Repérage et analyse de la reformulation paraphrastique dans les corpus oraux

Iris Eshkol-Taravella, Natalia Grabar

Repérage et analyse de la reformulation paraphrastique dans les corpus oraux

Iris Eshkol-Taravella[1] Natalia Grabar[2]
(1) CNRS UMR 7270 LLL, Université d'Orléans, 45100 Orléans, France
iris.eshkol@univ-orleans.fr
(2) CNRS UMR 8163 STL, Université Lille 3, 59653 Villeneuve d'Ascq, France
natalia.grabar@univ-lille3.fr

Résumé. Notre travail porte sur la détection automatique de la reformulation paraphrastique dans les corpus oraux. L'approche proposée est une approche syntagmatique qui tient compte des marqueurs de reformulation paraphrastique et des spécificités de l'oral. L'annotation manuelle effectuée par deux annotateurs permet d'obtenir une description fine et multidimensionnelle des données de référence. Une méthode automatique est proposée afin de décider si les tours de parole comportent ou ne comportent pas des reformulations paraphrastiques. Les résultats obtenus montrent jusqu'à 66,4 % de précision. L'analyse de l'annotation manuelle indique qu'il existe peu de segments paraphrastiques avec des modifications morphologiques (flexion, dérivation ou composition) ou de segments qui montrent l'équivalence syntaxique.

Abstract. Our work addresses the automatic detection of paraphrastic rephrasing in spoken corpus. The proposed approach is syntagmatic. It is based on paraphrastic rephrasing markers and the specificities of the spoken language. Manual annotation performed by two annotators provides fine-grained and multi-dimensional description of the reference data. Automatic method is proposed in order to decide whether sentences contain or not the paraphrases. The obtained results show up to 66.4% precision. The analysis of the manual annotations indicates that there are few cases in which paraphrastic segments show morphological modifications (inflection, derivation or compounding) or syntactic equivalence.

Mots-clés : Paraphrase, reformulation, corpus oral, marqueurs de reformulation paraphrastique.

Keywords: Paraphrase, reformulation, spoken corpus, markers of paraphrastic rephrasing.

1 Introduction

La langue naturelle propose des moyens variés pour exprimer une même idée de différentes manières (exemples (1) à (8)). La paraphrase est liée à d'autres notions, comme la synonymie (Hamon *et al.*, 1998), la variation (Daille *et al.*, 1996; Grabar & Zweigenbaum, 2000), ou la reformulation paraphrastique (Roulet, 1987; Rossari, 1990; Bouamor *et al.*, 2012). Nous nous arrêterons sur les notions de paraphrase et de reformulation dans notre travail. L'acception de ces phénomènes varie selon les courants linguistiques : si chez les générativistes la notion de paraphrase n'est pas acceptable car toute modification de forme (*e.g.* nombre, mode, diathèse) implique un changement sémantique notable (Chomsky, 1975), cette notion connaît actuellement une acception très large notamment grâce aux travaux de TAL. Le critère commun, sur lequel tout le monde semble s'accorder aujourd'hui en parlant de la paraphrase, est qu'il existe entre les expressions linguistiques en relation de paraphrase une équivalence sémantique, qui peut prendre toutefois des formes différentes.

Parmi les fonctionnalités de la paraphrase se trouvent des phénomènes intra-locuteur et inter-locuteur. La paraphrase peut aider à une facilitation de la compréhension du message par l'autre et contribuer au bon fonctionnement de la communication (François, 1990; Bouamor *et al.*, 2012). Elle peut aussi empêcher la clarté de la communication (Boucheron, 2000) (*e.g.* la littérature scientifique spécialisée perçue par les non spécialistes). La paraphrase contribue à la beauté de la langue car elle évite les répétitions et redondances. Pour que la paraphrase soit détectée, la similitude entre l'information évoquée et l'information répétée doit être reconnue par l'interlocuteur. Du point de vue de TAL, la paraphrase crée une réelle difficulté : qu'il s'agisse de la reconnaissance ou de la production, la paraphrase couvre un nombre important de catégories, qui mettent en oeuvre des mécanismes linguistiques fort variés. Nous nous arrêterons sur les notions de paraphrase et de reformulation paraphrastique. Nous présentons d'abord les travaux en linguistique qui se sont attachés à décrire la paraphrase et la reformulation paraphrastique (section 1.1), et ensuite les travaux de TAL qui visent à détecter la paraphrase de manière automatique (section 1.2). Nous précisons ensuite les objectifs poursuivis dans notre travail (section 1.3).

1.1 Description linguistique de la paraphrase et de la reformulation paraphrastique

1.1.1 Typologies linguistiques de la paraphrase

Il existe différentes manières de traiter et de décrire la paraphrase. Par exemple, elle peut faire référence à la situation d'énonciation et avoir une valeur contextuelle. Ainsi, deux types de paraphrases sont distingués (Culioli, 1976; Martin, 1976; Vezin, 1976; Fuchs, 1994) :

– La paraphrase situationnelle est une "définition en discours" qui détermine le sens d'un énoncé (ou d'une partie d'énoncé) par rapport au contexte énonciatif. Dans l'exemple (1), *depuis une centaine d'années* et *depuis 1900* reçoivent ainsi une valeur contextuelle.

(1) *Depuis une centaine d'années, les températures du globe tendent à augmenter.*
Depuis 1900, les températures du globe tendent à augmenter.

– La paraphrase linguistique, parfois appelée une "définition en langue", est liée aux classifications linguistiques existantes. Nous en donnons quelques exemples dans ce qui suit.

Si l'on essaie de faire une typologie des transformations linguistiques subies par les entités, les niveaux suivants de la langue peuvent être distingués (Melčuk, 1988; Vila *et al.*, 2011; Bhagat & Hovy, 2013) :

– la paraphrase morphologique, qui a pour condition le changement morphologique (flexion, nominalisation, adjectivation, composition, etc.), comme dans l'exemple (2) :

(2) *Pierre a enlevé son manteau.* ; *Pierre enlève son manteau.*

– la paraphrase lexicale, qui vise le changement au niveau lexical (synonymes, antonymes, mots plus génériques ou spécifiques), comme dans l'exemple (3) :

(3) *Pierre a enlevé son manteau.* ; *Pierre a enlevé sa veste. Pierre a enlevé son vêtement.*

– la paraphrase sémantique, qui couvre en général des segments allant au-delà du lexique (exemple (4)) :

(4) *Pierre a enlevé son manteau.* ; *Pierre s'est déshabillé.*

– la paraphrase syntaxique, qui réorganise la phrase (déplacement de composants, diathèse...), comme dans l'exemple (5) ;

(5) *En entrant dans la bibliothèque, Pierre a enlevé son manteau.*
Pierre a enlevé son manteau en entrant dans la bibliothèque.

