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

Edwide Antoine Natalia Grabar
CNRS, UMR 8163, F-59000 Lille, France
Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
edwige.anto@gmail.com, natalia.grabar@univ-lille3.fr

RÉSUMÉ

Les notions de domaines techniques, comme les notions médicales, présentent souvent des difficultés de compréhension par les non experts. Un vocabulaire qui associe les termes techniques aux expressions grand public peut aider à rendre les textes techniques mieux compréhensibles. L'objectif de notre travail est de construire un tel vocabulaire. Nous proposons d'exploiter la notion de reformulation grâce à trois méthodes : extraction d'abréviations, exploitation de marqueurs de reformulation et de parenthèses. Les segments associés grâce à ces méthodes sont alignés avec les terminologies médicales. Nos résultats permettent de couvrir un grand nombre de termes médicaux et montrent une précision d'extraction entre 0,68 et 0,98. Au total, plusieurs dizaines de milliers de paires sont proposés. Ces résultats sont analysés et comparés avec les travaux existants.

ABSTRACT

Exploitation of reformulations for the acquisition of expert/non-expert vocabulary.

Notions from technical areas, such as medicine, often present understanding difficulties for non-expert people. A vocabulary which associates technical terms with expressions used by lay people can help in making the technical texts easier to understand. The purpose of this work is to build such vocabulary. We propose to exploit the notion of reformulation through three methods : extraction of abbreviations, exploitation of reformulation markers, and of parentheses. The segments associated with these methods are then aligned with medical terminologies. Our results cover a large number of medical terms and show between 0.68 and 0.98 extraction precision. On the whole, several dozens of thousands of pairs are proposed. These results are analyzed and compared with the existing work.

MOTS-CLÉS : Reformulation, extraction d'information, terminologie médicale, langage profane.

KEYWORDS: Reformulation, information extraction, medical terminology, layman language.

1 Introduction

Dans tout 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 (McCray, 2005). Ceci les amène souvent à l'incapacité de prendre une décision face à un traitement, à comprendre les conséquences de la maladie et du traitement, ou tout simplement à comprendre leur maladie. Cette constatation devient plus importante car les patients ont un accès accru aux informations en ligne

(Zeng & Tse, 2006; Tran *et al.*, 2012) : de manière paradoxale, la disponibilité des informations accentue la barrière de l'incompréhension. Un patient, étant non expert du milieu médical, ne peut pas avoir les connaissances nécessaires pour comprendre les informations données par le corps médical, ni même pour rechercher les informations dont il a besoin. Trouver une information implique en effet 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). À cela s'ajoute le fait que le patient ne demande pas toujours les informations dont il a besoin à son médecin : 45% des patients tendent à utiliser internet, et seulement 16% se réfèrent à leur médecin (Zielstorff, 2003). S'il y a bien deux types de discours (patients *vs* 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).

L'objectif de notre travail consiste à acquérir un vocabulaire associant les termes spécialisés avec des expressions grand public pour faciliter la compréhension de termes médicaux par les non experts. Nous proposons d'exploiter la reformulation dans le discours médical, en utilisant des textes rédigés par des spécialistes ou bien issus de l'écriture collaborative, afin de garantir une plus grande fiabilité des informations extraites. Notre travail est principalement basé sur trois hypothèses :

– lorsqu'un professionnel de santé reformule un terme technique dans un texte grand public, cela indique qu'il s'agit d'une expression inconnue ou mal connue du public non expert ;
– l'acte de reformulation d'un terme technique par un professionnel médical peut permettre d'extraire ce terme ainsi que sa reformulation ;
– la reformulation est un phénomène langagier spontanément utilisé par les locuteurs dans différents types de textes et de discours.

Dans la suite du travail, nous présentons d'abord les travaux de l'état de l'art (section 2). Nous décrivons ensuite le matériel exploité (section 3) et la méthode proposée (section 4). Nous présentons les résultats obtenus et les discutons (section 5) et concluons avec des perspectives (section 6).

2 Contexte et Travaux de l'état de l'art

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). On parle alors d'*alphabétisation médicale* (ou *health literacy*), qui renvoie à la capacité de comprendre une information médicale, sa lecture et interprétation, ou bien la capacité de pouvoir rechercher les informations d'un domaine (Ratzan & Parker, 2000). Le degré d'alphabétisation médicale varie d'un individu à l'autre, et dépend de l'éducation et de la catégorie socio-professionnelle d'une personne, mais aussi de son histoire médicale (Zeng *et al.*, 2005a). Notons aussi que le vocabulaire profane médical est *dynamique* (Zeng & Tse, 2006) : il est aussi instable qu'un langage à proprement parler, et varie dans le temps et en fonction des personnes. En effet, à travers ses recherches, le patient tente de combler son manque de connaissances. De plus, ces notions peuvent être perçues différemment par rapport aux médecins : des termes comme *dépression* ou *paranoïa*, considérés comme des pathologies par les médecins, peuvent être utilisés par un non expert pour signifier un état de tristesse plus ou moins important (Zeng & Tse, 2006) et une notion plus ou moins péjorative, respectivement. De même, le terme *glycémie* est appréhendé comme *taux de sucre*, ce qui permet de l'adapter au niveau des connaissances du patient. On parle alors du processus de *traduction* (Zeng & Tse, 2006) qu'un non expert effectue cognitivement. Notons aussi que les

