Prédiction automatique de fonctions pragmatiques dans les reformulations

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

RÉSUMÉ

La reformulation participe à la structuration du discours, notamment dans le cas des dialogues, et contribue également à la dynamique du discours. Reformuler est un acte significatif qui poursuit des objectifs précis. L'objectif de notre travail est de prédire automatiquement la raison pour laquelle un locuteur effectue une reformulation. Nous utilisons une classification de onze fonctions pragmatiques inspirées des travaux existants et des données analysées. Les données de référence sont issues d'annotations manuelles et consensuelles des reformulations spontanées formées autour de trois marqueurs (*c'est-à-dire, je veux dire, disons*). Les données proviennent d'un corpus oral et d'un corpus de discussions sur les forums de santé. Nous exploitons des algorithmes de catégorisation supervisée et un ensemble de plusieurs descripteurs (syntaxiques, formels, sémantiques et discursifs) pour prédire les catégories de reformulation. La distribution des énoncés et phrases selon les catégories n'est pas homogène. Les expériences sont positionnées à deux niveaux : générique et spécifique. Nos résultats indiquent qu'il est plus facile de prédire les types de fonctions au niveau générique (la moyenne des F-mesures est autour de 0,80), qu'au niveau des catégories individuelles (la moyenne des F-mesures est autour de 0,40). L'influence de différents paramètres est étudiée.

ABSTRACT

Automatic prediction of pragmatic functions in reformulations.

Reformulations participate in structuring of discourse, especially in dialogues, and also contributes to the dynamics of the discourse. Reformulation is a significant act which has to satisfy precise objectives. The purpose of our work is to automatically predict the reason for which a speaker performs a reformulation. We use a classification with eleven pragmatic functions inspired by the existing work and by the data analyzed. The reference data are built through manual and consensual annotations of spontaneous reformulations introduced by three markers (*c'est-à-dire, je veux dire, disons*). The data are provided by spoken corpora and a corpus with forum discussions on health issues. We exploit supervised categorization algorithms and a set with several descriptors (syntactic, formal, semantic and discursive) for the prediction of the reformulation categories. The distribution of utterances and sentences is not homogeneous across categories. The experiments are positioned at two levels : general and specific. Our results indicate that it is easier to predict the types of functions at the general level (the average F-measure is around 0.80), than at the level of individual categories (the average F-measure is around 0.40). We study the influence of various parameters.

MOTS-CLÉS : Reformulation, apprentissage automatique, paraphrase, classification, fonction pragmatique.

KEYWORDS: Reformulation, machine learning, paraphrase, classification, pragmatic function.

1 Introduction

La reformulation consiste à reprendre ou redire quelque chose qui a déjà été dit. Elle est effectuée à la demande de l'interlocuteur ou par la volonté du locuteur. Cette notion est centrale dans notre travail. Les cadres applicatifs potentiels, qui impliquent la reformulation, concernent par exemple la recherche et l'extraction d'information, où il est nécessaire de détecter les segments équivalents afin d'augmenter le rappel, ou la traduction automatique, où il est nécessaire d'éviter des répétitions. L'objectif de notre travail est d'analyser les segments reformulés et surtout de prédire automatiquement la fonction pragmatique associée à chaque reformulation : étudier les raisons qui poussent les locuteurs à effectuer ces reformulations (donner une précision, définir, expliquer...). Une des utilités de ce travail consiste à travailler sur le phénomène de la reformulation et de la fonction pragmatique, qui sont des notions assez complexes à cerner et à décrire (section 1.2). Notre hypothèse est que le contenu des segments reformulés fournit les indices, qu'ils soient non linguistiques (*e.g.* taille des segments) ou linguistiques (*e.g.* lexicaux, syntaxiques, sémantiques, etc.), pour la prédiction des fonctions pragmatiques. Le coeur de notre travail et les expriences sont positionnés à deux niveaux :

– à un niveau général, selon le type de transformations linguistiques associées aux reformulations : ajout, suppression ou encore volume constant d'information ;
– à un niveau spécifique, exploitant les fonctions pragmatiques précises. Il s'agit de catégories comme la définition, l'explication, le résultat ou la précision, décrites dans la section 1.2.

Le travail présenté concerne la reformulation spontanée dans le discours oral et dans les discussions sur le web. La reformulation est introduite par trois marqueurs formés sur le verbe *dire* (*c'est-à-dire, je veux dire, disons*) dans la structure $S1$ *marqueur* $S2$.

Dans la suite de ce travail, nous décrivons d'abord les travaux de l'état de l'art (section 1.1) et cernons les fonctions pragmatiques (section 1.2). Nous présentons ensuite les données traitées (section 2) et les méthodes proposées (section 3). Finalement, nous présentons et discutons les résultats (section 4), et terminons avec les perspectives à ce travail (section 5).

1.1 Travaux de l'état de l'art

Travaux en linguistique de textes écrits. Dans la langue écrite, la reformulation est liée à plusieurs notions, plus ou moins proches :

– *Paraphrase.* La reformulation peut être vue comme la variante paraphrastique d'un segment linguistique dans laquelle des modifications formelles sont opérées (Neveu, 2004). Dans ce cas, la paraphrase apparaît comme le résultat d'une reformulation. La paraphrase est étudiée de différents points de vue : dans sa situation d'énonciation (Culioli, 1976; Flottum, 1995; Fuchs, 1994; Martin, 1976; Vezin, 1976) ; à travers les transformations linguistiques que subissent les segments paraphrasés à différents niveaux (Melčuk, 1988; Vila *et al.*, 2011; Bhagat & Hovy, 2013) ; en fonction de la taille d'entités paraphrasées (Flottum, 1995; Fujita, 2010; Bouamor, 2012).
– *Glose.* Ce terme, issu de la tradition philologique, désigne un commentaire sur un mot. Il impose au premier segment d'être une unité lexicale, alors que le deuxième segment correspond à la glose, souvent écrite en langage formel ou semi-formel (Authier-Revuz, 1995; Steuckardt, 2005).
– *Reprise.* Un terme plus générique, reprise, correspond à des pures et simples répétitions d'un segment textuel aux différents degrés de ses reformulations (Vion, 2006). La proximité sémantique entre les segments refomulés apparaît être un critère caractéristique de la reformulation.

