Etude de la paraphrase sous-phrastique en traitement automatique des langues

Houda Bouamor

RÉSUMÉ

La variabilité en langue est une source majeure de difficultés dans la plupart des applications du traitement automatique des langues. Elle se manifeste dans le fait qu'une même idée ou un même événement peut être exprimé avec des mots ou des groupes de mots différents ayant la même signification dans leur contexte respectif. Capturer automatiquement des équivalences sémantiques entre des unités de texte est une tâche complexe mais qui s'avère indispensable dans de nombreux contextes. L'acquisition *a priori* de listes d'équivalences met à disposition des ressources utiles pour, par exemple, améliorer le repérage d'une réponse à une question, autoriser des formulations différentes en évaluation de la traduction automatique, ou encore aider des auteurs à trouver des formulations plus adaptées.

Dans cette thèse, nous proposons une étude détaillée de la tâche d'acquisition de paraphrases sous-phrastiques à partir de paires d'énoncés sémantiquement liés. Nous démontrons empiriquement que les corpus parallèles monolingues, bien qu'extrêmement rares, constituent le type de ressource le plus adapté pour ce genre d'étude. Nos expériences mettent en jeu cinq techniques d'acquisition, représentatives de différentes approches et connaissances, en anglais et en français. Afin d'améliorer la performance en acquisition, nous réalisons la combinaison des paraphrases produites par ces techniques par une validation reposant sur un classifieur automatique à maximum d'entropie bi-classe. Un résultat important de notre étude est l'identification de paraphrases qui défient actuellement les techniques étudiées, lesquelles sont classées et quantifiées en anglais et français.

Nous examinons également dans cette thèse l'impact de la langue, du type du corpus et la comparabilité des paires des énoncés utilisés sur la tâche d'acquisition de paraphrases sous-phrastiques. Nous présentons le résultat d'une analyse de la performance des différentes méthodes testées en fonction des difficultés d'alignement des paires de paraphrases d'énoncés. Nous donnons, ensuite, un compte rendu descriptif et quantitatif des caractéristiques des paraphrases trouvées dans les différents types de corpus étudiés ainsi que celles qui défient les approches actuelles d'identification automatique.

Mots clés : Corpus monolingues, Acquisition de paraphrase, Classification automatique de paraphrase, Typologie de paraphrase

ABSTRACT

Language variation, or the fact that messages can be conveyed in a great variety of ways by means of linguistic expressions, is one of the most challenging and certainly fascinating features of language for Natural Language Processing, with wide applications in language analysis and generation. The term paraphrase is now commonly used to refer to textual units of equivalent meaning, down to the level of sub-sentential fragments. Although one can envisage to manually build high-coverage lists of synonyms, enumerating meaning equivalences at the level of phrases is too daunting a task for humans. Consequently, acquiring this type of knowledge by automatic means has attracted a lot of attention and significant research efforts have been devoted to this objective. In this thesis we use parallel monolingual corpora for a detailed study of the task of sub-sentential paraphrase acquisition. We argue that the scarcity of this type of resource is compensated by the fact that it is the most suited corpus type for studies on paraphrasing. We propose a large exploration of this task with experiments on two languages with five different acquisition techniques, selected for their complementarity, their combinations, as well as four monolingual corpus types of varying comparability. We report, under all conditions, a significant improvement over all techniques by validating candidate paraphrases using a maximum entropy classifier. An important result of our study is the identification of difficult-to-acquire paraphrase pairs, which are classified and quantified in a bilingual typology.

Keywords : Monolingual corpora, Paraphrase acquisition, Paraphrase automatic classification, Paraphrase typology

UNIVERSITÉ PARIS SUD
ÉCOLE DOCTORALE D'INFORMATIQUE
LIMSI-CNRS

THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE PARIS SUD

Spécialité : Informatique

par

Houda BOUAMOR

Titre :

