...des conférences enfin disons des causeries... Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux.

Natalia Grabar, Iris Eshkol-Taravella

...des conférences enfin disons des causeries...

Détection automatique de segments en relation de paraphrase dans les reformulations de corpus oraux

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

Résumé. Notre travail porte sur la détection automatique des segments en relation de 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. Les données de référence sont consensuelles. Une méthode automatique fondée sur l'apprentissage avec les CRF est proposée afin de détecter les segments paraphrasés. Différents descripteurs sont exploités dans une fenêtre de taille variable. Les tests effectués montrent que les segments en relation de paraphrase sont assez difficiles à détecter, surtout avec leurs frontières correctes. Les meilleures moyennes atteignent 0,65 de F-mesure, 0,75 de précision et 0,63 de rappel. Nous avons plusieurs perspectives à ce travail pour améliorer la détection des segments en relation de paraphrase et pour étudier les données depuis d'autres points de vue.

Abstract. Our work addresses automatic detection of segments with paraphrastic rephrasing relation in spoken corpus. The proposed approach is syntagmatic. It is based on paraphrastic rephrasing markers and the specificities of the spoken language. The reference data used are consensual. Automatic method based on machine learning using CRFs is proposed in order to detect the segments that are paraphrased. Different descriptors are exploited within a window with various sizes. The tests performed indicate that the segments that are in paraphrastic relation are quite difficult to detect. Our best average reaches up to 0.65 F-measure, 0.75 precision, and 0.63 recall. We have several perspectives to this work for improving the detection of segments that are in paraphrastic relation and for studying the data from other points of view.

Mots-clés : Corpus oraux, Paraphrase, Reformulation, Marqueur de reformulation paraphrastique, Apprentissage supervisé.

Keywords: Spoken Corpora, Paraphrase, Reformulation, Paraphrastic Reformulation Marker, Supervised Learning.

1 Introduction

La paraphrase joue un rôle important dans la langue. Ceci justifie en particulier la conception de la langue comme d'un système de paraphrasage par certains linguistes (Melčuk, 1988). Voilà quelques exemples de contextes où la paraphrase se relève importante :

– Dans les cours de langues, on demande souvent aux élèves de paraphraser des expressions ou des phrases afin de contrôler leur maîtrise de la langue, qu'elle soit maternelle ou étrangère ;
– De la même manière, il est possible de contrôler la compréhension d'une idée. Les premiers exercices de paraphrasage aurait ainsi apparus en effectuant l'exégèse des textes anciens : des textes sacrés (Bible, Coran, Tora) d'abord, et ensuite des textes théologiques, philosophiques ou scientifiques. Notons que la production d'explications ou de commentaires sur ces textes occupe toujours une place importante en philosophie, théologie et philologie des langues anciennes ;
– De manière plus naturelle, les locuteurs recourent à la paraphrase pour préciser leurs pensées et les transmettre au mieux à leurs interlocuteurs. Dans ce cas, la paraphrase découle de l'activité de reformulation. Notons que l'écriture relève également du paraphrasage : entre les différentes versions d'une oeuvre littéraire (Fuchs, 1982), d'un article de Wikipédia (Vila *et al.*, 2014) ou d'un article scientifique, les auteurs peuvent réécrire plusieurs fois leur texte avant de produire celui qui les satisfait enfin mieux.

En dehors des situations communes de la vie, la paraphrase joue aussi un rôle important pour différentes applications de TAL (Androutsopoulos & Malakasiotis, 2010; Madnani & Dorr, 2010; Bouamor *et al.*, 2012). L'objectif est alors de détecter les expressions linguistiques formellement différentes mais véhiculant une sémantique similaire ou proche :

– En recherche et extraction d'information, et dans les systèmes de questions-réponses, les paraphrases permettent d'augmenter la couverture des résultats grâce aux expressions équivalentes entre les requêtes ou les questions et les textes dans lesquelles les réponses doivent être trouvées. Par exemple, les paires {*infarctus du myocarde*; *crise cardiaque*} et {*maladie d'Alzheimer*; *maladie neurodégénérative*} contiennent des expressions différentes qui véhiculent cependant une sémantique identique ou proche. Si le système automatique dispose de telles connaissances, la couverture et la qualité de ses résultats peuvent être améliorées ;
– En traduction automatique, les paraphrases permettent d'éviter des répétitions lexicales et peuvent introduire ainsi une légèreté du texte cible. Par exemple, le segment original en anglais *Figure 10.2 shows money growth and output growth. There is a strong, but not absolute, link between money growth and output growth* est traduit en français de manière à éviter la répétition *money growth and output growth* : *Le graphique 10.2 montre une augmentation des fonds et de la production. Il existe entre ces éléments un lien étroit, bien que non absolu* (Scarpa, 2010). Différentes langues ont en effet une tolérance variable vis-à-vis des répétitions (Hatim & Mason, 1990; Baker, 1992) ;
– L'inférence textuelle (Dagan *et al.*, 2013) consiste à établir une relation entre deux segments textuels, appelés Texte et Hypothèse. L'inférence textuelle est une relation directionnelle, dans laquelle la vérité de l'Hypothèse peut être inférée à partir du sens du Texte, ou, en d'autres mots, il est possible de vérifier si l'Hypothèse est subsumée par le Texte. Par exemple, le Texte *The drugs that slow down or halt Alzheimer's disease work best the earlier you administer them* permet d'inférer que l'Hypothèse *Alzheimer's disease is treated by drugs* est vraie ; par contre, l'Hypothèse *Alzheimer's disease is cured by drugs* ne peut pas être inférée à partir de ce Texte. Dans cet exemple, le lien de paraphrase entre les paires {*administer drugs*; *treated by drugs*} et {*slow down or halt*; *cured by drugs*} permet justement d'établir ce lien d'inférence entre le Texte et l'Hypothèse.