patients utilisent souvent des termes plus génériques par rapport aux experts, ce qui est aussi le cas des dictionnaires généraux qui proposent des définitions des termes médicaux (Zeng & Tse, 2006). Lors de cette *traduction*, le patient risque alors de perdre une partie des informations que le médecin lui fournit. Il est donc important de disposer de vocabulaires qui proposent les équivalences grand public des termes techniques.

Reformulation et paraphrase. La reformulation correspond à l'action de redire d'une manière différente quelque chose qui a déjà été dit (Le Bot *et al.*, 2008), augmentant ainsi les chances de compréhension. La reformulation peut être effectuée à la demande d'un interlocuteur ou par décision du locuteur, conscient ou non que les informations puissent être difficiles à comprendre par d'autres. La reformulation peut se manifester par des schémas parfois liés par un marqueur (*e.g. c'est-à-dire*, *autrement dit*), les parenthèses, etc. Les segments reformulés n'introduisent pas toujours des équivalents sémantiques (Gulich & Kotschi, 1983), car il est possible d'avoir aussi des incises, des disfluences, des relations de causalité, etc. (Grabar & Eshkol-Taravella, 2015). En revanche, lorsque la reformulation fournit une équivalence sémantique, il est possible d'en extraire des paraphrases. La paraphrase est une notion répandue en langue et importante pour plusieurs domaines de recherche en TAL : la recherche d'information, la génération de texte, la synthèse de texte, la didactique des langues. Il reste cependant que cette notion est difficile à définir de différents points de vue, comme sa taille (Flottum, 1995; Fujita, 2010; Bouamor, 2012) et le type (Vila *et al.*, 2011; Bhagat & Hovy, 2013). Dans un domaine de spécialité, comme par exemple le domaine médical, nous supposons que les reformulations permettent de détecter les synonymes ou les paraphrases de termes techniques.

Détection automatique de paraphrases pour les termes médicaux. Nous présentons quelques travaux proposés pour la reconnaissance automatique de paraphrases dans la langue spécialisée, essentiellement en corpus comparables ou en corpus grand public :

- Un traducteur automatique de termes médicaux vers des expressions profanes et inversement a été proposé (McCray *et al.*, 1999). L'objectif principal de ce travail est d'introduire la ressource MEDLINE*plus* [1], organisée selon les sujets médicaux (cancers, obésité, etc). La méthode exploitée pour construire cette ressource n'est pas décrite. Un autre lexique patient en anglais (Consumer Health Vocabulary ou CHV) est développé (Zeng *et al.*, 2006) afin de favoriser la recherche d'information pour les patients. Il s'agit d'une initiative collaborative, impliquant des annotateurs humains, qui a pour but d'identifier les expressions profanes évoquant un terme médical et de les lier aux termes techniques. Là encore, l'accent principal est mis sur la description de la ressource. La ressource contient actuellement 141 213 termes uniques (ou 145 473 paires identifiant/terme uniques). Une majeure partie de ces termes sont identiques avec les termes techniques d'UMLS : après la normalisation de casse et d'ordre des mots, la suppression de ponctuation, tiret, pluriel, *-ing* finals et *of*, la ressource CHV propose 11 641 termes uniques (11 896 paires identifiant/terme uniques). Cette ressource a permis d'effectuer des tests avec le grand public : la compréhension de termes (Zeng *et al.*, 2005b) et de synonymes avec différents niveaux de technicité (Zeng *et al.*, 2005a). Cette série de travaux a également permis de relever les couples qui montrent une compréhension partielle (*e.g. dépression*) et les expressions profanes inexistantes dans le vocabulaire médical.
- Concernant les travaux effectués sur le français, des corpus monolingues comparables ont été étudiés afin d'en extraire des syntagmes utilisés par les experts et les non experts (Deléger &

1. https://www.nlm.nih.gov/medlineplus/

Zweigenbaum, 2008; Cartoni & Deléger, 2011). Les patrons morphosyntaxiques, les mesures de similarité ou les n-grammes permettent d'associer des syntagmes des deux corpus. Les auteurs montrent également que les médecins ont plutôt recours à des syntagmes nominaux (*traitement*) là où les non experts utilisent des syntagmes verbaux (*traiter*). La ressource obtenue, d'une taille réduite, n'est pas disponible.

