Reformulation à l'oral et dans le forum Web

Iris Eshkol-Taravella[1], et Natalia Grabar[2]
[1] CNRS UMR 7270 LLL, Université d'Orléans, 45100 Orléans, France
[2] CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France

Résumé. Le cadre du travail réalisé entre dans le domaine de la linguistique de corpus et du Traitement Automatique des Langues (TAL). Les recherches sont fondées sur trois corpus : deux corpus oraux (ESLO1 et ESLO2) et un corpus composé de discussions collectées sur un forum de Doctissimo. Le travail vise à détecter, analyser et comparer la reformulation introduite par trois marqueurs (c'est-à-dire, je veux dire et disons) dans le discours oral et le corpus écrit dialogique du Web. L'objectif est d'étudier la raison pour laquelle le locuteur décide de modifier un segment par un autre et de repérer un lien formel entre les deux segments reformulés aux différents niveaux linguistiques (morphologique, lexical, syntaxique, etc.).

Abstract. Reformulations in speech and in Web forum. The proposed work is related to the fields of Corpus Linguistics and Natural Language Processing. The work is based on three corpora: two spoken corpora (ESLO1 and ESLO2) and a corpus composed of discussions collected from Web forum, Doctissimo. We aim at detecting, analyzing and comparing the reformulations introduced by three markers (/in other words/, /that is to say/ and /let's say/) in spoken and dialogical discourse of Web languages. Our objective is to study the reason for which speakers decide to modify a given segment by another. We also want to detect formal links between the reformulated segments at various linguistic levels (morphological, lexical, syntactic, etc.).

1 Introduction

Le phénomène décrit dans cet article concerne les reformulations en français. Il s'agit d'un processus durant lequel le locuteur décide de reprendre un mot, un syntagme ou l'énoncé produit d'une manière différente sans le répéter. Ce processus peut être marqué par les unités spécifiques, communément appelées marqueurs de reformulation, comme *c'est-à-dire, autrement dit, en d'autres termes, etc.,* ou bien il peut être réalisé sans adjonction de marqueurs particuliers. Avant de présenter le travail effectué, l'état de l'art sur la notion de reformulation et sur les notions qui lui sont proches est développé. Le cadre du travail

réalisé entre dans le domaine de la linguistique de corpus et du Traitement Automatique des Langues (TAL). Les recherches sont fondées sur trois corpus : deux corpus oraux (ESLO1 et ESLO2) et un corpus composé de discussions collectées sur un forum de Doctissimo. Ces corpus sont présentés dans la deuxième section. La troisième section est consacrée à la méthodologie appliquée pour analyser le phénomène de reformulation dans les trois corpus. La méthodologie est fondée sur la modélisation du processus sous forme de marquage décrivant différents aspects linguistiques. Nous nous arrêtons plus particulièrement sur le jeu d'étiquettes exploitées, sur les principes de l'annotation effectuée et sur l'évaluation de cette annotation sous forme de calcul de l'accord inter-annotateur. Des expériences de détection automatique de la reformulation sont aussi présentées. L'analyse quantitative du phénomène étudié est développée dans la dernière partie de l'article.

2 État de l'art

Cette section est consacrée aux travaux dont l'objet d'études est la reformulation ou les phénomènes qui lui sont proches. Nous décrivons, en premier lieu, les travaux effectués plutôt sur l'écrit et nous présentons ensuite les recherches sur le discours oral.

2.1 Travaux sur les corpus écrits

La notion de reformulation est liée, en premier lieu, à celle de *paraphrase*. La notion de paraphrase est étudiée de différents points de vue par les chercheurs en linguistique. La paraphrase est analysée tout d'abord en tenant compte de la situation d'énonciation et a, par conséquent, une valeur contextuelle [13, 16, 17, 28, 40]. Elle est définie aussi à travers les transformations linguistiques que subissent les segments paraphrasés aux différents niveaux (morphologique, lexical, syntaxique, sémantique) [4, 29, 41]. Elle est décrite enfin en fonction de la taille des entités couvertes par la paraphrase : un mot, un ou plusieurs syntagmes, un ou plusieurs énoncé, etc. [16, 11]. Une autre notion proche étudiée sur les données écrites est *glose* [1, 36, 37]. Ce terme, issu de la tradition philologique, désigne un commentaire sur un mot. Il impose donc au premier segment d'être une unité lexicale, alors que le deuxième segment correspond à la glose. Vion (2006 : 11) définit le terme plus générique, *reprise*, qui va « de la pure et simple répétition d'un segment textuel aux différents degrés de ses reformulations ». L'auteur souligne aussi le critère de proximité sémantique en tant que critère caractéristique de la reformulation. Enfin, le projet ANNODIS, consacré à la constitution du corpus de référence annoté en structures discursives[1], distingue une relation rhétorique appelée *élaboration* qui semble se rapprocher du phénomène de reformulation.

2.2 Travaux sur le discours oral

De nombreux travaux sur l'oral s'intéressent à la reformulation qui est une caractéristique même de ce type de discours. Le discours oral diffère de l'écrit car on assiste aux étapes de son élaboration à la manière de brouillons qui précèdent la version finale de nos écrits [8]. L'écrit concrétise une version aboutie du discours quand l'oral le présente dans son processus avec ses hésitations, ses faux départs, ses ratures et ses reformulations. « […] le scripteur peut revenir sur ce qu'il a écrit, pour le corriger ou le compléter. A l'oral, […] toute erreur, tout raté ou mauvais départ ne peuvent être corrigés […] que par une reprise, une hésitation voire une rupture de construction qui laissent des traces dans le message même. » p. 30 [31].

[1] http://redac.univ-tlse2.fr/corpus/annodis/

Au début des années quatre-vingts, le terme *reformulation* apparaît dans les études linguistiques en Allemagne, France et Suisse [20, 21, 35]. Ce terme est utilisé dans le cadre des analyses des interactions verbales. Dans les années quatre-vingt-dix, Corinne Rossari (1990-1994) distingue deux types de reformulation : les *reformulations paraphrastiques*, qui instaurent une équivalence entre les segments reformulés et les *reformulations non paraphrastiques* qui opèrent un changement de perspective énonciative. Tout acte de reformulation dans le discours oral n'introduit donc pas toujours une paraphrase. De ce point de vue, deux catégories de marqueurs sont distinguées : les marqueurs de reformulation paraphrastique (MRP), comme *c'est-à-dire, autrement dit, je m'explique, en d'autres termes etc.* qui ont pour tâche principale d'établir une relation paraphrastique, et les marqueurs de reformulation non-paraphrastique, comme *en somme, en tout cas, de toute façon, enfin* etc.), qui ne montrent ce rôle que dans des contextes précis. En outre, les propriétés sémantiques des MRP permettent d'instaurer une relation de paraphrase même entre des segments qui n'entretiennent aucune équivalence sémantique constatable par ailleurs.

Les études syntaxiques sur le langage oral ont rapproché la notion de reformulation avec celle d'*énumération* ou encore de *répétition* [3, 10, 27]. Dans les trois cas, il s'agit d'un même procédé syntaxique : les éléments répétés, reformulés ou énumérés ont une même place syntaxique dans l'énoncé sur un axe paradigmatique. La distinction est pourtant possible grâce à l'utilisation d'indices formels, tels que les marqueurs d'énumération et de reformulation, ou bien des accords.

