LA NOTION DE REFORMULATION DANS LE DOMAINE DU TAL

Iris Eshkol-Taravella[1] et Natalia Grabar[2]

1. INTRODUCTION

Le procédé de reformulation occupe une place importante dans les recherches en linguistique en relation avec des études sur des interactions verbales depuis plus d'une trentaine d'années (Gülich & Kotschi 1983, 1987, Roulet 1987, Rossari 1990-97), à travers l'étude des marqueurs dans les corpus monolingues (Dostie 2004, Fløttum 1996, Khatchatourian 2008, Steuckardt 2005, Rossari 2005, Hwang 1993, Teston-Bonnard 2008) ou dans une perspective multilingue contrastive (Rossari 1993, Vassiliadou 2004, 2013), mais aussi du point de vue de l'acquisition du langage (Martinot 1994, 2000, Bernicot *et al.* 2006) ou de la didactique des langues (Besse 1985, Blondel 1996, Charolles & Coltier 1986, Kara 2007).

Le travail présenté dans cet article s'intéresse à la manière dont le procédé de reformulation est abordé dans le domaine du Traitement Automatique des Langues (TAL). Le TAL vise la conception d'outils capables de traiter automatiquement les données langagières en proposant des tâches comme la détection automatique d'une information recherchée dans un texte ou discours, l'annotation automatique du corpus, la correction orthographique, la traduction automatique, etc.

Dans les travaux en TAL, le repérage automatique des reformulations fait partie des tâches comme la détection du plagiat (Ferrero & Simac-Lejeune 2015), l'inférence textuelle (Androutsopoulos & Malakasiotis 2010, Dagan *et al.* 2013), la normalisation de langages contrôlés (Nasr

[1] Université Paris Nanterre, MoDyCo UMR 7114.

[2] CNRS, Université Lille, STL, UMR 8163.

1996), la recherche d'information et la traduction automatique (Madnani & Dorr 2010, Bouamor 2012). Des caractéristiques de recherches en TAL sur les reformulations seront présentées dans les sections 2 et 3. La notion de reformulation est prise dans ces travaux au sens très large. La section 4 sera consacrée à la synthèse de recherches menées par Iris Eshkol-Taravella & Natalia Grabar (2014-2017) sur l'analyse du procédé de reformulation dans les discussions à l'oral et sur le Web. Enfin, la section 5 montrera différentes applications où la détection de reformulations occupe une place importante. L'accent sera mis sur une tâche particulière visant la simplification de textes.

2. NOTION PARADIGMATIQUE : PARAPHRASE

Dans le domaine du TAL, la notion de reformulation est d'abord très proche de celle de paraphrase. C'est une notion étudiée aussi du point de vue paradigmatique fondé sur le fait qu'un mot, un syntagme ou une phrase peuvent être substitués par des segments sémantiquement et syntaxiquement équivalents.

2.1. Premiers travaux : Modèle « Sens ⇔ Texte »

L'origine du TAL peut être située aux États-Unis dans les années 50, où ont été effectués les premiers travaux en traduction automatique (TA) stimulés entre autres par l'apparition des premiers ordinateurs. Cependant, dès les années 60, le projet d'une traduction entièrement automatisée est sérieusement mis en doute (Bar-Hillel 1960) et l'accent est mis sur une approche déclarative où sont distingués la grammaire (la description linguistique) et les langages formels (qui rendent les informations linguistiques traitables par les ordinateurs). C'est l'approche qui est à la base de la théorie « Sens ⇔ Texte » proposée par Igor Melčuk *et al.* (1984, 1988, 1992, 2000).

L'objectif de cette théorie est de modéliser le langage naturel selon quatre niveaux de représentations : phonologique, morphologique, syntaxique et sémantique. La notion de paraphrase y occupe une place centrale. Elle est étroitement liée avec la notion de synonymie. Plus précisément, on obtient typiquement des phrases synonymes suite à différentes transformations : le passage d'une phrase active vers une phrase passive, le remplacement du prédicat verbal par le prédicat

nominal actualisé ou pas à l'aide d'un verbe support, etc., comme dans les exemples en (1).