Comme ces quelques exemples montrent, en fonction des objectifs poursuivis, la paraphrase requiert des phénomènes linguistiques plus ou moins nombreux. L'étendue des classifications de paraphrases proposées peut être ainsi plus ou moins couvrante, et aller de 25 catégories de paraphrases (Bhagat & Hovy, 2013) à 67 fonctions lexicales pour le paraphrasage (Melčuk, 1988). Le plus souvent, ces classifications focalisent sur un aspect donné, comme par exemple les moyens linguistiques mis en oeuvre (Melčuk, 1988; Vila *et al.*, 2011; Bhagat & Hovy, 2013), la taille de l'unité paraphrasée (Flottum, 1995; Fujita, 2010; Bouamor, 2012), les connaissances requises (Milicevic, 2007), le registre de langue. À notre connaissance, la seule classification multidimensionnelle est celle de (Milicevic, 2007) : elle couvre plusieurs des dimensions mentionnées. Dans notre travail précédent, nous avons également proposé de travailler sur une classification à plusieurs dimensions (Eshkol-Taravella & Grabar, 2014). Elle prend en compte les dimensions suivantes :

– la catégorie syntaxique des segments paraphrasés,
– le type de la relation lexicale entre ces segments (*e.g.* hyperonyme, synonyme, antonyme, instance, méronyme),
– le type de la modification lexicale (*e.g.* remplacement, suppression, ajout),
– le type de la modification morphologique (*i.e.* flexion, dérivation, composition),
– le type de la modification syntaxique (*e.g.* passif/actif),
– le type de la relation pragmatique liée aux fonctionnalités de la paraphrase et de la reformulation (*e.g.* définition, explication, précision, résultat, correction linguistique, correction référentielle, équivalence).

Dans notre travail, nous adoptons donc une acception large de la paraphrase.

2 Travaux existants en acquisition automatique de paraphrases

Plusieurs approches sont proposées pour la détection automatique de la paraphrase. De manière générale, ces approches reposent sur les propriétés paradigmatiques des mots et sur leur capacité de se substituer mutuellement dans un contexte donné. Ces approches dépendent du type de corpus exploités. Quatre types de corpus sont généralement distingués : corpus monolingues, corpus monolingues parallèles, corpus monolingues comparables, corpus bilingues parallèles.

Corpus monolingues. En corpus monolingues, deux types d'approches sont à noter :

– la similarité des chaînes d'édition permet de détecter les unités linguistiques (mots, syntagmes, etc) qui montrent des traits communs de surface et permettent de rapprocher deux chaînes comme {*When did Charle de Gaulle die ?*; *Charles de Gaulle died in 1970*} (Malakasiotis & Androutsopoulos, 2007),
– les méthodes distributionnelles permettent de détecter les unités qui apparaissent dans des contextes similaires, auquel cas ces unités montrent aussi des vecteurs contextuels ou syntaxiques similaires, et sont alors de bons candidats pour la paraphrase (*e.g.* {*Y is solved by X*; *Y is resolved in X*}) (Lin & Pantel, 2001; Pasça & Dienes, 2005).

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 corpus les plus utilisés est constitué des traductions en anglais de *20 000 lieux sous la mer* de Jules Verne. Lorsque les phrases de tels corpus sont alignées, il devient alors possible de les exploiter 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). Elles permettent d'extraire les paraphrases comme {*countless*; *lots of*}, {*undertone*; *low voice*}, {*shrubs*; *bushes*}, {*refuse*; *say no*}, {*dull tone*; *gloom*}, {*sudden appearance*; *apparition*} (Barzilay & McKeown, 2001).

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). En particulier, les entités nommées et les nombres font partie des repères pour l'extraction de paraphrases comme {*PERS1 killed PERS2*; *PERS1 let PERS2 die from loss of blood*} ou {*PERS1 shadowed PERS2*; *PERS1 kept his eyes on PERS2*} (Shinyama *et al.*, 2002).

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. Les traductions multiples d'une expression ou d'un mot peuvent alors correspondre aux paraphrases (Bannard & Callison-Burch, 2005; Callison-Burch *et al.*, 2008; Kok & Brockett, 2010). Par exemple, les paraphrases {*under control*; *in check*} peuvent être extraites parce qu'elles sont utilisées pour la traduction de *unter kontrolle* dans différents contextes (Bannard & Callison-Burch, 2005).

3 Objectifs

Dans notre précédent travail, nous avons commencé à étudier les reformulations paraphrastiques de l'oral, formées autour de trois marqueurs de reformulation paraphrastique : *c'est-à-dire*, *je veux dire* et *disons* dans l'exemple (1)). Nous avons en particulier proposé un guide pour l'annotation des reformulations paraphrastiques et l'avons testé lors de l'annotation des tours de parole de corpus *ESLO1* et *ESLO2*. Nous avons également défini et testé une méthode à base de règles pour distinguer, au sein d'un ensemble de tours de parole comportant un des trois marqueurs étudiés, les tours de parole qui comportent des reformulations paraphrastiques. Dans le travail actuel, nous poursuivons nos objectifs. D'une part, nos données de référence sont consensuelles et comportent plus d'exemples annotés. D'autres part, nous nous focalisons maintenant sur la détection automatique de segments faisant partie des reformulations paraphrastiques. Nous mettons en place pour ceci un système par apprentissage supervisé. La structure étudiée est de type : *segment1 marqueur-de-reformulation-paraphrastique segment2*. Les marqueurs de reformulation paraphrastique établissent alors le lien sémantique entre les deux segments en relation de paraphrase. Dans l'exemple (1), il s'agit du segment source *des conférences* et du segment cible *des causeries*. Il a été observé que, dans plusieurs cas, les segments impliqués n'ont aucun lien sémantique évident, mais, grâce au marqueur de reformulation et à la relation de paraphrase établie, ce lien peut apparaître (Gulich & Kotschi, 1983; Rossari, 1993). Notre objectif principal est donc de trouver les critères formalisables qui permettent de détecter automatiquement les segments en relation de paraphrase et d'établir ce lien sémantique entre ces segments. Il est possible que les corpus étudiés contiennent d'autres reformulations paraphrastiques introduites par d'autres phénomènes (d'autres marqueurs, l'intonation, les pauses...). Cependant, nous n'étudions pas ces autres reformulations paraphrastiques.

(1) *des conférences y en a assez souvent sur France culture enfin disons des causeries* [ESLO1_ENT_121_C]

Nous présentons d'abord les données linguistiques que nous traitons (section 4). Nous présentons ensuite la méthode (section 5), et les résultats obtenus (section 6). Nous terminons avec les orientations pour les travaux futurs (section 7).

4 Données linguistiques

Nous utilisons plusieurs types de données linguistiques : les marqueurs de reformulation paraphrastique (section 4.1), les corpus traités (section 4.2), les marqueurs de disfluence (section 4.3) et les données de référence grâce auxquelles le système automatique est créé et évalué (section 4.4).