Parmi les travaux récents sur l'oral, on peut citer le projet d'annotation multi-niveau de l'oral, Treebank Rhapsodie[2] qui voit dans une reformulation le procédé de corriger ou de préciser une précédente dénotation. Cette vision est expliquée dans [24]. Les auteurs utilisent la notion de reformulation pour présenter une typologie des *entassements* à l'oral qui, suite à d'autres travaux [5-7, 18], peuvent être utilisés pour établir des relations entre dénotations, créer de nouvelles dénotations, reformuler, exemplifier, préciser ou encore intensifier. Dans leur article, Sylvain Kahane et Paola Pietrandrea introduisent une notion de *reformulation dénotative.*

Le travail présenté ici concerne le discours oral, mais aussi le texte dialogique écrit que l'on peut rencontrer sur le Web dans les blogs et/ou forums. En nous inspirant des travaux sur l'oral, nous utilisons le terme de *reformulation* que nous définissons comme une activité du locuteur de modifier un segment, déjà produit dans son propre discours ou dans celui de son interlocuteur, par un autre segment. Cette modification peut être effectuée avec ou sans l'emploi d'un marqueur, mais elle garde un invariant sémantique permettant de la reconnaître et un lien sous-jacent entre les deux segments reformulés. Les cas observés dans les corpus étudiés dépassent largement le phénomène de la paraphrase qui présuppose une équivalence sémantique entre les expressions paraphrasées ou de la glose, qui est utilisée pour la définition d'une unité lexicale. Le phénomène étudié ici concerne les cas où le locuteur modifie le segment ou l'énoncé pour définir, exemplifier, préciser, expliquer, redire, dénommer ou encore synthétiser ce qu'il a dit précédemment. Cette acceptation large de la notion explique aussi pourquoi nous retenons ce terme.

Le travail vise à détecter, analyser et comparer la reformulation introduite par trois marqueurs (*c'est-à-dire, je veux dire* et *disons*) dans le discours oral et le discours écrit dialogique du Web. L'objectif est d'étudier la raison pour laquelle le locuteur décide de modifier un segment par un autre et de repérer un lien formel entre les deux segments reformulés aux différents niveaux linguistiques (morphologique, lexical, syntaxique, etc.).

[2] http://www.projet-rhapsodie.fr/

3 Données étudiées

3.1 Corpus oral ESLO

Le corpus oral sur lequel nous travaillons est le corpus ESLO (Enquêtes Sociolinguistiques à Orléans) [14] accessible sur le Web (http://eslo.huma-num.fr/). Il s'agit d'une grande masse de données orales transcrites et correspondant à deux enquêtes sociolinguistiques : l'une (ESLO1), effectuée par les chercheurs britanniques à Orléans dans les années 70, et l'autre (ESLO2), réalisée à partir de 2005 par les membres de l'équipe ESLO du Laboratoire Ligérien de Linguistique (LLL)[3]. Les deux corpus comprennent une gamme d'enregistrements variés (des reprises de contacts informelles comme des discussions entre amis, des enregistrements en micro caché, des conversations téléphoniques, des réunions publiques, des transactions commerciales, des repas de famille, des interviews de personnalités de la ville (monde politique, syndical, universitaire ou religieux, etc.).

Le sous-corpus du travail comprend 260 entretiens d'ESLO1 totalisant 2 349 829 occurrences de mots et 308 entretiens d'ESLO2 totalisant 1 412 891 occurrences de mots. Le choix des entretiens a été guidé, tout d'abord, par la présence représentative du phénomène étudié qui s'explique par la nature semi-directive de l'entretien. Les entretiens ont aussi un avantage majeur d'être tous transcrits, ce qui n'est pas le cas des autres enregistrements. Enfin, ESLO1 et ESLO2 permettent d'avoir les données comparables et structurées du point de vue diachronique et synchronique.

Le travail porte sur les transcriptions. Les fichiers audio ne sont donc pas traités[4]. 260 entretiens d'ESLO1 et 308 entretiens d'ESLO2 transcrits sont prétraités automatiquement. Les fichiers de transcription sont segmentés en tours de parole dans le but de reconstituer les énoncés :

- l'énoncé commence avec le changement de locuteur ;
- en cas de chevauchement, les segments correspondants sont associés aux énoncés de chacun des locuteurs impliqués et lorsqu'un locuteur continue de parler après un chevauchement, son tour de parole est prolongé d'autant.

Les énoncés contenant l'un des marqueurs *c'est-à-dire, je veux dire, disons* sont alors extraits. Nous obtenons ainsi 476 tours de parole dans 54 entretiens d'ESLO1 et 394 tours de parole dans 30 entretiens d'ESLO2 :

> *(1) eslo1_ENT_2_C :* ***par industriel je veux dire*** *euh j'ai* ***le côté commercial*** *oui*
>
> *(2) eslo2_ENT_6_C : voilà* ***financièrement je touche soixante-dix pour cent de ma retraite il garde trente pour cent c'est-à-dire tous les mois ma retraite elle est reniée de deux cent soixante-dix euros***

3.2 Corpus écrit dialogique : forum Doctissimo

Le deuxième corpus étudié provient du Web. Il s'agit du forum Doctissimo. Ce corpus de travail est constitué de discussions portant sur les problèmes cardiaques et les douleurs du dos. Ce corpus a été récupéré automatiquement. Il totalise 17 443 fils de discussion, 101 728 messages et plus de 7 millions de mots. 422 messages, contenant les marqueurs étudiés, sont extraits, comme par exemple :

[3] http://www.lll.cnrs.fr/

[4] Le travail lié au croisement des annotations effectuées et des informations prosodiques constitue une des perspectives de la recherche présentée.

> *(3) c'est surement l'inconscient; c'est ce qu'on appelle* ***psychosomatique c'est a dire que ton cerveau enregistre le stress dans la journéé et il s'en rappelle pendant la phase de repos****, donc le sommeil; bon courage et a trés bientot*

3.3 Corpus oral / corpus écrit dialogique

Les deux corpus sélectionnés semblent difficilement comparables. Tout d'abord, les contenus abordés dans les deux corpus sont différents. Dans le cas des entretiens d'ESLO, il s'agit d'interviews non formelles dans lesquelles les questions sur l'identité du locuteur ou encore sur sa vie à Orléans sont posées. Les sujets abordés sont assez variés mais il est rare qu'ils concernent le domaine médical ce qui est le cas du corpus du Web. Le forum Doctissimo est un site de discussions consacrées à la santé. Les deux corpus diffèrent aussi par le nombre d'erreurs orthographiques. Pour le corpus ESLO, même si le processus de transcription est manuel, on y observe peu d'erreurs de ce type. Ce fait peut être expliqué par le choix de la méthodologie de transcription adoptée par les chercheurs du LLL. Chaque enregistrement est transcrit en trois étapes successives, donnant lieu à trois versions différentes :

- transcription (A), première transcription rapide ;
- transcription (B) qui est la transcription (A) relue et corrigée par un deuxième transcripteur ;
- transcription (C), la transcription (B) relue et corrigée par un troisième transcripteur.

Le corpus du Web, de son côté, présente de nombreuses erreurs (marquées en gras dans les exemples qui suivent) :

> *(4) Bonjour à tous, j'ai 18 ans et depuis quelques temps j'ai mon coeur qui* ***d****'emballe de temps en temps, c'est à dire que je le* ***sent*** *battre fort et je* ***peut*** *voir ma poitrine se soulever...*

> *(5) Une dernière chose : ton coeur n'a pas besoin de toi pour fonctionner, je veux dire que si tu flippes* ***ca*** *ne* ***l'aid epas a*** *fonctionner mieux.*

Parmi les variations orthographiques les plus répandues, nous observons le non respect de l'accord, l'absence d'accents, les fautes de frappe ou encore les fautes causées par la correction automatique comme dans un exemple suivant :

> *(6) Mon coeur se* ***mec a claqier****, je veux dire par* ***la*** *bruit bizarre quand il est* ***a*** *l'effort et parfois au repos.*

où le locuteur mélange sur le clavier les deux touches *u* et *i* (ce qui transforme le verbe *claquer* à *claqier*), ne met pas d'accents sur *a* (l'adverbe *là*, la préposition *à*) et enfin soit se trompe de la touche en tapant le mot *mec*, soit ne fait pas attention à la correction automatique de ce mot (ce qui provoque l'apparition du verbe *met* à la place du *mec*).