(1) X et Y sont en relation de paraphrase
X est une paraphrase de Y
X paraphrase Y
On paraphrase X en/par Y
La paraphrase de X

La notion de synonymie est abordée du point de vue du « sens situationnel » sans tenir compte du « sens communicatif » ni du « sens rhétorique ». La synonymie et le paraphrasage sont pris au sens large : la synonymie approximative ou « la quasi-synonymie » sont acceptées également. L'approche de Melčuk a été exploitée dans les travaux sur la traduction automatique et la génération automatique de textes. Il s'agit d'une approche déductive fondée sur les compétences linguistiques des locuteurs de la langue. Le modèle proposé utilise ces compétences pour construire les règles nécessaires au fonctionnement des systèmes.

Le début des années 90 marque un tournant dans les recherches du TAL. C'est une conséquence de la disponibilité d'un volume croissant de données linguistiques au format numérique. Comme en témoigne l'un des pionniers de ces études : « […] la recherche basée sur corpus a vraiment décollé, non seulement comme un paradigme d'investigation linguistique reconnu mais comme une contribution clé pour le développement de logiciels de traitement du langage naturel » (Leech 1991 : 20). En particulier, les travaux sur des corpus écrits connaissent un essor.

2.2. Travaux sur les corpus écrits

Les travaux actuels dans le TAL sur les corpus écrits visent la détection automatique des paraphrases dans les textes, l'équivalence sémantique étant l'un des critères de leur reconnaissance. Pour atteindre cet objectif, il est nécessaire de créer des conditions qui favorisent la détection de paraphrases. Quatre types de corpus sur lesquels les chercheurs testent différentes méthodes peuvent être distingués :

– Corpus monolingues où les mots, les syntagmes, etc. sont transformés en vecteurs. Plus les vecteurs sont proches, c'est-à-dire plus le cosinus entre deux vecteurs est petit, plus ces deux mots ou syntagmes sont proches sémantiquement. Les vecteurs sont calculés par rapport à la fréquence et au contexte d'emploi de chaque unité

linguistique traitée. Ces méthodes s'appellent méthodes distributionnelles car elles exploitent la distribution de mots et de syntagmes dans le texte (Lin *et al.* 2001, Pasca *et al.* 2005). Selon cette méthode, les deux phrases en (2) peuvent être rapprochées.

(2) Y is solved by X
Y is resolved in X

La deuxième méthode exploite la similarité de chaînes d'édition (Malakasiotis & Androutsopoulos 2007) comme dans l'exemple (3), où les mots en italique sont des indicateurs du recoupement entre les deux phrases.

(3) When did *Charle de Gaulle die*?
*Charles de Gaulle die*d in 1970

– Corpus monolingues parallèles qui sont composés de différentes versions alignées de traductions d'une même œuvre dans une langue donnée (Och & Ney 2000, Barzilay *et al.* 2001, Ibrahim *et al.* 2003, Quirk *et al.* 2004). Dans ce cas, les traductions d'une unité donnée proposées dans différentes versions sont considérées pour être des paraphrases, comme en (4).

(4) Emma *cried*
Emma *burst in tears*

– Corpus monolingues comparables qui réunissent des textes couvrant le même événement mais tels qu'écrits par différents auteurs. Les articles de journaux qui couvrent un même événement sont typiquement de ce type. Les relations de paraphrases sont repérées grâce à l'alignement de phrases et de mots (Shinyama *et al.* 2002, Sekine 2005, Shen *et al.* 2006) et les méthodes distributionnelles décrites ci-dessus :

(5) PERS1 a tué PERS2, PERS1 et PERS2 sont morts de la perte du sang.

– Corpus bilingues parallèles qui réunissent les textes avec leurs traductions dans une autre langue. Ces corpus, alignés au niveau des phrases, sont aussi exploités pour la détection de paraphrases (Bannard *et al.* 2005, Callison *et al.* 2008, Kok *et al.* 2010).

3. NOTION SYNTAGMATIQUE : TRAVAUX SUR L'ORAL

3.1. Détection de disfluences

Dans les travaux sur les données orales, les reformulations sont étudiées en tant qu'éléments disfluents. Rappelons qu'une disfluence intervient lorsque le déroulement syntagmatique de l'énoncé est brisé (Blanche-Benveniste *et al.* 1990).

