# Reformulations avec et sans marqueurs : étude de trois entretiens de l'oral

Iris Eshkol-Taravella, Natalia Grabar

# Reformulations avec et sans marqueurs : étude de trois entretiens de l'oral

*Iris* Eshkol-Taravella[1], *Natalia* Grabar[2]

[1]MoDyCo – UMR 7114, Université Paris Nanterre, 200 avenue de la République, 92001 Nanterre, France
[2]STL – UMR 8163, CNRS, Université de Lille, F-59000 Lille, France

**Résumé.** La communication porte sur l'analyse des reformulations produites dans le discours oral. La reformulation est prise au sens large en tant que procédé de modification d'un segment par un autre mais elle garde toujours un lien sous-jacent entre les deux segments reformulés. Ce lien se manifeste à différents niveaux linguistiques : lexical, morphologique, sémantico-pragmatique. Généralement, les études sur la reformulation se fondent sur la présence de marqueurs. Le travail présenté montre que la reformulation peut être réalisée sans marqueurs également. Le segment source est interrompu et suivi par un segment reformulé. Cette interruption peut être « remplie » par un élément lexical : un marqueur qui introduit la reformulation (marqueur de reformulation « classique », marqueur de correction, marqueur d'exemplification, marqueur de conclusion), un élément disfluent (amorce, interjection, hésitation, marqueur discursif), un présentateur. Les raisons de la modification peuvent être nombreuses : correction, paraphrase, explication, définition, justification, conclusion, précision, dénomination, exemplification, etc. A partir de cette définition du procédé de reformulation, celui-ci est modélisé par un jeu d'étiquettes correspondantes selon lesquelles le corpus est annoté manuellement. La démarche est inductive et permet d'appréhender la reformulation sous un autre angle, ainsi que d'observer et de quantifier certaines de ses caractéristiques.

**Abstract. Reformulations with and without markers: study of three spoken interviews.** The communication addresses the analysis of reformulations in spoken language. Reformulations include a large set of situations, in which one segment is modified by another but still keeps the underlying link between these two segments. This link revealed at different linguistic levels: lexical, morphological, and semantic and pragmatical. Usually, the existing studies on reformulation rely on the presence of markers. The work presented here shows that reformulation can also be realized without markers. The source segment is then interrupted and followed be the reformulated segment. This interruption can be « filled » by lexical elements: marker which introduces the reformulation (« classical » reformulation marker, correction marker, exemplification marker, conclusion marker), dysfluence (primes, interjections, hesitations, discursive marker), presentator. Several reasons for the modification can exist: correction, paraphrase, explanation, definition, justification, conclusion, precision, denomination, exemplification, etc. Starting with this definition of reformulation, we model it with a set of tags, according to which the corpus is annotated manually. The approach is inductive and permits to apprehend the reformulation from another point of view, and to observe and quantify some of its characteristics.

## 1. Introduction

La reformulation est une des caractéristiques inhérentes du discours oral. Ce procédé occupe une place importante dans les travaux actuels en linguistique française.

La revue Corela (HS-18) a consacré un numéro[1] à l'étude des reformulations dans l'acquisition et l'enseignement du français langue étrangère et du français langue maternelle. En juin 2017, un colloque international intitulé « La reformulation à la recherche d'une frontière » a eu lieu à Uppsala. Ce colloque a réuni des chercheurs travaillant dans les disciplines et applications différentes comme la didactique, l'analyse de discours, la traduction, le traitement automatique des langues, etc. pour y débattre sur la définition de la notion de reformulation et sur la distinction entre la reformulation et les procédés proches comme exemplification, définition ou répétition. Un atelier consacré à la reformulation aura lieu à Genève en mars 2018. Il vise également à définir ce procédé et rassembler les linguistes travaillant dans divers cadres théoriques afin de réfléchir sur la notion de reformulation. Il s'agit donc d'une notion qui continue à préoccuper les linguistes et dont l'étude est fondée de plus en plus sur l'analyse de corpus.

Pour reconnaître la reformulation, les linguistes s'appuient le plus souvent sur la présence de marqueurs spécifiques. Ainsi, Rossari [30],[31],[32] introduit, dans les années 90, la distinction entre les marqueurs de la « reformulation paraphrastique », comme *c'est-à-dire*, *autrement dit*, etc. qui instaurent une relation d'équivalence sémantique entre les deux segments reformulés, et les marqueurs de la « reformulation non paraphrastique » comme *en somme*, *en bref*, etc. où cette équivalence sémantique est absente. Depuis, les marqueurs de reformulation paraphrastique restent au centre de plusieurs recherches [17, 40]. Les linguistes s'intéressent aux marqueurs particuliers : *c'est-à-dire* [14, 15, 39, 41], *disons* [34, 7], *je veux dire* [38], *cela dit* [33], *genre* [43], ou bien étudient les marqueurs en fonction de leur

[1] https://corela.revues.org/4029

appartenance à un groupe, comme par exemple les marqueurs dérivés du verbe *dire* [22, 36]. Notons que les marqueurs de reformulation peuvent également être analysés en tant que marqueurs discursifs [1, 10, 20, 29].

[21] constate que dans le discours oral, la reformulation peut être introduite par certains marqueurs discursifs comme *hum, bon, quoi*, mais aussi par une répétition d'un mot ou encore par des moyens (para)linguistiques : pause, allongement de syllabe, etc.

L'article s'inscrit dans le domaine de la linguistique de corpus avec l'application au traitement automatique des langues (TAL). L'hypothèse de départ est que la reformulation à l'oral peut être réalisée dans un énoncé avec mais aussi sans marqueurs. Pour vérifier cette hypothèse, trois transcriptions du corpus oral ESLO2[2] sont annotées en reformulations selon une typologie multidimensionnelle prédéfinie. Le corpus ainsi annoté a permis de faire des premières observations qui ont confirmé l'hypothèse et ont permis de quantifier et d'analyser les éléments présents entre les deux segments reformulés. Ces résultats sont également comparés avec les études antérieures des auteurs [11,12] portant sur les reformulations introduites par trois marqueurs *c'est-à-dire, je veux dire, disons.*

## 2. Données

Le corpus d'étude est composé de trois entretiens de l'oral conduits dans le cadre du corpus Enquêtes Sociolinguistiques à Orléans (ESLO2). Il s'agit des discussions en face à face entre un chercheur et un locuteur témoin à partir d'une trame d'entretien. Les questions portent sur le locuteur (son identité, son travail, sa famille) et sur sa vie à Orléans. Les métadonnées des trois entretiens sont présentées dans le Tableau 1.

**Tableau 1.** Métadonnées sur les entretiens

| Numéro d'entretien | 1005 | 1033 | 1053 |
|---|---|---|---|
| Durée | 01:02:00 | 00:56:00 | 00:41:00 |
| Nombre de mots | 13974 | 10306 | 8442 |
| Locuteur | Femme<br>Bac+5 et plus<br>Femme au foyer<br>âge : 45/55 | Femme<br>Bac<br>Commerçante<br>âge : 35/45 | Homme<br>Bac +2<br>Agent SNCF<br>police ferroviaire<br>âge : 35/45 |

La durée moyenne des entretiens est de 50 minutes. Le nombre de mots varie entre 8400 et 14000. Pour observer le procédé de reformulation d'une manière objective, les trois locuteurs de profils sociologiques différents mais d'un âge comparable (environ 40 ans) sont sélectionnés.

La transcription des entretiens est orthographique. Les fichiers de transcription ne comportent pas de signes de ponctuation, ni de majuscules au début d'énoncés pour éviter l'anticipation de l'interprétation car en ponctuant, le transcripteur « suggère une analyse avant de l'avoir faite » (p.142) [3].

## 3. Modélisation de reformulation

### 3.1. Définition de reformulation

Le procédé de reformulation est une caractéristique naturelle du langage humain. A l'écrit, il se présente en tant qu'un produit fini [18], alors qu'à l'oral la reformulation est une étape de production langagière, de l'élaboration du discours, d'où la présence de phénomènes comme hésitations, faux-départs, diverses formes de reprises, qu'on appelle les disfluences [2].

La modélisation de disfluences est proposée par [24] et reprise par [35]. Elle est composée de trois éléments :

[reparandum] * (phase d'édition) [repair]

- le « reparandum » désigne un segment que le locuteur souhaite modifier ;
- la « phase d'édition » est optionnelle, ce qui est marqué par les parenthèses. Elle suit le point d'interruption représenté par « * ». C'est pendant cette phase qu'apparaissent les disfluences ou les marqueurs, par exemple.
- le « repair » est un segment modifié.