4.1 Marqueurs de reformulation paraphrastique (MRP)

Nous exploitons trois marqueurs de reformulation paraphrastique (MRP) : *c'est-à-dire* (Gulich & Kotschi, 1983; Hölker, 1988; Beeching, 2007), *je veux dire* (Teston-Bonnard, 2008) et *disons* (Hwang, 1993; Petit, 2009; Saunier, 2012). 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 et étudié des trois. Ces trois marqueurs sont reconnus pour leur capacité d'introduire les reformulations paraphrastiques, mais ils peuvent également jouer d'autres rôles dans le discours, comme par exemple l'argumentation ou les disfluences.

4.2 Provenance et composition de 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. Il comprend 300 heures de parole (4 500 000 mots environ) et inclue 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/*).

4.3 Marqueurs de disfluence

Nous utilisons un ensemble de marqueurs de disfluence : *allez, allons, alors, là, enfin, euh, heu, bah, ben, hm, hum, hein, quoi, ah, oh, donc, bon, bè, eh*.

4.4 Données de référence

Comme observé dans la littérature, les marqueurs de reformulation paraphrastique peuvent apparaître dans des emplois qui ne sont pas toujours paraphrastiques (Gulich & Kotschi, 1983; Flottum, 1995; Rossari, 1993). L'objectif de l'annotation manuelle est de faire cette distinction et de marquer les segments en relation de paraphrase. Ainsi, l'exemple (1) contient une reformulation paraphrastique, alors que les exemples (2) et (3) ne contiennent pas de relations de paraphrase. Dans ce dernier cas, les marqueurs de reformulation paraphrastique peuvent ainsi être associés aux marqueurs discursifs ou aux disfluences. Lorsqu'un tour de parole contient une reformulation paraphrastique, les annotations plus précises, comme indiquées à la fin de la section 1, sont établies. Les segments en relation de reformulation paraphrastique peuvent être de différentes tailles et avoir différentes fonctions syntaxiques : noms, adjectifs, verbes, adverbes et pronoms ; groupes nominaux (exemple en (1)), verbaux (exemple en (12)), adjectivaux ou adverbiaux ; propositions (exemple en (10)) ; présentateur...

(2) *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]

(3) *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]

(4) *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]

Trois paires d'annotateurs ont participé dans la création des données de référence. Après une annotation indépendante, des séances de consensus ont été tenues. Le consensus porte sur la présence de la reformulation paraphrastique et sur les segments source et cible.

Tout marqueur de reformulation paraphrastique confondu, 611 tours de parole du corpus *ESLO1* et 498 tours de parole du corpus *ESLO2* de la partie *entretiens* sont analysés. Ces tours de parole contiennent 168 paraphrases du corpus *ESLO1* et 186 paraphrases du corpus *ESLO2*. Les tours de parole étudiés proviennent de 59 et 37 entretiens des corpus *ESLO1* et *ESLO2*, respectivement. Ces données de référence sont utilisées pour entraîner le système automatique à la détection des segments en relation de paraphrase et pour évaluer les résultats.

5 Méthode

La figure 1 présente le schéma général de la méthode, composée de plusieurs étapes : le pré-traitement des données (section 5.1), la détection de tours de parole avec les reformulations paraphrastiques (section 5.2). Le traitement continue avec les tours de paroles qui contiennent les reformulations paraphrastiques. Nous effectuons alors la détection des segments en relation de paraphrase (section 5.3) et l'évaluation de résultats (section 5.4). Dans le travail proposé, l'accent principal est mis sur la détection des segments en relation de paraphrase et l'évaluation de ces résultats avec les données de référence.

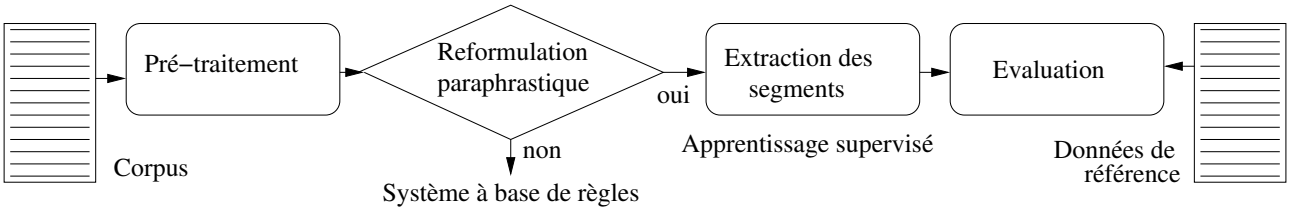

FIGURE 1 – Schéma général de la méthode.

5.1 Pré-traitement des données

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 d'origine 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, qui ont bénéficié de la correction orthographique supplémentaire par rapport aux versions A et B. Pour avoir des données comparables, nous avons sélectionné des entretiens dans les deux corpus. En l'absence de la ponctuation, nous avons reconstitué les tours de paroles en utilisant comme segmenteur :
– les tours de parole marqués dans la transcription par un changement de locuteur,
– et en traitant les chevauchements où deux locuteurs ou plus parlent en même temps.
Dans le cas des chevauchements, 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 tours de parole sont notre unité de travail. Lorsque les tours de parole se retrouvent sur plus d'un groupe de souffle, ces groupes sont séparés par les virgules dans nos sorties (exemple en (8)).

Les corpus avec les tours de parole reconstitués sont traités avec le chunker SEM (Tellier *et al.*, 2014) adapté à la langue orale. SEM détecte les chunks minimaux, comme présenté dans l'exemple (5) (même exemple qu'en (2)).

(5) *(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*

5.2 Distinction automatique entre les reformulations paraphrastiques et non paraphrastiques

Cette étape n'a pas été modifiée par rapport au travail précédent. Par soucis de clarté, nous en rappelons les principaux moments. L'objectif de cette étape est d'analyser les tours de parole qui contiennent les marqueurs de reformulation paraphrastique pour décider si, dans ces tours de parole, les marqueurs de reformulation paraphrastique introduisent la relation de reformulation paraphrastique ou non. Actuellement, cette étape est fondée sur des filtres communs à l'ensemble des marqueurs de reformulation paraphrastique :