D'une manière générale, on peut distinguer deux méthodes de la détection automatique de disfluences dans le corpus oral : méthodes symboliques fondées sur les règles implémentées décrivant le contexte d'apparition d'une telle ou telle disfluence (Constant & Dister 2010) et méthodes par apprentissage supervisé (Dutrey *et al.* 2014) où les algorithmes reconnaissent les disfluences sur la base d'un grand nombre d'exemples du même type. Les indices lexicaux et prosodiques dont peuvent tenir compte les algorithmes peuvent être déduits de manière automatique ou introduits par un expert humain.

3.2. Détection automatique de reformulations

Nous présentons ici une synthèse de travaux menés sur l'analyse outillée et le traitement automatique de reformulations dans les corpus non normalisés. Il s'agit de travaux effectués sur l'oral transcrit et les forums du Web considéré comme un corpus dialogique écrit (Eshkol-Taravella & Grabar 2016).

La méthodologie abordée tient compte du contexte et de la nature des données traitées et est fondée sur l'annotation manuelle suite à la modélisation multidimensionnelle élaborée préalablement. Les transcriptions de l'oral sont issues du corpus ESLO (Eshkol-Taravella *et al.* 2012). Il s'agit des entretiens tête-à-tête entre un chercheur et un habitant de la ville d'Orléans : 260 entretiens du corpus ESLO1 constitué dans les années 70 et 308 entretiens du corpus ESLO2 constitué dans les années 2010. ELSO2 est donc un corpus moderne. Les questions de l'entretien portent sur la vie des locuteurs dans Orléans et son agglomération. Quant au corpus du forum Web, il est composé de 17 443 fils de discussion (101 728 messages) provenant du site web Doctissimo. Il s'agit plus spécifiquement du forum dédié à l'hypertension.

Nous explicitons d'abord l'acception de la notion de reformulation. La modélisation du procédé est présentée par la suite, elle a permis d'effectuer une analyse quantitative et qualitative des reformulations

dans les corpus étudiés. Finalement, nous exposons une description synthétique du traitement automatique effectué visant la prédiction automatique des énoncés contenant la reformulation, des frontières de segments reformulés et des raisons pour lesquelles le locuteur ou le scripteur procèdent à la reformulation.

3.2.1. Définition

La reformulation est présente dans les données langagières que ce soit à l'écrit ou à l'oral. En élaborant son discours, le locuteur peut en effet recourir à la reformulation pour corriger, par exemple ses propos antérieurs. Levelt (1983) et Shriberg (1994) proposent un schéma du processus de correction qui comprend trois éléments [reparandum] (phase d'édition) [repair] :

– le « reparandum », un segment que le locuteur souhaite modifier ;
– la phase optionnelle d'édition ;
– le « repair » est un segment corrigé.

Le procédé de reformulation peut être vu comme un procédé plus large composé de trois éléments : segment source ou le segment qui sera reformulé ; une phase d'édition ou une phase optionnelle qui marque l'apparition d'éléments introduisant le segment reformulé ; et le segment reformulé :

[segment source] (phase d'édition) [segment reformulé]

À partir de ce modèle, le procédé de reformulation consiste dans la modification d'un segment source par un segment reformulé. Cette modification peut déclencher une « phase d'édition » indiquée par un marqueur, par exemple. Notons que la présence d'un marqueur n'est pas nécessaire pour que la reformulation ait lieu. Le corpus traité atteste des cas où la reformulation peut être réalisée sans aide d'un marqueur comme dans l'exemple suivant :

(6) je suis bien je me sens bien

La modification effectuée laisse des traces, un lien sous-jacent, « un invariant » (Martinot 1994) qui relie les deux segments reformulés et qui permet d'attester la reformulation et de reconnaître le procédé de

reformulation même si le marqueur de reformulation est absent. Ce lien sous-jacent est de nature multidimensionnel et peut se manifester à différents niveaux : morphologique, lexical, syntaxique, sémantique et pragmatique.