Le schéma peut être appliqué, selon nous, au procédé de reformulation où le « reparandum » renvoie vers le premier segment de discours, le segment qui sera reformulé, la « phase d'édition », si elle est présente, désigne l'apparition de marqueurs de reformulations, de disfluences ou d'autres éléments introduisant le segment reformulé, appelé « repair » dans le modèle de [24] et de [35]. Dans le cas où la « phase d'édition » est absente, le segment reformulé suit directement le segment source. Le procédé de reformulation peut donc être représenté de la manière suivante :

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

La reformulation est ainsi considérée dans notre travail comme un procédé de modification d'un segment, segment source, dans le discours par un autre, segment reformulé. Cette modification peut déclencher une « phase d'édition »

[2] eslo.huma-num.fr/

« signalée » d'une manière directe (présence de marqueurs de reformulation), d'une manière indirecte (présence de disfluences ou de marqueurs discursifs) mais peut également ne pas avoir cette phase du tout (aucun élément linguistique n'apparaît entre les deux segments reformulés). La présence d'un marqueur n'est donc pas nécessaire pour que la reformulation ait lieu et n'est pas un critère unique de sa reconnaissance.

Un autre critère est un lien sous-jacent existant entre les deux segments reformulés. Ainsi, [26] définit la reformulation comme la « reprise d'un énoncé antérieur qui maintient, dans l'énoncé reformulé, une partie invariante à laquelle s'articule le reste de l'énoncé » [26]. C'est cette « partie invariante », ou un lien sous-jacent, qui permet d'établir une relation entre les deux segments et qui permet de reconnaître le procédé de reformulation indépendamment de la présence de marqueurs de reformulation.

En outre, le locuteur peut recourir à la reformulation pour paraphraser ses propos antérieurs mais aussi pour les définir, expliquer, préciser, exemplifier, dénommer, conclure, justifier, corriger, paraphraser, etc. La reformulation est prise dans notre travail au sens large et dépasse largement la notion de « paraphrase » [16], [25], [27], [42] instaurant une relation d'équivalence sémantique entre les deux segments, de « glose »[36], [37] désignant un commentaire sur un mot ou de « reprise » [44] fondée sur la répétition lexicale. D'autre part, la reformulation est un exemple d'« élaboration » [28], un procédé plus large qui ne requiert pas obligatoirement le remplacement d'un segment par un autre. Le procédé de reformulation opère sur deux axes : syntagmatique et paradigmatique, alors que ceci n'est pas le cas d' « élaboration », car l'axe paradigmatique n'y est pas toujours présent. Le locuteur peut en effet procéder à l'« élaboration » de son discours sans produire une reformulation.

Suite à la définition du procédé de reformulation présentée ci-dessus, ce procédé est modélisé sous forme d'un jeu d'étiquettes décrit dans les travaux antérieurs [11,12]. Nous nous contentons ici d'en présenter une synthèse.

### 3.2. Annotation multidimensionnelle de reformulations

L'annotation est effectuée sous forme de balises XML et concerne tout d'abord les frontières des segments reformulés et des marqueurs s'ils sont présents. Plusieurs attributs sont distingués :

- relations lexicales : elles montrent le lien lexical (hyperonymie, hyponymie, méronymie, synonymie, antonymie, instance) entre les deux segments reformulés. Travaillant sur le langage oral spontané, les relations sont prises au sens large : elles dépassent les relations nominales et incluent également les relations par association.
- modifications morphologiques : elles indiquent les unités lexicales ayant la même racine. Trois cas sont distingués : dérivation, flexion et composition :
- relations sémantico-pragmatiques : il s'agit des raisons qui poussent le locuteur à utiliser la reformulation : correction linguistique, correction référentielle, définition, dénomination, exemplification, explication, justification, génération, opposition, paraphrase, précision, résultat.

L'annotation prend en compte le contexte et la nature du corpus.

Observons les exemples suivants qui montrent le résultat d'une telle annotation :

1) si Olivet est quand même bien desservi en transports euh <P1>ça reste un peu long</P1> <MR>c'est-à-dire que</MR> <P2 rel_pragm="explic">pour rentrer du lycée Sainte Croix notre fille Caroline peut mettre environ quarante-cinq minutes euh même pas loin d'une heure</P2> (ESLO2_1005)

Le locuteur remplace une proposition (P1) *ca reste un peu long* par une autre (P2) *pour rentrer du lycée Sainte Croix notre fille Caroline peut mettre environ quarante-cinq minutes euh même pas loin d'une heure*. Cette modification est introduite avec l'aide d'un marqueur de reformulation (MR) *c'est-à-dire*. Le locuteur procède à la reformulation pour mieux expliquer (explic) dans le deuxième segment la longueur du trajet qu'effectue sa fille.

La reformulation peut concerner trois segments à la suite :

2) <NP1>un métier</NP1> <NP2 rel_pragm="corr_ling" rel_lex="hypo">mon métier</NP2> <PRES>c'est</PRES> <P3 rel_lex="syno(mon métier/je travaille)" rel_pragm="prec">je travaille à la police ferroviaire</P3>(ESLO2_1053)

Dans cet exemple, le locuteur répond à la question posée sur sa profession. Il présente son métier en utilisant d'abord le groupe nominal indéfini *un métier*, ensuite un groupe nominal défini *mon métier* et enfin une expression plus étendue, l'énoncé *je travaille à la police ferroviaire*. Ce passage du générique vers le spécifique se manifeste par l'ajout de propriétés supplémentaires à la classe présentée par le groupe nominal indéfini, ce qui diminue l'extension de la classe et le rapproche d'une référence plus individualisante. Le locuteur commence par indiquer d'une manière générale qu'il a un métier, ensuite il précise qu'il s'agit de son métier et à la fin il détaille encore plus son métier et présente le lieu de son travail. La première modification est une correction linguistique (corr_ling) où le locuteur remplace le déterminant indéfini générique *un* par le déterminant possessif *mon* qui est plus spécifique. Cette modification est effectuée sans marqueur. La deuxième reformulation concerne le remplacement du groupe nominal (NP2) *mon métier* par une proposition (P3) *je travaille à la police ferroviaire* et elle est introduite par le présentateur (PRES) *c'est*. Par ce remplacement, le locuteur ajoute une précision (prec) sur son domaine d'activité. Dans ce contexte, où le locuteur répond à la question posée sur sa profession, les deux unités *mon métier* et *je travaille* peuvent être considérées comme des synonymes (syno).

L'exemple suivant montre le cas où les deux unités ont des relations au niveau morphologique :

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

Il s'agit d'une paraphrase (para), car les deux segments sont équivalents sémantiquement. Le locuteur modifie une flexion (flex) verbale : il remplace le plus-que-parfait du verbe *choisir* par son passé composé et ajoute le verbe *vivre.*

Enfin, les reformulations peuvent être introduites par des marqueurs :

4) <NP1>la police ferroviaire</NP1> <MR>c'est-à-dire que</MR> <PRES>c'est</PRES> <P2 rel_pragm="def" rel_lex="syno(la police ferroviaire/la police interne de la SNCF)">la police interne de la SNCF</P2>(ESLO2_1053)

Dans cet exemple, le locuteur modifie le groupe nominal (NP1) *la police ferroviaire* par une proposition (P2) *la police interne de la SNCF* pour le définir (def). On observe également une relation de synonymie (syno) entre deux groupes nominaux *la police ferroviaire* et *la police interne de la SNCF*. Deux éléments sont présents entre les segments reformulés : le marqueur de reformulation (MR) *c'est-à-dire que* et le présentateur (PRES) *c'est.*

Ces quelques exemples montrent que les reformulations mettent en jeu différents phénomènes et propriétés, et qu'elles peuvent être introduites par différents éléments. Dans la section qui suit, nous proposons une typologie de ces éléments fondée sur les données fournies par les corpus étudiés.

## 4. Typologie des éléments présents entre les deux segments

Les marqueurs de reformulation sont au centre de nombreux travaux sur la reformulation (voir la section 1) car ils restent un critère majeur et fiable de sa reconnaissance. En effet, l'étude de la reformulation dans les corpus est fondée souvent sur l'analyse des phrases ou des énoncés contenant les marqueurs ciblés et prédéfinis. Cependant, avec ce type d'approche, il existe le risque de passer « à côté » des reformulations introduites par des marqueurs non connus ou des reformulations réalisées sans marqueur.

En se fondant sur le schéma de [24],[35] décrit dans la section 3.1., [45], [46] introduisent la notion de « editing phrases ». Il s'agit des expressions comme *j'en sais pas, si tu veux, tu vois*, des interjections *ah, ben*, etc., des formes noyaux *non, ouais, oui, putain*, ou encore des marqueurs discursifs comme *bon, quoi* qui apparaissent durant la « phase d'édition » entre le « reparandum » et le « repair ». Les auteurs distinguent deux types de « editing phrases » : « backward looking » servant à la correction comme dans *elle a 19 pardon 29 ans* et « forward looking » qui marquent une continuité *elle a vous savez 29 ans*. Ils mentionnent également 23000 cas où « editing phrase » est absente dans le corpus Rhapsodie[3] analysé par [45].