La présence de cette variation orthographique n'est pas une seule caractéristique du corpus de forum. Provenant du Web, il contient les caractéristiques de l'écriture phonétique propre aux nouvelles formes de communications (langage sms, chats etc.) :

> *(7) TU n'as pas le même soin; et tous* ***ces*** *bien passées à la maison, par la je veux dire aucun faux mouvement, mais je n'ai pas eu la qualité du kiné et cela a servi a rien*

Pourtant, si l'on examine des transcriptions des corpus oraux, on peut y voir certaines similitudes avec le corpus du forum.

Les transcriptions de données orales ne contiennent pas de marques typographiques comme le point, la virgule, le point d'exclamation ou encore la majuscule en début d'énoncé pour éviter l'anticipation de l'interprétation [9]. Selon les auteurs, en ponctuant, le transcripteur « suggère une analyse avant de l'avoir faite » (1987: 142). La segmentation est faite soit sur une unité intuitive de type « groupe de souffle » repérée par le transcripteur humain, soit sur un « tour de parole », défini uniquement par le changement de locuteur. L'observation du corpus du Web a permis de constater de nombreux cas où les signes typographiques sont aussi manquants :

> *(8) salut tous le monde j'ai 34 ans et depuis 10 ans maintenants je souffre d'extrasystole elle viennent par periode dure quelque jours voir quelques semaine mais a chaque fois c'est la meme chose je vais en cardio et la c'est toujour la meme reponse ce n'est pas grave vous n'avez rien et hop on vous fout dehors en s'en excuserai presque de les avoir derangé je trouve qu'il n'y a pas assez de prise en charge des personnes comme nous et franchement j'en ai vraiment marre j'ai tous fait echo cardiaque r.a.s holter r.a.s on me dit que c'est le stress le stress de quoi il m'arive de les avoir meme quand tous va bien bref je ne sait plus quoi faire si vous avez les memes problemmes que moi c'est-à-dire des extrasystole ventriculaire et que vous avez trouvé le moyen de les calmer aidé moi s'il vous plait parceque la je n'en peut vraiment plus.*

Le corpus oral est caractérisé aussi par la présence d'éléments qui « brisent le déroulement syntagmatique » sans rien ajouter à la sémantique de l'énoncé [10], que l'on appelle les *disfluences* (hésitations, faux-départs, répétitions, autocorrections, amorce, etc.). A ces éléments, il faut ajouter les *marqueurs discursifs*, ces formes figées ou invariables qui peuvent constituer des énoncés à elles seules ou s'actualiser en différentes places d'un énoncé sans intégrer sa structure, (c'est-à-dire sans entrer en relation syntaxique avec un autre élément).

> *(9) eslo2_ENT_5_C : c'est une notion énorme* ***oui oui non mais c'est important oui c'est important*** *de pouvoir* ***euh*** *moi je pense que de communiquer on a eu cet exemple avec notre jeune fille au pair étrangère qui parlait très très bien le français* ***euh*** *je pense qu'elle aurait pas eu le parcours qu'elle a eu aujourd'hui si elle avait pas fait l'effort* ***de d'****apprendre notre langue comme elle l'a apprise et* ***euh*** *et c'est un facteur d'intégration* ***enfin*** *je veux dire ça reste* ***hm hm quand même*** *un sujet très à la mode* ***euh*** *on on ne peut pas* ***euh*** *s'intégrer dans une culture si on n'en parle pas la langue*

L'écriture sur les forums peut montrer aussi la présence de certains phénomènes de l'oral cités, comme par exemple les marqueurs discursifs :

> *(10) Il existe des techniques plus ou moins efficaces, […]des médicaments a vie et pour la vie, c'est-à-dire ils ne sont efficaces que lorsqu'ils sont pris tous les jours,[…] j´ai pas de problèmes cardiaques quotidiens, mon seul problèmes c´est de montées d´adrénalines trop fortes et le coeur qui s´emballe au quat de tour dès qu'il s´agit d´une situation nouvelle pour moi sans qu'elle soit réellement dangereuse : prendre la parole en public, faire un exposé etc... le trac* ***quoi****.*
>
> *(11) Et bien sûr en plaisantant je pourrais vous dire que dans les dérivés de la coumarine que vous avez cité, vous avez oublié "apegmone" ou encore "tromexane" et dans ceux de l'indanedione vous n'avez cité que le préviscan ... vous auriez pu rajouter "pindione", vous voyez ce que je veux dire ...* ***bon*** *ça c'est fait*

Suite à toutes ces observations, nous considérons que le forum comme Doctissimo est un cas de discours spontané libre et, peut, sur ce point, être comparé à un discours oral.

Nous parlerons donc de corpus écrit dialogique. L'observation du corpus du forum Doctissimo montre que le scripteur, s'exprimant d'une manière très libre, laisse souvent les énoncés non finis et/ou avec les erreurs et n'utilise pas toujours les signes typographiques de segmentation. On constate également la présence des marqueurs discursifs fréquents à l'oral. Les raisons en peuvent être nombreuses : le besoin du locuteur de s'exprimer rapidement sur un sujet qui le touche plus particulièrement, les souffrances et/ou les contraintes imposées au locuteur par la maladie et/ou les douleurs (ce qui est le cas souvent des utilisateurs du forum Doctissimo), qui rendent plus difficile le processus de frappe sur le clavier et/ou limitent ses fonctions visuelles etc.

3.4 Trois marqueurs étudiés

Le travail porte sur les reformulations introduites avec trois marqueurs : *c'est-à-dire, je veux dire* et *disons*, tous formés à partir du même verbe *dire*.

Ces trois marqueurs ont été largement étudiés dans la littérature. Selon les travaux de [2, 20], le marqueur *c'est-à-dire* est capable d'instaurer la relation de paraphrase entre des énoncés non équivalents sémantiquement. D'une manière générale, il est utilisé essentiellement pour corriger, reformuler ou argumenter ce qui a été énoncé avant. Il peut par ailleurs désigner la justification, l'hésitation et la conclusion. Dans ce dernier cas, il peut être substituable par *donc.*

Le marqueur *disons,* quant à lui, note souvent une rupture en mettant ainsi fin au niveau coénonciatif précédent. Par conséquent, il existe une analogie entre *disons* et *enfin* en tant que moyen de rectification [22]. Selon [30], ce marqueur semble impossible à être supprimé de l'énoncé parce que le segment, qu'il introduit, exprime une nuance sémantique différente.

Teston-Bonnard (2008) a étudié les propriétés syntaxiques spécifiques de *je veux dire* dans le corpus oral pour savoir si cette expression marque toujours la reformulation. Plusieurs emplois possibles de *je veux dire* sont proposés. Deux d'entre eux (verbes « recteurs faibles » et « parenthèses ») se paraphrasent par *autrement dit, c'est-à-dire* et *je reprends.*

Les corpus sur lesquels nous avons travaillé ont permis d'observer aussi les différents cas d'emplois de ces marqueurs. Observons les trois exemples où l'unité lexicale *disons* est utilisée à des fins différentes :

> *(12) il y a énormément de vieilles familles euh bourgeoises … **ils sont souvent disons moins aisés** que les familles d'ouvriers les familles d'employés*
>
> *(13) nous avons une expression chez nous … **nous disons** que les gens qui gardent ces choses ont euh une mentalité d'écureuil*
>
> *(14) **basée** euh **sur le capitalisme** enfin la société française **disons** euh euh **basée sur les valeurs** euh euh **erronées***

Dans l'exemple (12), *disons* est un marqueur discursif, il est inséré au milieu d'un syntagme *ils sont moins aisés.* Il peut être supprimé sans que le sens de l'énoncé soit modifié. Dans l'exemple (13), il s'agit d'un verbe *dire* à la première personne du pluriel du présent de l'indicatif. C'est seulement dans le dernier exemple (14) que *disons* joue le rôle d'un marqueur de reformulation et ce sont ces cas qui font l'objet de notre étude.