3.2.2. Lien sous-jacent multidimensionnel

La dimension morphologique concerne les unités lexicales présentes dans les deux segments et ayant la même racine ou la même base. On peut distinguer trois cas : la flexion, la dérivation et la composition. Les unités lexicales peuvent également avoir des relations lexicales d'hyperonymie, d'hyponymie, de méronymie, de synonymie, d'antonymie et d'instance. Du point de vue de la syntaxe, deux cas sont distingués : structure active et structure passive. Huit catégories reflétant le lien au niveau sémantico-pragmatique entre les deux segments et une raison pour laquelle le locuteur ou le scripteur procède à la reformulation sont proposées : correction, définition, dénomination, exemplification, explication, justification, paraphrase, précision, résultat.

Tous ces niveaux sont décrits dans un jeu d'étiquettes pré-établi et marqués dans le corpus à l'aide des balises au cours de l'annotation manuelle.

Ainsi, dans l'exemple :

(7) on fait ce que l'on appelle <NP1>*un carton*</NP1> <MR>*c'est-à-dire*</MR> le le <NP2 rel_lex="hypero(carton/dessin)" rel_pragm="prec">*ce dessin-là agrandi*</NP2> mais à la grandeur de la fenêtre

le locuteur remplace le groupe nominal (NP1) « un carton » par un autre (NP2) « ce dessin-là agrandi » à l'aide d'un marqueur (MR) « c'est-à-dire ». Le lien observé entre les deux segments se manifeste d'abord au niveau lexical car le nom « carton » est un hyperonyme (hypero) de « dessin ». Au niveau sémantico-pragmatique, le locuteur remplace un groupe nominal indéfini par un groupe nominal introduit par un démonstratif pour préciser (prec) ses propos antérieurs.

L'exemple suivant :

(8) <P1>*on avait choisi Olivet*</P1> nos enfants étaient scolarisés là donc euh voilà <P2 modif_morph="flex(avait/a)" relpragm="para">*on a choisi de vivre à Olivet*</P2> ce qu'on ne regrette pas

est un cas où les deux segments gardent un lien au niveau de la flexion (flex). Il s'agit donc d'une relation de nature morphologique. De plus, le

locuteur effectue une autre modification morphologique : il remplace le plus-que-parfait « avait choisi » par le passé composé « a choisi ». Avec ces modifications morphologiques, le sens des deux segments reste identique. Il s'agit donc d'une paraphrase (para).

(9) en général même avec <NP1>*l'accent un peu de travers*</NP1> <MR>*je veux dire*</MR> <NP2 rel_lex="hypo(l'accent un peu de travers/l'accent)" modif_lex="suppr(un peu de travers)" rel_pragm="res">*l'accent*</NP2> l'orthographe oui alors là

L'exemple (9) montre la reformulation qui sert à indiquer un résultat (res). Le locuteur synthétise ses propos à l'aide d'un marqueur de reformulation (MR) « je veux dire » en supprimant « un peu de travers ». Les deux segments sont unis par le lien lexical d'hyponymie : le groupe nominal « l'accent un peu de travers » est une sorte de « l'accent ».

(10) En fait, avez-vous eu depuis votre RVA mécanique <NP1>*des interventions "bénignes"*</NP1> <MR>*c'est à dire*</MR> <NP2 rel_pragm="exempl">*des dents à extraires par exemple*</NP2> et quels ont étés des différents problèmes !!

Enfin, dans l'exemple (10) l'utilisateur du blog Doctissimo donne un exemple (exempl) « des interventions bénignes ». La reformulation est introduite par le marqueur (MR) « c'est-à-dire ».

3.2.3. Annotation manuelle

L'annotation manuelle a été effectuée par deux experts. L'accord inter-annotateur calculé sur la présence de reformulations introduites à l'aide de trois marqueurs (« c'est-à-dire », « je veux dire », « disons ») est de 61 % dans le corpus ESLO1, de 53 % dans le corpus ESLO2 et de 80 % dans le corpus de forum. Selon la grille de Landis & Koch (1977), il s'agit d'un accord modéré pour ESLO2 et d'un accord fort pour ESLO1 et le forum.