Le travail présenté ici s'intéresse aussi aux éléments pouvant apparaître durant la « phase d'édition ». Son objectif est d'étudier les éléments dont le locuteur se sert pour rompre le flux de parole et introduire la reformulation et de proposer leur typologie. La reformulation est prise au sens large, en tant que procédé de modification d'un segment par un autre pour expliquer, définir, corriger, justifier, préciser, dénommer, exemplifier, paraphrases, conclure, etc. les propos. Notons que ce travail est exploratoire et est fondé sur les observables du corpus oral analysé. Il s'agit d'une démarche inductive.

Quatre cas sont observés dans le corpus. Tout d'abord, comme il a été indiqué précédemment, la reformulation peut être réalisée sans marqueurs et sans aucune « editing phrase ». Dans ce cas, les deux segments se suivent. La reformulation peut être introduite par un marqueur et ces cas ont été largement étudiés dans la littérature linguistique (voir la section 1). Par ailleurs, le locuteur peut interrompre son discours pour le reprendre après une hésitation, pause, interjection, etc. : ce sont les cas de disfluences caractéristiques du discours oral. Enfin, il est observé aussi que les éléments présentateurs peuvent se trouver entre les deux segments reformulés.

### 4.1. Marqueurs de reformulation

Les marqueurs de reformulation regroupent quatre classes :

- marqueurs de reformulation « classiques » (MR) : il s'agit des marqueurs comme *c'est-à-dire, disons, autrement dit, etc.*
- marqueurs d'exemplification (MRE) : ce sont des marqueurs *par exemple, genre, style, etc.* qui introduisent le deuxième segment contenant des exemples, comme dans :

5) et puis euh tout ce qui est émissions euh hm <MRE>style</MRE> Envoyé Spécial euh Zone Interdite non (ESLO2_1053)
6) alors y a eu y a eu des des grands moments euh <MRE>type</MRE> euh visite du parc justement pour euh validation de de choses très tout à fait particulières (ESLO2_1005)
7) tous les produits bah <MRE>genre</MRE> la crème le beurre euh en direct (ESLO2_1033)

- marqueurs de conclusion ou concluants : ce sont des marqueurs de discours *donc, enfin, de toutes façons, etc.* qui sont utilisés pour synthétiser ou finaliser ce qui a été dit précédemment ou pour donner une conséquence de ce qui a été dit :

8) il vous fait faire du rameur euh du vélo euh il nous emmène faire du footing euh <MRCONC>voilà</MRCONC> pendant une heure il vous fait faire du sport (ESLO2_1033)
9) je fais pas mal de tennis <MRCONC>donc</MRCONC> je joue à l'USMO (ESLO2_1005)
10) la gare ferme <MRCONC>de toute façon</MRCONC> y aura plus y a plus personne dans la gare (ESLO2_1053)

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

- marqueurs de correction (MRCOR) : ces marqueurs sont utilisés pour corriger une erreur. Le locuteur considère qu'il s'est trompé et procède à la reformulation pour rectifier l'erreur commise.

11) les scouts accompagnent Jeanne d'Arc dans Orléans donc le choix a été fait depuis très longtemps de l'accompagner le huit mai de pas être présent sur les fêtes le huit le premier mai <MRCOR>pardon</MRCOR> et de ne pas être présent sur les fêtes le huit (ESLO2_1005)

### 4.2. Présentateurs

Le corpus annoté a permis d'observer les cas où les éléments appartenant à la classe de présentateurs (*c'est, voici, voilà, on a, il y a, etc.)* se trouvent entre les deux segments reformulés. Les raisons de leur apparition sont variées : l'introduction de la définition (exemple 4), de la précision (exemple 2) ou de l'explication, comme dans l'exemple qui suit :

12) la campagne normande […] très très belle <PRES>c'est</PRES> vallonné (ESLO2_1033)

### 4.3. Disfluences

Les disfluences sont les éléments qui rompent le flux de la parole, brisent le déroulement syntagmatique [2] et marquent ainsi « des énoncés en cours d'élaboration » [8]. Il peut s'agir des hésitations *euh, hum,* des répétitions de mots ou de segments identiques, des amorces de mots ou de segments, des interjections et même des marqueurs discursifs. Ce groupe d'éléments se rapproche de la notion de « editing phrases » de [45],[46].

Dans le cas de reformulation, le segment en cours d'élaboration est interrompu, ce qui se manifeste par l'apparition d'un ou des éléments disfluents, et il est repris par un autre mais le lien sous-jacent au niveau lexical, morphologique et/ou sémantico-pragmatique est toujours présent ce qui permet d'attester le procédé de reformulation.

Quatre types de disfluences sont distingués suite à l'analyse du corpus :

- amorces (DA) : nous entendons par amorces des mots ou des séquences de mots entamés mais pas terminés par le locuteur.

Certaines amorces de mots sont marquées dans les transcriptions d'ESLO par un tiret, comme c'est le cas de l'exemple (13) où le syntagme *ils doiv-* n'est pas finalisé par le locuteur et se trouve entre deux segments reformulés :

13) alors et vos vos enfants <DA>ils doiv-</DA> <DMD>enfin</DMD> vos filles (ESLO2_1005)

Lorsque le deuxième segment commence par une répétition d'un mot, nous avons considéré ces répétitions comme des amorces :

14) vous vous en occupez toute seule <DI>hein</DI> <DA>c'est c'est</DA> c'est vous qui gérez ça hm (ESLO2_1033)

- hésitations (DH) : les hésitations sont exprimées souvent à l'oral par des mots comme *euh, hum.* Elles peuvent se placer dans n'importe quel endroit de l'énoncé y compris entre les deux segments reformulés :

15) le prix <DA>des</DA> <DH>euh</DH> le tarif (ESLO2_1053)

16) on marche beaucoup en se promenant dans les rues <DH>euh</DH> en ville (ESLO2_1033)

- interjections (DI) : ce sont des mots qui expriment un bruit, un cri, une émotion ou un sentiment. Il s'agit souvent d'unités lexicales monosyllabiques comme *hein* (l'exemple 14) ou *oh, ah* :

17) on n'y va pas <DI>oh</DI> <DMD>non</DMD> <DA>c'est pas</DA> c'est pas tellement notre truc ça (ESLO2_1005)

- marqueurs discursifs (DMD) : cette classe de marqueurs, très présents dans le discours oral, regroupe des unités lexicales très variées comme *bon, bien, quoi, comment dire, mais, enfin* (l'exemple 13) *etc.*, ou encore certaines formes noyaux comme *oui, non* (l'exemple 17). Il s'agit des mots qui changent leur catégorie et leur fonctionnement à l'oral. Selon [8], « Toute forme peut potentiellement devenir une interjection. On assiste alors à une recatégorisation grammaticale […], le phénomène par lequel un mot ayant une classe grammaticale dans le lexique peut, en discours, changer de classe » (p. 350). Contrairement aux amorces, répétitions et hésitations, les marqueurs discursifs ne sont pas toujours considérés comme les disfluences [4, 5]. Néanmoins, dans ce travail, les marqueurs discursifs placés entre les deux segments reformulés sont annotés comme faisant partie des disfluences. Tout comme d'autres disfluences, ces éléments peuvent être modifiés ou supprimés sans que le sens de l'énoncé soit affecté.

Les exemples (13)-17) montrent également qu'un élément disfluent est souvent suivi par un autre élément disfluent. Dans les exemples (13) et (15), les amorces *il doiv-* et *des* sont accompagnées respectivement par un marqueur discursifs *enfin* et une hésitation *euh.* Dans l'exemple (14), une interjection *hein* est suivie par une amorce *c'est c'est,* alors que dans l'exemple (17), on retrouve trois éléments de suite : une interjection *oh,* un marqueur discursif *non* et une amorce *c'est pas.*

## 5. Méthodologie

### 5.1. Prétraitement automatique

Avant de procéder à l'annotation manuelle, les fichiers de transcription sont prétraités et segmentés automatiquement. L'objectif principal consiste à reconstituer les énoncés. Par exemple, il est considéré qu'en cas de chevauchements, les segments chevauchés font partie des énoncés de chacun des interlocuteurs.

### 5.2. Procédure de l'annotation manuelle

Les trois entretiens sont annotés ensuite selon le jeu d'étiquettes décrit dans les sections 3 et 4. Contrairement aux travaux antérieurs, toutes les reformulations avec et sans marqueurs sont annotées dans le corpus.

Les deux entretiens (1005, 1053) sont annotés par deux annotateurs ensemble. Le troisième entretien (1033) est annoté par deux annotateurs séparément. Ceci a permis de calculer l'accord inter-annotateur. Une version consensuelle est obtenue ensuite grâce aux discussions entre les annotateurs.

### 5.3. Accord inter-annotateur

L'accord inter-annotateur est calculé entre les deux versions d'un même corpus annoté par deux annotateurs séparément. Deux mesures sont utilisées : le kappa de Cohen [6] et le kappa de Fleiss [13]. La première est dédiée aux annotations effectuées par deux annotateurs, alors que la deuxième peut aussi être utilisée en cas d'intervention de plusieurs annotateurs.