1. Si le marqueur de reformulation paraphrastique est placé en début ou en fin de tours de parole, alors il est considéré que le contexte n'est pas suffisant et que ce tour de parole ne comporte pas de paraphrase ;
2. Si le marqueur de reformulation paraphrastique est entouré des marqueurs discursifs (comme ceux présenté dans la section 4.3), d'euh d'hésitation, d'interjections (*ben hm ouais*), d'amorces (*s-*), etc. répétés, il est considéré que le marqueur de reformulation paraphrastique 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 ;
3. Si le marqueur de reformulation paraphrastique apparaît dans un contexte lexical spécifique (emploi de *nous* devant *disons*), ou si le marqueur de reformulation paraphrastique apparaît dans des suites argumentatives (*e.g. par contre, mais, en revanche, au contraire*), ce tour de parole ne comporte pas de paraphrase ;
4. Si le marqueur de reformulation paraphrastique apparaît à l'intérieur d'une locution, comme *indépendamment de* ou *plus ou moins grossiers* (exemples (2) et (3)), alors il est considéré que ce contexte ne comporte pas de paraphrase. Ce test est effectué sur les sorties du chunker (exemple (5)). Pour vérifier que la locution existe dans la langue, nous interrogeons un moteur de recherche généraliste et analysons les fréquences attestées sur la Toile. Ces fréquences donnent une indication générale et relative sur les segments testés. Chaque segment est testé de trois manières (exemple (5)) : 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 marqueur de reformulation paraphrastique, 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 à des seuils donnés pour qu'il soit considéré que ce segment comporte une paraphrase. Dans le cas où les fréquences sont plus élevées que le seuil, ce test indique que la locution existe dans la langue et qu'il s'agit en effet de disfluence.

L'application et le test de ces filtres ont montré que la meilleure combinaison consiste à donner la priorité aux filtres 1, 3 et 4, et d'attribuer moins d'importance au filtre 2 car les disfluences peuvent apparaître autour des paraphrases, comme *disons* dans l'exemple (1). Dans ce cas, la précision atteint jusqu'à 66,4 %.

5.3 Extraction automatique de segments en relation de reformulation paraphrastique

Cette étape correspond à l'apport principal de notre travail. Les segments en relation de paraphrase sont détectés avec un système d'apprentissage supervisé. En disposant des données de référence et en travaillant sur le paraphrasage syntagmatique, nous exploitons les fonctionnalités des CRF telles qu'elles sont implémentées dans Wapiti (Lavergne *et al.*, 2010) pour créer notre système. Les données de référence sont divisées aléatoirement en deux ensembles : un ensemble d'entraînement et un ensemble de test, composés de 60 % et 40 % des tours de parole, respectivement.

5.3.1 Les catégories à détecter

L'objectif principal est de détecter deux segments en relation paraphrastique : le segment source, qui est repris ultérieurement dans le texte, et le segment cible, qui propose une nouvelle expression linguistique pour l'idée déjà exprimée par le segment source. Les catégories à détecter sont donc les suivantes :

1. **M** : marqueur de reformulation paraphrastique,
2. **SEG1** : segment source, qui apparaît avant le marqueur de reformulation paraphrastique,
3. **SEG2** : segment cible, qui apparaît après le marqueur de reformulation paraphrastique,
4. **O** : tout autre token dans les tours de parole (out).

form	*POS*	*chunkBI*	*chunk*	*heu*	*num*	*dmf*	*stem*	*MRP*	*ref.*
...									
la	DET	I-PP	PP	N	28	MIL	la	O	O
cuisson	NC	I-PP	PP	N	29	MIL	cuisson	O	O
rapide	ADJ	I-PP	PP	N	30	MIL	rapid	O	O
quoi	PROWH	I-PP	PP	EUH	31	MIL	quoi	O	O
des	DET	B-NP	NP	N	32	MIL	de	O	SEG1
morceaux	NC	I-NP	NP	N	33	MIL	morceau	O	SEG1
nobles	ADJ	I-NP	NP	N	34	MIL	nobl	O	SEG1
ce	PRO	I-NP	NP	N	35	MIL	ce	O	O
qu’	PROREL	B-NP	NP	N	36	MIL	qu’	O	O
ils	CLS	B-NP	NP	N	37	MIL	il	O	O
appellent	V	B-VN	VN	N	38	MIL	appellent	O	O
quoi	PROWH	B-NP	NP	EUH	39	MIL	quoi	O	O
c’	CLS	B-NP	NP	N	40	MIL	c’	M	M
est	V	B-VN	VN	N	41	MIL	est	M	M
à	P	B-PP	PP	N	42	MIL	à	M	M
dire	VINF	I-PP	PP	N	43	MIL	dir	M	M
les	DET	B-NP	NP	N	44	MIL	le	O	SEG2
rosbifs	ADJ	I-NP	NP	N	45	MIL	rosbif	O	SEG2
les	DET	I-NP	NP	N	46	MIL	le	O	SEG2
biftecks	NC	I-NP	NP	N	47	MIL	bifteck	O	SEG2
et	CC	B-CONJ	CONJ	N	48	MIL	et	O	SEG2
tout	PRO	B-NP	NP	N	49	MIL	tout	O	SEG2
ça	PRO	B-NP	NP	N	50	MIL	ça	O	SEG2
...									

TABLE 1 – Un extrait de données annotées, avec une reformulation paraphrastique.

5.3.2 Les unités et les descripteurs

Les unités traitées sont les tours de parole. Chaque mot des tours de parole est décrit avec un ensemble de descripteurs, comme exemplifié dans le tableau 1. Nous présentons les descripteurs et justifions leur choix :

– *form* : forme de chaque mot graphique tel qu’il apparaît dans le texte. La forme correspond à l’information de base directement disponible dans le texte ;
– *POS* : étiquette morpho-syntaxique calculée par SEM. L’étiquette morpho-syntaxique peut être indicative d’un lien syntaxique entre les segments à mettre en relation, et peut ainsi faciliter la détection de la relation de paraphrase ;
– *chunksBI* : chunk de SEM avec le marquage de debut et de fin. Les chunks marquent la segmentation des tours de parole en segments syntaxiques. Cette information peut aider à délimiter les frontières des segments en relation de paraphrase ;
– *chunks* : chunk de SEM sans le marquage de debut et de fin. Par rapport à *chunksBI*, il s’agit d’un descripteur simplifié car il ne s’agit plus de détecter le début et la suite des segments syntaxiques, mais juste de faire la distinction entre les différents segments syntaxiques. Nous pensons que ce descripteur peut être aussi intéressant pour la tâche visée ;
– *heu* : marquage des disfluences exprimées par des marqueurs spécifiques considérés dans notre travail (section 4.3). Nous pensons que les marqueurs de disfluences peuvent également aider à détecter les segments en relation de paraphrase car, dans ce cas, étant employés avec les marqueurs de reformulation, les marqueurs de disfluences peuvent souligner l’acte de reformulation ;
– *num* : numéro du mot depuis le début de chaque tour de parole. Ce descripteur permet surtout de marquer les mots apparaissant au début des tours de parole et donc de limiter potentiellement la taille des segments à détecter ;
– *début/milieu/fin* : position relative de chaque mot graphique où les positions de début et de fin correspondent à 20 % des tokens du début et de la fin des tours de parole, respectivement. Le reste des mots fait partie de la position du milieu. Ce descripteur est similaire au descripteur précédent *num* ;
– *stem* : mot racinisé avec le raciniseur `Snowball` (Porter, 1980) implémenté dans le module Perl `Lingua::Stem`. Nous pensons que les mots racinisés peuvent aider à établir un lien entre les segments en relation de paraphrase au cas où ces segments comportent des mots formellement similaires. Bien qu’étant assez brutal avec les mots, `Snowball` permet de les réduire à des chaînes de caractères plus facilement comparables ;