– Un autre travail a proposé de détecter des paraphrases en français pour les composés néoclassiques, très fréquents dans la langue médicale (Grabar & Hamon, 2015). La méthode exploite l'analyse morphologique de Dérif (Namer, 2009), traduit les racines grecques et latines vers le français (*myocardiaque : myo=muscle, cardia=cœur*), et recherche ensuite les syntagmes qui contiennent les mots de la décomposition des termes néoclassiques (*muscle du cœur*). Ce travail ne couvre que des composés néoclassiques et ne traite pas les constructions syntaxiques (*infarctus du myocarde*).

3 Matériel

3.1 Les terminologies médicales

Nous exploitons (1) la partie en français de l'UMLS (Unified Medical Language System) (Lindberg *et al.*, 1993), qui est un ensemble de ressources terminologiques, et (2) la terminologie SNOMED International (Systematized Nomenclature of Medicine) (Côté *et al.*, 1993) diffusée par ASIP santé [2]. Ces deux terminologies sont complémentaires et totalisent 323 964 termes. Chaque terme est associé à au moins un type sémantique et un groupe sémantique. Par exemple, *infarctus* et *AVC* sont des pathologies, *aspirine* un produit chimique. Notons que l'UMLS contient également les vocabulaires *MedlinePlus* (McCray, 1989) et CHV (Zeng *et al.*, 2006), mais disponibles en anglais uniquement.

3.2 Corpus

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

Le corpus de développement est issu du forum *masante.net* [3], créé et modéré par des médecins. Ce site permet aux utilisateurs de poser des questions médicales, auxquelles deux professionnels de santé répondent. Nous exploitons la partie composée des réponses de médecins car nous supposons qu'ils peuvent expliquer des termes médicaux potentiellement compliqués et qu'ils fournissent des informations fiables. Le corpus contient 6 139 réponses, totalisant 315 362 occurrences.

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, totalisant 15 235 219 occurrences. Le corpus contient des informations encyclopédiques sur plusieurs notions médicales. Les contributeurs ont en général une bonne connaissance des sujets abordés et, de plus, ils écrivent de manière collaborative. Leur objectif est de présenter les notions techniques et de les rendre accessibles au grand public.

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

3.3 Ressources linguistiques

Mots vides. Les mots vides sont les mots grammaticaux tels que *de, et, à, ou, etc*, les auxiliaires *est, a*, et certains adverbes *tout, plusieurs*. Notre liste contient 111 mots vides.

Ressources morphologiques. Ces ressources permettent d'assurer une normalisation des termes. Elles comportent 163 823 paires de mots couvrant les dérivations {*aorte*; *aortique*} et les flexions {*aortique*; *aortiques*}. Elles sont issues des travaux existants (Grabar & Zweigenbaum, 2000; Zweigenbaum & Grabar, 2003) et ont été complétées à partir de nos corpus.

4 Méthodes

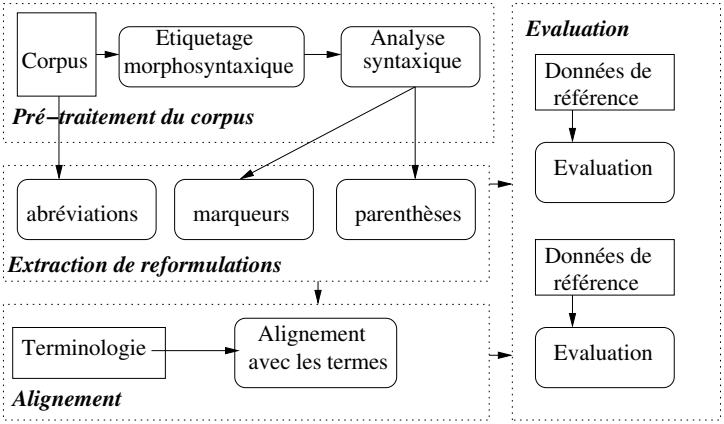


Fig. 1 – Schéma général de notre approche

La figure 1 présente le schéma général de notre approche, qui se compose de quatre étapes : (1) pré-traitement de corpus (section 4.1), (2) extraction de reformulations (section 4.2), (3) alignement avec la terminologie médicale (section 4.3), et (4) évaluation des résultats (section 4.4). Pour l'extraction de reformulations, qui est l'étape principale de l'approche, nous traitons trois cas de figure, tous très fréquents dans les documents médicaux, afin de détecter :

– les siglaisons, souvent difficiles à comprendre, et de leurs formes étendues (section 4.2.1), dans les structures *forme étendue (abréviation)* et *abréviation (forme étendue)* ;
– les reformulations avec trois marqueurs très fréquents dans les corpus *c'est-à-dire, autrement dit, encore appelé* (section 4.2.2), dans la structure *concept marqueur reformulation* ;
– les reformulations parenthésées (section 4.2.3), dans la structure *concept (reformulation)*.