- *Description.* Dans des études littéraires, la reformulation est liée avec la notion de description (Magri-Mourgues, 2013). Trois types sont alors distingués :
  - reformulation par addition : lorsque le même référent a plusieurs dénominations. Dans les exemples cités, l'objet est décrit de manière assez extensive en expliquant sa fonction, alors que les reprises anaphoriques assurent la cohésion et la cohérence au texte ;
  - reformulation par substitution (ou reformulation corrective) : la seconde occurrence tend à effacer la première, dans un mouvement de correction. Dans les exemples présentés, les deux dénominations successives ({*charmilles*; *dentelles*}, {*la neige*; *la glace*}) sont liées par une relation d'analogie, qui, activant certains sèmes communs, rendent la juxtaposition cohérente ;
  - reformulation par superposition : la reformulation se limite au cadre phrastique sans établir de hiérarchie entre les unités successives. Il s'agit de la traduction inter-linguale : {*djahels*; *ignorants*}, {*cuevas*; *habitations troglodytes*}, {*chanteuses*; *oualems*}, {*danseuses*; *ghavasies*}.
- *Élaboration.* Le projet Annodis, consacré à la constitution du corpus de référence annoté en structures discursives, distingue une relation rhétorique appelée élaboration qui semble se rapprocher du phénomène de reformulation (Péry-Woodley *et al.*, 2009).

Travaux en linguistique du discours parlé. La langue orale diffère cependant de l'écrit car on assiste à son élaboration à la différence du produit final que l'on peut trouver dans l'écrit (Blanche-Benveniste *et al.*, 1991). En effet, l'écrit concrétise une version définitive du discours quand l'oral le présente dans son processus avec ses hésitations, ses faux départs, ses ratures et ses reformulations. De nombreux travaux s'intéressent à la reformulation dans l'oral car elle en constitue une des caractéristiques fondamentales. Plusieurs points de vue sont possibles, mais la reformulation y est fortement associée aux disfluences et auto-réparations :

- Le terme de reformulation est utilisé dans le cadre des analyses d'interactions verbales (Gülich & Kotschi, 1987; Roulet, 1987). Deux types de reformulations sont distingués : les reformulations paraphrastiques, qui instaurent une équivalence entre les segments reformulés, et les reformulations non paraphrastiques, qui opèrent un changement de perspective énonciative (Rossari, 1990). Comme tout acte de reformulation dans l'oral n'introduit pas toujours une paraphrase, deux catégories de marqueurs sont distinguées : les marqueurs de reformulation paraphrastique (MRP), comme *c'est-à-dire, je veux dire, je m'explique, en d'autres termes etc.* qui ont pour tâche principale d'établir une relation paraphrastique, et les marqueurs de reformulation non-paraphrastique, comme *en somme, en tout cas, de toute façon, enfin, etc.*, qui montrent ce rôle dans des contextes précis. En outre, les propriétés sémantiques des MRP permettent d'instaurer une relation de paraphrase même entre des segments qui n'entretiennent aucune équivalence sémantique constatable.
- Les études syntaxiques sur le langage oral ont rapproché le phénomène de reformulation avec celui d'énumération ou de répétition (Levelt, 1983; Blanche-Benveniste *et al.*, 1991; Benzitoun, 2004). Dans tous les cas, il s'agit d'un même procédé syntaxique : les éléments répétés, reformulés ou énumérés ont une même place syntaxique dans l'énoncé sur un axe paradigmatique. La distinction est pourtant possible grâce à l'utilisation des indices formels, tels que les marqueurs d'énumération et de reformulation, ou bien des accords.
- La reformulation peut aussi être associée au procédé de correction ou de précision, comme c'est le cas dans l'annotation multi-niveau de l'oral dans le corpus Treebank Rhapsodie (Kahane & Pietrandrea, 2012). Les auteurs utilisent la notion de reformulation pour présenter la typologie des entassements à l'oral qui, suite à d'autres travaux (Blanche-Benveniste *et al.*, 1991; Blanche-Benveniste, 1995; Bilger, 1999; Guénot, 2006), peuvent être utilisés pour établir des relations entre dénotations, créer de nouvelles dénotations, reformuler, exemplifier, préciser ou encore intensifier. Les auteurs introduisent également la notion de reformulation dénotative.

Travaux de TAL. Dans les travaux de TAL, la reformulation dans le corpus écrits est très souvent associée à la paraphrase, qui est vue comme le résultat de la reformulation. Ceci rappelle les positions de certains linguistes. Les questions de recherche posées en TAL concernent d'une part la détection automatique des paraphrases (Barzilay & McKeown, 2001; Shinyama *et al.*, 2002; Bannard & Callison-Burch, 2005; Malakasiotis & Androutsopoulos, 2007; Lin & Pantel, 2001) et d'autre part leur utilisation dans différentes applications, comme par exemple la détection de plagiat (Ferrero & Simac-Lejeune, 2015), l'inférence textuelle (Androutsopoulos & Malakasiotis, 2010; Dagan *et al.*, 2013), la normalisation de langages contrôlés (Nasr, 1996), la recherche d'information et la traduction automatique (Madnani & Dorr, 2010; Bouamor, 2012).

Dans les travaux de TAL qui portent sur les corpus oraux, la reformulation est proche des disfluences et les travaux existants concernent souvent sa détection automatique. Mentionnons par exemple les méthodes à base de règles et de patrons élaborés manuellement (Bouraoui & Vigouroux, 2004; Constant & Dister, 2010) et une méthode par apprentissage supervisé (Dutrey *et al.*, 2014) pour la détection de disfluences. Dans les travaux sur l'oral, la détection de reformulations est liée à la réparation des énoncés, en vue de les nettoyer, et à leur reconstitution. Il s'agit également de pouvoir décrire et formaliser les reformulations et disfluences dans ce type de données.