Le corpus annoté manuellement a permis de faire quelques observations quant à la distribution du lien multi-dimensionnel dans les reformulations annotées dans les deux corpus. Les deux segments gardent les relations au niveau lexical (Figure 1). La relation lexicale la plus fréquente est la synonymie (environs 35 %). La fonction semantico-pragmatique la plus fréquente est la précision. Il a été constaté qu'il existe un lien entre la fonction sémantico-pragmatique et la taille des segments reformulés :

– si le segment 2 est plus long que le segment 1 : le locuteur *précise, définit, explique* ou *exemplifie* ses propos ;
– si le segment 1 est plus long que l segment 2 : le locuteur *conclut* ou *dénomme* ce qui a été dit ;
– si les deux segments ont la taille identique : le locuteur *paraphrase* ou fait les *corrections.*

La reformulation vise souvent à apporter de nouvelles informations par rapport au segment source. Dans 60 % des cas, le segment 2 est plus long que le segment 1. Très peu de segments (entre 0,07 % et 0,10 % dans les corpus oraux et entre 0,13 % et 0,14 % dans le corpus de forum) ont la même taille. La paraphrase se trouve dans 7 % à 10 % des reformulations. La reprise des mêmes bases ou racines est rare dans les reformulations étudiées : 8 % dans ESLO1, 10 % dans ESLO2 et 4 % dans le corpus de forum. La reprise de mots identiques apparaît dans 14 % de reformulations du corpus de forum et dans 40 % de reformulation des corpus oraux.

Au niveau de la syntaxe, dans 60 % des cas, il existe une équivalence syntaxique entre les éléments en relation de reformulation.

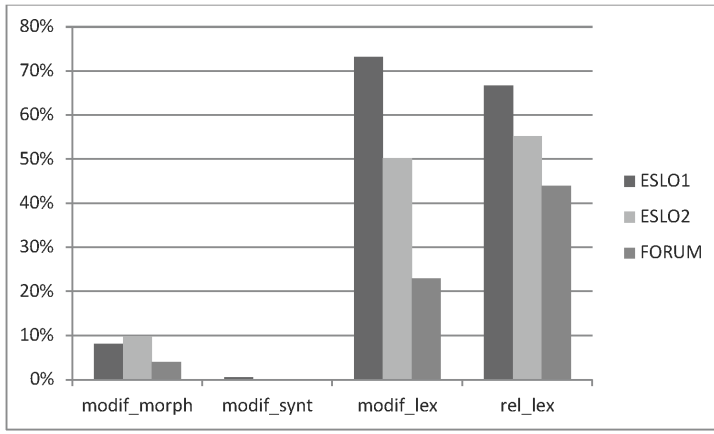

Figure 1. Distribution du lien multi-dimensionnel dans les trois corpus.

Il a été observé des cas dans lesquels le locuteur ou le scripteur se servent des reformulations pour donner leurs propres définitions ou opinions comme dans les exemples suivants :

(11) un film de de valeur assez assez restreint disons un film euh américain
(12) c'est ce qu'on appelle psychosomatique c'est a dire que ton cerveau enregistre le stress dans la journée et il s'en rappelle pendant la phase de repos, donc le sommeil

3.2.4. Traitement automatique

Trois types de traitements automatiques ont été effectués :

– prédiction des énoncés contenant la reformulation (Eshkol-Taravela & Grabar 2014) ;
– prédiction des frontières de segments reformulés (Grabar & Eshkol-Taravella 2015) ;
– prédiction des fonctions pragmatiques (Grabar & Eshkol-Taravella 2016).

La prédiction des énoncés contenant la reformulation exploite des règles tenant compte de la position et du contexte des marqueurs, de la présence des disfluences et de la vérification si le marqueur fait partie d'un syntagme comme dans l'exemple :

(13) leur façon de choisir la viande dans ce qu'ils achètent et caetera *indépendamment disons de leurs origines* de classe,

où « disons » joue le rôle d'une sorte de disfluences qui peut apparaître à n'importe quel endroit de l'énoncé et n'introduit donc pas la reformulation.