L'accord sur la présence d'une reformulation dans un énoncé est calculé. Il montre que dans 922 énoncés de l'entretien 1033, 56 énoncés contiennent une reformulation selon l'annotateur 1 et 69 reformulations selon l'annotateur 2. Les annotateurs sont d'accord sur 31 énoncés qui contiennent une reformulation et sur 828 énoncés où la reformulation est absente. Le kappa de Cohen est 0,459 et le kappa de Fleiss 0,46.

On peut comparer cet accord avec celui obtenu suite à l'annotation de reformulations introduites par trois marqueurs *c'est-à-dire, je veux dire, disons*, qui montre un Kappa de Cohen de 0.526. Suivant la grille standard [23], il s'agit d'un accord modéré. Dans le travail actuel, l'accord reste modéré mais est légèrement inférieur. Ceci peut être expliqué par le fait que la tâche actuelle est plus difficile : la reformulation est recherchée dans les contextes avec et sans marqueurs. L'absence de marqueur conduit l'annotateur à rechercher un lien sous-jacent entre les deux segments reformulés. A titre de comparaison, dans le projet Annodis[4] (consacré à l'annotation d'un corpus écrit en structures discursives), le kappa de Kohen est environ de 0,6 [19]. Annoter les transcriptions de l'oral, non ponctuées, avec la présence de beaucoup de disfluences est une tâche plus difficile, ce que montre l'accord inter-annotateur calculé.

## 6. Résultats quantitatives

Le corpus annoté a permis de faire des observations quantitatives concernant les reformulations réalisées avec et sans marqueurs, la distribution de différents éléments présents entre les deux segments reformulés et les liens au niveau lexical, morphologique et sémantico-pragmatique entre les deux segments. Il a permis également de comparer ces résultats avec ceux obtenus dans les travaux précédents portant sur la reformulation introduite à l'aide de trois marqueurs *c'est-à-dire, je veux dire,* et *disons.*

Chaque observation est effectuée deux fois : dans le corpus composé de trois entretiens ensemble pour dégager une tendance générale et dans les trois entretiens séparément pour voir si cette tendance se vérifie dans chaque entretien ou dépend du locuteur et de la situation d'enregistrement.

### 6.1. Observations générales

En moyenne, 14% des énoncés contiennent les reformulations. La répartition des reformulations dans chaque entretien varie de 10% à 20%.

Dans la majorité des cas, un énoncé contient une reformulation qui concerne un segment mais il existe des cas où la modification porte sur plusieurs segments comme dans l'exemple (2).

Un même segment dans un énoncé peut aussi faire partie de plusieurs reformulations. Seize cas de ce type sont observés dans le corpus, comme :

18) non <P1>la gare ferme</P1> <MRCONC>de toute façon</MRCONC> <DA>y aura plus</DA> <P2 rel_pragm= "res">y a plus personne dans la gare</P2>
19) non la gare ferme de toute façon <PRES1>y aura plus</PRES1> <PRES2>y a plus</PRES2> personne dans la gare

Nous considérons qu'un même énoncé, dans les deux exemples ci-dessus, contient deux reformulations distinctes. Dans le premier cas, le locuteur remplace une proposition (P1) *la gare ferme* par une autre (P2) *y a plus personne dans la gare* pour montrer le résultat (res) ou la conséquence de l'événement annoncé dans le premier segment. Deux éléments se trouvent entre les deux segments reformulés : un concluant (MRCONC) *de toute façon* et une amorce (DA) *y aura plus.* Dans le deuxième exemple (19), il s'agit de la correction linguistique portant sur le temps du verbe (le futur du verbe *avoir* est remplacé par le présent), aucun élément n'apparaît entre les deux segments reformulés. Pour permettre une meilleure annotation, les énoncés contenant plusieurs reformulations distinctes sont dupliqués.

### 6.2. Marqueurs et d'autres éléments présents entre les deux segments

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

### *6.2.1. Reformulations avec vs sans marqueurs*

Dans 84% des énoncés contenant la reformulation, celle-ci est réalisée sans marqueurs et seulement dans 16% des cas, la reformulation est introduite à l'aide d'un marqueur. Précisons que les cas sans marqueurs sont de trois types : les deux segments sont séparés par un élément disfluent ou une série de disfluences, les deux segments sont séparés par un présentateur et enfin aucun élément n'apparaît entre les deux segments reformulés.

Si l'on calcule la distribution des reformulations avec et sans marqueur dans chaque entretien séparément, cette tendance est confirmée : dans l'entretien 1033, les marqueurs ne présentent que 5 % des cas de reformulation, dans l'entretien 1053, ils apparaissent dans 14 % des cas et dans l'entretien 1005, ils sont utilisés dans 28 % des cas. En relation avec le Tableau 1, il est possible de remarquer que plus le niveau d'étude est élevé, plus la personne semble utiliser les marqueurs pour introduire les reformulations : le niveau bac du locuteur de l'entretien 1033, le niveau bac + 2 du locuteur de l'entretien 1053, et le niveau bac + 5 du locuteur de l'entretien 1005. Cette observation doit cependant être vérifiée sur un échantillon plus important de locuteurs.

### *6.2.2. Quatre structures observées*

Quatre structures de reformulation sont observées dans le corpus :
- SEG1 Marqueur SEG2
- SEG1 Disfluences SEG2
- SEG1 Présentateur SEG2
- SEG1 SEG2[5]

La première structure est sans doute la plus étudiée dans la littérature linguistique. Il s'agit d'un segment reformulé et introduit à l'aide d'un marqueur paraphrastique ou non paraphrastique. La deuxième structure est la plus fréquente, comme cela montrent les résultats décrits dans cette section : les deux segments sont séparés par une série de disfluences. La troisième structure est peu fréquente mais elle est observée dans le corpus étudié et concerne l'apparition d'un présentateur entre les deux segments. Enfin, la quatrième structure représente les cas où les deux segments n'ont aucun élément linguistique entre eux.

Observons la distribution de ces quatre structures dans le corpus (Fig. 1). Dans la majorité des cas (66 % à 74 %), le locuteur place un élément disfluent entre les deux segments. D'une manière générale, seulement 5 % de présentateurs introduisent une reformulation. Comme il a été constaté dans la section 6.2.1., environs 16 % de reformulations sont réalisées dans le corpus à l'aide d'un marqueur. Enfin, dans 13 %, aucun élément n'apparaît entre les deux segments reformulés.

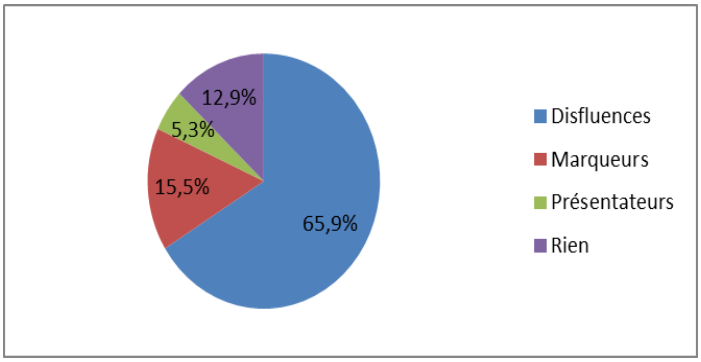


**Fig. 1.** Distribution de quatre types d'éléments présents entre les deux segments reformulés dans le corpus

Si l'on observe cette distribution dans chaque entretien séparément (Fig. 2), la tendance d'une présence majoritaire des disfluences se vérifie. La distribution entre les trois autres types d'éléments varie en fonction des interviews. L'apparition du présentateur entre les deux segments reformulés représente 11 % dans l'entretien 1053 et de 2 % à 3 % dans les entretiens 1005, 1033. Dans l'entretien 1053, les cas où la reformulation ne contient aucun élément entre les deux segments sont absents. Ces cas sont pourtant fréquents dans l'entretien 1033. Les reformulations dans l'entretien 1005 contiennent plus de marqueurs que dans d'autres entretiens. Ces spécificités pourraient dépendre du profil sociologique de locuteurs et des thématiques abordées au cours des interviews.

[5] Nous ne comptons pas ici les cas où le locuteur insère les segments qui n'ont aucun rapport avec les segments reformulés comme une proposition *nos enfants étaient scolarisés là* dans l'exemple (3).