1.2 Fonctions pragmatiques de la reformulation

Reformuler est un acte qui est toujours significatif et qui poursuit des objectifs précis. C'est ce que nous appelons la fonction pragmatique de la reformulation, à savoir le rôle de la reformulation spontanée que l'on peut observer dans l'oral ou dans les discussions sur le web. La reformulation met en relation deux segments : le segments reformulé $S1$ et le segment qui contient la refomulation $S2$. Dans notre étude, la reformulation est établie grâce à des marqueurs formés sur le verbe *dire* (*c'est-à-dire, je veux dire, disons*) au sein de la structure *S1 marqueur S2*.

Nous distinguons plusieurs fonctions pragmatiques entre $S1$ et $S2$, qui sont inspirée des typologies proposées dans la littérature (Gülich & Kotschi, 1987; Hölker, 1988; Beeching, 2007; Kanaan, 2011) et motivées par nos données. Des exemples provenant de nos corpus sont également présentés :

- *Définition* : un terme dans $S1$ est défini dans $S2$. Il s'agit souvent de termes techniques spécialisés. La définition est neutre et précise, dont l'objectif est de faire comprendre une notion technique :
  - *la TVA je n'ai jamais bien compris ce que c'était eh ben c'est-à-dire c'est une taxe c'est une taxe à la valeur qu'on ajoute et la taxe est toujours comptée sur la valeur qui est ajoutée autant de fois comme la marchandise est ah bon , est transaxée oui et c'est une taxe qui est ajoutée à la transaction encore une taxe sur la va- sur la marge bénéficiaire* [eslo1-011]
  - *des rumeurs euh disons qu'en fait il y a des événements qui se passent et après on s- on s'en fait tu vois euh je te dis un truc tu le répètes à quelqu'un au bout de trois personnes ça a déjà pas forcément même voire pas du tout le même sens* [eslo2-12]
  - *avec une ETO c'est à dire une echographie tansoesophagienne (une écho ou le palpeur est introduit dans l'estomac) [forum]*
- *Explication* : le locuteur explique quelque chose à son interlocuteur ($S2$ explique $S1$). Pour vérifier la fonction, on peut remplacer le marqueur par *parce que*. L'explication est similaire à la définition tout en étant moins formelle. De plus, elle porte sur des situations. Cette relation est proche du lien de cause à effet dans les annotations Annodis (Péry-Woodley *et al.*, 2009).
  - *ce garçon je sais bien qu'il ne peut pas se marier avec euh c'est-à-dire qu'il aurait pu avtrouver une jeune fille euh qui fasse sa licence euh dans un milieu comme le nôtre* [eslo1-010]

  - *on a apprécié un peu plus le voyage c'est-à-dire qu'on ouais hm hm hm on a rencontré voilà des des des des vrais gens on va dire qui sont pas intéressés que par notre pognon qui ouais ouais ouais pour le coup eux nous ont invité donc nous ont ouvert leur porte et hm hm hm hm c'est même eux qui nous ont euh donné des choses quoi et on a bien pu discuter* [eslo2-10]
  - *j'ai entendu parler (sur le net) des bêtabloquants, or il parait que céest des médicaments a vie et pour la vie, c'est-à-dire ils ne sont efficaces que lorsquéils sont pris tous les jours* [forum]
- *Exemplification* : Le locuteur donne des exemples dans $S2$ d'une entité mentionnée dans $S1$. Ainsi, $S2$ peut comporter des entités nommées ou des énumérations :
  - *des morceaux nobles ce qu'ils appellent quoi c'est à dire les rosbifs les biftecks et tout ça* [eslo1-001]
  - *y a un peu de règles c'est-à-dire que euh oui en règle générale effectivement on regarde pas la télé le soir quand on a classe le lendemain surtout quand on est en sixième on est enc- on a encore besoin de dormir* [eslo2-5]
  - *2 heures plus tard, elle a eu tous les symptomes d'un avc c'est à dire perte de parole, hémiplégie, fièvre...* [forum]
- *Justification* : le locuteur justifie quelque chose (des événements, des actes) à son interlocuteur. Dans ce cas, $S2$ propose une justification de $S1$:
  - *la langue française est plus difficile disons on peut pas dire la plus difficile des langues européennes mais c'est difficile* [eslo1-007]
  - *ça c'est c'est un peu connu c'est à dire c'est alors j'ai pas à le dire pas vraiment* [eslo2-21]
  - *Je voudrais mieux comprendre pour mieux pouvoir aider ....merci pour votre aide, bien évidement vous n'êtes absolument pas obligés de me répondre si vous ne le souhaitez pas, mais disons que votre vécu m'apporterai un plus* [forum]
- *Précision* : c'est une fonction assez large. Elle marque la volonté du locuteur d'ajouter une information dans le but d'éclaircir $S1$. Elle ressemble à la relation *élaboration* dans les annotations Annodis (Péry-Woodley *et al.*, 2009) et donne plus de détails sur l'événement décrit dans $S1$:
  - *je lis oui l'Equipe depuis l'âge de dix-sept ou dix-huit ans c'est-à-dire avant c'était l'Auto mais j'ai toujours ah oui toujours toujours toujours toujours lu l'Equipe* [eslo1-045]
  - *les aînés partent eux aussi de manière moins systématique c'est-à-dire que les aînés partent pas forcément tous les ans mais souvent* [eslo2-5]
  - *La trinitrine m'a été prescrite vendredi dernier, c'est à dire depuis une semaine* [forum]
- *Dénomination* : il s'agit de l'attribution d'un nom à une entité unique mentionnée dans $S1$, ce qui la différencie de *exemplification* où l'existence d'autres entités du même type est présupposée :
  - *en particulier c'est l'endroit où en somme ça s'est produit le plus au début c'est-à-dire à Nanterre* [eslo1-058]
  - *depuis le début c'est à dire que j'ai fait euh [...] depuis soixante-dix-sept* [eslo2-16]
  - *depuis qu'on m'avait changer de traitement c'est a dire le nebilox* [forum]
- *Résultat* : le locuteur résume ou bien indique la conséquence de $S1$. Le marqueur marque une conclusion ou une conséquence par rapport à $S1$, qui peut être implicite ou explicite. Le marqueur peut être remplacé par exemple par *pour en somme, en bref, en conclusion, en résumé, donc*. Cette relation est proche du lien de cause à effet dans le corpus Annodis (Péry-Woodley *et al.*, 2009) :
  - *avec l'accent un peu de travers je veux dire l'accent* [eslo1-002]
  - *quand je rentrais le soir à la maison que mes enfants me demand- me faisaient une petite demande ou une petite crise je me disais non non mais attendait là on va on va recadrer tout ça euh non enfin je veux dire on a un un super décalage* [eslo2-2]
  - *A ma sortie, j'ai retrouvé pratiquement l'usage de ma jambe, de mon bras et ma main gauche, disons que je pouvais être autonome* [forum]