4.1 Pré-traitement de corpus

Les corpus sont étiquetés et analysés syntaxiquement avec Cordial (Laurent *et al.*, 2009). L'analyse syntaxique permet de délimiter les syntagmes et sert de base pour l'extraction de reformulations. Dans le tableau 1, nous présentons un extrait de sorties de Cordial pour la phrase *Vous devez les faire brûler par un gastroentérologue spécialisé, c'est-à-dire un proctologue*. Les champs exploités sont les

formes, les lemmes, et les informations syntaxiques : étiquetage syntaxique (*POS, POSMT*, groupes syntaxiques (*GS, type GS*) et propositions (*Prop*).

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.

4.2 Trois méthodes pour l’extraction de reformulations

4.2.1 Extraction des siglaisons

Nous traitons les sigles avec un algorithme de l’état de l’art (Schwartz & Hearst, 2003), qui fonctionne sur le corpus brut et détecte deux types de structures : (1) *forme étendue/concept (abréviation)* et (2) *abréviation (forme étendue/concept)*. Contrairement au travail d’origine, nous prenons en compte les majuscules car elles marquent mieux les abréviations. La distinction entre ces deux patrons est effectuée selon le nombre de mots entre parenthèses : si un seul mot se trouve entre parenthèses, alors il s’agit du premier patron et vice versa. Ensuite, les mots proches de l’abréviation sont parcourus et comparés avec les lettres de l’abréviation : pour le patron 2, il s’agit de mots entre les parenthèses ; pour le patron 1, il s’agit de mots avant les parenthèses parcourus en sens inverse. Trois cas de figures peuvent se présenter :

– toutes les lettres de l’abréviation sont appariées : il s’agit alors d’une *extraction complète* ;
– une partie des lettres est appariée : il s’agit alors d’une *extraction incomplète* ;
– aucune lettre de l’abréviation n’est appariée.

4.2.2 Extraction des reformulations avec marqueurs

La méthode exploite trois notions, que nous introduisons sur l’exemple *un gastroentérologue spécialisé, c’est-à-dire un proctologue* : (1) le marqueur de reformulation *c’est-à-dire* ; (2) le concept

ou segment reformulé, à gauche du marqueur (*un gastroentérologue spécialisé*) ; et (3) la reformulation, à droite du marqueur (*un proctologue*). Dans cette structure, nous nous attendons à avoir le terme plus spécialisé à gauche du marqueur, comme dans *une sécrétion de colostrum, c'est-à-dire une gouttelette venue des canaux galactophores* du tableau 2. Cependant, le terme plus spécialisé peut également se trouver à droite du marqueur, comme dans *Vous devez les faire brûler par un gastroentérologue spécialisé, c'est-à-dire un proctologue* du tableau 1. La désambiguïsation (complexification ou simplification) doit être effectuée manuellement par la suite.

Nous avons considéré trois niveaux d'informations syntaxiques de Cordial :
- *Catégories syntaxiques* (champs *POS* et *POSMT* du tableau 1) : sont difficiles à exploiter car présentent peu de régularités et demandent à définir des patrons syntaxiques spécifiques ;
- *Groupes syntaxiques* (champ *GS* du tableau 1) : sont des informations exploitables car les groupes de mots qui précèdent et suivent le marqueur appartiennent généralement au même groupe syntaxique (groupes 8/F et 16/D dans le tableau 1). Cependant, certains segments peuvent s'étendre sur plusieurs syntagmes, comme dans *une sécrétion de colostrum, c'est-à-dire une gouttelette venue des canaux galactophores*, où la reformulation *une gouttelette venue des canaux galactophores* correspond à deux syntagmes (17 et 20 dans le tableau 2).
- *Propositions* (champ *Prop* du tableau 1) sont aussi intéressants à exploiter et permettent d'aller au-delà des groupes syntaxiques et d'obtenir des extractions probablement plus complètes.

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 – Analyse syntaxique d'un extrait de la phrase *une sécrétion de colostrum, c'est-à-dire une gouttelette venue des canaux galactophores*.

Nous testons deux versions de la méthode pour l'extraction de reformulations, la *méthode par groupes syntaxiques* et la *méthode par groupes propositionnels*, alors que le repérage des concepts est toujours effectué avec les syntagmes. Nous commençons par la détection de phrases contenant les marqueurs de reformulation. Ensuite, pour extraire les segments en relation de reformulation, nous nous repérons par rapport au marqueur : nous parcourons en sens inverse les mots situés avant le marqueur et, s'ils appartiennent au même syntagme, nous les retenons comme le concept ; nous parcourons ensuite les mots situés après le marqueur, que nous considérons comme la reformulation s'ils appartiennent au même syntagme ou à la même proposition. Les tests préliminaires indiquent que la *méthode par groupes propositionnelles* est plus performante : elle offre des résultats avec des reformulations plus complètes et dépasse de 20 % la *méthode par groupes syntaxiques*. C'est donc la méthode par groupes propositionnelles qui est exploitée dans la suite du travail.