4 Annotation des reformulations

4.1 Objectifs

Pour étudier le phénomène de reformulation dans les corpus, nous avons choisi la méthode fondée sur l'annotation multidimensionnelle. Les informations annotées concernent : (1) les deux segments reformulés, (2) les liens formels aux niveaux lexical, morphologique, syntaxique existant entre les unités qui composent ces segments, (3) les modifications que le premier segment subit au cours du processus de reformulation et (4) la raison pour laquelle, selon nous, le locuteur reformule ce qui a été dit. Cette annotation manuelle répond à plusieurs objectifs. Tout d'abord, elle permet de faire une première distinction entre l'emploi avec et sans la reformulation. De même, le corpus annoté manuellement est un corpus de référence par excellence qui permet de procéder à l'analyse du phénomène annoté. En ce qui concerne le traitement automatique, le corpus annoté manuellement peut servir de modèle pour l'apprentissage automatique [18] et peut permettre l'évaluation du module de détection automatique du phénomène développé [15].

4.2 Convention et format de l'annotation

L'annotation proposée est multidimensionnelle. Elle porte d'une part sur les deux segments reformulés, mais aussi sur la relation établie par le marqueur de manière générale. Elle est effectuée sous forme de balises : l'information contenue dans les balises désigne les catégories syntaxiques des segments et les attributs indiquent les informations sur le processus et le fonctionnement de la reformulation établie grâce au marqueur dans chaque cas. Les corpus sont annotés selon les conventions prédéfinies. Selon les spécifications, les informations annotées sont de différentes natures (syntaxique, lexicale, morphologique et pragmatique) :

- Nous annotons la catégorie syntaxique (*N, A, V, Prep*…) ou type de constituant syntaxique (*NP, VP, AP, PP*) d'un segment ce qui permet de vérifier une éventuelle équivalence syntaxique entre les deux segments reformulés ;
- Nous indiquons les modifications syntaxiques, comme la modification passif/actif ;
- Nous annotons les cas où il existe un lien d'hyperonymie, d'hyponymie, de synonymie, d'antonymie, de méronymie etc. entre les unités lexicales composant les deux segments ;
- Nous ajoutons à cela les modifications lexicales effectuées dans le processus de reformulation : remplacement d'un segment ou d'une partie d'un segment par un autre, suppression d'un/des unités lexicales, ajout d'unités lexicales dans le deuxième segment ;
- Le lien morphologique entre les deux segments est indiqué par l'attribut « modifications morphologiques » et concerne la flexion, la dérivation et la composition ;
- Les énoncés avec la reformulation contiennent une indication concernant la fonction pragmatique, l'attribut « relations pragmatiques », c'est-à-dire la raison pour laquelle le locuteur remplace un segment dans son discours par un autre. Ces raisons peuvent être nombreuses : définition, explication, exemplification, précision, dénomination, résultat, correction linguistique, correction référentielle, paraphrase [21, 24].

Observons les exemples suivants :

*(15) eslo1_ENT_149_C|pendant nous avons fait grève à la Régie Renault euh de <NP1>**Saint Jean de la Ruelle**</NP1><MRP>**c'est-à-dire**</MRP> <NP2 rel-lex= « mer(Saint Jean de la Ruelle/Orléans) » rel-pragm= « coref »>**Orléans**</NP2> parce que c'est ça fait partie d'Orléans* »

*(16) eslo1_ENT_002_C|on fait ce que l'on appelle <NP1>**un carton**</NP1> <MDR>**c'est-à-dire**</MDR> le le <NP2 rel-lex=«hyper(carton/dessin)» modif-lex= «remplacement(un/ce…-là) remplacement(carton/dessin) ajout(Adj=agrandi)» rel-pragm=«prec»>**ce dessin-là agrandi** </NP2> mais à la grandeur de la fenêtre*

*(17) eslo1_ENT_121_C|euh <VP1>**démocratiser l'enseignement**</VP1> <MRP>**c'est-à-dire** </MRP> <VP2 rel-lex= « syn(démocratiser/permettre à tout le monde) syn(enseignement/ faculté) » modif-lex= « ajout(rentrer à) » rel-pragm= « explic »>**permettre à tout le monde de rentrer en faculté**</VP2>*

(17)forum|Donc <P1>cela demande beaucoup d'investissement</P1>, <MRP>c'est à dire</MRP> <P2 rel_lex="syno(cela demande/il faut)" rel_pragm="prec">il faut tout gérer, les aides à domicile, le changement des médecins, etc...</P2>

Dans l'exemple (15), le locuteur décide de remplacer le lieu *Saint Jean de la Ruelle* par *Orléans* en utilisant le marqueur *c'est-à-dire* pour permettre à son interlocuteur, un anglophone, de mieux comprendre de quelle endroit il s'agit. Les deux entités nommées entretiennent une relation de méronymie : *Saint Jean de la Ruelle* fait partie de l'agglomération d'*Orléans*. En ce qui concerne la relation pragmatique, nous indiquons une correction référentielle déclenchée par une supposée incompréhension de l'interlocuteur. Dans l'exemple (16), le groupe nominal *un carton* est remplacé par un autre groupe nominal *ce dessin-là agrandi.* Plusieurs modifications ont été mises en œuvre pour effectuer cette correction afin de préciser l'information fournie. L'unité lexicale *carton* est remplacée par son hyperonyme *dessin,* on passe on passe d'une référence indéfinie *un* à un déictique *ce…-là*, et on ajoute l'adjectif qualificatif *agrandi*. Enfin, dans l'exemple (17), le locuteur remplace une proposition *cela demande beaucoup d'investissement* par une autre *il faut tout gérer, les aides à domicile, le changement des médecins, etc...* pour préciser à son interlocuteur ce qu'il entend par *beaucoup d'investissement*. Cette précision est suivie par le remplacement de *cela demande* par son synonyme *il faut.*

4.3 Processus de l'annotation

L'annotation des corpus, les corpus oraux (ESLO1et ESLO2) et le corpus de forum, est effectuée manuellement en respectant le format XML. La méthodologie choisie pour annoter les deux corpus est la suivante : les deux annotateurs annotent d'abord le même corpus séparément, ensuite ils annotent le même corpus ensemble pour arriver à un consensus en cas de discordance. Trois versions d'annotations du même corpus sont ainsi obtenues, la dernière étant considérée comme la version définitive.

Plusieurs paires d'annotateurs participent à la tâche. Les auteurs de l'article assurent les annotations. Par ailleurs, une partie de corpus est annotée par des étudiants, dans le cadre du séminaire de recherches consacré à l'enrichissement des corpus. La même méthodologie d'annotation est alors appliquée. La version consensuelle des annotations des étudiants est revue et corrigée si nécessaire par les auteurs de l'article.

4.4 Évaluation

L'annotation manuelle est évaluée en comparant les versions de corpus annotés provenant de différents annotateurs. En ce qui concerne les données annotées par les étudiants, la version finale des étudiants et la version corrigée sont comparées. L'accord inter-annotateur sur les jugements de l'existence de la relation de reformulation est ensuite calculé avec le kappa de Cohen [12].

