. L'ensemble de ces extractions et alignements fournit une bonne base pour la constitution du vocabulaire qui associe les termes d'experts avec les expressions des non experts. Ce type de vocabulaire peut être directement utilisé pour la simplification de documents médicaux et de santé.

Par ailleurs, nous avons également participé dans la constitution d'un corpus de transcriptions de conversations téléphoniques du SAMU d'Arras. Le corpus collecté actuellement est de taille modeste et n'a pas pu être utilisé dans notre travail. Cependant, nous faisons des proposition méthodologiques pour son utilisation ultérieure.

**Expérience personnelle.** Ce stage proposait un travail riche, dans un domaine encore peu exploité. La recherche documentaire m'a permis d'étendre mes connaissances sur un sujet qui m'attire particulièrement. J'ai donc pu découvrir les travaux de recherche qui se font dans le milieu médical au croisement de plusieurs domaines de recherche : linguistique informatique et traitement automatique des langues, création et utilisation de

terminologies, utilisation et développement d'outils automatiques. Par ailleurs, l'analyse de données réelles m'a permise de mener une réflexion sur la manière de les exploiter au mieux afin d'obtenir des résultats attendus. C'était un travail gratifiant, car minutieux et demandant réflexion. La création d'outils informatiques pour effectuer différentes tâches m'a permise de mettre en œuvre mes connaissances en développement informatique. Par ailleurs, l'amélioration des méthodes, l'analyse des résultats obtenus, leur évaluation et l'analyse de celle-ci m'ont permise d'apprendre à avoir du recul face aux résultats.

## Références


Abdaoui, A., Azé, J., Bringay, S., Grabar, N. & Poncelet, P. (2014). Predicting medical roles in online health fora. In *SLSP*, pp. 247–258.

Bannard, C. & Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *ACL*, pp. 597–604.

Barras, C., Geoffrois, E., Wu, Z. & Liberman, M. (1998). Transcriber : a free tool for segmenting, labeling and transcribing speech. In *Conference on Language Resources and Evaluation (LREC)*, pp. 1373–1376.

Bhagat, R. & Hovy, E. (2013). What is a paraphrase ? *Computational Linguistics*, **39**(3), 463–472.

Bot, M.-C. L., Schuwer, M. & Élisabeth Richard (dir.) (2008). *La reformulation : Marqueurs linguistiques – Stratégies énonciatives*. Rennes : Rivages linguistiques.

Bouamor, H., Max, A. & Vilnat, A. (2012). Étude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques. *TAL*, **53**(1), 11–37.

Boubé, N. & Tricot, A. (2010). *Qu'est-ce-que rechercher de l'information ? : état de l'art*. Villeurbanne : Presses de l'Enssib.

Bove, R. (2008). *Analyse syntaxique automatique de l'oral : étude des disfluences*. Thèse de doctorat, Université d'Aix-Marseille I, Marseille.

Boyé, M., Tran, T. & Grabar, N. (2014). Nlp-oriented contrastive study of linguistic productions of alzheimer and control people. In L. . Springer, Advances in Natural Language Processing, Ed., *POLTAL*, pp. 412–424.

Cartoni, B. & Deléger, L. (2011). Découverte de patrons paraphrastiques en corpus comparable : une approche basée sur les n-grammes. In *TALN*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

Côté, R. A., Rothwell, D. J., Palotay, J. L., Beckett, R. S. & Brochu, L. (1993). *The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International*. Northfield : College of American Pathologists.

Dale, E. & Chall, J. (1948). A formula for predicting readability. *Educational research bulletin*, **27**, 11–20.

Delbecque, T., Jacquemart, P. & Zweigenbaum, P. (2005). Utilisation du réseau sémantique de l'umls pour la définition de types d'entités nommées médicales. *CORIA*, pp. 101–118.

Deléger, L. & Zweigenbaum, P. (2008). Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA 2008*, pp. 146–50.

Grabar, N. & Hamon, T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. In *TALN 2015*, Caen, France. To appear.

Grabar, N. & Zweigenbaum, P. (2000). A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, pp. 310–314.

Gulich, E. & Kotschi, T. (1983). Les marqueus de la reformulation paraphrastique. *Cahiers de linguistique française*, **5**, 305–351.