4.2.3 Extraction de reformulations parenthésées

L'extraction des reformulations marquées par des parenthèses est aussi effectuée à partir de corpus analysés syntaxiquement par Cordial. Le déroulement de la méthode est le suivant :

– les mots entre parenthèses sont détectés et considérés comme la reformulation,
– le concept est extrait en parcourant les mots qui précèdent les parenthèses et en utilisant les codes syntaxiques (même approche que dans la section 4.2.2) ;
– des filtres, à base de marqueurs spécifiques, permettent d'éliminer certains candidats (essentiellement des précisions temporelles, des énumérations, des incises et des oppositions) ;
– les reformulations parenthésées portant sur les siglaisons ne sont pas traitées car elles sont prises en charge par une méthode dédiée (section 4.2.1).

4.3 Alignement de segments extraits avec les termes médicaux

L'objectif de l'alignement des segments extraits avec les termes médicaux est double :
– vérifier la pertinence des extractions : si au moins un des segments extraits se trouve dans la terminologie, alors il s'agit certainement d'une reformulation concernant une notion médicale ;
– associer les segments extraits aux termes médicaux et rendre possible l'exploitation de la ressource constituée dans des contextes de recherche d'information, d'indexation, etc.

L'alignement est effectué avec une méthode qui permet de comparer les segments extraits avec les termes de la terminologie, en effectuant la désaccentuation, une normalisation morphologique grâce aux ressources (section 3.3) et en supprimant les mots vides. Cette méthode permet de contrôler le taux d'alignement entre le segment extrait et le terme. Nous pouvons donc exploiter ces deux paramètres pour optimiser les alignements. Lorsque des alignements complets (100 % pour le segment et 100 % pour le terme) sont disponibles, ils sont proposés en priorité : {*AINS : ains.C0003211/C-60300*; *anti inflammatoires non stéroïdiens : anti inflammatoires steroidiens.C0003211*}. En revanche, lorsque les alignements complets ne sont pas disponibles, les alignements partiels sont proposés comme candidats : *les réactions de défense : defense/T-540A0, defense hote/F-C0480*.

4.4 Évaluation

Nous constituons des jeux de référence, en suivant des guides d'annotation, pour évaluer l'extraction de reformulations et pour évaluer l'alignement avec les termes. Deux annotateurs sont impliqués dans l'annotation manuelle du corpus de développement et une séance de consensus est effectuée à la fin. Nous utilisons le Kappa de Cohen (Cohen, 1960) pour calculer l'accord inter-annotateur. L'accord est situé entre 0 et 1, où 1 correspond à l'accord parfait entre les annotateurs (Landis & Koch, 1977).

Les phrases contenant les reformulations sont annotées : les concepts sont balisés <*C*>*concept*</*C*>, les marqueurs <*M*>*marqueur*</*M*>, et les reformulations <*R*>*reformulation*</*R*>. Les propositions d'alignement du corpus de développement avec les taux de recouvrement 40(segment)-40(terme), sont validées par deux annotateurs. Les alignements complets et partiels sont considérés.

Les accords inter-annotateur se trouvent dans le tableau 3 : pour les extractions et les alignements. Pour les extractions, l'accord sur les mots 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 mots. L'accord concernant les reformulations avec parenthèses est modéré, tandis que les deux autres accords sont presque parfaits. Le taux modéré avec les parenthèses est sans doute causé par une plus grande complexité du jugement sur leur pertinence. L'accord inter-annotateur des alignements indique 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	*Mots*	
Abréviations	0,661	0,967	0,208
Marqueurs	0,24	0,816	0,714
Parenthèses	0,651	0,575	0,817

TABLE 3 – Accord inter-annotateur des extractions et des alignements pour chaque méthode, au niveau des phrases et des mots pour les extractions de reformulations.

Nous utilisons trois mesures classiques d'évaluation : précision P, rappel R et F-mesure F. La performance des méthodes d'extraction est effectuée au niveau des deux segments : concept et reformulation. L'évaluation prend en compte les frontières de ces segments. Elle est effectuée avec le script d'évaluation de la tâche 3 de la campagne DEFT 2015 [4]. L'évaluation est effectuée sur les segments extraits exacts (les frontières sont alors respectées) et inexacts (des recouvrements entre les extractions et la référence sont acceptés). La qualité de l'alignement est évaluée par rapport aux données de référence, en prenant en compte les propositions correctes et incorrectes. Cette évaluation, effectuée sur le corpus de développement pour lequel les données de référence sont créées, permet de définir les seuils optimaux d'alignement qui sont ensuite appliqués au corpus de test.

5 Résultats

Nous présentons et discutons les résultats de l'extraction (section 5.1) et de l'alignement (section 5.2), et proposons une comparaison entre nos méthodes et avec l'état de l'art (section 5.3).