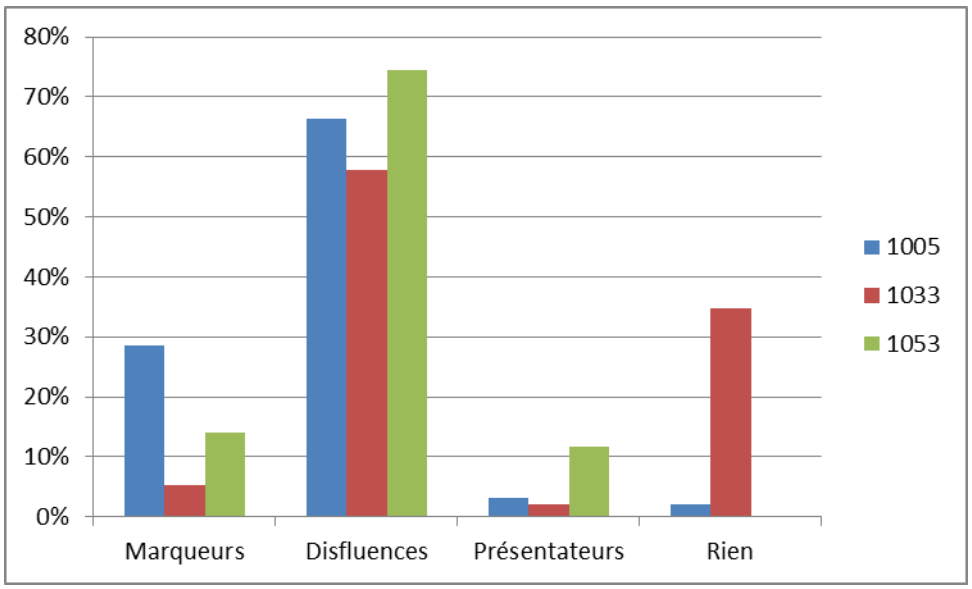


**Fig. 2.** Distribution de quatre types d'éléments présents entre les deux segments reformulés dans chaque entretien

Ces observations quantitatives permettent de conclure que l'hypothèse de départ se vérifie dans le corpus étudié. La reformulation à l'oral est effectivement réalisée dans la majorité des cas sans marqueurs mais le locuteur peut insérer entre les deux segments reformulés un ou plusieurs éléments disfluents. Les cas où les présentateurs sont placés entre les deux segments sont rares. Il faut noter également les exemples fréquents, surtout dans l'entretien 1033, où aucun élément n'apparaît entre les deux segments reformulés.

### *6.2.3. Marqueurs de reformulation*

Quatre types de marqueurs sont distingués dans ce travail : marqueurs de reformulation « classiques » (MR), marqueurs d'exemplification (MRE), marqueurs concluants (MRCONC) et marqueurs de correction (MRCOR). D'une manière générale, la distribution de ces marqueurs dans le corpus montre que 41 % des marqueurs sont les MR, 34 % sont les MRCONC, 18 % appartiennent aux MRE et environs 7 % sont les MRCOR (Fig. 3).

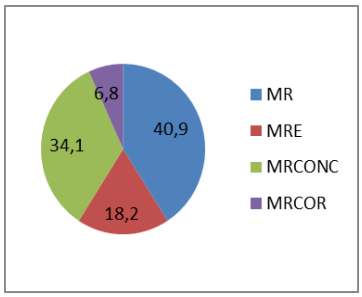


**Fig. 3.** Distribution des marqueurs de reformulation dans le corpus

Si l'on calcule cette distribution dans chaque entretien séparément (Fig. 4), on constate que l'emploi majoritaire des marqueurs de reformulation « classiques » ne se vérifie pas dans l'entretien 1033 où ces marqueurs ainsi que les marqueurs de correction sont absents : les marqueurs d'exemplifications constituent 60 % des cas et les marqueurs concluants 40 % des cas. Dans les deux autres entretiens, la distribution entre les marqueurs suit la tendance générale.

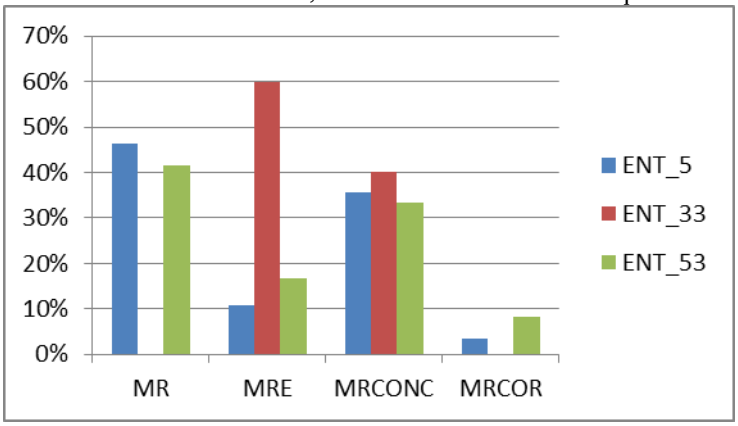


**Fig. 4.** Distribution des marqueurs dans les trois entretiens

### *6.2.4. Disfluences*

Quatre types de disfluences sont distingués dans ce travail : amorces (DA), hésitations (DH), interjections (DI) et marqueurs discursifs (DMD).

Leur distribution dans le corpus (Fig. 5) montre que les marqueurs discursifs et les hésitations sont les plus fréquents (environs 35 %), les amorces apparaissent entre les deux segments dans 20 % des cas, les interjections sont les moins fréquentes (7,6 %).

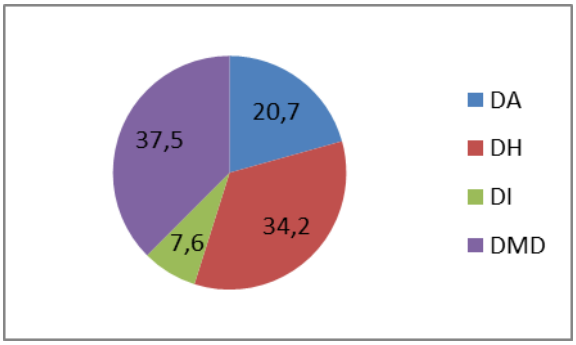


**Fig. 5.** Distribution des disfluences présentes entre les deux segments reformulés dans le corpus

Dans 23 % des cas, un élément disfluent est suivi par un autre. Cette tendance semble se vérifier dans tous les entretiens analysés : 24,6% dans l'entretien 1005, 21,8% dans l'entretien 1033 et 23,4% dans l'entretien 1053.

Après avoir extrait les reformulations contenant les disfluences, il est constaté que l'ordre dans lequel les disfluences apparaissent semble être aléatoire. Quelques tendances les plus fréquentes sont :

- DA DH ou DA DMD : les amorces sont le plus souvent suivies par les hésitations et les marqueurs discursifs ;
- DH DMD : les hésitations sont suivies par les marqueurs discursifs ;
- DMD DA ou DMD DH ou DMD DMD : les marqueurs discursifs peuvent être suivis par les amorces, les hésitations ou d'autres marqueurs discursifs.

Les plus longues séries de disfluences attestées dans le corpus contiennent jusqu'à sept éléments :

- DI DMD DI DA DMD DA DMD
- DMD DMD PRES DA DMD DH DMD
- DA MR DMD DH MR DH DA

Il est également observé que seuls les marqueurs discursifs peuvent se suivre, comme dans les exemples suivants :

20) votre travail <DMD>justement</DMD> <DMD>en fait</DMD> cette occupation principale (1005)
21) mais dans les restaurants <DMD>oui</DMD> <DMD>mais</DMD> on va aussi au restaurant (1053)

Dans ce travail, les marqueurs discursifs est une classe extrêmement hétérogène qui regroupe les marqueurs de discours comme *donc, enfin, en même temps, mais, etc.*, les formes noyaux comme *oui, non, merde, etc.*, mais aussi les mots qui changent leur catégorie et leur fonctionnement à l'oral [8] comme *bon, quoi, voyez-vous, etc.* Leur combinaison est donc moins surprenante.

### *6.2.5. Présentateurs*

D'une manière générale, les présentateurs sont peu présents entre les deux segments reformulés.

L'analyse de chaque interview montre que l'apparition d'un présentateur est plus fréquente (11,6 %) dans l'entretien 1053 et ne représente que 2 % ou 3 % dans les entretiens 1005 et 1033.

## 6.3. Liens formalisés entre les deux segments

L'annotation effectuée porte également sur les segments reformulés et sur les relations que les unités lexicales des deux segments peuvent entretenir entre elles.

Trois niveaux sont annotés :

- le premier (rel_lex) concerne le lexique et indique si les unités ont les liens d'hyperonymie, d'hyponymie, de synonymie, d'antonymie, de méronymie ou d'instance ;
- le deuxième (modif_morph) montre si les unités ont la même racine et marque trois possibilités : dérivation, flexion et composition ;
- le troisième niveau (rel_pragm) signale la raison de la reformulation effectuée et s'inscrit dans le domaine de la sémantique et de la pragmatique. Plusieurs raisons sont distinguées : correction linguistique, correction référentiel, définition, dénomination, exemplification, explication, généralisation, justification, opposition, paraphrase, précision, résultat.

Observons la distribution de ces liens dans le corpus. Presque 66 % de toutes les reformulations contiennent des relations au niveau lexical, dans 12 % des cas, les unités des deux segments sont liées morphologiquement et 22 % de reformulations restent sans liens morphologiques et lexicaux attestés. Cette tendance se vérifie dans chaque entretien pris séparément (Fig. 6).