Tableau 1. Accord inter-annotateur sur l'ensemble des corpus

Corpus	*Accord*	*Interprétation*
ESLO1	0,617	Accord fort
ESLO2	0,526	Accord modéré
Forum (entre les chercheurs)	0,732	Accord fort
Forum (étudiants/chercheurs)	0,836	Accord presque parfait

Dans le Tableau 1, nous présentons les accords inter-annotateurs pour les corpus. Les interprétations de l'accord suivent la grille standard [26]. L'accord est modéré avec le corpus ESLO2, qui présente en effet des structures discursives complexes. L'accord est fort pour les corpus ESLO1 et forum annoté par les chercheurs. Lorsqu'il s'agit de la reprise des annotations des étudiants par les chercheurs, l'accord est presque parfait, ce qui montre que c'est une tâche assez consensuelle sur les données écrites. Dans l'ensemble d'énoncés et de phrases comportant les marqueurs étudiés, les annotateurs reconnaissent entre 17 et 27 % des reformulations dans le corpus ESLO1, entre 26 et 33 % des reformulations dans le corpus ESLO2, et entre 60 et 63 % des reformulations dans le corpus forum.

5 Détection automatique des énoncés contenant la reformulation

La détection automatique des énoncés contenant la reformulation est effectuée avec une méthodologie proposée pour traiter les corpus oraux [15]. Elle est étendue au traitement du corpus de forum. La méthodologie est présentée dans la

Fig. ***1***. L'annotation manuelle des reformulations a permis de définir des règles de désambiguïsation automatique. Le traitement automatique consiste alors à décider si, autour d'un marqueur, il existe une relation de reformulation ou non.

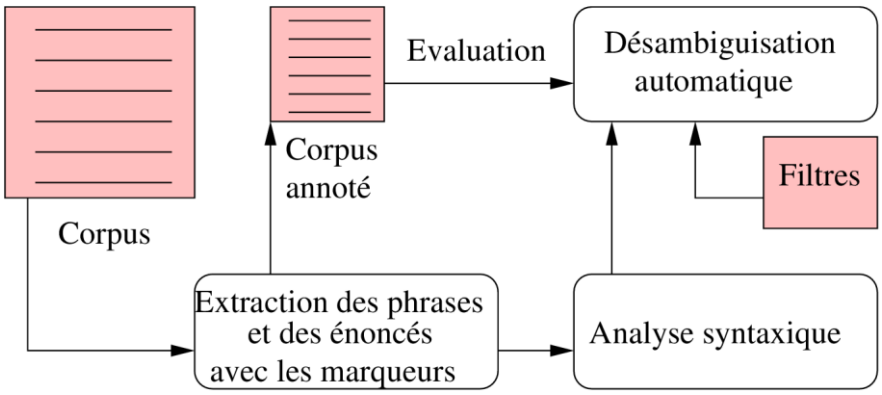

Fig. 1. Schéma général pour la détection automatique des reformulations

Les tours de paroles et les messages de discussion sont prétraités par le chunker SEM[5] adapté à l'oral car le modèle a été appris sur l'échantillon d'ESLO [38] et par l'analyseur syntaxique Cordial [25]. Les sorties de ces outils sont utilisées dans le processus de désambiguïsation automatique pour distinguer les emplois avec et sans reformulation. Plusieurs filtres, qui vérifient le contexte d'emploi des marqueurs, sont appliqués :

- Lorsque *disons* suit le pronom personnel *nous,* il est considéré qu'il s'agit d'un verbe *dire* conjugué au présent de l'indicatif ;
- Lorsque l'un des marqueurs se trouve dans des suites argumentatives (e.g. *par contre, mais, en revanche, au contraire, cependant*), il est considéré qu'il s'agit de l'introduction d'une nouvelle information par opposition à ce qui est annoncé précédemment et non de la reformulation ;
- Lorsque le marqueur est placé en début ou en fin d'énoncé, le contexte n'est pas jugé suffisant pour établir une reformulation ;
- Lorsque l'un des marqueurs se retrouve au milieu d'un syntagme syntaxique, il est considéré qu'il est employé dans le discours en tant que marqueur discursif. Observons les syntagmes où est inséré le marqueur *disons* :

  (18) une personne ***habitant disons à cinq kilomètres de de chez vous*** (V disons GPREP)

  je ***vocalise disons une demi-heure*** (V disons GN)

  la porte *euh* ***disons de cette pièce*** (GN disons PREP GN)

  les jeunes disons qui se marient (GN disons PrREL V)

  c'est ***un chant disons assez vulgaire*** (GN disons ADJ)

  on n'a pas été ***suivi disons par les parents*** (V disons par GN)

  comment vous ***avez disons choisi*** (VAUX disons PP)

Les règles de désambiguïsation sont implémentées pour pouvoir décider automatiquement si un tour de parole ou une phrase comprenant un des trois marqueurs contiennent ou non une reformulation. Pour détecter les emplois où il s'agit d'un marqueur discursif, et non d'un marqueur de reformulation, c'est-à-dire lorsque le marqueur en question apparaît à l'intérieur d'une locution, un moteur de recherche généraliste est interrogé. La fréquence importante d'un syntagme sur la Toile montre le degré de son figement. Le module de la détection automatique a été évalué. La précision atteint 63% pour le corpus ESLO1 et 66% pour ESLO2.

6 Analyse linguistique des résultats

L'annotation effectuée a permis de vérifier et de soumettre des hypothèses concernant le phénomène de la reformulation introduite à l'aide des trois marqueurs *c'est-à-dire, je veux dire* et *disons* dans les trois corpus étudiés.

6.1 ESLO1 et ESLO2

L'annotation manuelle du corpus ESLO montre que le marqueur *c'est-à-dire* est le plus fréquent dans le corpus oral. *C'est-à-dire* et *disons* représentent 30% de cas dans ESLO1 et

[5] http://www.lattice.cnrs.fr/sites/itellier/SEM.html

16% dans ESLO2 en tant que marqueurs de reformulation (MR). Le marqueur *je veux dire* fonctionne comme un MR entre les deux segments dans 14% (ESLO1) et 19% (ESLO2) des cas. Il est donc moins utilisé dans ESLO1 et plus fréquent que les deux autres marqueurs dans ESLO2. Faut-il y voir une évolution diachronique de la langue telle que ce marqueur reprendrait la fonction de MRP des deux autres marqueurs. Une autre explication est liée au contexte d'enregistrement d'ESLO2 qui favorise mieux la parole spontanée et incite l'interviewé à parler d'une manière moins formelle. La structure du marqueur *je veux dire* (le pronom personnel à la première personne *je* et le verbe modal *vouloir*) présuppose ainsi une implication plus forte de l'énonciateur symptomatique d'un changement en cours.

L'analyse du corpus annoté a permis de constater le parallélisme syntaxique entre l'entité source et l'entité reformulée mentionné déjà par les travaux précédents [20, 34]. Dans la majorité des cas (60 %), il existe une équivalence syntaxique entre les éléments en relation de reformulation. On pourrait expliquer ce phénomène aussi par l'aspect subjectif de l'annotation manuelle. En effet, la décision sur la délimitation des frontières des segments reformulés revient aux annotateurs qui peuvent être « influencés » par une catégorie syntaxique du premier segment.