5.1 Résultats d'extraction

	Corpus de développement			*Corpus de test*		
	Abrév.	*Marqueurs*	*Parenthèses*	*Abrév.*	*Marqueurs*	*Parenthèses*
nb occ.	75	96	312	88 762	2 757	100 103
nb types/ref.	42	96	305	8 106	2 710	92 971

TABLE 4 – Nombre d'extractions pour chaque méthode, sur les corpus de développement et de test.

Une première remarque est que les reformulations peuvent correspondre à différentes situations : définitions, explications, synonymes, etc. (Grabar & Eshkol-Taravella, 2015). Du moment où la reformulation permet d'expliciter le terme reformulé, elle est considérée comme intéressante. Dans le tableau 4, nous indiquons les résultats quantitatifs des extractions. Globalement, nous voyons que le corpus de test, étant beaucoup plus grand, fournit plus d'extractions, que les reformulations avec les parenthèses sont les plus productives, et que les abréviations sont les plus redondantes.

Avec l'extraction des abréviations, nous observons différents cas :

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

– correct : {*ESF*; *Editions Sociales Françaises*}, {*CD26*; *cluster de différenciation 26*} ;
– partiel correct : {*CHUM*; *Université Montréal*}, {*CHU*; *hôpital universitaire*} ;
– partiel incorrect : {*SEPP*; *plus*}, {*NFS*; *faire sang*}.
Plusieurs extractions sont partielles : 33 dans le corpus de développement et 52 165 dans le corpus de test. Cependant, certaines d'entre elles peuvent correspondre aux formes étendues correctes : {*CIV*; *communication interventriculaire*}, où *I* et *V* représentent *inter*ventriculaire ; {*TG*; *thyroglobuline*}, où *T* et *G* représentent *thyroglobuline*. Il y a beaucoup de doublons car les formes étendues montrent en général un bon degré de figement, même si quelques abréviations peuvent en recevoir plusieurs : {*TOC*; *trouble obsessionnel compulsif*} et {*TOC*; *troubles obsessionnels compulsifs*}. Certaines abréviations sont correctement extraites en anglais : {*PYLL*; *potential years life lost*}.

Les extractions avec les marqueurs offrent très peu de doublons, car les reformulations sont assez libres (*des canaux galactophores : qui fabriquent le lait de la femme, qui sécrètent le lait*). Comme dans un travail existant (Grabar & Eshkol-Taravella, 2015), le marqueur *c'est-à-dire* est nettement plus productif que les autres marqueurs. Le marqueur *encore appelé* introduit plus facilement des synonymes {*troubles fonctionnels intestinaux*; *colopathie fonctionnelle*}. Les reformulations ne sont pas nécessairement des définitions universelles ({*un bruxisme*; *des mouvement automatiques des mâchoires*}, {*du faisceau nerveux pyramidal*; *qui commande les mouvements des membres*}), mais peuvent aussi dépendre du contexte : {*des effets antagonistes*; *que le Lopressor inhibe l'effet du Bricanyl*}.

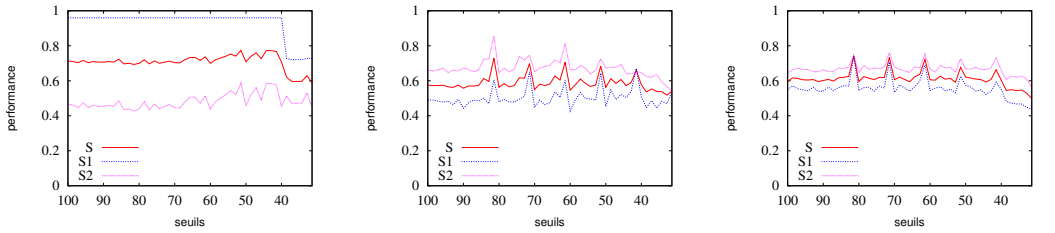
Nous observons très peu de redondance au sein des extractions avec les parenthèses, car ici aussi l'expression est libre : {*un proctologue*; *c'est souvent un gastroentérologue spécialisé dans les lésions anales*} et {*un proctologue*; *gastroentérologue*}. Beaucoup d'extractions concernent des entités nommées ({*Hôpital des Peupliers*; *Paris*}. Certaines propositions non pertinentes restent difficiles à filtrer, comme {*énergétique*; *carence plutôt liée au marasme*}.