Les deux autres traitements ont été développés avec des méthodes d'apprentissage automatique supervisé. Il s'agit de reproduire l'annotation manuelle à partir du corpus de référence annoté manuellement. La prédiction des frontières de segments reformulés est réussie dans 44 % des reformulations. En ce qui concerne la catégorisation de la nature sémantico-pragmatique des reformulations, trois catégories (suppression d'information, ajout d'information, volume d'information comparable) ont pu être détectées avec la précision de 80 %. La reconnaissance des huit catégories individuelles (définition, dénomination, exemplification, explication, justification, paraphrase, précision, résultat) montre une performance qui ne dépasse pas 40 %.

3.3. Applications

3.3.1. État de l'art

Le TAL est un domaine appliqué dans le sens où les chercheurs conçoivent des algorithmes et programmes et les testent dans le cadre de différentes applications informatiques pour traiter les données linguistiques. Plusieurs de ces applications peuvent être plus efficaces lorsque la notion de reformulation est prise en compte. Par exemple, les reformulations peuvent augmenter la couverture des résultats grâce aux expressions équivalentes en recherche et extraction d'informations et dans les systèmes question-réponse. Elles peuvent également être utilisées dans la traduction automatique pour éviter des répétitions lexicales (Madnani & Dorr 2010, Bouamor 2012). Reconnaître automatiquement les reformulations est également un objectif très important pour la détection du plagiat (Barron-Cedeño *et al*. 2013, Iyer & Singh 2005, Oberreuter & Velásquez 2013). Ainsi, l'application développée par Ferrero & Simac-Lejeune (2015) recherche des correspondances de concepts et de mots clefs. Elle peut reconnaître les reformulations suivantes :

(14) En cinquante *ans*, grâce à *des efforts considérables dans la* recherche et *l'élaboration de la fusion*, la performance des plasmas a été multipliée par 10'000.
(15) En une cinquantaine *d'années*, grâce à *un immense effort de* recherche, la performance des plasmas *produits par les machines de fusion* a été multipliée par 10000.
(16) La performance des plasmas produits par les machines de fusion a été multipliée par 10,000 grâce à un immense effort de la recherche *bien que cela ait pris une cinquantaine d'années.*

Une autre application possible est la reconnaissance de l'inférence textuelle. L'objectif de cette application consiste à déterminer automatiquement si un segment de texte (H) peut être déduit d'un autre segment de texte (T) (Dagan *et al.* 2005, 2013, Androutsopoulos & Malakasiotis 2010) comme dans l'exemple (17).

(17) T : « Amine a 40 degrés de fièvre, sa mère l'a pris immédiatement à l'hôpital ».
H : « Amine est malade ».

Dans cet exemple, la phrase H peut être considérée comme une reformulation de la phrase T. Il y a donc une relation d'implication entre les deux.

3.3.2. Simplification de textes

La simplification de textes a pour objectif d'effectuer automatiquement une transformation d'un texte d'origine afin d'en produire un texte plus simple et plus accessible. Ce type d'application s'adresse à la population qui peut avoir des difficultés de lecture et de compréhension (enfants, étrangers, personnes non ou mal-alphabétisées, personnes avec des maladies comme l'aphasie ou la trisomie, personnes non spécialisées par rapport à un domaine scientifique et technique, etc.). Le plus souvent, la simplification est effectuée aux niveaux syntaxique et lexicale. La simplification syntaxique permet de produire une phrase plus simple suite aux transformations syntaxiques qui allègent sa structure, comme dans l'exemple (18), où les segments en italique sont supprimés lors de la simplification :

(18) Un archipel est un ensemble d'îles *relativement* proches les unes des autres. Le terme « archipel » vient du grec ancien « Archipelagos », littéralement « mer principale » *(de « archi » : « principal » et « pélagos » : « la haute mer »). En effet*, ce mot désignait originellement la mer Égée, caractérisée par son grand nombre d'îles *(les Cyclades, les Sporades, Salamine, Eubée, Samothrace, Lemnos, Samos, Lesbos, Chios, Rhodes, etc.)*.
Un archipel est un ensemble de plusieurs îles, proches les unes des autres. Le mot « archipel » vient du grec « archipelagos », qui signifie littéralement « mer principale » et désignait à l'origine la mer Égée, caractérisée par son grand nombre d'îles.