Les études sur les paraphrases et reformulation citées dans la section 2, convergent vers l'existence du lien formel entre les segments reformulés. Ce lien est modélisé à travers les transformations morphologiques, lexicales et syntaxiques. Une annotation multidimensionnelle tenant compte de ces aspects a fait la démonstration du contraire :

- Les modifications morphologiques ne représentent que 10% des reformulations annotées ;
- Nous n'avons observé qu'un seul exemple de modification syntaxique (actif/passif) dans le corpus annoté ESLO1 ;
- Les modifications faites au niveau lexical représentent 57% de toutes les reformulations annotées dans ESLO1 et 30% dans ESLO2 ;
- En ce qui concerne les relations lexicales, elles occupent 75% des cas annotés dans ESLO1 et 58% dans ESLO2. Du point de vue diachronique, nous observons donc un recul des liens lexicaux entre les éléments contenus dans les deux segments reformulés.

En conclusion, nous pouvons constater, à l'intérieur du corpus oral analysé, qu'il existe très peu de repères formels pour détecter des segments en relation de reformulation dans ce type de constructions. Cette remarque est valable à différents niveaux (morphologique, lexical et syntaxique).

D'autres observations moins significatives peuvent être ajoutées. Parmi les catégories syntaxiques, les plus fréquentes sont les propositions suivies des groupes nominaux. Dans un corpus oral, reformuler des énoncés entiers semble « naturel ». La relation lexicale la plus souvent annotée est la synonymie (*ça marchait/y avait pas de rejet, quelque chose de potable/quelque chose euh de correct*, etc.), ce qui semble aussi logique. Le lien de méronymie (*les camps scouts/les plus jeunes, papier/bibliothèque*, etc.) entre les éléments des deux segments augmente dans ESLO2. Nous avons considéré comme méronymie les cas du rapport /*partie vs tout*/ en y ajoutant également les liens par association. La conduite pas très formelle de l'entretien dans ESLO2 peut favoriser la présence de ces liens associatifs entre les éléments. En ce qui concerne les modifications lexicales, suite à leur analyse quantitative, le locuteur paraît préférer faire la substitution d'un mot ou d'un sous-segment (l'exemple 16) plutôt que d'ajouter de nouveaux mots (les exemples 16 et 17*)* ou d'en supprimer du premier segment (*l'accent un peu de travers/l'accent, un autre mot même euh même un mot euh affreux/un autre mot,* etc.).

6.2 Forum

Les variations orthographiques fréquentes dans l'écriture sur les forums rendent le traitement et l'analyse de ce corpus difficile. Ce phénomène concerne aussi les marqueurs étudiés. Ainsi, le marqueur *c'est-à-dire* est écrit de quatre façons différentes dans le corpus : *c'est-à-dire*, *c'est à dire*, *cad*, *c'est a dire*. De même manière, il a été observé une variation dans l'emploi de la majuscule : *disons* vs *Disons*. Toutes ces particularités doivent être prises en compte dans le traitement automatique de ce corpus mais aussi dans l'analyse quantitative.

Suite à l'annotation manuelle du phénomène de reformulation dans le corpus de forum, on peut constater que le marqueur *c'est-à-dire* est le plus fréquent. L'équivalence entre les catégories syntaxiques se vérifie dans 37 % de cas.

L'hypothèse de l'existence du lien formel au niveau linguistique entre les deux segments reformulés a été vérifiée aussi dans le corpus de forum. Selon l'annotation, les modifications morphologiques ne représentent que 4% des reformulations annotés. Dans neuf cas annotés, sept montrent une dérivation (*névralgies/nerf, intercostales/côtes*, etc.). Les modifications syntaxiques sont absentes. Comme dans le corpus ESLO, ce sont les relations lexicales qui sont les plus nombreuses. Elles occupent 44% des cas annotés. Les modifications faites au niveau lexical représentent 23% de toutes les reformulations annotées. Il s'agit le plus souvent des remplacements d'une partie ou d'une unité lexicale par une autre. Parmi les relations lexicales, la plus répandue est la synonymie (35%) (*je suis chez moi/je suis immobile, fin/mort,* etc.), suivie de l'hyperonymie (27%) (*troubles/extrasystoles, humeur/être déprimée, agressive,* etc.), de la méronymie (16%) (*psychosomatique/cerveau, bras/main, vendredi/semaine*, etc.) et de l'instance (14%) (*traitement/nebilox, le centre caridologie du nord/la clinique de saint denis,* etc.).

En ce qui concerne les relations pragmatiques, nous constatons que la personne s'exprimant dans le cadre du forum sur le Web utilise le processus de la reformulation le plus souvent pour préciser ce qui a été dit (37%) (les exemples 4 et 7). La relation la moins fréquente est la justification (3%) :

> *(19) Il y a aussi le fait que* ***je n'arrive pas à faire confiance en le diagnostic de ce cardiologue (disons qu'il n'avait pas l'air de rouler sur l'or et mon cerveaux ne perd pas de temps pour en tirer toutes sortes de conclusions plus ou moins stupides****), enfin bref je suis en plein doute*

6.3 ESLO1 et ESLO2 vs Forum

L'analyse comparative fondée sur l'annotation multidimensionnelle des trois corpus a permis de faire les hypothèses et les observations suivantes.

Observons la Fig. ***2*** qui montre la distribution des relations formelles entre les deux segments reformulés.

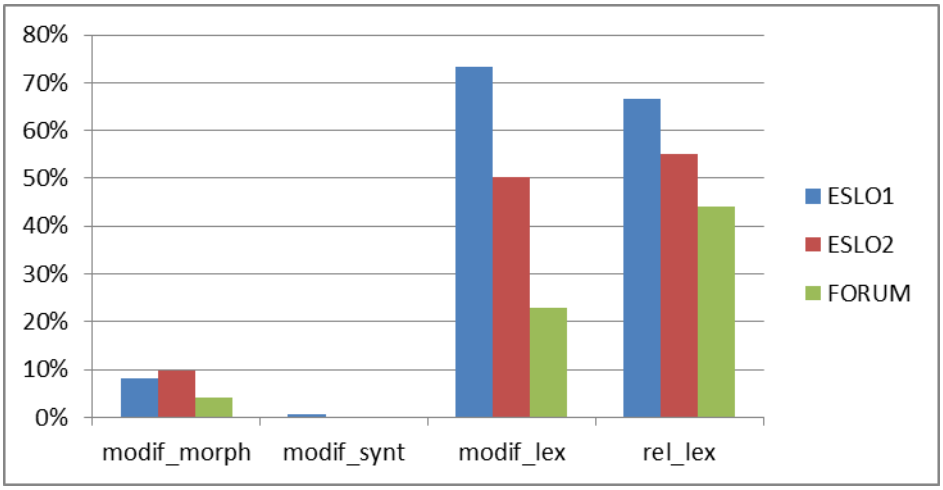

Fig. 2. Distribution des liens formels entre les segments reformulés dans trois corpus

Les trois corpus montrent une tendance générale : le lien formel entre les deux segments reformulés se manifeste plutôt au niveau lexical. Les relations lexicales qui varient entre 45% et 65% sont les plus représentées, suivies des modifications lexicales 23%-72%. Ce lien est très représenté dans ESLO1, sa fréquence diminue progressivement avec ESLO2 et le forum. Cela peut être expliqué par la nature et le cadre de constitution du corpus. ESLO1 est le plus formel, il s'agit d'interviews dirigées par un locuteur anglophone. ESLO2 est un corpus oral contemporain : les cadres des entretiens menés par les membres de l'équipe LLL, sont plus libres et les questions sont moins directives. Le corpus de forum est un corpus issu du Web où les utilisateurs s'expriment d'une manière très libre sans aucune contrainte. Les modifications morphologiques sont peu représentées dans les trois corpus. Enfin, les transformations syntaxiques sont quasiment absentes dans les trois corpus.