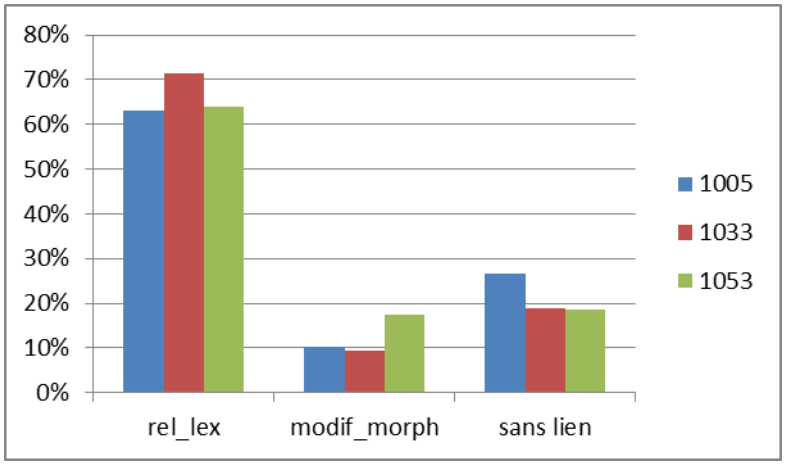


**Fig. 6.** Distribution des liens lexicaux et morphologiques dans chaque entretien

Les observations quantitatives portent, en premier lieu, sur le corpus annoté dans le cadre de ce travail où toutes les reformulations avec et sans marqueurs sont prises en compte. En deuxième lieu, ces résultats sont comparés avec ceux obtenu dans les travaux précédents portant sur les reformulations introduites par trois marqueurs *c'est-à-dire*, *je veux dire* et *disons.*

### *6.3.1. Liens lexicaux*

Les liens lexicaux sont présents dans environs 66 % des reformulations. Notons qu'on peut avoir dans une reformulation plusieurs liens lexicaux, c'est-à-dire que plusieurs unités peuvent être liées lexicalement. Il est constaté que dans 47 % des cas, il s'agit des liens de synonymie (syno) et dans 27% des cas c'est la méronymie (mero). Les autres relations semblent être moins présentes (Fig. 7).

L'analyse de chaque entretien pris individuellement (Fig. 8) montre que, malgré la différence entre les locuteurs, c'est la relation de synonymie (syno) qui domine parmi les relations lexicales annotées. La méronymie (mero) occupe la deuxième place, sa fréquence relative varie de 12 % à 23 % selon les entretiens (elle est plus fréquente dans l'entretien 1005). Les autres relations sont moins fréquentes et varient dans les trois entretiens. Les instances sont peu présentes dans les trois corpus : leur fréquence est au-dessous de 10%. Les liens d'antonymie sont absents dans l'entretien 1005 et très peu présents dans les deux autres entretiens. Les hyponymes apparaissent dans 20% des cas dans l'entretien 1053 et les hyperonymes dans 15% des cas dans l'entretien 1033.

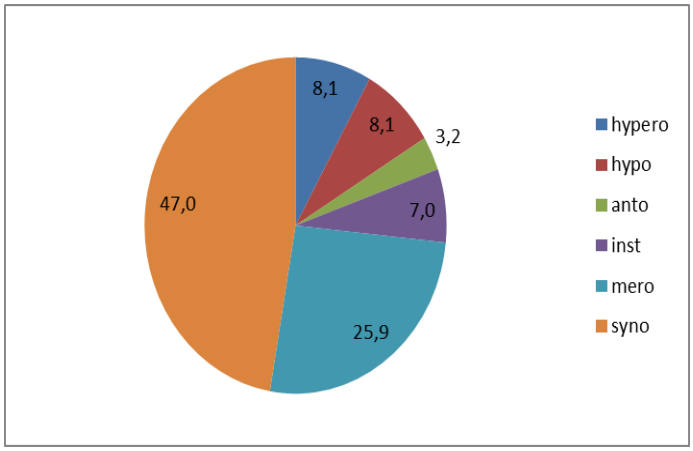


**Fig. 7.** Distributions de liens lexicaux dans le corpus

### *6.3.2. Liens morphologiques*

Les unités lexicales appartenant aux deux segments reformulés peuvent entretenir une ou plusieurs relations morphologiques. Dans la majorité des cas, où il y a le lien au niveau morphologique, cela concerne la flexion (flex) 62 %. La dérivation (deriv) n'apparaît que dans environ 35 % des cas. La composition (compos) est presque absente, un seul cas est annoté dans ce corpus. Ce fait peut être expliqué par le contenu abordé au cours des entretiens qui n'a pas favorisé l'apparition de mots composés. La tendance est la même si l'on observe ce type de liens dans les trois entretiens séparément (Fig. 9).

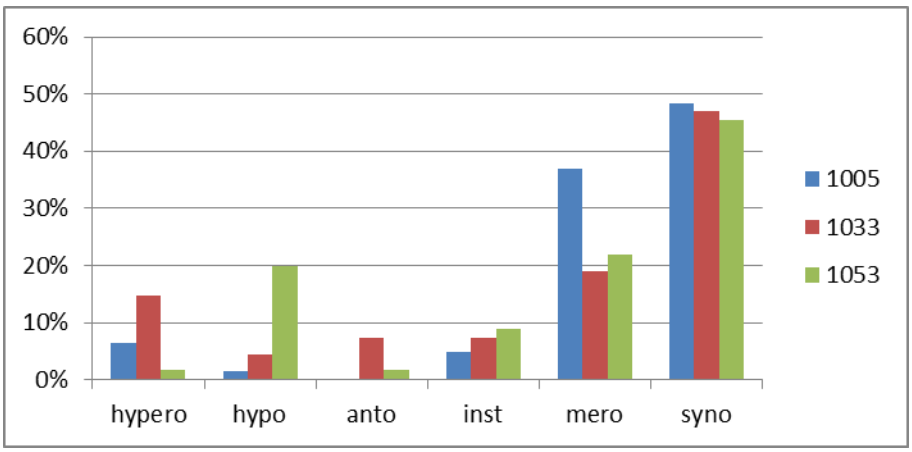


**Fig. 8.** Distributions de liens lexicaux dans trois entretiens

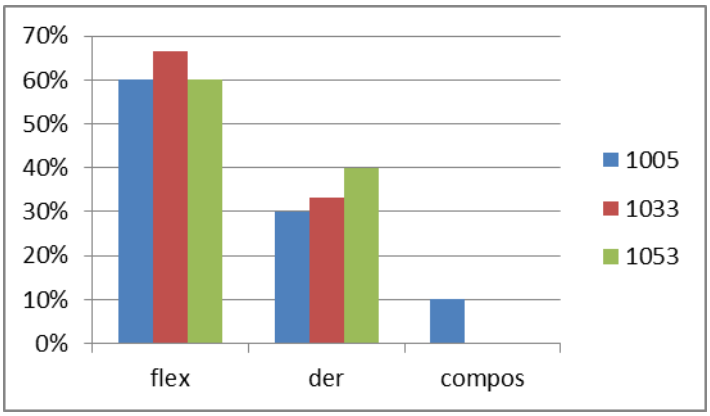


**Fig. 9.** Distributions de liens morphologiques dans trois entretiens

### *6.3.3. Relations sémantico-pragmatiques*

La raison la plus fréquente de la reformulation, dans le corpus observé, est la paraphrase (para) avec 33 %, ce qui semble être pertinent par rapport au processus étudié. La précision (prec) apparaît dans presque 17 % des cas et la correction linguistique (corr_ling) dans 12 % des cas. L'exemplification (exempl), le résultat (res) et l'explication (explic) sont présents dans environs 8 % des reformulations annotées. Les autres raisons de reformulation (corrections référentielles (corr_ref), dénominations (denom), définitions (def), générations (gener), justifications (justif), oppositions (oppos) sont peu fréquentes dans le corpus, ce qui peut être expliqué par le contenu abordé pendant les entretiens et par le profil de chaque locuteur (Fig. 10).