- *Correction linguistique* : $S2$ apporte une correction linguistique (article, nombre...) de $S1$. En général, il s'agit d'une correction faite à l'initiative du locuteur :
  - *des artisans euh hm hm hm hm hm alors c'est-à-dire artisans* [eslo2-13]
- *Correction référentielle* : $S2$ apporte une correction de lieu, de temps, etc. Il s'agit également d'une correction faite à l'initiative du locuteur :
  - *jusqu'à seize ans oui oui bien bon c'est-à-dire euh dans le primaire privé* [eslo1-010]
  - *j'habitais rue Lazare Carnot c'est à dire donc au sud de la Source* [eslo2-16]
- *Paraphrase* : $S1$ répète l'information de $S2$, mais d'une autre manière, et on ne voit pas de différence entre les deux. Cette fonction concerne aussi les répétitions identiques :
  - *quelque chose de potable disons quelque chose euh de correct* [eslo1-007]
  - *toujours les mêmes c'est-à-dire euh tous ceux qu'on connait* [eslo2-4]
  - *Il n'a acune maladie (je veux dire qu'il ne prend aucun médicament* [forum]
- *Opposition* : $S1$ reprend l'information de $S2$ sous forme négative :
  - *elle était incapable de rien faire elle au point de vue vendeuse c'est-à-dire elle elle est pas mauvaise euh elle est agréable au point de vue clientèle elle a été incapable de passer son certificat d'études* [eslo1-045]

Comme nous voyons, les emplois de marqueurs de reformulation observés dans les corpus étudiés dépassent largement le phénomène de la paraphrase, qui présuppose une équivalence sémantique entre les expressions paraphrasées, et couvrent un ensemble de situations très large.

2 Données traitées et exploitées

Nous travaillons avec plusieurs types de données : (1) deux types de corpus (deux corpus ESLO et les forums de discussions médicales (section 2.1)), (2) les segments en relation de reformulation (section 2.2) obtenus suite à l'annotation manuelle consensuelle des corpus, et (3) plusieurs ressources linguistiques (section 2.3).

2.1 Corpus

ESLO. Les corpus ESLO (Enquêtes Sociolinguistiques à Orléans) (Eshkol-Taravella *et al.*, 2012) : *ESLO1* et *ESLO2* sont des corpus oraux de la langue française. *ESLO1*, la première enquête sociolinguistique à Orléans, a été réalisée en 1968-1971. Ce corpus comprend 300 heures de parole (4 500 000 mots environ) et inclut 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 [1].

Forum de discussion. Le corpus *forum* est collecté sur le forum de discussions *Hypertension* de Doctissimo [2]. Ce corpus fournit 12 588 fils de discussion contenant 67 652 messages et 6 788 361 occurrences de mots. Les messages de ce corpus sont écrits par des internautes, qui ont besoin de s'exprimer sur leurs maladies. Il s'agit des écrits non normés, qui peuvent contenir des erreurs

1. http://eslo.tge-adonis.fr/
2. http://forum.doctissimo.fr/sante/hypertension-problemes-cardiaques/liste_sujet-1.htm

d'orthographe et de syntaxe, et d'autres éléments linguistiques non conventionnels (abréviations spécifiques, émoticônes...).

2.2 Segments en relation de reformulation

Nous disposons de 4 120 énoncés ou phrases comportant les trois marqueurs de reformulation étudiés. Ces énoncés et phrases proviennent des corpus analysés : *ESLO1*, *ESLO2* et *forum*. Dans le corpus *forum*, une phrase correspond à une séquence linguistique séparée par la ponctuation forte. Dans les corpus oraux, la segmentation en énoncés est faite en fonction des groupes de souffle, des tours de parole mais prend également en compte les chevauchements où les locuteurs parlent en même temps. Ainsi, en cas de chevauchement, l'énoncé du locuteur, qui continue de parler après ce chevauchement, continue. Ces phrases et énoncés sont annotés par deux annotateurs indépendants et sont soumis ensuite à des séances de consensus. L'accord inter-annotateur est calculé avec le kappa de Cohen (Cohen, 1960). Dans le tableau 1, nous indiquons les accords constatés sur la décision quant à la présence de reformulations, et leur interprétation standard (Landis & Koch, 1977). Il s'agit d'un accord fort et modéré. Lorsque l'accord inter-annotateur est calculé au niveau des fonctions pragmatiques, il est extrêmement faible : 0,127 sur *ESLO1* et 0,0211 sur *ESLO2*.

Corpus	*Accord*	*Interprétation*
ESLO1	0,617	Accord fort
ESLO2	0,526	Accord modéré
Forum	0,784	Accord fort

TABLE 1 – Accord inter-annotateur sur la présence de reformulations dans les énoncés et phrases.

Fonction	*ESLO1*	*ESLO2*	*ESLO*	*forum*	*total*
cor-ling	-	2 (0)	2 (0)	-	2 (0)
cor-ref	5 (3)	1 (0)	6 (2)	-	6 (1)
def	16 (10)	14 (8)	30 (9)	41 (16)	71 (12)
denom	2 (1)	3 (1)	5 (1)	24 (9)	29 (5)
exempl	29 (18)	15 (9)	44 (13)	21 (8)	65 (11)
explic	26 (16)	16 (9)	42 (13)	25 (10)	67 (11)
justif	1 (0)	8 (5)	9 (3)	8 (3)	17 (3)
oppo	2 (1)	-	2 (0)	-	2 (0)
para	14 (9)	18 (10)	32 (10)	20 (8)	52 (9)
prec	47 (29)	54 (31)	101 (30)	88 (34)	189 (32)
res	19 (12)	43 (25)	62 (18)	32 (12)	94 (16)
total	161	174	335	259	594

TABLE 2 – Distribution de phrases et énoncés selon les fonctions pragmatiques et les corpus. *ESLO=ESLO1+ESLO2 ; total=ESLO+forum*. Les pourcentages sont indiqués entre les parenthèses.