Nous effectuons une évaluation des extractions par rapport aux données de référence construites sur le corpus de développement (tableau 5). Comme l'unité d'évaluation est une phrase contenant une extraction, les trois mesures d'évaluation sont identiques. Nous pouvons voir que les abréviations montrent des performances élevées, en appariement exact et inexact. En revanche, avec les marqueurs et parenthèses, il existe une grande différence entre les appariements exacts et inexacts : il peut en effet être difficile de régler les frontières des segments (la prise en compte ou non de mots grammaticaux, d'adjectifs...) en se basant sur les informations syntaxiques. L'évaluation inexacte indique également que les extractions proposées contiennent souvent les propositions attendues ou bien chevauchent avec elles. Ces résultats montrent aussi que l'extraction de segments en relation de reformulation, formée autour de marqueurs, est beaucoup plus difficile avec les corpus oraux, où la F-mesure dépasse rarement 0,35 pour l'extraction de segments reformulés (Grabar & Eshkol-Taravella, 2015).

Précision	*Abréviations*			*Marqueurs*			*Parenthèses*		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
exacte	0.74	0.74	0.74	0.24	0.24	0.24	0.23	0.23	0.23
inexacte	0.94	0.94	0.94	0.98	0.98	0.98	0.68	0.68	0.68

TABLE 5 – Précision, rappel et F-mesure des extractions pour chaque méthode.

5.2 Résultats d'alignement



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

Fig. 2 – Précision des alignements avec la terminologie, corpus de développement.

Pour définir les taux d'alignement optimaux, nous nous basons sur les alignements de référence. La figure 2 montre les résultats, avec la précision en abscisse et les seuils d'alignement en ordonnée. Entre les deux seuils d'alignement de segments, se trouvent les seuils d'alignement 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), etc. 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. Dans ce dernier cas, un terme (*période d'ovulation*) n'est pas aligné en entier, en revanche ses composants sont alignés séparément (*periode/C0332311.T079.CONC, ovulation/C0029965.T042.PHYS*). Nous remarquons que le segment 2 *S2* est souvent plus facile à aligner que le segment 1 *S1*, sauf pour les abréviations qui sont souvent absentes des terminologies. Pour les abréviations, les meilleures précisions sont entre les seuils 50-100 et 40-100. Ces seuils optimaux sont donc appliqués lors des alignements des extractions du corpus de test.

	Corpus de développement			*Corpus de test*		
	Abrév.	*Marq.*	*Par.*	*Abrév.*	*Marq.*	*Par.*
nb occurrences	75	96	312	88 762	2 757	100 103
alignements totaux	11	5	38	154	42	3 738
aignements partiels	44	37	123	1 634	557	25 708
non alignés	20	54	150	6 318	1 937	60 928

TABLE 6 – Nombre d'alignements totaux et partiels sur les deux corpus (développement et test).

Dans le tableau 6, nous indiquons le nombre d'alignements sur les deux corpus, en exploitant les seuils optimaux. Ici encore, le corpus de test et l'extraction par parenthèses fournissent le plus d'alignements. Nous observons différents types d'alignement :

– complet correct : {*syndrome polyalgique idiopathique diffus*; *syndrome polyalgique idiopathique diffus.C0016053.T047.DISO*} ;
– variation morpho-syntaxique : {*troubles fonctionnels intestinaux*; *troubles gastrointestinaux fonctionnels/C0559031.T047.DISO*} et {*troubles fonctionnels intestinaux*; *troubles gastro intestinaux fonctionnels/C0559031.T047.DISO*} ;

– partiel : {*semaines amenorrhee*; *amenorrhee/C0002453.T047.DISO*} ;
– compositionnel : *cause/C0085978.T078.CONC* et *pus/C0034161.T031.ANAT* pour *cause de pus* ;
– incorrect : {*LCR*; *ph lcr/C0853364*}, {*liquide cerebro*; *regime liquide/C-F2300*}.

5.3 Comparaison entre les méthodes et avec l'état de l'art

Méthode	*Type de termes*	*Nb. extractions*	*Précision*
Abréviations	abréviations	42, 8 106	0,74/0,94
Marqueurs	tout type	96, 2 710	0,24/0,98
Parenthèses	tout type	305, 92 971	0,23/0,68
Définitions (Grabar & Hamon, 2015)	tout type	1 028	0,52, 0,68
Morphologie (Grabar & Hamon, 2015)	composés	1 128	0,76, 0,86
(Deléger & Zweigenbaum, 2008)	morpho-syntaxique	65, 82	0,67, 0,60
(Cartoni & Deléger, 2011)	morpho-syntaxique	109	0,66
(Schwartz & Hearst, 2003)	abréviations	785	0,95

TABLE 7 – Comparaison entre les méthodes avec avec l'état de l'art.

Le tableau 7 propose une comparaison entre les méthodes et leur comparaison avec l'état de l'art : le type de termes, le nombre d'extractions et la performance de ces méthodes. La comparaison entre nos trois méthodes indique qu'elles sont complémentaires car seulement 7 extractions sont identiques, dont {*en ambulatoire*; *sans hospitalisation*}, {*microcytaires*; *de petite taille*}, {*une jaunisse*; *ictère*}. Par rapport à l'état de l'art, la deuxième partie du tableau indique les indices sur d'autres travaux existants. Nous pouvons voir que les méthodes proposées sont performantes : elles proposent un nombre d'extractions très élevé et montrent une précision fiable par rapport aux travaux existants.