La simplification lexicale se concentre sur la substitution de mots complexes par leurs équivalents plus simples à comprendre et convenant au contexte. Par exemple, dans l'exemple (19), le terme médical « desmorrhexie » est remplacé par son équivalent « rupture des ligaments » :

(19) L'intervention chirurgicale s'impose, surtout en cas de fractures multiples ou lors de *desmorrhexie* sacro-iliaque.
L'intervention chirurgicale s'impose, surtout en cas de fractures multiples ou lors de *rupture des ligaments* sacro-iliaques.

C'est donc pour la simplification lexicale qu'il est nécessaire de disposer de ressources spécifiques pour pouvoir effectuer une telle substitution. Il s'agit non seulement de détecter les paraphrases mais aussi de contrôler que ces paraphrases sont en effet plus faciles à comprendre.

Nous nous intéressons principalement à la simplification de textes de spécialité qui comportent très souvent des termes techniques.

Plusieurs travaux ont été effectués pour constituer un lexique de simplification grâce à l'exploitation d'informations morphologiques (Grabar & Hamon 2016), de définitions (Grabar & Hamon 2016), d'expansions d'abréviations (Antoine & Grabar 2017) et de reformulations (Antoine & Grabar 2017). Nous présentons ici brièvement l'expérience qui exploite les reformulations. Le travail est effectué dans deux corpus écrits provenant d'un forum de discussion modérés par les médecins et une encyclopédie collaborative en ligne. L'exploitation de reformulations est motivée par différentes raisons :

1) la reformulation d'un terme technique dans un texte grand public indique qu'il s'agit d'une expression inconnue ou mal connue du public ;
2) l'acte de reformulation introduit par les marqueurs offre des indices formels sur la reformulation ;
3) la reformulation est un phénomène langagier utilisé spontanément par les locuteurs.

Le même schéma de reformulation que celui indiqué auparavant en §3.2.1 est exploité : [segment source] (phase d'édition) [segment reformulé]. Ce schéma permet de mettre en relation un terme technique (souvent « segment source ») et sa reformulation moins technique (souvent « segment reformulé »). La « phase d'édition » est instanciée par un des trois marqueurs considérés (« c'est-à-dire », « autrement dit » et « encore appelé »).

Une étape importante consiste à délimiter le segment source et le segment reformulé dans le texte. Nous exploitons pour ceci l'analyse syntaxique qui est effectuée avec Cordial (Laurent *et al*. 2009). Dans le tableau 1 est présenté un exemple de phrase analysée syntaxiquement. Plusieurs niveaux d'informations syntaxiques sont disponibles : la catégorie syntaxique (CS), le groupe syntaxique (GS), le type de groupe syntaxique (type GS), et la proposition (Prop). Lors de la détection automatique, le marqueur de reformulation, qui correspond à « c'est-à-dire » dans l'exemple, est le déclencheur. L'information sur le groupe syntaxique (GS) pour la détection du segment source, est exploitée ensuite : tout le segment qui précède le marqueur et qui fait

partie du même GS est extrait. Cela correspond à « par un proctologue » comme segment source. Un principe similaire est suivi pour la détection du segment reformulé : ce segment doit suivre le marqueur mais en revanche c'est l'information sur la proposition (Prop) qui est exploitée car nous retenons le segment qui va jusqu'à la fin de la proposition. Cela permet d'extraire « un gastroentérologue spécialisé » comme segment reformulé.

Tableau 1. Exemple d'analyse syntaxique de la phrase avec une reformulation « Vous devez les faire brûler par un proctologue, c'est-à-dire un gastroentérologue spécialisé ».