Parmis les 4 120 emplois de marqueurs, 594 occurrences introduisent les reformulations. Dans le tableau 2, nous indiquons la distribution de ces emplois selon les corpus et les fonctions pragmatiques. Le corpus *ESLO* est composé des deux corpus oraux *ESLO1* et *ESLO2* ; le corpus *total* est composé

de tous les corpus disponibles (*ESLO* et *forum*). Il s'agit des données que nous proposons d'étudier et que nous voulons prédire automatiquement. Nous voyons que la fonction pragmatique la plus utilisée est *précision*. Les fonctions les plus rares sont *correction linguistique* et *opposition*, certainement parce qu'elles ne sont pas souvent marquées par les marqueurs étudiés. Leur pertinence peut être reconsidérée. D'autres fonctions ne sont pas distribuées de la même manière dans les corpus. *Définition* est très fréquente dans les discussions sur le web, qui traitent les questions médicales : la présence de termes médicaux et de leurs définitions y est importante. *Exemplification* et *explication* sont très utilisées dans *ESLO1* et *ESLO2*, sans doute parce que l'intervieweur est d'origine anglaise. *Justification* a la même distribution dans *ESLO2* et *forum* : les discussions sont plus libres dans ces deux corpus, alors que les conditions d'enregistrement dans *ESLO1* sont plus formelles et, comme il a été remarqué ci-dessus, les intervieweurs ont été d'origine anglaise. *Paraphrase* est employée d'une manière plus au moins comparable dans les trois corpus. *Dénomination* est peu présente dans le corpus oral, contrairement au *forum*. Dénommer un médicament, un traitement sont les cas observés dans le corpus du web. Ces remarquent montrent que la nature des corpus et le contexte de leur production guident et influencent l'utilisation du processus de reformulation chez le locuteur.

2.3 Ressources linguistiques

Nous exploitons plusieurs types de ressources : (1) une liste de mots vides ; (2) les marqueurs de disfluence ; (3) les clusters distributionnels de mots générés à partir de nos corpus ; (4) un lexique d'hyponymes ; (5) les marqueurs lexicaux qui peuvent être associés aux fonctions pragmatiques.

Mots vides. Les mots vides (n=69) correspondent surtout aux mots grammaticaux du français. Ils sont utilisés pour alléger les traitements et pour se concentrer sur les mots non grammaticaux.

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*.

Clusters de mots. Les clusters distributionnels de mots sont générés à partir des corpus de notre travail : *ESLO1*, *ESLO2*, *ESLO* (la fusion de *ESLO1* et *ESLO2*), *forum* et tous les corpus pris ensemble (*total*). Les corpus sont segmentés, la casse est réduite vers les minuscules, les mots vides sont éliminés. Les clusters sont générés en exploitant les algorithmes de clusterisation (Brown *et al.*, 1992; Liang, 2005). Il s'agit d'un clustering hiérarchique agglomeratif basé sur l'information distributionnelle des mots. Au sein d'un cluster, les mots sont reliés sémantiquement car ils apparaissent dans des contextes similaires. Nous générons des ressources distributionnelles avec 200 à 600 clusters.

Hyponymes. Un lexique d'hyponymes est extrait automatiquement à partir de la ressource Wiktionary [3] en français. La structure des articles de Wiktionary est exploitée pour extraire les libellés de l'entrée et de ses hyperonymes. Le lexique contient 12 161 couples {*hypéronyme*; *hyponyme*}, comme par exemple {*lexique*; *dictionnaire*}, {*armée*; *légion*}, {*disque*; *CDROM*}, {*période*; *année*}. Les mots, qui se trouvent dans un même couple, ont un lien sémantique fort.

Marqueurs lexicaux. Nous utilisons un petit ensemble de marqueurs (n=17), qui sont associés aux fonctions pragmatiques. Trois types de marqueurs sont distingués : (1) les marqueurs introductoires (*e.g. voilà, c'est, ce sont*), qui peuvent marquer les définitions ; (2) les marqueurs de cause (*e.g. c'est pourquoi, parce que, car*), qui peuvent apparaître avec *résultat* ; (3) les marqueurs d'exemplification (*e.g. exemple, comme, entre autre*), qui peuvent apparaître avec la fonction *exemplification*.

3. https://fr.wiktionary.org/wiki/Wiktionnaire

3 Méthodes

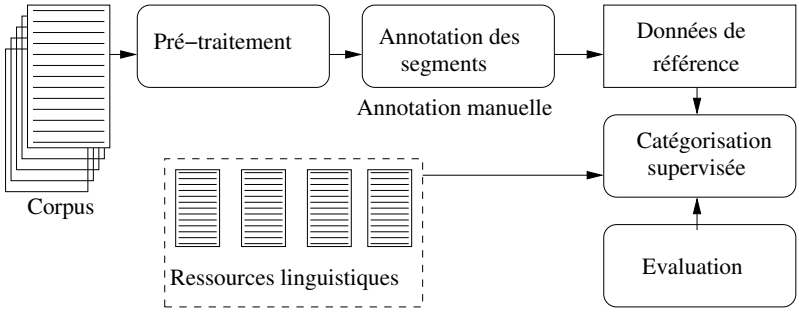


Fig. 1 – Schéma général de la méthode.

La figure 1 présente le schéma général de la méthode. Les étapes principales sont : (1) le prétraitement et la création des données de référence ; (2) la catégorisation supervisée des segments pour prédire leur fonction pragmatique ; et (3) l'évaluation. Nous effectuons la catégorisation supervisée en exploitant la plateforme Weka (Manning & Schütze, 1999) et plusieurs des algorithmes dans leur configuration standard : `J48` et `REPTree` (Quinlan, 1993), `RandomForest` (Breiman, 2001), `SMO` (Platt, 1998), `DecisionTable` (Kohavi, 1995), `OneR` (Manning & Schütze, 1999). Nous décrivons les données de référence, les catégories et les descripteurs utilisés, et les modalités de l'évaluation.