– *MRP* : marquage du marqueur de reformulation paraphrastique. Ce marqueur correspond également à l'information disponible directement dans le texte. De plus, comme c'est lui qui permet d'établir la relation de paraphrase, son rôle pour la détection des segments peut être important.

5.3.3 Les patrons pour la gestion des descripteurs et des contextes

Les patrons sous `Wapiti` indiquent quels descripteurs il faut exploiter, comment les combiner, dans quel contexte les étudier, etc. Nous appliquons trois séries de patrons pour tester plusieurs hypothèses :

1. utilisation de la forme seule : dans une fenêtre de 3 à 12 mots avant et après le token courant la forme est considérée, de même que la combinaison *forme/MRP*. Il s'agit des informations directement disponibles dans le texte ;
2. la taille de la fenêtre autour du mot courant a une influence : pour l'ensemble de descripteurs, nous faisons varier la largeur de la fenêtre allant de 3 à 12 mots avant et après un token donné ;
3. les combinaisons de différents descripteurs ont une influence sur les résultats : dans la fenêtre de 7 mots avant et après le token courant, nous faisons varier les combinaisons de descripteurs.

5.4 Évaluation

L'évaluation est effectuée par rapport aux données de référence. Nous calculons la macro-précision, le macro-rappel et la macro-F-mesure au niveau des catégories (Sebastiani, 2002) et les moyennes globales. La baseline correspond à l'exploitation de la forme seule dans une fenêtre de 7 mots avant et après le token courant, avec la combinaison *forme/MRP*. Il s'agit des informations disponibles dans le texte directement et de la largeur de fenêtre moyenne dans nos expériences.

6 Résultats et Discussion

6.1 Trois séries d'expériences

Dans l'ensemble de test, les catégories *SEG1* et *SEG2* représentent 8,9 % et 11,2 %, respectivement, au niveau des tokens.

Les tableaux 2 à 4 présentent les résultats obtenus pour les trois séries d'expériences. Pour chaque expérience, nous indiquons les valeurs de la précision, du rappel et de la F-mesure pour les catégories visées : *O*, *SEG1* et *SEG2*. De manière générale, le marqueur est la catégorie la plus facile à reconnaître (toujours proche de 1,0), tandis que les segments en relation de paraphrase restent plus difficiles à détecter. Nous calculons aussi la moyenne des performances.

	O			*SEG1*			*SEG2*		
Taille de la fenêtre	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
E1 (3 mots)	0,81	0,97	0,88	0,61	0,13	0,21	0,50	0,13	0,20
E2 (4 mots)	0,81	0,95	0,88	0,51	0,18	0,27	0,39	0,13	0,20
E3 (5 mots)	0,81	0,97	0,88	0,57	0,13	0,21	0,45	0,12	0,19
E4 (6 mots)	0,81	0,97	0,88	0,59	0,11	0,19	0,46	0,12	0,19
E5 (7 mots) - Baseline	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
E6 (8 mots)	0,81	0,98	0,88	0,62	0,12	0,20	0,51	0,11	0,18
E7 (9 mots)	0,81	0,97	0,88	0,59	0,11	0,19	0,49	0,14	0,22
E8 (10 mots)	0,81	0,97	0,88	0,61	0,13	0,21	0,48	0,15	0,22
E9 (11 mots)	0,82	0,96	0,88	0,61	0,20	0,30	0,51	0,16	0,24
E10 (12 mots)	0,81	0,97	0,88	0,61	0,13	0,21	0,46	0,15	0,22

TABLE 2 – Utilisation de la forme seule, dans une fenêtre allant de 3 à 12 mots.

Le tableau 2 montre les résultats lorsque seules les formes et la combinaison *forme/MRP* sont utilisées. La différence entre les expériences concerne la largeur de la fenêtre qui varie de 3 à 12 mots à gauche et à droite du token donné. Cette configuration montre une moyenne autour de 0,60 : la précision moyenne est autour de 0,70 tandis que le rappel est autour de 0,58. Les segments en relation de paraphrase sont reconnus avec une précision assez élevée mais un rappel très bas (le

plus souvent, moins de 0,20). La différence entre les expériences n'est pas très importante. La baseline est bien placée. Les meilleures moyennes de F-mesure sont observées avec les fenêtres de 11 et 4 mots.

	O			*SEG1*			*SEG2*		
Taille de la fenêtre	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
E1 (3 mots)	0.84	0.77	0.81	0.27	0.29	0.28	0.34	0.51	0.40
E2 (4 mots)	0.82	0.58	0.68	0.24	0.17	0.20	0.20	0.64	0.30
E3 (5 mots)	0.84	0.37	0.52	0.40	0.23	0.30	0.16	0.81	0.26
E4 (6 mots)	0.84	0.40	0.54	0.20	0.61	0.31	0.22	0.68	0.34
E5 (7 mots)	0.83	0.46	0.60	0.27	0.50	0.35	0.16	0.55	0.25
E6 (8 mots)	0.86	0.45	0.59	0.13	0.70	0.21	0.42	0.29	0.35
E7 (9 mots)	0.85	0.61	0.71	0.16	0.63	0.25	0.40	0.23	0.29
E8 (10 mots)	0.87	0.29	0.43	0.18	0.70	0.29	0.21	0.72	0.33
E9 (11 mots)	0.87	0.35	0.50	0.26	0.64	0.37	0.19	0.74	0.30
E10 (12 mots)	0.88	0.51	0.65	0.15	0.70	0.25	0.45	0.48	0.46

TABLE 3 – Utilisation de l'ensemble des descripteurs dans une fenêtre allant de 3 à 12 mots.