Henry, S. & Pallaud, B. (2004). Amorces de mots et répétitions dans les énoncés oraux. *Recherches sur le français parlé*, **18**, 201–229.

Jacquemin, C. (1994). Recycling terms into a partial parser. In *Proceedings of the fourth conference on Applied natural language processing*, Stuttgart, Germany.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., ans N Bertoldi, M. F., Cowan, B., Shen, W., Mora, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. & Herbst, E. (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, Prague, République Tchèque.

Landis, J. & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Laurent, D., Nègre, S. & Séguéla, P. (2009). Apport des cooccurrences à la correction et à l'analyse syntaxique. In *TALN*.

Lindberg, D., Humphreys, B. & McCray, A. (1993). The unified medical language system. *Methods Inf Med*, **32**(4), 281–291.

Madnani, N. & Dorr, B. J. (2010). Generating phrasal and sentential paraphrases : A survey of data-driven methods. *Computational Linguistics*, **36**, 341–387.

McCray, A. T. (1989). The UMLS semantic network. In *Proceedings of the 13$^{th}$ Annual SCAMC*, pp. 503–507, Washington.

McCray, A. T., Loane, R. F., Browne, A. C. & Bangalore, A. K. (1999). Terminology issues in user access to web-based medical information. In *Proceedings of the AMIA Symposium*, pp. 107 : American Medical Informatics Association.

Namer, F. (2000). FLEMM : un analyseur flexionnel du français à base de règles. *Traitement automatique des langues (TAL)*, **41**(2), 523–547.

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

NLM (2001). *Medical Subject Headings*. National Library of Medicine, Bethesda, Maryland. *www.nlm.nih.gov/mesh/meshhome.html*.

Och, F. & Ney, H. (2003). A systematic comparison of various statistical alignment models. In *Computational Linguistics*.

Pallaud, B. (2002). Les amorces de mots comme faits antonymiques en langage oral. *Recherches sur le français parlé*, **17**, 79–102.

Pang, B., Knight, K. & Marcus, D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. In *Proceedings of NAACL-HLT*, pp. 102–109, Edmonton, Canada.

Petrov, S. & Klein, L. B. R. T. D. (2006). Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australie.

Ratzan, S. & Parker, R. (2000). *Introduction. In : National Library of Medicine current bibliographies in medicine : health literacy, Selden CR, Zorn M, Ratzan SC, Parker RM,*. U.S. Department of Health and Human Services : NLM Pub No CMB 200-1. Bethesda, MD : National Institutes of Health.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, pp. 44–49, Manchester, UK.

Schwartz, A. S. & Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Pacific Symposium on Biocomputing*, pp. 451–456.

Vergely, P., Condamines, A., Fabre, C., Josselin-Leray, A., J Rebeyrolle, J. & Tanguy, L. (2009). Analyse linguistique des interactions patient/médecin. *Actes éducatifs de soins*, **92**(5).

Zeng, Q. & Tse, T. (2006). Exploring and developing consumer health vocabularies. *JAMIA*, **13**, 24–29.

Zeng, Q. T., Kim, E., Crowell, J. & Tse, T. (2005a). A text corpora-based estimation of the familiarity of health terminology. In *ISBMDA 2006*, pp. 184–92.

Zeng, Q. T., Tse, T., Crowell, J., Divita, G., Roth, L. & Browne, A. C. (2005b). Identifying consumer-friendly display (CFD) names for health concepts. In *AMIA 2006*, pp. 859–63.

Zeng, Q. T., Tse, T., Divita, G., Keselman, A., Crowell, J. & Browne, A. C. (2006). Exploring lexical forms : first-generation consumer health vocabularies. In *AMIA 2006*, pp. 1155–1155.

Zielstorff, R. D. (2003). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*.

Zweigenbaum, P. & Grabar, N. (2003). Corpus-based associations provide additional morphological variants to medical terminologies. In *American Medical Informatics Association (AMIA)*.