Données de référence. Les données de référence sont obtenues suite à des annotations manuelles et consensuelles des phrases et énoncés. Le tableau 2 présente ces données, selon les fonctions pragmatiques et les corpus. Deux types de corpus sont traités : les corpus oraux *ESLO* et le corpus de discussions sur le web *forum*. Nous pouvons voir qu'entre ces deux types de corpus, les reformulations sont distribuées de manière homogène, tandis que nous y observons une surreprésentation de plusieurs fonctions, comme *précision*, *résultat*, *définition*, *explication* et *exemplification*.

Catégories. Les catégories correspondent aux fonctions pragmatiques du tableau 2. Comme trois catégories sont très peu peuplées (*correction linguistique, correction référentielle, opposition*), nous faisons également des expériences avec les huit catégories les plus fréquentes sur l'ensemble de données. De plus, une autre expérience est positionnée à un niveau plus général, selon le volume d'information fournie lors de la reformulation et mesuré par la taille des segments :
– suppression d'information dans $S2$ par rapport à $S1$: *résultat, dénomination* ;
– ajout d'information par rapport à $S1$: *définition, exemplification, explication, justification, précision* ;
– volume d'information comparable : *paraphrase, correction linguistique, correction référentielle, opposition*.
Cette typologie ressemble à celle proposée dans un travail existant (Magri-Mourgues, 2013), mais nous distinguons en plus la suppression d'information dans $S2$. Cela nous permet d'effectuer des expériences à deux niveaux : niveau générique avec trois catégories (ajout, suppression et stabilité du volume d'information), et niveau spécifique avec huit catégories.

Descripteurs. Nous utilisons plusieurs descripteurs pour cerner la nature des fonctions pragmatiques. Les valeurs de tous les descripteurs sont transformées en valeurs numériques :
– la longueur des segments $S1$ et $S2$, en mots et en caractères,
– la différence de longueur des segments $S1$ et $S2$, en mots et en caractères,
– l'équivalence entre les catégories syntaxiques des deux segments,
– si la catégorie syntaxique des deux segments est un groupe nominal ou une proposition,
– la présence des segments ou de leurs mots dans les mêmes clusters : tous les mots, tous les mots sauf les mots

identiques, tous les mots sauf les mots vides, tous les mots sauf les mots vides et identiques. Les nombres et les pourcentages de mots partagés sont calculés. Nous utilisons plusieurs ensembles de clusters : ils sont calculés sur différents corpus (*ESLO1*, *ESLO2*, *ESLO*, *forum* et tous les corpus pris ensemble (*total*)) et avec des nombres différents de clusters à générer (nous retenons 300 et 600 clusters dans l'analyse des résultats),
– la présence de marqueurs de disfluence dans les segments,
– la présence de nombres dans les segments,
– la présence de marqueurs lexicaux spécifiques de exemplifications, de cause et de structures introductoires,
– la présence des segments ou de leurs mots dans les couples reliés par la relation d'hyperonymie.

Comme nous voyons, ces descripteurs se positionnent à différents niveaux : formel, syntaxique, sémantique et discursif. Ces descripteurs sont calculés automatiquement, en exploitant ou non des ressources linguistiques.

Évaluation. L'évaluation est effectuée avec des mesures classiques en TAL : précision, rappel et F-mesure. Nous présentons les résultats de cette évaluation telle que calculés par la plateforme Weka. Par ailleurs, nous effectuons une validation croisée à 10 plis : les données sont partagées en 10 ensembles et, à chaque itération, un ensemble sert à effectuer l'entraînement alors que les autres ensembles servent pour le test. L'évaluation finale correspond à la moyenne des évaluations de chaque itération.

4 Résultats

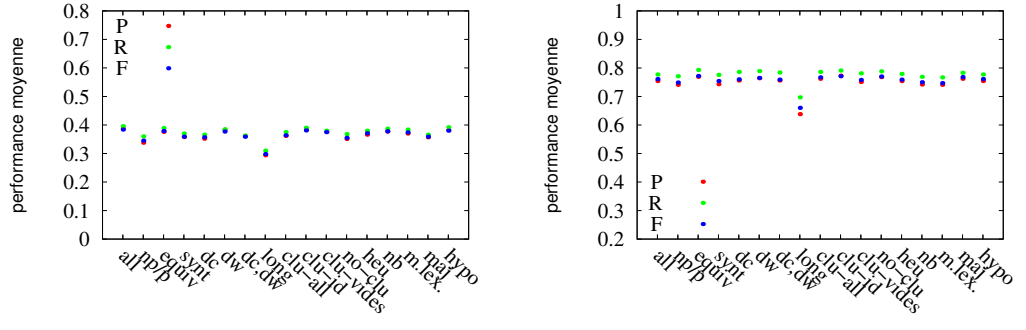
Fonct.	J48			REPTree			RandomForest			SMO			DecisionTable		
	P	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
def	0,28	0,30	0,29	0,34	0,17	0,22	0,39	0,31	0,34	0,13	0,01	0,03	0,24	0,17	0,20
denom	0,27	0,24	0,26	0,00	0,00	0,00	0,17	0,10	0,13	0,40	0,07	0,12	0,25	0,03	0,06
exempl	0,21	0,22	0,21	0,19	0,11	0,14	0,32	0,26	0,29	0,44	0,26	0,33	0,13	0,03	0,05
explic	0,23	0,22	0,23	0,31	0,13	0,19	0,28	0,31	0,30	0,00	0,00	0,00	0,00	0,00	0,00
justif	0,00	0,00	0,00	0,00	0,00	0,00	0,50	0,11	0,19	0,00	0,00	0,00	0,00	0,00	0,00
para	0,24	0,23	0,24	0,37	0,19	0,25	0,33	0,23	0,27	0,20	0,02	0,04	0,32	0,23	0,27
prec	0,33	0,34	0,33	0,39	0,66	0,49	0,41	0,52	0,46	0,37	0,89	0,53	0,35	0,66	0,46
res	0,48	0,48	0,48	0,49	0,64	0,55	0,55	0,60	0,57	0,66	0,47	0,56	0,52	0,60	0,55
moyen.	0,30	0,31	0,30	0,33	0,38	0,33	0,39	0,40	0,38	0,32	0,40	0,31	0,28	0,36	0,30
equal	0,35	0,17	0,23	0,25	0,04	0,07	0,44	0,19	0,27	0,00	0,00	0,00	0,00	0,00	0,00
minus	0,53	0,58	0,55	0,61	0,59	0,60	0,56	0,52	0,54	0,65	0,26	0,37	0,60	0,49	0,54
plus	0,84	0,88	0,86	0,83	0,92	0,87	0,83	0,90	0,87	0,78	0,98	0,87	0,81	0,94	0,87
moyen.	0,74	0,76	0,75	0,74	0,79	0,76	0,75	0,78	0,76	0,69	0,77	0,70	0,71	0,78	0,74