Le tableau 3 montre les résultats lorsque l'ensemble des descripteurs est utilisé dans une fenêtre variant de 3 à 12 mots. Par rapport au tableau 2, nous observons plus de différences dans les résultats. La reconnaissance des segments en relation de paraphrase est meilleure grâce à l'amélioration du rappel. Par contre, les performances de la catégorie *O* diminuent, ce qui mène aussi à la diminution de moyennes globales. Vu la différence entre les expériences, nous supposons que la portée optimale de chaque descripteur est différente et la largeur optimale de sa fenêtre doit être adaptée. Pour ceci, des tests complémentaires doivent être effectués.

	O			*SEG1*			*SEG2*		
Différentes combinaisons	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Baseline	0,82	0,94	0,88	0,47	0,13	0,21	0,37	0,18	0,24
E1	0.84	0.94	0.89	0.47	0.42	0.44	0.70	0.17	0.27
E2 (E1 + 7 mots(chunk/MRP))	0.83	0.36	0.50	0.12	0.59	0.20	0.27	0.52	0.36
E3 (E2 + 7 mots(POS/MRP))	0.85	0.19	0.32	0.36	0.43	0.39	0.15	0.93	0.26
E4 (E3 + 7 mots(stem/MRP))	0.87	0.36	0.51	0.20	0.65	0.31	0.21	0.69	0.32
E5 (E4 + 7 mots(début))	0.83	0.46	0.60	0.27	0.50	0.35	0.16	0.55	0.25
E6 (E5 - 7 mots(POS))	0.84	0.48	0.61	0.23	0.52	0.32	0.22	0.66	0.33
E7 (E5 - 7 mots(chunk))	0.84	0.26	0.40	0.20	0.60	0.30	0.18	0.78	0.30
E8 (E5 - 7 mots(heu))	0.85	0.39	0.53	0.24	0.45	0.32	0.20	0.81	0.32
E9 (E5 - 7 mots(stem))	0.85	0.32	0.47	0.31	0.28	0.29	0.15	0.82	0.26
E10 (E5 - 7 mots(MRP))	0.83	0.42	0.56	0.17	0.47	0.25	0.22	0.69	0.34

TABLE 4 – Différentes combinaisons de descripteurs dans une fenêtre de 7 mots. L'expérience *E1* correspond à la combinaison de descripteurs : dans une fenêtre de 7 mots (POS, chunk, heu, stem, MRP), dans une fenêtre de 1 mot (chunk/MRP, POS/MRP, stem/MRP). Les autres combinaisons dérivent de *E1*.

Le tableau 4 montre les résultats obtenus avec l'exploitation de différentes combinaisons de descripteurs dans une fenêtre de 7 mots. Ces différentes expériences dépassent la baseline. L'expérience *E1* correspond à la combinaison de descripteurs : dans une fenêtre de 7 mots (POS, chunk, heu, stem, MRP), dans une fenêtre de 1 mot (chunk/MRP, POS/MRP, stem/MRP). Les autres combinaisons dérivent de *E1*. Les expériences qui montrent la meilleure reconnaissance des segments en relation de paraphrase (F-mesure supérieure à 0,30) sont *E4, E6* et *E8*. Il est difficile de faire des généralisations, mais l'ajout de nouveaux descripteurs, de même que la suppression de POS et heu, semblent être bénéfiques. Le marqueur de reformulation paraphrastique est considéré comme le token pivot car, dans notre approche, c'est ce marqueur qui a le potentiel d'établir la relation de paraphrase. Avec ces expériences, la précision et le rappel sont assez équivalents, ce qui permet d'obtenir de meilleures performances globales par rapport à la baseline. La reconnaissance des segments paraphrasés restent cependant difficile.

En résumé, nous pouvons faire plusieurs observations :

– les marqueurs de reformulation sont toujours bien reconnus,
– les positions O out sont aussi assez bien reconnues,
– la détection des segments cible et source restent difficile et montre des performances variables selon les expériences,
– parmi les meilleures expériences, nous avons la baseline (utilisation des formes dans une fenêtre de 7 mots et de la combinaison *forme/MRP*) et les expériences basées sur différentes combinaisons du tableau 4,
– parmi les meilleures fenêtres, nous avons 4, 7 et 11 mots.

6.2 Discussion

Concernant les 3 séries d'expériences proposées, nous pouvons remarquer que :
– l'exploitation de la forme seule est assez efficace du point de vue des valeurs moyennes. Cette série d'expériences est surtout intéressante car la précision est élevée par comparaison aux autres expériences ;
– les combinaisons ciblées de l'expérience présentée dans le tableau 4 montrent l'intérêt d'effectuer de telles combinaisons. Avec certains tests, nous obtenons des valeurs de précision et de rappel assez élevées. Nous devons cependant effectuer plus de tests pour optimiser les résultats ;
– le rappel reste le point faible des expériences actuelles. L'amélioration principale de la méthode concerne cet aspect.

Dans d'autres expériences, non présentées ici, nous avons traité séparément les corpus *ESLO1* et *ESLO2*, et les annotations de deux annotateurs $A1$ et $A2$. Il s'agit des données de référence sans le consensus. Les ensembles d'entraînement et de test étaient également indépendants. Les modèles créés sur chacun des ensembles de données (*ESLO1*/$A1$, *ESLO1*/$A2$, *ESLO2*/$A1$, *ESLO2*/$A2$) ont été appliqués sur d'autres ensembles. L'objectif était de voir quelle est la portabilité de ces modèles. Ces expériences montrent que :
– il est plus facile de détecter les segments paraphrasés dans le corpus *ESLO1*, quel que soit le modèle appliqué (corpus ou annotateur). En effet, le corpus *ESLO2* comporte des tours de parole beaucoup plus longs, ce qui rend la détection des segments paraphrasés plus compliquée dans ce corpus. Cette situation peut s'expliquer par la méthodologie mise en œuvre dans la constitution du corpus ESLO2 où l'entretien mené est beaucoup moins formel et l'intervieweur laisse plus de liberté à son interlocuteur ;
– les modèles liés aux annotateurs montrent aussi des performances variables, mais qui se ressentent moins par comparaison avec l'influence des corpus ;
– assez systématiquement, la détection des segments cible *SEG2* est plus aisée que la détection des segments source *SEG1*. En relation avec cette observation, une hypothèse est que dans le flux des énoncés, après l'apparition du marqueur de reformulation, il est plus évident de se rendre compte que le *SEG2* doit être retrouvé.