# A Guide d'annotation des reformulations

## A.1 Annotation des abréviations et de leur formes étendues

Pour les abréviations : Considérées comme un concept à reformuler, les abréviations sont balisées avec la lettre *C* pour *concept* : <C>ABREVIATION</C>

Pour les définitions : Nous avons deux sortes de balises, selon si la reformulation est complète ou non :

- si celle-ci est complète, c'est-à-dire que chaque mot de l'abréviation est défini tel que *virus respiratoire syncitial (VRS)*, elle est balisée ainsi : <Rspe>signification complète de l'abréviation</Rspe>. Le *spe* signifie *spécifique*.
- si la reformulation est incomplète mais générique, elle est balisée ainsi : <Rgen>signification de l'abréviation</Rgen>

Les reformulations des concepts sont balisées dans leur ensemble, et non mot par mot. Les mots vides sont ainsi pris en compte s'il y en a. En revanche, nous n'y insérons pas les mots vides avant et après : *un <Rgen>examen des urines</Rgen>* et non : *<Rgen>un examen des urines</Rgen>*

Nous ne gardons que les concepts accompagnées de leur reformulations : nous ignorons donc les concepts dépourvus de reformulation.

Nous ne gardons que les phrases comprenant les patrons suivants :

- reformulation en plusieurs mots + (ABREVIATION)
- ABREVIATION + (reformulation en plusieurs mots)

Nous ignorons les structures telles que : *les transaminases (ALAT, ASAT ou anciennement appelées SGPT, SGOT)* ou encore *une MST (on dit IST)*, car celles-ci ne contiennent pas de réelle définition.

Des cas tels que : *VIH (virus du sida)* sont à considérer comme génériques car nous y trouvons tout de même un mot correct, mais ils seront plutôt traités dans le programme d'extraction de reformulation par parenthèses.

Afin de vérifier la définition d'une abréviation, vous pouvez consulter le lien suivant : *https ://fr.wikipedia.org/wiki/Liste_d'abréviations_en_médecine*

## A.2 Annotation des reformulations introduites par des marqueurs (c'est-à-dire, autrement dit, encore appelé)

Ce type de reformulation est divisé en trois parties : le concept, le marqueur, la reformulation. Nous avons donc une balise pour chaque type :

- pour le concept : <C>concept</C>
- pour le marqueur : <M>marqueur</M>
- pour la reformulation : nous distinguons les reformulations génériques (valables de façon générale) et contextuelles (spécifiques à la situation donnée du corpus) :
  - exemple de reformulation générique : <C>un progestatif</C>, <M>c'est-à-dire</M> <Rgen>qu'il favorise la muqueuse utérine</Rgen>
  - exemple de reformulation contextuelle : <C>pas en faveur de lésions organisées</C>. <M>Autrement dit</M>, <Rspe>votre examen ne montre pas

d'anomalie très évocatrice</Rspe>

Nous tachons de préserver une cohérence dans les annotations concept/reformulation, concernant les éléments morphosyntaxiques. Si un déterminant se trouve dans le concept, nous le gardons aussi dans la reformulation s'il y en a un. Par exemple :
*<C>un gastroentérologue spécialisé</C>, <M>c'est-à-dire</M> <Rgen>un proctologue</Rgen>.*
et non : *un <C>gastroentérologue spécialisé</C>, <M>c'est-à-dire</M> <Rgen>un proctologue</Rgen>.*
ou encore : *<C>un gastroentérologue spécialisé</C>, <M>c'est-à-dire</M> un <Rgen> proctologue</Rgen>.*

Dans certains cas, des exemples peuvent être ajoutés, faisant partie de la reformulation, car il ajoute une explication supplémentaire. Nous considérons la reformulation comme *générique*. Par exemple :
*<C>une enzyme protéolytique</C>, <M>c'est à dire</M> <Rgen>qui digère les protéines comme le fait le suc pancréatique</Rgen>.*

Dans d'autres cas, un ajout d'information rend la reformulation contextuelle si on le considère faisant partie de celle-ci. Deux solutions s'offre à nous :

- nous ne prenons que la partie générique, faisant office de définition suffisante pour le concept. Par exemple : *<C>un blepharospasmme</C>, <M>c'est-à-dire</M>, <Rgen>un mouvement saccadé</Rgen> de la paupière* ; ce type de phénomène peut apparaître sur tous les muscles du corpus humain, et non spécifiquement sur la paupière.
- nous considérons l'information supplémentaire comme faisant partie de la reformulation, ce qui la rend spécifique. Par exemple : *<C>un blepharospasmme</C>, <M>c'est-à-dire</M>, <Rspe>un mouvement saccadé de la paupière</Rspe>*

Des problèmes peuvent survenir dans le corpus, comme des phrases coupées. Si les informations sont trop incomplètes, nous ne prenons pas ces phrases en considération dans notre annotation. Par exemple : *Cher(e) nacer, La sinusite est une inflammation chronique des sinus, c'est-à-dire, des c Bien cordialement.* Ici, nous ne pouvons pas dire si la reformulation est générique ou spécifique.