– la paraphrase mixte (*e.g.*, lexico-syntaxique, lexico-sémantique, etc.), qui concerne les modifications opérant simultanément à plusieurs niveaux (Bouamor *et al.*, 2012).

La notion de paraphrase peut aussi être décrite en fonction de la taille d'entités couvertes par la paraphrase (Flottum, 1995; Fujita, 2010; Bouamor, 2012) :

– la paraphrase lexicale, se situant au niveau d'un mot (exemples (6)) :

(6) {*bouquin*, *livre*}, {*bâtiment*, *maison*}

– la paraphrase sous-phrastique, avec laquelle la synonymie se fait au niveau des syntagmes, des tournures de phrases ou des fragments de textes (exemples (7)) :

(7) *Il a envie de*, *Il aimerait bien* ; *X ne doute pas de Y*, *X est sûr de Y*

– la paraphrase phrastique, où plusieurs segments sont en relation de paraphrase tandis que la sémantique de la phrase est préservée (exemples (8)) :

(8) *Comment vont vos enfants ?* ; *Comment se portent vos gamins ?* ; *Les gosses vont bien ?*

Les classifications existantes de la paraphrase focalisent souvent sur un aspect donné, décrit avec plus ou moins de finesse : 67 fonctions lexicales pour le paraphrasage (Melčuk, 1988), 25 catégories de paraphrases (Bhagat & Hovy, 2013). À notre connaissance, la seule classification multidimensionnelle est celle de (Milicevic, 2007), avec les dimensions suivantes :

– type de connaissances mis en jeu pour la production de paraphrases,
– modifications de sens impliquées,
– types de moyens d'expression utilisés (cette dimension est proche d'autres classifications existantes),

– exactitude du lien paraphrastique,
– mode de production.

Notons que la notion de paraphrase peut également couvrir d'autres dimensions :
– registre de langue (les équivalences inter-discours (Elhadad & Sutaria, 2007; Deléger & Zweigenbaum, 2008), les niveaux soutenu ou parlé de la langue, etc.),
– la langue (les équivalences inter-langues ou les traductions (Fuchs, 1982; Milicevic, 2007)).

Nous parlons de la paraphrase au sens large, tout en la réservant aux expressions d'une seule langue. Nous considérons ainsi que la paraphrase peut être utilisée non seulement pour reformuler mais aussi pour décrire, exemplifier, préciser ou expliquer une idée exprimée auparavant par un locuteur.

1.1.2 Reformulation paraphrastique à l'oral

La reformulation est propre autant à la langue soutenue, comme celle des articles scientifiques, qu'à la langue parlée, bien qu'elle montre des différences dans les deux cas (Flottum, 1995; Rossari, 1992). Ainsi, dans l'écrit, c'est le produit fini qui se présente au destinataire (Hagège, 1985), alors que l'oral l'exhibe dans les étapes de son élaboration. Il est en effet commun de trouver dans la langue orale des traces de sa propre production (*e.g.* hésitations, faux-départs, formes diverses de reprises) à la manière de brouillons qui précèdent la version finale des écrits (Blanche-Benveniste *et al.*, 1991). De manière générale, il est considéré que la reformulation est une activité du locuteur qui s'appuie sur un segment déjà produit dans son propre discours ou dans celui de son interlocuteur, avec ou sans l'emploi d'un marqueur, afin d'en modifier certains aspects (lexical, syntaxique, sémantique, pragmatique) tout en gardant un invariant permettant de reconnaître l'opération ainsi mise en place (Gülich & Kotschi, 1987; Kanaan, 2011). Pour ces différentes raisons, tout acte de reformulation dans le discours oral n'introduit pas toujours une paraphrase (Rossari, 1990). De ce point de vue, on distingue deux catégories de marqueurs : les marqueurs de reformulation non-paraphrastique (*e.g. en somme, en tout cas, de toute façon, enfin*, etc.) et les marqueurs de reformulation paraphrastique (ou MRP), comme *c'est-à-dire, autrement dit, je m'explique, ça veut dire, en d'autres termes* (Rossari, 1990, 1993). Les critères, qui permettent de détecter la reformulation paraphrastique sont (Gulich & Kotschi, 1983; Rossari, 1993) :
– trois critères phonétiques : répétition du contour intonatif de la phrase ; réduction de la vitesse de débit ; et articulation remarquablement nette des deux syllabes qui terminent l'énoncé doublon ;
– parallélisme syntaxique entre l'entité source et l'entité paraphrasée ;
– présence d'un MRP, bien qu'il soit possible d'avoir une relation de paraphrase sans marqueur. Parmi les MRP, les auteurs distinguent ceux qui ont pour tâche principale d'établir une relation paraphrastique (*e.g. c'est-à-dire, autrement dit*) et ceux qui ne montrent ce rôle que dans des contextes précis.

Les MRP fournissent un marquage formel de liens paraphrastiques entre deux segments : segment source et segment cible (ou paraphrasé). Les propriétés sémantiques des MRP permettent d'instaurer une relation de paraphrase même entre les segments qui n'entretiennent aucune équivalence sémantique visible par ailleurs (Rossari, 1993).

1.2 Description et détection de la paraphrase en TAL

Deux états de l'art récents sur les méthodes pour la détection automatique de la paraphrase (Madnani & Dorr, 2010; Androutsopoulos & Malakasiotis, 2010) montrent l'intérêt important réservé à ces méthodes et ressources dans le domaine du TAL. Les approches proposées pour la détection automatique de la paraphrase dépendent du type de corpus exploités et reposent généralement sur les propriétés paradigmatiques des mots (leur capacité de se substituer mutuellement) :