6 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 non techniques. Nous exploitons pour cela la reformulation dans le discours médical écrit. Nous proposons 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)* et extraction de reformulations marquées par des parenthèses. Les méthodes sont réglées sur le corpus de développement et ensuite appliquées au corpus de test. Nous nous basons ainsi sur la spontanéité de la reformulation dans la langue. Selon les méthodes, la précision exacte varie entre 0,23 et 0,74, la précision inexacte varie entre 0,68 et 0,98. Les abréviations montrent la précision la plus élevée et les parenthèses la moins élevée. Ces trois méthodes sont complémentaires.

Nous avons plusieurs perspectives pour ce travail. Actuellement, nous considérons que les concepts se situent immédiatement avant le marqueur ou les parenthèses, alors que ces segments peuvent être éloignés, comme dans *l'infection virale dont elle peut-être atteinte, c'est-à-dire, surtout la grippe*. Nous pouvons exploiter d'autres marqueurs, tels que *l'équivalent de*, *ou encore*, *ou*. Grâce au corpus annoté, une méthode par apprentissage supervisé peut être exploitée. Au sein de la ressource acquise,

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. Ce type de vocabulaire peut être utilisé pour la simplification de documents médicaux et de santé.

Remerciements

Ce travail a été effectué dans le cadre du projet *EQU (Éthique Qualité Urgence)* financé par l'appel Thématique de l'établissement de l'université Lille 3.

Références

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

BOUAMOR H. (2012). *Étude de la paraphrase sous-phrastique en traitement automatique des langues*. Thèse de doctorat, Université Paris Sud, Paris.

BOUBÉ N. & TRICOT A. (2010). *Qu'est-ce-que rechercher de l'information ? : état de l'art*. Villeurbanne.

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.

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

FLOTTUM K. (1995). *Dire et redire. La reformulation introduite par "c'est-à-dire"*. Thèse de doctorat, Hogskolen i Stavanger, Stavanger.

FUJITA A. (2010). Typology of paraphrases and approaches to compute them. In *CBA to Paraphrasing & Nominalization*, Barcelona, Spain. Invited talk.

GRABAR N. & ESHKOL-TARAVELLA I. (2015). ...des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux. In *TALN 2015*.

GRABAR N. & HAMON T. (2015). Extraction automatique de paraphrases grand public pour les termes médicaux. In *TALN 2015*, Caen, France. 14 p.

GRABAR N. & ZWEIGENBAUM P. (2000). A general method for sifting linguistic knowledge from structured terminologies. *JAMIASUP*, p. 310–314.

GULICH E. & KOTSCHI T. (1983). Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française*, **5**, 305–351.

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*.

LE BOT M.-C., SCHUWER M. & ÉLISABETH RICHARD (DIR.) (2008). *La reformulation : Marqueurs linguistiques – Stratégies énonciatives*. Rennes : Rivages linguistiques.

LINDBERG D., HUMPHREYS B. & MCCRAY A. (1993). The unified medical language system. *Methods Inf Med*, **32**(4), 281–291.

MCCRAY A. (2005). Promoting health literacy. *J of Am Med Infor Ass*, **12**, 152–163.

MCCRAY A., LOANE R., BROWNE A. & BANGALORE A. (1999). Terminology issues in user access to web-based medical information. In *AMIA Symposium 1999*, p. 107.

MCCRAY A. T. (1989). The UMLS semantic network. In *Proceedings of the 13th Annual SCAMC*, p. 503–507, Washington.

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

RATZAN S. & PARKER R. (2000). *Introduction. In : National Library of Medicine current bibliographies in medicine : health literacy*. U.S. Department of Health and Human Services : NLM Pub No CMB 200-1. Bethesda, MD : National Institutes of Health.

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

TRAN T., DASSE P., LETELLIER L., LUBJINKOWIC C., THERY J. & MACKOWIAK M. (2012). Les troubles du langage inauguraux et démence : étude des troubles lexicaux auprès de 28 patients au stade débutant de la maladie d'Alzheimer. In F. NEVEU, V. MUNI TOKE, P. BLUMENTHAL, T. KLINGER, P. LIGAS, S. PRÉVOST & S. TESTOND-BONNARD, Eds., *Congrès Mondial de Linguistique Française*, p. 1659–1672. SHS Web of Conferences (1).

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).

VILA M., ANTÒNIA MART M. & RODRÍGUEZ H. (2011). Paraphrase concept and typology. a linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, **46**, 83–90.

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*, p. 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*, p. 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*, p. 1155–1155.

ZIELSTORFF R. D. (2003). Controlled vocabularies for consumer health. *Journal of Biomedical Informatics*, **36**(4-5), 326–33.

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