En cas de doute sur l'annotation du concept (quoi prendre, quoi ne pas considérer comme faisant partie du concept), il est peut être utile, dans certains cas de procéder à une substitution. Par exemple : *la muqueuse hyperplasiée, autrement dit, trop importante.* Nous pouvons remplacer *hyperplasiée* par *trop importante* et non *la muqueuse hyperplasiée* entièrement. Seule *hyperplasiée* serait considéré comme le concept.

## A.3 Annotation des reformulations par parenthèses

Les reformulations par parenthèses se divisent en deux parties, comme pour les abréviations : le concept, généralement accolé à la parenthèses ouvrante, et la reformulation, comprise dans les parenthèses. Les balises sont semblables aux autres annotations :

- concept : <C>concept</C>
- reformulation : <Rgen>(reformulation)</Rgen> pour les reformulation génériques (générales) ou <Rspe>(reformulation)</Rspe> pour les reformulations

spécifiques (ou contextuelles à la situation).

Notons que nous prenons en compte les parenthèses comme faisant partie de la reformulation. Elles jouent le rôle de délimitation de la reformulation, même si des informations supplémentaires peuvent y être intégrées (comme dans l'exemple : *(trop de globules rouges dans vos urines car la norme est de 10000)*).

Certains parenthésages ne sont pas des reformulations par équivalence mais ont plutôt un rôle d'exemple. Pour cela, nous utilisons la balise suivante : *<C>des laxatifs doux</C> <Ex>(type polyethylene glycol)</Ex>*. Elles sont reconnaissables grâce à des termes tels que *comme, type, par exemple*, des énumérations, *etc* ou encore aux points de suspension. Les noms de traitements ou de maladie peuvent être aussi considérés comme des exemples. D'autres font office de précisions. Nous Introduisons donc une dernière balise comme ceci : *<C>des traitements</C> <Pre>(progestatifs)</Pre>* Les précisions ne jouent pas le rôle de définition, contrairement au reformulations *spécifiques*, mais affinent l'information du concept.

Concernant les concepts, nous tâchons de garder les groupes nominaux, c'est-à-dire un nom et son déterminant. Par exemple :
*<C>des prélèvements</C> <Rgen>(biopsie)</Rgen>*
Et non : *des <C>prélèvements</C> <Rgen>(biopsie)</Rgen>*

Attention : tout parenthésage n'est pas nécessairement une reformulation ; il faut donc distinguer les reformulations et les autres types de parenthésages.

- Nous ignorons tout ce qui ne sera pas considéré comme une reformulation, une précision ou un exemple. Par exemple, les parenthèses commençant par *et, y compris, surtout*, qui sont des ajouts d'informations.
- Les phrases commençant pas *qui, en, mais, ou.*
- Celles contenant un point d'interrogation ou d'exclamation, s'ils expriment respectivement une recherche d'information de la part du médecin ou un commentaire. Ou tout autre signe montrant ce type de phrases (mot interrogatif par exemple).
- Celles contenant des données temporelles telles que *3 jours, dans une semaine, pendant 3 mois, plusieurs fois, rarement*, etc.
- Ou encore les phrases contenant des signes tels que *%* (trop précis), les mots *pharmacien* et *médecin*, inclus dans des phrases types *demandez conseil à votre pharmacien.*
- Les causalités : *vous n'êtes pas contaminé (vous êtes guéri d'une contamination ancienne).* Mais certains termes peuvent aider tels que *parce qu, car, puisqu'.*

Notes générales :

- Les phrases contenant plusieurs parenthèses à annoter sont à dupliquer (une annotation par phrase)
- Des recherches sur Google peuvent être effectuées pour résoudre un doute.