1. *Corpus monolingues.* En corpus monolingues, la similarité des chaînes d'édition (Malakasiotis & Androutsopoulos, 2007) et les méthodes distributionnelles sont le plus souvent utilisées. Dans ce dernier cas, si les unités linguistiques (mots, syntagmes, etc) ont des vecteurs similaires, elles sont alors de bons candidats pour la paraphrase (Lin & Pantel, 2001; Pasça & Dienes, 2005) ;
2. *Corpus monolingues parallèles.* Lorsqu'un texte dans une langue est traduit plus d'une fois dans une autre langue, les traductions de ce texte permettent de constituer un corpus monolingue parallèle. Un des plus utilisés est constitué des traductions en anglais de *20 000 lieux sous la mer* de Jules Verne. L'exploitation de tels corpus est notamment possible grâce aux méthodes d'alignement de mots (Och & Ney, 2000). Différentes méthodes ont été proposées pour l'exploitation de tels corpus (Barzilay & McKeown, 2001; Ibrahim *et al.*, 2003; Quirk *et al.*, 2004) ;
3. *Corpus monolingues comparables.* Les corpus monolingues comparables contiennent typiquement des textes produits indépendamment sur un même événement, comme par exemple les articles de presse qui couvrent l'actualité. La cohérence thématique de ces textes d'un côté et les méthodes distributionnelles ou bien l'alignement de phrases

comparables de l'autre côté permettent d'induire les relations de paraphrase entre les segments de texte (Shinyama *et al.*, 2002; Sekine, 2005; Shen *et al.*, 2006) ;

4. *Corpus bilingues parallèles.* Les corpus bilingues parallèles, qui contiennent typiquement la traduction d'un texte dans une autre langue, peuvent aussi être utilisés pour la détection de la paraphrase. Dans cette situation, les traductions multiples d'une expression ou d'un mot peuvent correspondre aux paraphrases (Bannard & Callison-Burch, 2005; Madnani *et al.*, 2008; Callison-Burch *et al.*, 2008; Kok & Brockett, 2010).

1.3 Objectifs

L'objectif que nous poursuivons dans notre travail concerne la détection de reformulations paraphrastiques. L'originalité du travail proposé consiste en points suivants :
– Le corpus de travail est un corpus oral, très peu exploité jusqu'ici pour détecter les paraphrases (Bouamor *et al.*, 2012) ;
– La méthode choisie pour détecter les reformulations paraphrastiques dans un corpus monolingue est une approche syntagmatique et non distributionnelle réservé à ce type de matériel (Madnani & Dorr, 2010) ;
– Une annotation multidimensionnelle de la paraphrase est proposée. Elle permet de créer les données de référence ;
– La distinction automatique entre les reformulations paraphrastiques et non-paraphrastiques est effectuée.
Nous décrivons d'abord les données exploitées (section 2) et les méthodes proposées (section 3). Nous présentons et discutons ensuite les résultats dans la section 4, et terminons avec des perspectives de recherches (section 5).

2 Données linguistiques

2.1 Corpus

Nous travaillons avec les corpus ESLO (Enquêtes Sociolinguistiques à Orléans) (Eshkol-Taravella *et al.*, 2012) : *ESLO1* et *ESLO2*. *ESLO1*, la première enquête sociolinguistique à Orléans, a été réalisée en 1968-1971 par des professeurs de français de l'University of Essex, Language Centre, Colchester (Royaume-Uni), en collaboration avec des membres du B.E.L.C. (Bureau pour l'étude de l'enseignement de la langue et de la civilisation françaises de Paris). Le corpus *ESLO1*, constitué à Orléans mais archivé ensuite de manière fragmentaire ailleurs, est revenu dans les années 1990 au LLL (Laboratoire Ligérien de Linguistique). Le laboratoire a mis au format standard ce corpus d'enquêtes sociolinguistiques comprenant 300 heures de parole (4 500 000 mots environ) incluant une gamme d'enregistrements variés. En prenant en compte l'expérience d'*ESLO1* et l'évolution des cadres théoriques et méthodologiques de la constitution et de l'exploitation de grands corpus oraux à visée variationniste, une nouvelle enquête *ESLO2* a été entamée en 2008. À terme, *ESLO2* comprendra plus de 350 heures d'enregistrements afin de former avec *ESLO1* un corpus de plus de 700 heures et d'atteindre les dix millions de mots. Les corpus *ESLO1* et *ESLO2* sont accessibles en ligne (*http ://eslo.tge-adonis.fr/*).

Pour avoir des données comparables dans les deux corpus, nous avons sélectionné 260 entretiens d'*ESLO1* totalisant 2 349 829 occurrences de mots et 308 entretiens d'*ESLO2* totalisant 1 412 891 occurrences de mots. Les fichiers transcrits d'ESLO respectent deux principes : l'adoption de l'orthographe standard et le non-recours à la ponctuation de l'écrit. 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 locuteurs. Nous avons utilisé les versions C de transcription. Ces fichiers de transcription n'ont pas été corrigés et ont été pris comme tels.

2.2 Marqueurs de reformulation paraphrastique (MRP)

Nous exploitons trois MRP : *c'est-à-dire*, *je veux dire* et *disons*. Le point commun entre eux est qu'ils sont formés à partir du même verbe *dire*. Le marqueur *c'est-à-dire* est le plus lexicalisé des trois et semble être le plus étudié. Les propriétés qu'on lui reconnaît sont les suivantes (Gulich & Kotschi, 1983; Hölker, 1988; Beeching, 2007) :
– il est utilisé dans les monologues et dans les dialogues, à l'écrit et à l'oral ;
– les éléments liés ne peuvent pas être échangés car ils n'ont pas d'égalité entre eux ;
– *c'est-à-dire* peut commuter avec *à savoir, en réalité, autrement dit, en d'autres termes* et *donc* ;
– ce marqueur peut instaurer la relation de paraphrase entre des énoncés non équivalents sémantiquement ;
– les trois fonctions prototypiques de ce marqueur sont : corrigeante, reformulante et argumentative, mais il peut aussi marquer la conclusion (il est alors substituable par *donc*), la justification ou l'hésitation.

Les caractéristiques du marqueur *disons* ont été montrées dans (Hwang, 1993) :
– il est assez proche de *je dirais, je veux dire* ;
– il existe une analogie entre *eh bien* et *disons* du point de vue énonciatif, car ils marquent tous les deux une rupture : en mettant fin au niveau coénonciatif précédent, le locuteur signale l'ouverture d'un plan énonciatif différent et égocentré ;
– il existe une analogie entre *disons* et *enfin* en tant que moyens de rectification.
(Saunier, 2012) distingue six pôles pour décrire les différents sens de *disons* : *à peu près*, *en bref*, *ou plutôt*, *plus précisément*, *oui et non* et *ni la chèvre ni le chou*. Pour (Petit, 2009), il est impossible de supprimer *disons* de l'énoncé parce que le deuxième segment exprime une nuance sémantique différente. Souvent, il est lié à la recherche d'un terme plus adéquat. En ce qui concerne le marqueur *je veux dire*, (Teston-Bonnard, 2008) identifie plusieurs *je veux dire* et démontre ainsi que deux des statuts syntaxiques repérés (verbes recteurs faibles et parenthèses) se paraphrasent par *autrement dit, c'est-à-dire, je reprends*. Comme le montrent les recherches citées ci-dessus, ces trois éléments peuvent avoir des fonctions différentes. Dans notre travail, nous nous intéressons à eux en tant que marqueurs de reformulation paraphrastique.