TABLE 3 – Performances de différents algorithmes dans la prédiction des fonctions pragmatiques : huit et trois catégories, l'ensemble de corpus, clusters générés sur l'ensemble de corpus.

Dans le tableau 3, nous indiquons les performances de différents algorithmes. Il s'agit de l'expérience pour la prédiction de huit (partie haute) et trois (partie basse) catégories, respectivement. Tous les corpus (*ESLO* et *forum*) sont traités ensemble, la ressource distributionnelle est également générée sur l'ensemble de corpus, avec 600 clusters, tous les descripteurs sont utilisés. Nous pouvons voir que `RandomForest` optimise la prédiction des catégories et ne se focalise pas sur certaines d'entre elles. Notons que ceci est aussi le cas avec `J48`, mais les moyennes obtenues avec `J48` sont un peu moins bonnes. Avec `RandomForest`, nous avons une moyenne de 0,38 pour les huit catégories, et 0,76 pour les trois catégories. Nous continuons la présentation des résultats avec `RandomForest` et le même paramétrage que dans le tableau 3.

Différents paramètres et descripteurs influencent les résultats :

– la figure 2 indique la moyenne des performances (précision, rappel, F-mesure) obtenues lorsque différents descripteurs sont supprimés de l'ensemble de descripteurs. Globalement, les performances restent proches de celles obtenues avec l'ensemble total des descripteurs (l'expérience avec tous les descripteurs *all* apparaît en première position sur la courbe) : 0,4 avec huit catégories et 0,8 avec trois catégories. Notons cependant que la suppression de certains descripteurs (équivalence des catégories syntaxiques *équiv*, informations sur les clusters (*clu-all, clu-in, clu-vides, no-clu*), caractères majuscules *maj*) est bénéfique pour la prédiction des trois catégories (figure 3(b)), alors qu'avec huit catégories l'ensemble total des descripteurs est toujours plus efficace que lorsque la suppression de certains d'entre eux est effectuée (figure 3(a)). La suppression des informations sur la longueur de $S1$ et $S1$ *long* conduit toujours à une détérioration importante ;
– avec l'ensemble total de descripteurs, le descripteur le plus efficace est celui qui indique la différence de longueur en caractères entre $S1$ et $S2$. Avec ce descripteur utilisé seul, la F-mesure globale est 0,28 et 0,73, avec huit et trois classes respectivement, en gardant les paramètres du tableau 3. D'autres descripteurs liés à la longueur de $S1$ et $S2$ sont aussi importants. Lorsque ces descripteurs sont supprimés, c'est la présence de marqueurs de disfluence qui est retenue comme le meilleur descripteur ;
– la ressource distributionnelle ne montre pas d'influence entre le corpus *total* et le corpus *forum*. En revanche, avec *ESLO2*, il est préférable d'avoir les ressources distributionnelles générées sur le même corpus ou bien sur les deux corpus *ESLO* : la nature et le contenu du corpus oral *ESLO2* restent sans doute spécifiques ;
– la figure 3 indique la reconnaissance des fonctions pragmatiques dans trois corpus (*ESLO*, *forum* et *total*). Cette figure reprend en partie les données du tableau 3 pour le corpus *total*. Avec huit catégories, nous voyons que la fonction *résultat* est la mieux reconnue dans tous les corpus. *précision* et *définition* montrent une prédiction moins bonne mais également stable selon les corpus. *paraphrase* est assez bien reconnue dans *forum*, mais plus faiblement dans les autres corpus, alors que *exemplification*, *explication* et *justification* sont assez bien reconnus dans les corpus *ESLO*. Avec trois catégories, la catégorie *plus* est la mieux reconnue. Ces observations doivent être liées avec le volume de données de référence dans chaque corpus et catégorie : *résultat*, *précision* et, par conséquent *plus*, sont les catégories les plus peuplées.

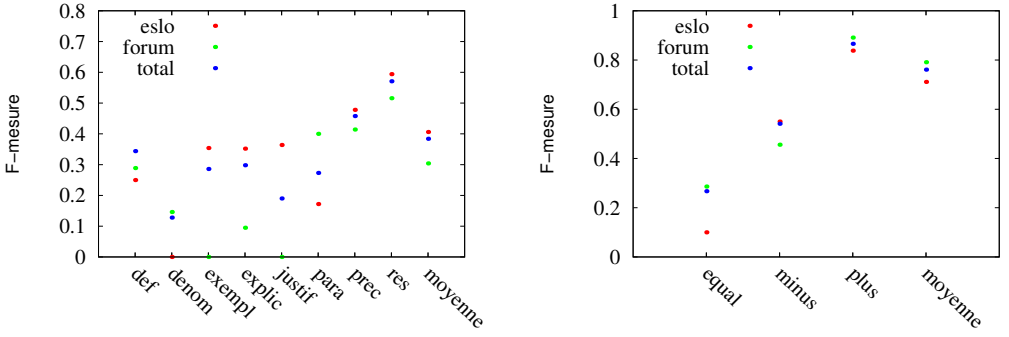


(a) Descripteurs supprimés, 8 catégories

(b) Descripteurs supprimés, 3 catégories

Fig. 2 – Élimination de descripteurs différents à chaque expérience.