La distribution des relations lexicales dans les trois corpus (voir Fig. ***3***) montre une fréquence importante d'un lien de *synonymie* entre les unités lexicales des segments reformulés. Le lien d'*antonymie* est le moins fréquent. Ces deux observations semblent logiques vue la nature du phénomène étudié. Au niveau hiérarchique le locuteur a une tendance d'utiliser un hyperonyme dans le premier segment. Cette tendance est confirmée par la relation pragmatique la plus fréquente : *précision* (voir Fig. ***4***). Le locuteur reformule son énoncé pour rendre ses propos plus précis et plus clairs. Le lien de méronymie est très fréquent dans ESLO2 pour des raisons liées, selon nous, aux conditions d'enregistrements de ce corpus.

En ce qui concerne la distribution des relations pragmatiques dans les trois corpus (voir Fig. ***4***), c'est la *précision* avec une fréquence relativement élevée (entre 26% et 34%) qui occupe la première place. Deux fonctions : *correction référentielle* (l'exemple 15) et *justification* (l'exemple 19), sont peu présentes dans les trois corpus. La *paraphrase* a une fréquence quasi similaire dans les trois corpus et varie entre 7% et 10% :

> *(20) forum|C'est un appareil qui prend automatiquement la tension pendant* ***24 heures c'est-à-dire jour et nuit***
>
> *(21) eslo2_ENT_12|****en vélo*** *enfin* ***je veux dire*** *euh ou* ***en deux roues***

Les autres fonctions ont une distribution plus aléatoire. Certaines de ces observations peuvent être expliquées par la nature des corpus. C'est le cas de la *dénomination*

(22) c'etait surtout depuis qu'on m'avait changer de ***traitement c'est-à-dire le nebilox***

et de la *définition*

(23) ***la maladie de bouveret, c'est à dire un fil conducteur en trop au niveau du coeur et de temps en temps il y a court circuit et ça disjoncte***

fréquentes dans le corpus de forum qui traite de la santé. En effet, les utilisateurs dénomment ou définissent souvent dans leurs messages les termes médicaux.

D'une manière générale, l'annotation des relations pragmatiques a permis de distinguer trois processus fonctionnels de la reformulation : le locuteur (i) ajoute une nouvelle information (*explication, précision, exemplification, justification et définition*), (ii) répète la même information mais d'une autre façon, dans ce cas il est possible de supprimer le marqueur et de changer les segments de place sans modifier le sens de l'énoncé (*paraphrase*) et (iii) synthétise (*résultat*) ou dénomme (*dénomination*) ce qui vient d'être dit. Ce constat est confirmé par l'association entre une relation pragmatique annotée et la taille des segments calculée en nombre de mots. Nous distinguons ainsi trois cas :

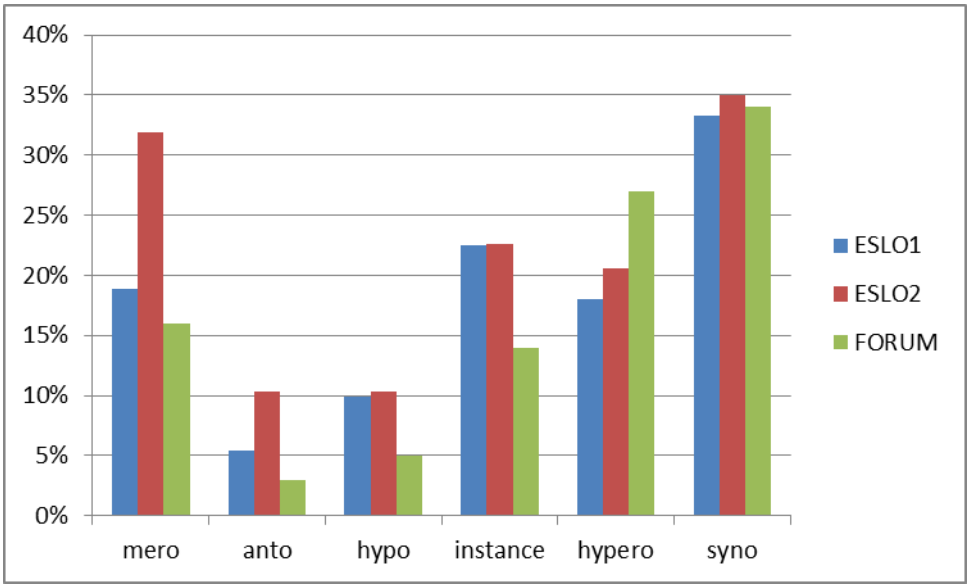

Fig. 3. Distribution des relations lexicales dans les trois corpus

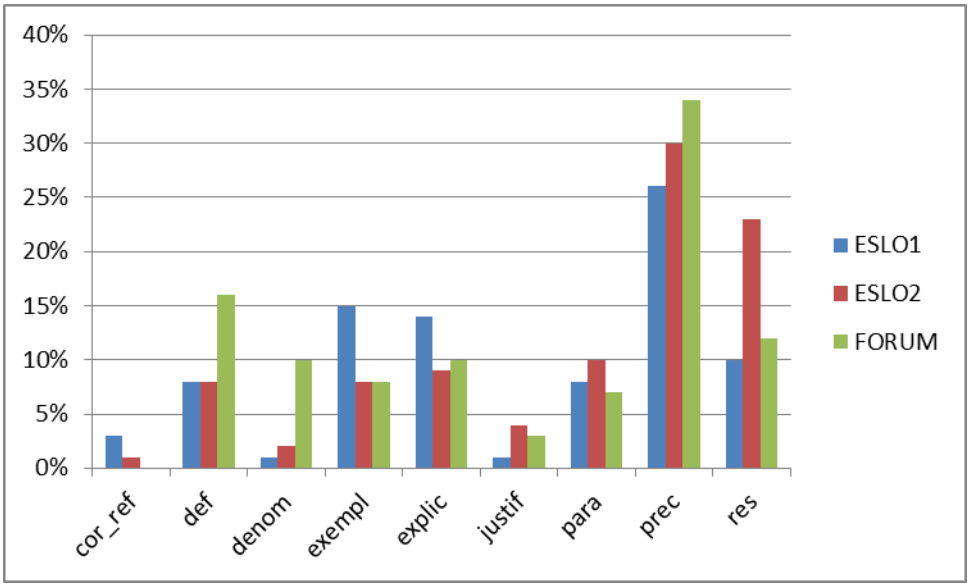

Fig. 4. Distribution des relations pragmatiques dans trois corpus

- le segment 2 est plus long que le segment 1 : le locuteur précise, définit, explique ou exemplifie ses propos (les exemples 1, 4-7, 17, 23) ;
- le segment 1 est plus long que le segment 2 : le locuteur conclut, raccourcit ou synthétise ce qui a été dit (l'exemple 2) ou il donne le nom à ce qui a été annoncé précédemment (l'exemple 22) ;
- les deux segments sont équivalents : il s'agit de la paraphrase « pure », de la correction linguistique et référentielle (les exemples 15, 20, 21).

Ce lien entre la fonction pragmatique de la reformulation et la différence de longueur entre les segments reformulés pourrait être mobilisé comme critère de classification automatique de la reformulation.

7 Conclusion

Le présent article aborde le phénomène de la reformulation à travers l'analyse de trois corpus : corpus oral ESLO1, enregistré dans les années 70 et transcrit maintenant par l'équipe du laboratoire LLL, corpus oral contemporain ESLO2, comparable à ESLO1, et le corpus de forum.

La méthodologie utilisée est fondée sur la modélisation du phénomène à travers les étiquettes marquant les différents aspects linguistiques et décrivant la modification que subit un segment ou un énoncé au cours de la reformulation. Les trois corpus ont été annotés avec le même jeu d'étiquettes. L'annotation effectuée est donc multidimensionnelle. Le module de détection automatique des énoncés et des tours de paroles contenant la reformulation a été proposé et testé. Ce module ainsi que l'annotation manuelle ont été évalués.