3 Méthodologie pour la détection de paraphrases

Les tours de parole avec les trois MRP sont extraits des deux corpus et pré-traités (section 3.1). La méthode proposée est fondée sur le traitement manuel (section 3.2) et automatique (section 3.3) de ces corpus. Nous effectuons également une analyse et évaluation des résultats (section 3.4).

3.1 Préparation des corpus

Une des plus grandes difficultés du traitement automatique des transcriptions de l'oral est l'absence de marques formelles de segmentation. Si à l'écrit les signes de ponctuation remplissent bien cette fonction, à l'oral ce sont les éléments paralinguistiques qui marquent le début et la fin de l'énoncé : la pause, l'intonation, etc. Pour résoudre en partie ce problème, nous avons utilisé comme segmenteur un tour de parole marqué dans la transcription par un changement de locuteur. La difficulté s'est posée alors pour les cas de chevauchement où les deux locuteurs parlent en même temps. Dans ces situations, 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 continue. Les corpus sont ensuite traités avec le chunker SEM (Eshkol *et al.*, 2014) adapté à la langue orale. SEM détecte les chunks minimaux, comme présenté dans l'exemple (9) (même exemple qu'en (12)).

(9) *(est/V)VN (-ce/CLS)NP (que/CS)CONJ (vous/CLS)NP (remarquez/V)VN (une/DET différence/NC sensible/ADJ)NP (entre/P vos/DET différents/ADJ clients/NC dans/P leur/DET façon/NC de/P choisir/VINF)PP (la/DET viande/NC) NP (dans/P ce/PRO)PP (qu'/PROREL ils/CLS)NP (achètent/V)VN (et/CC)CONJ (caetera/V)VN (,/CLS)NP (indépendamment/V disons/VPP)VN (de/P leurs/NC)PP (oui/I)IntP (origines/NC)NP (de/P classe/NC ,/ADJ)PP*

3.2 Annotation manuelle des reformulations paraphrastiques

L'annotation manuelle a pour objectif de distinguer entre les reformulations paraphrastiques et les reformulations non paraphrastiques, mais aussi de proposer une annotation plus fine. Pour les contextes de reformulation paraphrastique, l'annotation est focalisée sur les deux segments (des mots ou des segments plus grands) mis en relation de reformulation autour d'un MRP, mais aussi sur la relation établie par le MRP de manière générale. L'annotation est effectuée en suivant plusieurs dimensions, dont certaines sont inspirées par les classifications existantes (section 1.1.1) :

1. *catégorie syntaxique* : les segments mis en relation sont annotés par leur catégorie syntaxique (N, A, V, Prep...) ou leur type de constituant syntaxique (NP, VP, AP, PP). Cela permet également de voir s'il existe une équivalence syntaxique entre les deux segments. Les segments mis en relation ne sont pas définis sur une base syntaxique (*e.g.* des chunks), mais sur le critère sémantique de paraphrase qui se trouve à la base de cette relation ;
2. chaque relation est annotée avec plusieurs arguments évocateurs des classifications existantes de la paraphrase (section 1.1.1). S'il existe plusieurs éléments paraphrasés, ils sont tous annotés de cette manière :
   – *rel-lex* : type de la relation lexicale entre deux éléments paraphrasés : hyperonyme, synonyme, antonyme, instance, méronyme ;
   – *modif-lex* : type de la modification lexicale : remplacement, suppression, ajout ;

  - *modif-morph* : type de la modification morphologique : flexion, dérivation, composition ;
  - *modif-synt* : type de la modification syntaxique : passif/actif... ;
3. *rel-pragm* : type de la relation pragmatique. Cette relation est liée aux fonctionnalités de la paraphrase ou de la reformulation. Notre typologie est inspirée des typologies proposées dans la littérature (Gülich & Kotschi, 1987; Kanaan, 2011). Parmi ces fonctionnalités, nous distinguons : définition, explication, exemplification, précision, dénomination, résultat, correction linguistique, correction référentielle, équivalence.

Les exemples (10) et (11) montrent le résultat de cette annotation (les annotations sont en bleu, les références des fichiers *ESLO* entre les crochets). Nous pouvons ainsi voir que les segments {*Saint Jean de la Ruelle*, *Orléans*} (exemple (10)) et {*démocratiser l'enseignement*, *permettre à tout le monde de rentrer en faculté*} (exemple (11)) sont en relation de paraphrase, tout en mettant en jeu des mécanismes linguistiques et pragmatiques différents.

(10) *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="cor-ref">Orléans</NP2> parce que c' est ça fait partie d' Orléans* [ESLO1_ENT_149_C]

(11) *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>* [ESLO1_ENT_121_C]

3.3 Détection automatique de reformulations paraphrastiques

Le traitement automatique principal consiste à décider si, autour d'un MRP, il existe une relation de reformulation paraphrastique ou non. Pour ceci, nous mettons en place plusieurs filtres qui sont communs à tous les MRP :
- Si le MRP est placé en début ou en fin de TdP, alors il est considéré que ce TdP ne comporte pas de paraphrase ;
- Si le MRP est entouré des marqueurs discursifs (*donc, enfin, quoi...*), euh d'hésitation, interjections (*ben hm ouais*), amorces (*s-*), etc. répétés, il est considéré que le MRP fait partie des disfluences de l'oral (une accumulation d'éléments qui brisent le déroulement syntagmatique (Blanche-Benveniste *et al.*, 1991)) et n'introduit pas la paraphrase ;
- Si le MRP apparaît dans un contexte lexical spécifique (emploi de *nous* devant *disons*), ou si le MRP apparaît dans des suites argumentatives (*e.g. par contre, mais, en revanche, au contraire*), ce TdP ne comporte pas de paraphrase ;
- Si le MRP apparaît à l'intérieur d'une locution, comme *indépendamment de* ou *plus ou moins grossiers* (exemples (12) et (13) de la section 4.2), alors il est considéré que ce contexte ne comporte pas de paraphrase. Ce test est effectué sur les sorties du chunker (exemple (9)). Pour vérifier que la locution existe dans la langue, nous interrogeons un moteur de recherche généraliste et analysons les fréquences attestées sur la Toile. À notre avis, l'usage de la Toile fournit des informations plus complètes que celles que l'on peut trouver dans des corpus de référence. Chaque segment est testé de trois manières (exemple (9)) : avec un (*(caetera)VN (indépendamment)VN (de leurs)PP*), deux (*(et)CONJ (caetera)VN (indépendamment)VN (de leurs)PP (origines)NP*) ou trois chunks *(achètent)VN (caetera)VN (indépendamment)VN (de leurs)PP (origines)NP (de classe)PP* à droite et à gauche du MRP, excepté les disfluences et la ponctuation. La taille maximale du segment est empiriquement limitée à sept mots. La fréquence moyenne de ces segments doit être inférieure au seuil entre 10 et 6 000 pour que ce segment soit considéré comme une paraphrase. Dans le cas de fréquences plus élevées que le seuil, ce test indique que la locution existe dans la langue et qu'il s'agit en effet de disfluence.