Une analyse des matrices de confusion entre les huit catégories indique que certaines fonctions sont très proches et souvent confondues. La fonction *précision* est confondue souvent avec d'autres fonctions. L'explication provient de la nature de la fonction même. *Précision* semble être une catégorie assez large qui peut contenir *explication*, *définition*, *exemplification*, *dénomination* et demande donc des contraintes plus formelles à spécifier. Une autre raison est la fréquence très importante de cette fonction dans le corpus annoté par rapport aux autres fonctions, ce qui peut favoriser sa reconnaissance automatique. Notons aussi que la catégorie *dénomination* cause de très nombreuses confusions : la plupart de ses instances sont catégorisées ailleurs.



(a) 8 catégories (b) 3 catégories

Fig. 3 – Performance de reconnaissance des functions pragmatiques dans chaque corpus.

5 Conclusion et Perspectives

Nous proposons d'étudier les reformulations dans les corpus oraux et les corpus contenant les discussions du web. Nous nous concentrons sur les fonctions pragmatiques des reformulations : la raison pour laquelle un locuteur effectue une reformulation. Nous avons constitué une classification avec onze fonctions (*e.g. définition, exemplification, résultat, paraphrase, correction linguistique*). Notre objectif est d'étudier et de prédire ces fonctions, grâce à l'analyse du contenu des segments $S1$ et $S2$ mis en relation par trois marqueurs (*c'est-à-dire, je veux dire, disons*). L'exploitation des données de référence consensuelles et d'algorithmes d'apprentissage supervisé permet d'effectuer des expériences à deux niveaux : (1) au niveau générique, avec les catégories selon que l'information est ajoutée, supprimée ou constante, nous obtenons des performances autour de 0,80 ; (2) au niveau spécifique des catégories individuelles, nous obtenons des performances autour de 0,40. Quelques descripteurs (ceux liés à la longueur des segments et aux disfluences) jouent un rôle important.

Prédire et apprendre l'information de nature pragmatique est extrêmement difficile. Ce travail est donc exploratoire et permet de constater les différents points qui doivent être pris en compte dans les travaux qui vont suivre. Du point de vue linguistique, il serait important de reconsidérer certaines fonctions : *correction linguistique*, *opposition*, *précision*. Cette dernière devrait être affinée avec plus de critères formels.

Références

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

AUTHIER-REVUZ J. (1995). *Ces mots qui ne vont pas de soi : boucles réflexives et non-coïncidences du dire*. Paris : Larousse.

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.

BENZITOUN C. (2004). L'annotation syntaxique de corpus oraux constitue-t-elle un problème spécifique ? In *RECITAL 2004*.

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

BILGER M. (1999). Coordination : analyses syntaxiques et annotations. *Recherches sur le français parlé*, **15**, 255–272.

BLANCHE-BENVENISTE C. (1995). Le semblamble et le dissemblable en syntaxe. *Recherches sur le français parlé*, **13**, 7–33.

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.

BOURAOUI J.-L. & VIGOUROUX N. (2004). Analyse des erreurs de performance et des stratégies correctives dans le dialogue oral spontané : apports à l'étude des pathologies du langage. *Revue Parole*, **29-30**, 121–152.

BREIMAN L. (2001). Random forests. *Machine Learning*, **45**(1), 5–32.

BROWN P., DESOUZA P., MERCER R., DELLA PIETRA V. & LAI J. (1992). Class-based n-gram models of natural language. *Computational Linguistics*, **18**(4), 467–479.

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

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

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

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

DUTREY C., CLAVEL C., ROSSET S., VASILESCU I. & ADDA-DECKER M. (2014). A CRF-based approach to automatic disfluency detection in a French call-centre corpus. In *International Speech Communication Association Conference (INTERSPEECH 2014)*, p.5̃.

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 des Langues*, **52**(3), 17–46.

FERRERO J. & SIMAC-LEJEUNE A. (2015). Détection automatique de reformulations – correspondance de concepts appliquée à la détection de plagiat. In *EGC 2015, RNTI-E-28*, p. 287–298.

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

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.

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.

GUÉNOT M. (2006). La coordination considérée comme un entassement paradigmatique : description, formalisation et intégration. In P. MERTENS, C. FAIRON, A. DISTER & P. WATRIN, Eds., *TALN 2006*, p. 178–187.

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

KAHANE S. & PIETRANDREA P. (2012). La typologie des entassements en français. In *CMLF 2012*, p. 1809–1828.

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.

KOHAVI R. (1995). The power of decision tables. In *Proceedings of the European Conference on Machine Learning*, p. 174–189 : Springer Verlag.

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

LEVELT W. (1983). Monitoring and self-repair in speech. *Cognition*, (14), 41–104.

LIANG P. (2005). *Semi-Supervised Learning for Natural Language*. Master, Massachusetts Institute of Technology, Boston, USA.

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.

MAGRI-MOURGUES V. (2013). *Reformulation et dialogisme dans le récit de voyage*, In O. GANNIER, Ed., *Echos des voix, échos des textes, Classiques Garnier*.

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.

MANNING C. D. & SCHÜTZE H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA : MIT Press.

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.

NASR A. (1996). *Un modèle de reformulation automatique fondé sur la théorie sens-texte : application aux langues contrôlées*. Thèse de doctorat, Université Paris 6.

NEVEU F. (2004). *Dictionnaire des sciences du langage*. Paris : Colin.

PLATT J. C. (1998). Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning* : MIT Press.

PÉRY-WOODLEY M.-P., ASHER N., ENJALBERT P., BENAMARA F., BRAS M., FABRE C., FERRARI S., HO-DAC L.-M., LE DRAOULEC A., MATHET Y., MULLER P., PRÉVOT L., REBEYROLLE J., TANGUY L., VERGEZ-COURET M., VIEU L. & WIDLÖCHER A. (2009). ANNODIS : une approche outillée de l'annotation de structures discursives. In *TALN 2009*.

QUINLAN J. (1993). *C4.5 Programs for Machine Learning*. San Mateo, CA : Morgan Kaufmann.

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

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

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

STEUCKARDT A. (2005). *Les marqueurs formés sur dire*, In A. STEUCKARDT & A. NIKLAS-SALMINEN, Eds., *Les Marqueurs de glose*, p. 51–65.

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.

VION R. (2006). Reprise et modes d'implication énonciative. *La Linguistique*, **2**(42), 11–28.