De manière générale, le travail avec les données de référence consensuelles entre les annotateurs et la fusion des annotations provenant des corpus *ESLO1* et *ESLO2* montre un effet légèrement bénéfique sur les résultats. Par ailleurs, nous avons aussi observé que la prise en compte de la catégorie *M*, même si sa détection est évidente, permet d'améliorer la détection des segments en relation de paraphrase de quelques points.

6.3 Analyse des erreurs

L'analyse des sorties indique que très souvent les segments sont détectés, mais avec des frontières différentes que celles définies par les données de référence. En (6) et (7), nous présentons deux exemples différents : en A se trouvent les tours de parole tels qu'annotés dans les données de référence, en B se trouvent les résultats de la détection automatique. Les segments en relation de paraphrase sont en bleu et surlignés. Nous pouvons voir que les segments détectés automatiquement sont plus larges. Les segments proposés par le modèle sont aussi acceptables et des hésitations quant à la taille de ces segments étaient aussi présentes lors de l'annotation manuelle. Dans de nombreux cas, les sorties sont intéressantes et utilisables au moins comme une base d'annotation. Cependant, comme l'évaluation des résultats de CRF dépend de la détection des frontières, cela diminue les chiffres de l'évaluation automatique des résultats. Notons aussi que la taille des segments détectés augmente avec l'agrantissement de la fenêtre (la série d'expériences du tableau 3).

(6) *A. et cinq kilomètres c'est-à-dire j'avais quatre kilomètres à faire quatre et quatre huit je faisais huit kilomètres tous les jours et à pieds ah oui* [ESLO1_ENT_011_C]
B. et cinq kilomètres c'est-à-dire j'avais quatre kilomètres à faire quatre et quatre huit je faisais huit kilomètres tous les jours et à pieds ah oui [ESLO1_ENT_011_C]

(7) *A. et et vous par exemple approximativement vous combien de fois euh quelle est la fréquence avec laquelle vous regardez le dictionnaire c'est à dire une fois par mois une fois par an une fois par euh , oh* [ESLO1_ENT_047_C]
B. et et vous par exemple approximativement vous combien de fois euh quelle est la fréquence avec laquelle vous regardez le dictionnaire c'est à dire une fois par mois une fois par an une fois par euh , oh [ESLO1_ENT_047_C]

Dans notre modèle, nous cherchons à établir une relation de paraphrase dans la séquence *SEG1 MRP SEG2*, où le marqueur de reformulation paraphrastique fait le lien entre les deux segments. Pourtant, la détection des segments *SEG1* et *SEG2* peut être dissociée et seulement un des deux segments trouvé, comme dans les exemples (8) et (9). Il semblerait donc que le marqueur de reformulation paraphrastique n'est pas le seul élément qui marque la paraphrase. D'autres descripteurs utilisés peuvent donc aussi participer dans l'établissement de cette relation. Par ailleurs, comme noté plus haut, le segment cible est détecté plus facilement que le segment source.

(8) *A. n'est-ce pas , alors quand toutes les pièces sont coupées on le on le met en plomb c'est-à-dire qu'on prend un tout petit plomb et on met on rassemble toutes les pièces* [ESLO1_ENT_002_C]
B. n'est-ce pas , alors quand toutes les pièces sont coupées on le on le met en plomb c'est-à-dire qu'on prend un tout petit plomb et on met on rassemble toutes les pièces [ESLO1_ENT_002_C]

(9) *A. oui enfin par industriel je veux dire euh j'ai le côté commercial* [ESLO1_ENT_002_C]
B. oui enfin par industriel je veux dire euh j'ai le côté commercial [ESLO1_ENT_002_C]

Un autre type d'erreurs peut être observé avec les marqueurs, qui sont fusionnés avec les segments paraphrasés. Notons que cette erreur apparaît avec le marqueur *disons*, qui est le moins grammaticalisé dans le rôle de marqueur de reformulation paraphrastique. Notons aussi que cette fusion s'accentue avec l'agrandissement de la fenêtre dans laquelle les tokens et les descripteurs sont analysés (la série d'expériences du tableau 3).

Finalement, les énoncés avec plusieurs reformulations paraphrastiques (comme en (10)) sont mal gérées actuellement.

(10) *s- y en a sûrement mais si vous voulez euh je s- sincèrement je crois que <P1>il faut pas mettre les inconvénients en exergue</P1> enfin <MRP>je veux dire</MRP> <P2>on raisonne pas comme ça</P2> <MRP>je veux dire</MRP> euh <P3>on est heureux là où on est</P3> et je pense qu'on serait oui on essayerait de trouver euh des moyens d'être heureux donc euh y a sûrement des inconvénients mais sincèrement on les cherche pas hein et donc on les trouve pas*

7 Conclusion et Travaux futurs

Nous avons proposé un travail sur la détection automatique de segments en relation de paraphrase. Ce travail a deux points originaux principaux :
– la paraphrase étudiée est formée de manière syntagmatique dans une structure particulière de type : *SEG1 MRP SEG2*,
– le travail est effectué sur un corpus de l'oral, où les traces de reformulations sont fréquentes et peuvent être observées grâce à l'emploi de marqueurs spécifiques.
Nous avons mis en place un système d'apprentissage qui repose sur les CRF et une série de descripteurs, et de leurs combinaisons. Les descripteurs sont étudiés dans une fenêtre plus au moins grande. Parmi les meilleures expériences, nous trouvons la baseline, qui consiste en l'utilisation des formes dans une fenêtre de 7 mots et de la combinaison *forme/MRP*, et les expériences basées sur différentes combinaisons (*e.g. chunk/MRP*, *pos/MRP*, *stem/MRP*) dans une fenêtre de 7 mots. Les meilleurs résultats atteignent une moyenne allant jusqu'à 0,65 de F-mesure, 0,75 de précision et 0,63 de rappel. Nous observons qu'il est aisé de reconnaître les marqueurs de reformulation paraphrastique. Par contre, les segments en relation de paraphrase restent plus difficiles à détecter. Il s'agit surtout de la difficulté à détecter correctement leurs frontières. Le défi principal consiste en amélioration du rappel.