3.4 Analyse et évaluation

L'évaluation est effectuée de deux manières :
- pour l'annotation manuelle, nous calculons l'accord inter-annotateur pour les jugements sur l'existence de la relation de paraphrase. Comme deux annotateurs ont participé à cette tâche, nous appliquons le kappa de Cohen (Cohen, 1960). Le protocole d'annotation a été mis en place et ajusté sur une partie du corpus *ESLO1*, tandis que l'évaluation et l'accord inter-annotateur sont calculés sur d'autres TdP du corpus *ESLO1* et sur la partie *entretiens* du corpus *ESLO2* ;
- pour la détection automatique de relations paraphrastiques, elle est évaluée par rapport aux annotations manuelles. Nous calculons la précision des résultats.

L'analyse des résultats porte sur une étude de la fréquence des relations et de la répartition des différents attributs de manière générale et en fonction des MRP. L'accent principal est mis sur l'existence de relations paraphrastiques, mais aussi sur l'équivalence syntaxique entre les segments mis en relation et sur l'existence de modifications morphologiques (flexion, dérivation ou composition), qui peuvent donner des indications formelles de paraphrasage.

4 Résultats et Discussion

4.1 Préparation des corpus

	ESLO1	*ESLO2*
nombre de fichiers de transcription	260	308
taille de corpus (occ de mots)	2 349 829	1 412 891
taille moyenne des fichiers de transcription	9 037,80	4 587,31
nombre de TdP	166 602	70 707
taille moyenne des TdP	14,10	19,98
c'est-à-dire	1 849	594
je veux dire	285	291
disons	1 068	183
total de TdP avec les MRP	3 202	1 068
taille des TdP avec les MRP (minimale)	1	1
taille des TdP avec les MRP (maximale)	6 382	1 050
taille des TdP avec les MRP (moyenne)	62,88	86,34

TABLE 1 – Différentes informations sur les données : taille des corpus, nombre et taille moyenne des TdP, nombre de TdP avec les trois MRP étudiés, taille des TdP avec les MRP.

Le tableau 1 présente la taille des corpus en nombre de mots et indique différentes informations sur les extractions effectuées : nombre et taille moyenne des TdP, nombre de TdP avec les trois MRP étudiés, taille des TdP avec les MRP. Ces chiffres montrent que la taille moyenne des TdP est entre 14 et 19 mots : de nombreux TdP sont en effet de taille minimale, avec un ou deux mots seulement. Le marqueur *c'est-à-dire* est toujours le plus fréquent, avec plus de la moitié des TdP contenant un MRP. Rappelons que *c'est-à-dire* établit principalement une relation paraphrastique même entre les TdP ayant une équivalence sémantique faible (Gulich & Kotschi, 1983). Quant à *disons*, il est très fréquent dans le corpus *ESLO1* mais beaucoup moins dans *ESLO2*. Il est possible que ce soit dû à l'évolution diachronique de la langue : d'autres mots ont pu reprendre cette fonction discursive. En ce qui concerne la taille moyenne des TdP avec les MRP, elle est assez élevée (62,88 dans *ESLO1* et 86,34 dans *ESLO2*). Ces TdP peuvent en effet comporter des paraphrases et montrer la genèse et la précision des idées (Hagège, 1985; Blanche-Benveniste *et al.*, 1991) de la part des locuteurs. Nous pouvons aussi observer que la taille maximale des TdP peut aller jusqu'à 1 050 dans *ESLO2* et 6 382 dans *ESLO1*.

4.2 Annotation manuelle des paraphrases

	ESLO1					*ESLO2*				
	A1		*A2*		*accord*	*A1*		*A2*		*accord*
	oui (%)	*non (%)*	*oui (%)*	*non (%)*		*oui (%)*	*non (%)*	*oui (%)*	*non (%)*	
c'est-à-dire	96 (33)	193 (67)	66 (23)	223 (77)	249	74 (37)	124 (63)	65 (32)	137 (68)	162
je veux dire	16 (25)	49 (75)	8 (12)	57 (88)	57	47 (34)	91 (66)	27 (20)	110 (80)	107
disons	18 (15)	104 (85)	8 (7)	115 (93)	106	10 (18)	45 (82)	9 (16)	46 (84)	46
total de TdP	130 (27)	346 (73)	82 (17)	395 (83)	412	131 (33)	260 (67)	101 (26)	293 (74)	315

TABLE 2 – Jugements sur la relation de paraphrase dans les contextes avec les MRP : pour deux annotateurs et leur accord.

Tout MRP confondu, 476 TdP du corpus *ESLO1* et 394 TdP du corpus *ESLO2* de la partie *entretiens* sont analysés (54 et 30 entretiens respectivement). Cette annotation permet de proposer un premier jeu de données de référence et un guide d'annotation. Le tableau 2 montre les résultats des annotations par les deux annotateurs impliqués. Les annotateurs reconnaissent entre 17 et 27 % de contextes paraphrastiques dans le corpus *ESLO1* et entre 26 et 33 % de contextes paraphrastiques dans le corpus *ESLO2*. L'annotateur *A1* a tendance à accepter plus de contextes comme paraphrastiques, ce qui montre l'aspect subjectif de ce type d'annotation. La perception de paraphrase varie en effet d'un annotateur à l'autre. L'accord entre les annotateurs est de 0,617 pour *ESLO1*, ce qui correspond à un accord substantiel, et de 0,526

pour *ESLO2*, ce qui correspond à un accord modéré (Landis & Koch, 1977). Il s'agit d'un niveau d'accord assez important, surtout lorsque l'on travaille avec des données linguistiques qui peuvent introduire la subjectivité dans leur perception. Comme annoncé dans la littérature, ces MRP peuvent apparaître dans des emplois paraphrastiques et non-paraphrastiques. Dans les exemples (12) et (13), qui ne contiennent pas de relations de paraphrase, les MRP peuvent ainsi être associés aux marqueurs discursifs faisant partie des disfluences.