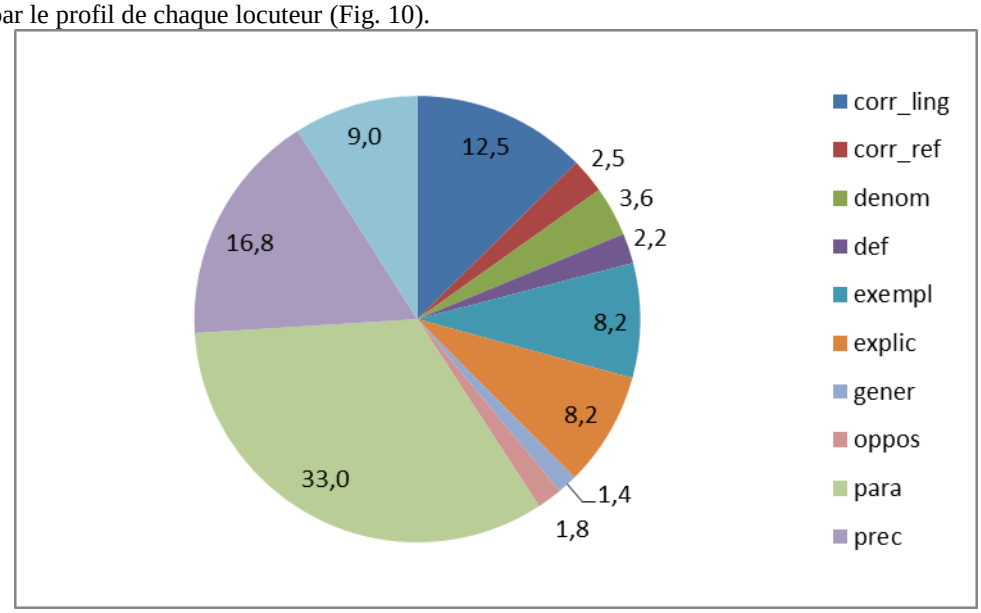


**Fig. 10.** Distributions de relations sémantico-pragmatiques dans le corpus

Comme il a été observé dans le corpus global, la relation la plus fréquente dans les trois entretiens pris séparément est la paraphrase (para) suivie de précision (prec) et de correction linguistique (corr_ling) (Fig. 11). D'autres types de

relations sémantico-pragmatiques sont repartis d'une manière plus aléatoire selon les entretiens et leur apparition est sans doute liée avec les thématiques abordées.

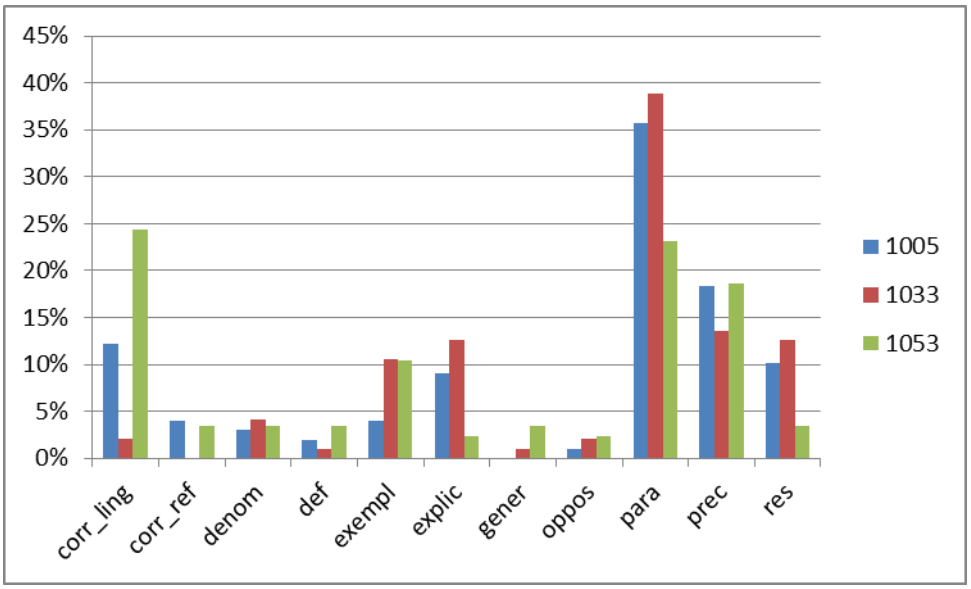


**Fig. 11.** Distribution de relations sémantico-pragmatiques dans trois corpus

### *6.3.4. Comparaison avec les résultats des travaux précédents*

Le travail présenté dans le cadre de cet article fait suite aux recherches effectuées ces dernières années sur les reformulations dans les entretiens transcrits de l'oral. Les travaux précédents ont été limités à l'analyse des reformulations introduites par trois marqueurs dérivés du verbe *dire* : *c'est-à-dire, je veux dire, disons.* Le travail actuel est plus large car il concerne toutes les reformulations produites dans le corpus, qu'elles soient réalisées avec ou sans marqueur. Dans les deux cas, la méthodologie reste identique : elle est fondée sur l'annotation manuelle selon le schéma multidimensionnel préétabli. La différence par rapport aux autres travaux concerne essentiellement deux points :
- une annotation plus large est effectuée, ce qui permet d'observer toutes les reformulations et leur fonctionnement ;
- la prise en compte, dans le schéma d'annotation, de tous les éléments se trouvant entre les deux segments reformulés.

La comparaison des résultats porte sur l'annotation des segments et de leurs attributs. Plus précisément, il s'agit de relations lexicales, morphologiques et sémantico-pragmatiques entre les deux segments reformulés.

Plusieurs tendances se voient confirmées :
- le lien au niveau lexical est le plus présent : il représente 55 % dans le travail précédent et 66 % dans le travail actuel ;
- c'est la synonymie qui domine parmi les relations lexicales annotées : elle représente 35 % dans les travaux antérieurs et 47 % dans les travaux actuels ;
- la méronymie occupe la deuxième place du point de vue de sa fréquence : 32 % dans les travaux antérieurs et 26 % dans le travail actuel. Notons que la méronymie inclut la relation partie vs tout mais aussi les liens associatifs.
- les modifications morphologiques représentent environ 10 % ;

Les relations sémantico-pragmatiques sont plus difficiles à comparer car elles dépendent beaucoup de la nature des entretiens et des thématiques abordées.

## 7. Conclusion

Le travail présenté concerne le procédé de reformulation. Il est pris au sens large en tant que procédé de modification d'un segment par un autre avec et sans marqueurs. La reformulation garde toujours un lien sous-jacent, « une partie invariante » [26] entre les deux segments reformulés. Ce lien se manifeste à différents niveaux linguistiques : lexical, morphologique et sémantico-pragmatique. Le segment source est interrompu et suivi par un segment reformulé. Cette interruption peut être « remplie » par un élément lexical : un marqueur qui introduit la reformulation (marqueur de reformulation « classique », marqueur de correction, marqueur d'exemplification, marqueur de conclusion), un élément disfluent (amorce, interjection, hésitation, marqueur discursif) ou un présentateur. Les raisons de la reformulation peuvent être nombreuses : correction, paraphrase, explication, définition, justification, conclusion, précision, dénomination, exemplification, etc. A partir de cette définition du procédé de reformulation, celui-ci est modélisé par un jeu d'étiquettes correspondantes selon lesquelles le corpus est annoté manuellement.

L'analyse effectuée grâce à l'annotation montre que la reformulation est réalisée à l'oral sans marqueurs dans 80% des cas. Les éléments, qui apparaissent fréquemment entre les deux segments reformulés, sont de nature disfluente. Ces éléments correspondent à une caractéristique inhérente du discours oral : leur présence lors du procédé de reformulation n'est donc pas étonnante.

Lorsqu'un marqueur de reformulation est absent au cours de la reformulation, l'annotateur humain se repère par rapport aux indices (« aux traces ») du lien formel entre les deux segments. Sans ces indices, il est impossible d'attester la reformulation. Par conséquent, l'annotation manuelle n'est pas homogène, ce que confirme l'accord inter-annotateur sur le jugement de la présence de la reformulation dans l'énoncé qui montre un Kappa de 0,46.

Le travail présenté est un travail exploratoire. Annoter manuellement les données conformément à la définition de la reformulation ci-dessus est une tâche de longue haleine, c'est pourquoi le corpus étudié ne comprend que trois entretiens (32722 mots). Cependant, il montre que la modélisation et les conventions d'annotation définies sont pertinentes grâce à la comparaison des résultats quantitatifs avec ceux obtenus lors de nos travaux précédents réalisés sur un corpus beaucoup plus important (3 762 720 mots).

L'objectif du travail est d'étudier le procédé de reformulation à partir du corpus. Le point de départ est un corpus brut qui est annoté selon la modélisation préalablement effectuée. La modélisation est fondée aussi sur l'observation manuelle du corpus. Cette démarche inductive permet d'appréhender la reformulation sous un autre angle ainsi que d'observer et de quantifier certaines de ses caractéristiques. Plusieurs apports du travail peuvent être cités :

- d'une manière générale, en moyenne, 15% des énoncés produits à l'oral contiennent la reformulation et dans 84% de ces énoncés, la reformulation est réalisée sans marqueurs
- si la reformulation n'est pas introduite par un marqueur (marqueurs « classique » de reformulation, marqueur de correction, marqueur d'exemplification, marqueur de conclusion), trois cas sont observés : les deux segments se suivent, les éléments disfluents et enfin les présentateurs apparaissent entre les deux segments reformulés ;
- le lien entre les deux segments se manifeste surtout au niveau lexical (66%) ;
- les relations lexicales les plus fréquentes sont la synonymie et la méronymie, alors que la raison la plus fréquente de la reformulation est la paraphrase. Les liens au niveau morphologique sont peu présents ;
- plus le niveau d'étude du locuteur est élevé, plus la personne semble utiliser les marqueurs pour introduire les reformulations. Cette observation doit cependant être vérifiée sur un échantillon plus important de locuteurs.