Nous avons plusieurs perspectives à ce travail. Tout d'abord, le travail actuel peut être amélioré de plusieurs points de vue. Nous pouvons ainsi tester d'autres classifieurs, comme les réseaux à longues mémoires court-terme (LSTM) (Schmidhuber, 1997) indiqués par un des relecteurs, et améliorer le traitement des énoncés avec plusieurs reformulations paraphrastiques. Nous pouvons aussi tester d'autres descripteurs et leurs combinaisons pour améliorer la détection des segments en relation de paraphrase. Cela concerne surtout le rappel, mais les performances globales devraient être améliorées également. Une fois stabilisé, le modèle d'apprentissage peut être utilisé pour pré-annoter d'autres tours de paroles et préparer

ainsi les données à valider et corriger par les annotateurs humains. Nous pensons que cela peut faciliter l'annotation humaine et la création des données de référence plus importantes. Nous voulons aussi tester les modèles générés avec les trois marqueurs traités ici sur les tours de parole qui contiennent d'autres marqueurs (*e.g. ça veut dire, j'allais dire, notamment, autrement dit, en d'autres termes, en d'autres mots*). De la même manière, les modèles générés peuvent être testés sur d'autres corpus. Ainsi, pouvons-nous étudier si les reformulations paraphrastiques introduites par différents marqueurs montrent des régularités communes. Lorsque les données de référence le permettent, nous prévoyons de tester les marqueurs de reformulation séparément. Cela permettra d'étudier la généricité des modèles et de la reformulation paraphrastique d'une autre manière.

D'autres perspectives consistent à appréhender ces données d'autres points de vue. Par exemple, nous pouvons prendre en compte et analyser conjointement les éléments prosodiques et acoustiques associés aux différents tours de parole. Notre hypothèse est que les tours de parole avec les reformulations paraphrastiques montrent des différences par rapport aux tours de parole avec les marqueurs étudiés mais n'introduisant pas les reformulations paraphrastiques. De cette manière, le filtrage entre ces deux types de tours de parole peut reposer sur ce critère aussi. L'étape de distinction entre les tours de parole paraphrastiques et non devrait évoluer et reposer sur des mécanismes plus robustes qu'un système à base de règles.

Nous avons aussi commencé à explorer les reformulations paraphrastiques, introduites par les mêmes marqueurs ou bien par des marqueurs différents, dans les corpus écrits (les forums de discussion et la presse journalistique). Une comparaison entre ces types de corpus et les modes de reformulation correspondent à une perspective intéressante. Par ailleurs, dans les corpus oraux d'autres paraphrases peuvent être introduites par d'autres moyens que les marqueurs de reformulation paraphrastique.

Parmi d'autres perspectives, se trouvent par exemple :
- l'exploitation de `Distagger` (Constant & Dister, 2010) pour améliorer la détection des disfluences et des répétitions,
- la création des annotations consensuelles concernant les autres paramètres (relation lexicale, morphologique, syntaxique et pragmatique),
- l'exploitation des propriétés inhérentes des segments annotés (et détectés automatiquement) pour induire ces autres paramètres. Par exemple, en (11), le segment source est beaucoup plus long que le segment cible, ce qui semble être typique de la relation pragmatique *résultat*. Tandis que dans l'exemple (12), la situation est contraire : le segment cible est plus développé que le segment source. Nous pensons que cette situation peut être typique de relation pragmatique comme *définition* ou *explication*.
- le traitement des relations de paraphrase entre différents tours de parole, alors qu'actuellement nous le faisons au sein d'un même tour de parole uniquement,
- l'étude de l'emploi des reformulations paraphrastiques en croisant les annotations avec les critères sociologiques des locuteurs.

(11) *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]

(12) *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]

Finalement, nous voudrions exploiter les données acquises dans notre travail pour les applications TAL, comme la recherche d'information ou l'inférence textuelle. Lorsque les entretiens sont effectués avec des spécialistes de domaines (*e.g.* boulangers, métiers médicaux, peintres, imprimeurs, artistes, commerçants), les personnes peuvent proposer les paraphrases ou définitions pour les termes techniques, comme dans les exemples en (13). De telles connaissances peuvent être exploitées pour effectuer la simplification lexicale des textes (Carroll *et al.*, 1998; Candido, Jr. *et al.*, 2009).

(13) {*on le met en plomb*; *on prend un tout petit plomb et on met on rassemble toutes les pièces*}
{*le ballottage*; *il reste deux personnes en présence*}
{*l'appareil digestif*; *foie vésicule pancréas et cætera*}
{*la médecine interne*; *la médecine spécialisée mais de toutes les maladies générales*}
{*le franglais*; *d'employer des mots comme snack bar*}

Remerciements

Nous remercions les relecteurs pour leurs remarques, qui nous ont permis de prendre plus de recul, d'améliorer la qualité du papier et de prévoir d'autres pistes dans les futurs travaux.

Références

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

BAKER M. (1992). *In Other Words : A Coursebook on Translation*. London, UK : Routledge.

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.

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

CANDIDO, JR. A., MAZIERO E., GASPERIN C., PARDO T. A. S., SPECIA L. & ALUISIO S. M. (2009). Supporting the adaptation of texts for poor literacy readers : a text simplification editor for Brazilian Portuguese. In *EdAppsNLP '09 Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 34–42.

CARROLL J., MINNEN G., CANNING Y., DEVLIN S. & TAIT J. (1998). Practical simplification of english newspaper text to assist aphasic readers. In *AAAI-98 Workshop on Integrating Artifficial Intelligence and Assistive Technology*, p. 7–10.

CONSTANT M. & DISTER A. (2010). *Automatic detection of disfluencies in speech transcriptions*, In C. S. PUBLISHING, Ed., *Spoken Communication*, p. 259–272.

DAGAN I., ROTH D., SAMMONS M. & ZANZOTTO F. (2013). *Recognizing Textual Entailment*. Milton Keynes, UK : Morgan & Claypool Publishers.

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.

ESHKOL-TARAVELLA I. & GRABAR N. (2014). Repérage et analyse de la reformulation paraphrastique dans les corpus oraux. In *TALN 2014*.

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

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

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

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

HATIM B. & MASON I. (1990). *Discourse and the Translator*. London, UK : Longman.

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.

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

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *ACL*, p. 504–513.

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.

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.

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.

PORTER M. (1980). An algorithm for suffix stripping. *Program*, **14**(3), 130–137.

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

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*.

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.

SCARPA F. (2010). *La Traduction spécialisée. Une approche professionnelle à l'enseignement de la traduction*. Ottawa, Canada : University of Ottawa Press. Language Arts & Disciplines.

SCHMIDHUBER S. H. J. (1997). Long short-term memory. *Neural computation*, **7**(8), 1735–1780.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

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.

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

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.

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.

VILA M., RODRÉGUEZ H. & MARTÍ M. (2014). Relational paraphrase acquisition from wikipedia : The WRPA method and corpus. *Natural Language Engineering*, p. 1–35.