(12) *est-ce que vous remarquez une différence sensible entre vos différents clients dans leur façon de choisir la viande dans ce qu'ils achètent et caetera indépendamment <MRP>disons</MRP> de leurs oui origines de classe* [ESLO1_ENT_001_C]

(13) *mais il y a des termes qui sont plus ou moins euh euh <MRP>disons</MRP> euh grossiers qui sont employés plus ou- plus facilement euh par certaines couches de la société en fonction des fréquentations des uns ou des autres quoi* [ESLO1_ENT_003_C]

	ESLO1				*ESLO2*			
	A1		*A2*		*A1*		*A2*	
	oui	*non*	*oui*	*non*	*oui*	*non*	*oui*	*non*
c'est-à-dire (%)	33	67	22	78	37	63	32	68
je veux dire (%)	25	75	12	88	34	66	20	80
disons (%)	15	85	7	93	18	82	6	94

TABLE 3 – Pourcentage des constructions paraphrastiques autour des MRP.

Le tableau 3 indique le pourcentage des constructions paraphrastiques et non-paraphrastiques autour des MRP. Pour les deux annotateurs, *c'est-à-dire* est le plus grammaticalisé de ce point de vue car il introduit le plus de relations de paraphrase, tandis que *disons* est le moins grammaticalisé. *je veux dire*, qui a la position intermédiaire, est plus proche de *c'est-à-dire*. Concernant *disons*, nous pensons qu'il est le plus ambigu des trois MRP car d'une part il peut signifier le verbe *dire* et donc, en quelque sorte, signifier l'emploi contraire à la paraphrase où il marque le début d'une nouvelle idée : le locuteur introduit alors quelque chose de nouveau comme dans *disons que...* D'autre part il peut être employé en tant que marqueur discursif, associé aux *expressions stéréotypées* fonctionnant en tant que adverbe, conjonction, interjection, etc. (Gulich & Kotschi, 1983), ou disfluences (exemples (12) et (13)).

Dans la majorité des cas (plus de 70 %), il n'existe pas d'équivalence syntaxique entre les éléments en relation de paraphrase (comme dans les exemples (14) et et (15)). Notons que cet aspect dépend du choix des annotateurs. Ainsi, dans l'exemple (14), au lieu de la proposition *les gens me semblent plus plus affables* il est aussi possible de choisir le syntagme adjectival *plus affables* ou l'adjectif *affables*. Un autre aspect intéressant de l'annotation est lié aux modifications morphologiques observables entre les segments en relation de paraphrase. Nous pouvons ainsi voir que de telles modifications sont annotées pour environ dix relations paraphrastiques par corpus, tout MRP confondus. En voilà quelques exemples : {*achat*, *achète*}, {*connais*, *connu*}, {*pourrait*, *pouviez*}, {*client*, *clientèle*}, {*manoeuvres*, *manuel*}, {*aller*, *vais*}. Cela indique qu'il existe très peu d'accroches formelles pour détecter les segments en relation de paraphrase dans ce type de constructions. Les modifications syntaxiques, comme par exemple le changement de la voix active en voix passive, ne sont quasiment pas présentes, avec seulement un exemple au sein des 54 et 30 entretiens annotés respectivement dans *ESLO1* et *ESLO2*. En ce qui concerne les modifications lexicales, nous observons surtout le remplacement d'un sous-segment par un autre. Comme cela a été noté dans la littérature (Gulich & Kotschi, 1983; Rossari, 1993), dans plusieurs cas, nous rencontrons effectivement des segments, qui n'ont aucun lien sémantique évident, mais, grâce à un MRP et à la relation de paraphrase établie, ce lien peut apparaître (exemples en (16) ou (17)).

(14) *je préfère mieux le le nord de la France franchement le département du Nord et le département du Pas-de-Calais où <P1>les gens me semblent plus plus affables</P1> <MRP>disons</MRP> euh <PP2 rel-lex="syn" rel-pragm="explic">avec qui j' ai on a plus facilement des des rapports agréables</PP2>* [ESLO1_ENT_003_C]

(15) *y a le euh le le plus grand goup- groupe et puis euh ce qu'on appelle <NP1>toujours les mêmes</NP1> <MRP>c'est-à-dire</MRP> euh <P2 rel-lex="syn" rel-pragm="equiv">tous ceux qu'on connait</P2> quoi* [ESLO2_ENT_1004_C]

(16) *des conférences y en a assez souvent sur France culture enfin <MRP>disons</MRP> des causeries* [ESLO1-

_ENT_121_C]

Parmi les relations lexicales, les plus fréquentes sont les relations de synonymie et d'hypéronymie, suivies par les instances dans le cas des entités nommées, et l'équivalence et le résultat. En fonction des relations pragmatiques assignées, nous pouvons distinguer trois fonctions effectuées par les MRP (dans l'ordre décroissant de fréquence dans les corpus) :
– la possibilité d'ajouter une nouvelle information, notée par des relations pragmatiques d'explication, de précision, d'exemplification et de définition. Cette fonction peut être rapprochée des fonctions corrigeante, reformulante et argumentative notées dans la littérature (Gulich & Kotschi, 1983; Hölker, 1988). Dans tous ces cas, il s'agit de rendre l'énoncé plus riche et clair, comme dans les exemples (14) et (11) ;
– l'établissement de la relation d'équivalence : redire la même chose, mais avec d'autres moyens linguistiques. Ce qui est intéressant avec l'équivalence, mais aussi avec la définition, est que, contrairement à ce qui a été observé dans la littérature (Gulich & Kotschi, 1983; Petit, 2009), il est possible de supprimer le MRP et d'échanger les segments de place sans que cela modifie la sémantique de l'énoncé (exemples (15) et (16)) ;
– avec la relation *résultat*, nous pouvons observer le phénomène inverse à l'explication : le deuxième segment peut être réduit par rapport au premier et en proposer une synthèse (exemple (17)).