Forme	***Lemme***	***CS***	***GS***	***Type GS***	***Prop***
Vous	vous	PPER2P	1	S	1
devez	devoir	VINDP2P	2	V	1
les	le	PPER3P	3	C	2
faire	faire	VINF	4	D	2
brûler	brûler	VINF	5	V	3
par	par	PREP	8	F	3
un	un	DETIMS	8	F	3
proctologue	proctologue	NCMS	8	F	3
,	,	PCTFAIB	-	-	3
c'	ce	PDS	11	N	3
est	est	ADV	-	N	3
-à	à	PREP	14	I	3
-dire	dire	VINF	14	I	3
un	un	DETIMS	16	D	3
gastroentérologue	gastroentérologue	NCMS	16	D	3
spécialisé	spécialisé	ADJMS	16	D	3
.	.	PCTFORTE	-	-	-

Au préalable, un des corpus (corpus de forum) est annoté manuellement pour y marquer le segment source, le marqueur de reformulation et le segment reformulé. Cette annotation est nécessaire pour évaluer les extractions effectuées automatiquement.

Notre méthode a permis d'effectuer plusieurs extractions à partir des deux corpus analysés : 96 à partir du corpus de forum et 2 757 à partir du corpus d'encyclopédie, pour un total de 2 853 paires segment source/segment reformulé. Comme les reformulations sont effectuées

de manière libre par les locuteurs, il y a très peu de doublons dans les extractions et nous obtenons 2 806 paires différentes.

La précision et la complétude des résultats sont calculées de deux manières : avec le calcul strict (les frontières des deux segments doivent être reproduites exactement) nous obtenons 24 %, alors qu'avec le calcul lâche (les frontières des deux segments peuvent être différentes de celles annotées manuellement, ce qui permet d'inclure ou d'exclure les déterminants, les adverbes, les prépositions, etc.) nous obtenons 98 %. Cette évaluation indique que les résultats obtenus automatiquement montrent une bonne précision et complétude et qu'ils peuvent être exploités très facilement pour la simplification de textes techniques. Dans la série qui suit, nous présentons quelques exemples extraits automatiquement.

(20) *des canaux galactophores* c'est-à-dire *sécrètent le lait*
(21) *erratiques* c'est-à-dire qu*'ils changent de d'aspect et d'endroit*
(22) par *une lithiase* c'est-à-dire *un caillou*
(23) *clivage du moi* c'est-à-dire comme *une opposition entre le moi et la réalité*
(24) *au gré de la désintégration radioactive du 18 F* c'est-à-dire avec *une demi-vie d'environ*
(25) *un trouble de l'identité sexuelle* c'est-à-dire qu*'ils s'identifient à un genre ne correspondant pas à leur sexe biologique*

Deux difficultés principales ont été rencontrées :

- mauvaise détection de frontières, comme dans « une toxi-infection, c'est-à-dire, qu'elle peut », où le segment reformulé ne comporte pas toute la reformulation. Ce type d'erreurs peut provenir de la structure de la phrase et de la difficulté d'en obtenir une analyse syntaxique correcte ;
- absence d'équivalence sémantique, comme dans « en 10 ans autrement dit sur 64 millions de personnes », où les deux segments ne correspondent pas à une reformulation. Ce type d'erreurs peut être dû à l'ambiguïté des marqueurs.

Malgré les quelques cas d'erreurs détectés, les résultats obtenus sont intéressants, facilement exploitables et soulignent encore l'utilité des reformulations pour le TAL.

4. CONCLUSION

Les reformulations sont au centre de plusieurs travaux de recherche en TAL (Traitement Automatique des Langues) car elles peuvent apporter des informations nécessaires pour différentes applications (recherche et extraction d'information, traduction automatique, implication textuelle, simplification, etc.). Dans ces travaux, la notion de reformulation est souvent prise au sens large et se trouve considérée selon les besoins spécifiques des applications.

Dans cet article, nous avons d'abord passé en revue les travaux menés en TAL en détection automatique de reformulations et de paraphrases, et les différents types de corpus écrits exploités pour ceci. Nous avons présenté ensuite nos travaux sur les reformulations effectuées dans les corpus oraux et le corpus dialogique du web. Nous avons ensuite décrit une expérience qui exploite les reformulations dans les corpus écrits. Elle poursuit l'objectif d'acquérir automatiquement un lexique qui peut servir à la simplification lexicale des textes de spécialité.