En se fondant sur les corpus annotés, l'analyse quantitative du phénomène a été effectuée. Il a été observé qu'il existe très peu de liens formels entre les deux segments reformulés. Le lien formel le plus fréquent concerne les relations lexicales. La catégorie syntaxique qui subit le plus la reformulation est le groupe nominal. On constate une

équivalence syntaxique entre les deux segments reformulés qui peut s'expliquer par le processus de l'annotation manuelle. La raison pour laquelle le locuteur procède à la reformulation semble être la précision, même si d'autres fonctions pragmatiques ont été annotées. Il a été observé aussi un lien entre la fonction pragmatique et la taille des segments reformulés.

Plusieurs perspectives du travail sont envisagées. En premier lieu, le croisement des annotations effectuées et des critères sociologiques sur les locuteurs d'ESLO sera réalisé. L'objectif est de vérifier l'existence du lien entre le processus de la reformulation et le profil sociologique du locuteur. La prise en compte des informations prosodiques et temporelles est envisagée en deuxième lieu. Enfin, le filtrage automatique des reformulations et la détection de segments en relation de reformulation sera améliorée et testée sur l'ensemble des corpus disponibles.

Références

1. J. Authier-Revuz, *Ces mots qui ne vont pas de soi : boucles réflexives et non-coïncidences du dire*. Paris : Larousse, (1995).
2. K. Beeching, La co-variation des marqueurs discursifs bon, c'est-à-dire, enfin, hein, quand même, quoi et si vous voulez : une question d'identité ? *Langue française*, **154(2)**, p. 78-93, (2007).
3. C. Benzitoun, L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? *RECITAL 2004*, Fès, 21 avril 2004, (2004).
4. R. Bhagat, E. Hovy, What is a paraphrase ? *Computational Linguistics*, **39(3)**, p. 463-472, (2013).
5. M. Bilger, Coordination: analyses syntaxiques et annotations. *Recherches sur le français parlé*, **15**, p. 255-272, (1999).
6. C. Blanche-Benveniste, Un modèle d'analyse syntaxique "en grilles" pour les productions orales. *Anuario de Psicologia*, **47**, p. 11-28, (1990).
7. C. Blanche-Benveniste, Le semblable et le dissemblable en syntaxe. *Recherches sur le français parlé*, **13**, p. 7-33, (1995).
8. C. Blanche-Benveniste, Les études sur l'oral et le travail d'écriture de certains poètes contemporains. *Langue française, L'oral dans l'écrit,* vol. **89**, pp. 52-71, (1991).
9. C. Blanche-Benveniste, C. Jeanjean, *Le français parlé, transcription et édition*. Paris : Didier érudition, (1987).
10. C. Blanche-Benveniste, M. Bilger, C. Rouget, K. Van den Eynde, *Le français parlé : études grammaticales*. Paris : éditions du CNRS, (1990).
11. H. Bouamor, *Étude de la paraphrase sous-phrastique en traitement automatique des langues*. Thèse de doctorat, Université Paris Sud, Paris, (2012).
12. J. Cohen, A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20(1)**, p. 37-46, (1960).
13. A. Culioli, *Notes du séminaire de DEA*, Paris, (1976).
14. I. Eshkol-Taravella, O. Baude, D. Maurel, L. Hriba, C. Dugua, I. Tellier, Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *TAL*, **52(3)**, p. 17-46, (2012).
15. I. Eshkol-Taravella, N. Grabar, Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. *TALN2014*, (2014).

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

17. C. Fuchs, *Paraphrase et énonciation*. Paris : Orphys, (1994).

18. N. Grabar, I. Eshkol-Taravella, ...des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux. *TALN2015*, (2015).

19. M.-L. Guénot, La coordination considérée comme un entassement paradigmatique: description, formalisation et intégration. In Mertens, P., C. Fairon, A. Dister & P. Watrin (éds), *Cahiers du Cental*, **2:1**, *Verbum ex machina, Actes de la 13ème Conférence sur le Traitement Automatique des Langues*, Leuven, Belgique, 10-13 April 2006, 1, p. 178-187, (2006).

20. E. Gulich, T. Kotschi, Les marqueurs de la reformulation paraphrastique. *Cahiers de linguistique française*, **5**, p. 305-351, (1983).

21. E. Gülich, T. Kotschi, Les actes de reformulation dans la consultation La dame de Caluire. In P. Bange, (ed.), *L'analyse des interactions verbales. La dame de Caluire : une consultation*, P Lang, Berne, pp. 15–81, (1987).

22. Y. Hwang, Eh bien, alors, enfin et disons en français parlé contemporain. *L'Information Grammaticale*, **57**, p. 46-48, (1993).

23. S. Kahane, P. Pietrandrea, La typologie des entassements en français. *CMLF2012*, (2012).

24. L. Kanaan, *Reformulations, contacts de langues et compétence de communication : analyse linguistique et interactionnelle dans des discussions entre jeunes Libanais francophones*. Thèse de doctorat, Université d'Orléans, Orléans, (2011).

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

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

27. W. J. M. Levelt, *Monitoring and self-repair in Speech, Cognition*, **14**, p. 41-104, (1983).

28. R. Martin, *Inférence, antonymie et paraphrase*. Paris : Klincksieck, (1976).

29. I. Melčuk, Paraphrase et lexique dans la théorie linguistique sens-texte in lexique et paraphrase. *Lexique*, **6**, p. 13-54, (1988).

30. M. Petit, *Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs.* Thèse de doctorat, Université d'Orléans, Orléans, (2009).

31. M. Riegel, J.-C. Pellat, R. Rioul, *Grammaire Méthodique du Français*. PUF, Paris, (1994).

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

33. C. Rossari, De l'exploitation de quelques connecteurs reformulatifs dans la gestion des articulations discursives. *Pratiques*, **75**, p. 111-124, (1992).

34. C. Rossari, *Les opérations de reformulation : analyse du processus et des marques dans une perspective contrastive français-italien*. Berne : Peter Lang, (1994).

35. E. Roulet, Complétude interactive et connecteurs reformulatifs. *Cahiers de linguistique française*, **8**, p. 111-140, (1987).

36. A. Steuckardt, A. Niklas-Salminen, *Le mot et sa glose*. Publications de l'Université de Provence, (2003).

37. A. Steuckardt, Les marqueurs formés sur dire. In A. Steuckardt & A. Niklas-Salminen (éds), *Les Marqueurs de glose*, Aix-en-Provence : Publications de l'Université de Provence, p. 51-65, (2005).

38. I. Tellier, I. Eshkol-Taravella, Y. Dupont, I. Wang, Peut-on bien chunker avec de mauvaises étiquettes POS ? Actes de *TALN2014*, (2014).

39. S. Teston-Bonnard, Je veux dire est-il toujours une marque de reformulation ? In M. L. Bot, M. Schuwer, E. Richard, Eds., *Rivages linguistiques. La Reformulation. Marqueurs linguistiques. Stratégies énonciatives*, p. 51-69. Rennes : PUR, (2008).

40. L. Vezin, Les paraphrases : étude sémantique, leur rôle dans l'apprentissage. *L'année psychologique*, **76(1)**, p. 177-197, (1976).

41. M. Vila, M. Antònia, H. Rodríguez, Paraphrase concept and typology, a linguistically based and computationally oriented approach. *Procesamiento del Lenguaje Natural*, **46**, p. 83-90, (2011).

42. R. Vion, Reprise et mode d'implication énonciative. *La linguistique*, **42**, p. 11-28, (2006).