(17) *voilà <P1>le côté très betonné voilà c'est pas ils ont pas développé les les logements étudiants suffisamment ils ont pas développé l'off- l'offre culturelle euh en même temps</P1> donc enfin <MRP>je veux dire</MRP> voilà <P2 rel-pragm="res">c'est mort</P2>* [ESLO2_ENT_1012_C]

4.3 Détection automatique de reformulations paraphrastiques

	ESLO1		*ESLO2*	
	A1	A2	A1	A2
filtres	40,5	40,5	37,7	37,8
filtres + fréquences (>6000)	25,8	25,9	18,7	18,9
filtres + fréquences prioritaires (>6000)	63,0	63,0	66,4	66,3

TABLE 4 – Évaluation de la détection automatique des reformulations paraphrastiques (précision).

Les résultats sur la détection automatique des reformulations paraphrastiques se trouvent dans le tableau 4. L'évaluation est effectuée en termes de précision. Notons que les résultats sont cohérents entre les deux annotateurs dans les deux corpus. Quant au jugement sur la présence de reformulations paraphrastiques, nous observons que les filtres prenant en charge les disfluences orales permettent d'atteindre jusqu'à 40 % de précision. L'ajout de filtres supplémentaires (fréquences sur la Toile) aux filtres de disfluences détériore les résultats : nous pouvons perdre jusqu'à 18 %. Par contre, lorsque les fréquences sont prioritaires sur les filtres de disfluences, les résultats sont améliorés et peuvent atteindre jusqu'à 66,4 % de précision. Dans cette configuration, si les fréquences satisfont nos critères, nous considérons qu'un TdP peut contenir une paraphrase même si le MRP se trouve dans un contexte de disfluences de l'oral. Les résultats s'améliorent avec l'augmentation du seuil. La seuil maximum testé est 6 000, l'augmentation est observée jusqu'au seuil 4 500.

Nos recherches montrent également qu'il existe des schémas plus compliqués que ceux décrits dans les travaux antérieurs (Rossari, 1990) et ceux pris actuellement en charge par notre système :
– la relation de paraphrase peut aussi se construire sur plus d'un TdP : lorsque le locuteur est interrompu sans chevauchements et lorsqu'il continue son discours plus loin. Actuellement, nous ne traitons pas ce type de situations car la détection de paraphrases est effectuée au sein d'un même TdP ;
– nous avons distingué deux situations selon que les segments en relation de la paraphrase sont contigus ou distants par rapport au MRP. Ainsi, dans l'exemple (18), les segments sont distants. Comme noté, ils peuvent être séparés du MRP par des disfluences ou bien par d'autres segments à contenu. Nous avons essayé de prendre en compte la possibilité d'avoir des disfluences intercalées, ce qui améliore en effet les résultats globaux ;
– le MRP peut aussi se trouver non pas entre mais après les deux segments en relation de paraphrase (exemple en (19)).

(18) *<PP1>jusqu'à seize ans</PP1> oui oui bien bon <MRP>c'est-à-dire</MRP> euh <PP2 rel-lex="syn" rel-pragm="cor-ref">dans le primaire privé</PP2> n'est-ce pas* [ESLO1_ENT_010_C]

(19) *et elle travaille euh dans des instituts d' enfants euh plus ou moins adaptés enfin plutôt inadaptés <MRP>disons </MRP> et en plus de ça elle a une clientèle personnelle ici* [ESLO1_ENT_003_C]

5 Conclusion et perspectives

Nous avons proposé un travail sur la détection de reformulations paraphrastiques dans des corpus monolingues oraux du français (corpus *ESLO1* et *ESLO2*). Une des originalités du travail consiste à prendre en compte les spécificités de l'oral, que ce soit grâce à la reconstitution des tours de parole, à la considération des disfluences dans le contexte ou à l'utilisation d'outils de TAL adaptés à l'oral (Eshkol *et al.*, 2014). Un autre aspect original est que nous abordons la détection de reformulations paraphrastiques avec une approche syntagmatique, alors que la plupart des approches existantes pour la détection de paraphrases exploitent les propriétés paradigmatiques de la langue. Notre travail repose sur une utilisation combinée de l'annotation manuelle et d'un traitement automatique des données. L'annotation permet de produire un premier jeu de données de référence et de faire plusieurs observations sur les relations de paraphrase, en particulier grâce à une annotation multidimensionnelle et fine. Elle a servi aussi à l'évaluation de la méthode automatique et peut être exploitée dans l'apprentissage automatique. L'accord inter-annotateur obtenu est de 0,617 et 0,526 pour les corpus *ESLO1* et *ESLO2* respectivement. Rappelons aussi que nous adoptons une notion large de la paraphrase (Melčuk, 1988; Bhagat & Hovy, 2013), qui peut clarifier, expliciter, développer ou résumer l'idée exprimée auparavant par un locuteur. Les traitements automatiques proposent une série de critères pour distinguer entre les contextes paraphrastiques et non-paraphrastiques. La précision obtenue est 66,4 %. Si ces critères sont élaborés et testés sur des corpus oraux et au sein d'une approche syntagmatique, nous pensons qu'ils peuvent être transposables à d'autres corpus.

Par rapport aux travaux existants, l'analyse des corpus proposée ici confirme les constatations faites par les chercheurs :
– la reformulation n'est pas toujours paraphrastique et les MRP n'introduisent pas toujours des relations de paraphrase (Rossari, 1990), les MRP pouvant effectivement assumer d'autres rôles dans la langue ;
– les MRP peuvent instaurer la relation de paraphrase entre les segments qui n'entretiennent aucune équivalence sémantique visible (Gulich & Kotschi, 1983; Rossari, 1993).

Pour d'autres constatations faites dans la littérature (l'équivalence syntaxique entre les segments, la possibilité d'échanger les segments de place, la possibilité de supprimer le MRP), nous avons proposé de nouvelles observations.

Plusieurs perspectives peuvent être proposées pour continuer ce travail. Tout d'abord, l'implication d'un autre annotateur et des séances de conciliation entre les annotateurs peuvent permettre d'obtenir des données de référence plus consensuelles. De même, d'autres MRP peuvent être considérés et une analyse comparative plus détaillée entre les emplois des MRP étudiés peut être effectuée. Pour la détection automatique de reformulations paraphrastiques, il est nécessaire d'améliorer la détection des répétitions et de tester une approche par apprentissage automatique pour le repérage de relations paraphrastiques et de segments en relation de paraphrase. D'autres indices encore peuvent être utilisés pour la détection de relations de paraphrase, en particulier ceux fournis par les informations les informations paralinguistiques disponibles dans les transcriptions. Une autre perspective consiste à traiter les relations de paraphrase entre différents tours de parole, alors qu'actuellement nous le faisons au sein d'un même tour de parole uniquement. Nous voulons aussi comparer la reformulation paraphrastique telle qu'elle est effectuée à l'écrit et à l'oral : ce processus est d'une part similaire, car il consiste à éclaircir et faciliter la transmission et la compréhension de l'information, mais d'autre part, il est aussi différent quant à son processus cognitif (Hagège, 1985; Blanche-Benveniste *et al.*, 1991). Comme nous l'avons indiqué, les deux corpus exploités, tout en étant créés avec le même type de situation d'enregistrement d'entretiens semi-guidés, ont été constitués avec 40 ans de différence. Cela présente la possibilité de mener une analyse diachronique des MRP. Nous pouvons aussi étudier l'emploi des MRP en croisant les annotations avec les critères sociologiques des locuteurs. Finalement, nous envisageons de diffuser les données de référence constituées auprès des chercheurs.