En perspective, la détection automatique des reformulations avec et sans marqueur est envisagée. Pour cela, il est nécessaire d'annoter manuellement plus de données. Ce travail est en cours. Le schéma d'annotation sera appliqué aux autres corpus dont le corpus de forums de discussion en ligne. Le corpus ainsi annoté nous servira de corpus de référence pour l'apprentissage automatique et pour une étude du fonctionnement des reformulations dans le discours.

## Références


1. 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)
2. C. Blanche-Benveniste, M. Bilger, C. Rouget, K. Van Den Eynde, Le français parlé. *Études grammaticales*, Paris, CNRS Éditions, (1990).
3. C. Blanche-Benveniste, C. Jeanjean, *Le français parlé, transcription et édition*, Paris, Didier érudition, (1987).
4. B. Boula de Mareüil, G. Adda, M. Adda-Decker, C. Barras, B. Habert, P. Paroubek, Une étude quantitative des marqueurs discursifs, disfluences et chevauchements de parole dans des interviews politiques, *TIPA*, Travaux interdisciplinaires sur la parole et le langage, (2013) [URL : http://tipa.revues.org/830]
5. R. Bove, C. Chardenon, J. Véronis, Prise en compte des disfluences dans un système d'analyse syntaxique automatique de l'oral. In P.Mertens, A.Dister, P.Watrin Eds. *TALN2006*, *Verbum ex machina*, p.103-111, (2006).
6. J. Cohen, A coefficient of agreement for nominal scales. Educational and Psychological Measurement, **20(1)**, p. 37-46, (1960).
7. J. Delahaie, Dis, dis donc, disons : du verbe au(x) marqueur(s) discursif(s), *Langue française*, **2, 186**, p. 31-48, (2015).
8. A. Dister, *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL*. Thèse de doctorat, Université de Louvain, (2007).
9. C. Dutrey, S. Rosset, M. Adda-Decker, C. Clavel, I. Vasilescu, Disfluences dans la parole spontanée conversationnelle : détection automatique utilisant des indices lexicaux et acoustiques, *XXXe Journées d'Étude sur la Parole (JEP'14)*, Le Mans, France, p. 366-373, (2014).
10. G. Dostie, *Pragmaticalisation et marqueurs discursifs. Analyse sémantique et traitement lexicographique*, Bruxelles, Duculot, (2004)
11. I. Eshkol-Taravella, N. Grabar, Reformulation à l'oral et dans le forum Web, *CMLF2016*, Tours, France, SHS Web of Conferences **27**, (2016).
12. I. Eshkol-Taravella, N. Grabar, Reformulation en tant que procédé multidimensionnel. In La reformulation : à la recherche d'une frontière, Houda Landolsi, Coco Norén, Maria Svensson (dir.), Studia Romanica, **87**, Acta Universitatis Upsaliensis, Suède, (à paraître).
13. J. L. Fleiss, Measuring nominal scale agreement among many raters, *Psychological Bulletin*, **76(5)**, p. 378-382, (1971).
14. K. Fløttum, *La reformulation introduite par c'est-à-dire*, Stavanger, Høgscolesenteret i Rogaland, (1994)
15. K. Flottum, *Dire et redire. La reformulation introduite par "c'est-à-dire"*, Thèse de doctorat, Hogskolen i Stavanger, Stavanger, (1995).
16. C. Fuchs, *Paraphrase et énonciation*. Paris : Orphys, (1994).

17. E. Gülich, T. Kotschi, Les marqueurs de la reformulation paraphrastique, *Cahiers de linguistique française*, **5**, p. 305-351, (1983).
18. C. Hagège, *L'homme de paroles. Contribution linguistique aux sciences humaines*. Paris, Fayard, (1985).
19. L-M. Ho-Dac, L'expérience ANNODIS, *Consortium « Corpus écrits », annotations de haut niveaux*. [https://groupes.renater.fr/wiki/corpus-ecrits/_media/public/presentation_mai_ho_dac.pdf] (consulté le 10/3/2018)
20. Y. Hwang, Eh bien, alors, enfin et disons en français parlé contemporain. *L'Information Grammaticale*, **57**, 46-48, (1993).
21. T. Jeanneret, Pourquoi reformuler et comment le faire? *Tranel*, **18**, p. 67-81, (1992).
22. E. Khatchatourian, Les marqueurs de reformulation formés à partir du verbe dire, in La reformulation : marqueurs linguistiques, stratégies énonciatives, Marie-Claude Le Bot, Martine Schuwer, Richard Élisabeth (dir.), Rennes, PUR, p. 19-33, (2008).
23. J. Landis, G. Koch, The measurement of observer agreement for categorical data, *Biometrics*, **33**, p. 159-174, (1977).
24. W. J. M. Levelt, Monitoring and Self-repair in Speech, *Cognition*, **14**, p. 41-104, (1983)
25. R. Martin, *Inférence, antonymie et paraphrase*. Paris : Klincksieck, (1976).
26. C. Martinot, *La reformulation dans des productions orales de définitions et explications. (Enfants de maternelle)*, Thèse de doctorat, Université Paris VIII, (1994).
27. I. Melčuk, Paraphrase et lexique dans la théorie linguistique sens-texte in lexique et paraphrase. *Lexique*, **6**, p. 13-54, (1988).
28. L. Prévot, L. Vieu, N.Asher. Une formalisation plus précise pour une annotation moins confuse : la relation d'élaboration d'entité. *Journal of French Language Studies*, **19(2),** p. 207-289, (2009)
29. M. Petit, *Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs*, Thèse de doctorat, Université d'Orléans, (2009).
30. C. Rossari, Projet pour une typologie des opérations de reformulation, *Cahiers de linguistique française*, **11**, p. 345-359, (1990).
31. C. Rossari, De l'exploitation de quelques connecteurs reformulatifs dans la gestion des articulations discursives, *Pratiques*, **75**, p. 111-124, (1992).
32. C. Rossari, *Les opérations de reformulation. Analyse du processus et des marques dans une perspective contrastive français-italien*, Berne, Peter Lang, (1994).
33. C. Rossari, Cela dit : un marqueur de prise de conscience, *Langues et langage*, **12**, p. 87-101, (2005).
34. E. Saunier, Disons : un impératif de dire? Remarques sur les propriétés du marqueur et son comportement dans les reformulations, *L'Information Grammaticale*, **132**, p. 25-34, (2012).
35. E. E. Shriberg, *Preliminaries to a Theory of Speech Disfluencies*. Thèse de doctorat, Berkeley University of California, (1994).
36. A. Steuckardt, Les marqueurs formés sur dire, in Les marqueurs de glose, Agnès Steuckardt, Aïno Niklas-Salminen (dir.), Presses de l'Université de Provence, p. 51-65, (2005).
37. A. Steuckardt, A. Niklas-Salminen, *Le mot et sa glose*. Publications de l'Université de Provence, (2003).
38. 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).
39. H. Vassiliadou, *Les connecteurs c'est-à-dire (que) en français et ðilaði en grec. Analyse syntaxique et sémantico-pragmatique*, Thèse de doctorat, Université de Strasbourg, (2004).
40. H. Vassiliadou, Quand les voies de la reformulation se croisent pour mieux se séparer : à savoir, autrement dit, c'est-à-dire, en d'autres termes, in La reformulation : marqueurs linguistiques, stratégies énonciatives, Marie-Claude Le Bot, Martine Schuwer, Richard Élisabeth (dir.), Rennes, PUR, p. 35-50, (2008).
41. H. Vassiliadou, La formation de c'est-à-dire (que) et de ses correspondants dans les langues romanes : quelques remarques, in Actes del *26é Congrés de Lingüística i Filologia Romàniques*, Emili Casanova Herrero, Cesáreo Calvo Rigual (dir.), Berlin, W. de Gruyter, Tome III, p. 453-464, (2013).
42. L. Vezin, Les paraphrases : étude sémantique, leur rôle dans l'apprentissage. *L'année psychologique*, **76(1)**, p. 177-197, (1976).
43. J. Vigneron-Bosbach, Genre en français, like en anglais : des marqueurs de reformulation ? Actes du colloque *La reformulation : à la recherche d'une frontière*, 8-9 juin 2017, Université d'Uppsala, (à paraître).
44. R. Vion, Reprise et mode d'implication énonciative. *La linguistique*, **42**, p. 11-28, (2006).
45. Y. Tian, C. Beyssade, Y. Mathieu, J. Ginzburg, Editing Phrases, *SemDial 2015 - goDIAL*, The 19th Workshop on the Semantics and Pragmatics of Dialogue, Gothenburg, Sweden, (2015).
46. Y. Tian, T. Maruyama, J. Ginzburg, Filled Pauses and Self Addressed Questions, *Journal of Psycholinguistic Research*, **46:4**, (2017).