Remerciements. Nous remercions Yoann Dupont pour son aide dans l'adaptation du logiciel SEM à nos corpus oraux et les relecteurs pour leur aide dans l'amélioration de la qualité du papier.

Références

ANDROUTSOPOULOS I. & MALAKASIOTIS P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, **38**, 135–187.

BANNARD C. & CALLISON-BURCH C. (2005). Paraphrasing with bilingual parallel corpora. In *ACL*, p. 597–604.

BARZILAY R. & MCKEOWN L. (2001). Extracting paraphrases from a parallel corpus. In *ACL*, p. 50–57.

BEECHING K. (2007). 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), 78–93.

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

BLANCHE-BENVENISTE C., BILGER M., ROUGET C. & VAN DEN EYNDE K. (1991). *Le français parlé. Études grammaticales*. Paris : CNRS Éditions.

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

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

BOUCHERON S. (2000). La langue de l'un, et celle de l'autre : l'entre parenthèses comme aire de reformulation. In *Répétition, Altération, Reformulation*, p. 113–118. Besançon : Presses Universitaires Franc-Comtoises.

CALLISON-BURCH C., COHN T. & LAPATA M. (2008). Parametric : An automatic evaluation metric for paraphrasing. In *COLING*, p. 97–104.

CHOMSKY N. (1975). *Reflections on language*. New-York, USA : Pantheon books.

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

CULIOLI A. (1976). *Notes du séminaire de DEA, 1983-84*. Paris.

DAILLE B., HABERT B., JACQUEMIN C. & ROYAUTÉ J. (1996). Empirical observation of term variations and principles for their description. *Terminology*, **3**(2), 197–257.

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

ELHADAD N. & SUTARIA K. (2007). Mining a lexicon of technical terms and lay equivalents. In *BioNLP*, p. 49–56.

ESHKOL I., TELLIER I., DUPONT Y. & WANG I. (2014). Peut-on bien chunker avec de mauvaises étiquettes pos ? In *TALN 2014*.

ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C. & TELLIER I. (2012). Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *Traitement Automatique de Langues*, **52**(3), 17–46.

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

FRANÇOIS F. (1990). La communication inégale. Heurs et malheurs de l'interaction verbale. In *Actualités pédagogiques et psychologiques*. Neuchâtel-Paris : Delachaux & Niestlé.

FUCHS C. (1982). *La paraphrase*. Paris : PUF.

FUCHS C. (1994). *Paraphrase et énonciation*. Paris : Orphys.

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

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

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

GÜLICH E. & KOTSCHI T. (1987). 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. 15–81. Berne : P Lang.

HAGÈGE C. (1985). *L'homme de paroles. Contribution linguistique aux sciences humaines*. Paris : Fayard.

HAMON T., NAZARENKO A. & GROS C. (1998). A step towards the detection of semantic variants of terms in technical documents. In *International Conference on Computational Linguistics (COLING-ACL'98)*, p. 498–504, Université de Montréal, Montréal, Quebec, Canada.

HÖLKER K. (1988). *Zur Analyse von Markern*. Stuttgart : Franz Steiner.

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

IBRAHIM A., KATZ B. & LIN J. (2003). Extracting structural paraphrases from aligned monolingual corpora. In *International Workshop on Paraphrasing*, p. 57–64.

KANAAN L. (2011). *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.

KOK S. & BROCKETT C. (2010). Hitting the right paraphrases in good time. In *NAACL*, p. 145–153.

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

LIN D. & PANTEL L. (2001). Dirt - discovery of inference rules from text. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 323–328.

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

MADNANI N., RESNIK P., DORR B. & SCHWARTZ R. (2008). Applying automatically generated semantic knowledge : A case study in machine translation. In *NSF Symposium on Semantic Knowledge Discovery, Organization and Use*, p. 60–61.

MALAKASIOTIS P. & ANDROUTSOPOULOS I. (2007). Learning textual entailment using SVMs and string similarity measures. In *ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, p. 42–47.

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

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

MILICEVIC J. (2007). *La paraphrase : Modélisation de la paraphrase langagière*. Peter Lang.

OCH F. & NEY H. (2000). Improved statistical alignment models. In *ACL*, p. 440–447.

PASÇA M. & DIENES P. (2005). Aligning needles in a haystack : Paraphrase acquisition across the Web. In *IJCNLP*, p. 119–130.

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

QUIRK C., BROCKETT C. & DOLAN W. (2004). Monolingual machine translation for paraphrase generation. In *EMNLP*, p. 142–149.

ROSSARI C. (1990). Projet pour une typologie des opérations de reformulation. *Cahiers de linguistique française*, **11**, 345–359.

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

ROSSARI C. (1993). *Les opérations de reformulation. Analyse du processus et des marques dans une perspective contrastive français-italien*, In P. LANG, Ed., *Sciences pour la communication*.

ROULET E. (1987). Complétude interactive et connecteurs reformulatifs. *Cahiers de linguistique française*, **8**, 111–140.

SAUNIER E. (2012). Disons : un impératif de dire ? Remarques sur les propriétés du marqueur et son comportement dans les reformulations. *L'Information Grammaticale*, **132**, 25–34.

SEKINE S. (2005). Automatic paraphrase discovery based on context and keywords between NE pairs. In *International Workshop on Paraphrasing*, p. 80–87.

SHEN S., RADEV D., PATEL A. & ERKAN G. (2006). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *ACL-COLING*, p. 747–754.

SHINYAMA Y., SEKINE S., SUDO K. & GRISHMAN R. (2002). Automatic paraphrase acquisition from news articles. In *Proceedings of HLT*, p. 313–318.

TESTON-BONNARD S. (2008). 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.

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

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