ÉTUDE DE LA PARAPHRASE SOUS-PHRASTIQUE EN TRAITEMENT AUTOMATIQUE DES LANGUES

soutenue publiquement le 11 Juin 2012

JURY

Rapporteur	Yves LEPAGE	Prof. à l'Université Waseda
Rapporteur	Emmanuel MORIN	Prof. à l'Université de Nantes
Examinateur	Philippe LANGLAIS	Prof. à l'Université de Montréal
Examinatrice	Adeline NAZARENKO	Prof. à l'Université Paris-Nord
Examinateur	François YVON	Prof. à l'Université Paris Sud
Directrice	Anne VILNAT	Prof. à l'Université Paris Sud
Co-directeur	Aurélien MAX	Maître de Conf. à l'Université Paris Sud

لا تخجل من السؤال عن شيء تجهله فخير لك ان تكون جاهلا
مرة من ان تظل على جهلك طول العمر

يوسف السباعي

The question of whether machines can think is about
as relevant as the question of whether submarines can swim.

— Edsger Dijkstra, 1984 —

À mes parents Messaoud et Mélika Bouamor,
À mes soeurs Wafa, Dhouha, Yosra et Manel,
À mes frères Noureddine, Mohamed Ali et Alaeddine,
À ma petite nièce Dorra,
À tous les Bouamor

REMERCIEMENTS

Je tiens à remercier Anne Vilnat, ma directrice de thèse, pour sa bonne humeur, sa disponibilité, ses encouragements, son soutien moral et son encadrement.

En parlant de l'encadrement, je remercie Aurélien Max, mon co-directeur, pour son suivi continue, ses conseils utiles et les discussions animées que l'on a eu pendant ces quatre années de thèse. Ses conseils et suggestions m'ont permis de prendre les bonnes décisions et mener à bien ce travail.

Je voudrais remercier mes rapporteurs Yves Lepage et Emmanuel Morin pour l'intérêt qu'ils ont porté à mon travail ainsi que les remarques et les suggestions qu'ils m'ont faites.

Je souhaite aussi remercier les autres membres de jury, Adeline Nazarenko, Philippe Langlais et François Yvon qui m'a fait l'honneur de présider mon jury de thèse.

Je voudrai aussi remercier les personnes avec qui j'ai eu l'occasion de travailler, mes chers co-auteurs : Delphine Bernhard, Camille Dutrey et Gabriel Illouz.

Je remercie Xavier Tannier qui était à la fois : mon "officemate", mon psy et mon prof de Java.

Mais cette thèse s'est également inscrite dans un environnement humain particulièrement chaleureux. Je pense à toutes ces personnes grâce à qui je ne me suis jamais senti seule, personnes qui m'ont accompagné au long de ces années de travail. Je remercie en particulier Nèdè pour son bon café, son aide et son soutien moral.

Je conclurai donc en remerciant tous amis, je ne cite pas les noms pour ne pas faire de jaloux. La liste est tellement longue ... Merci à toutes et à tous.

Aussi, je tiens à remercier ma soeur Dhouha pour s'être occupée de moi et pour avoir supporté mes interminables plaintes et pleurnichements.

J'aimerais aussi remercier tous les membres du LIMSI dans leur ensemble pour leur bonne humeur et les discussions agréables qu'on a eu.

Last and not least, je tiens à remercier mes parents Messaoud et Mélika pour tout ce qu'ils m'ont inculqué et appris et sans qui je ne serai pas arrivé là. Un grand merci du fond du coeur à mon frère Noureddine sans qui je ne serai pas venu faire cette thèse en France. Un très grand merci aussi à ma soeur Wafa, à ma tante Yosra, à ma belle-soeur Manel et mon petit coeur Dorra. Un grand merci à toute ma famille.

Et je termine enfin en remerciant tous ceux que j'ai pu oublier.

TABLE DES MATIÈRES

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION

MOTIVATIONS ET PROBLÉMATIQUE

Les langues en usage aujourd'hui renferment un large éventail de phénomènes linguistiques et de caractéristiques faisant l'objet d'études très diverses. Parmi ces caractéristiques, nous nous intéressons à *la variabilité* qui se manifeste dans le fait qu'une même idée, ou un même événement, peut être exprimée sous plusieurs formes. La variété de formes d'expression reflète la richesse du vocabulaire des langues et définit un phénomène linguistique considéré comme un défi majeur dans le domaine du Traitement Automatiques des Langues (TAL) : la paraphrase.

Les locuteurs d'une langue utilisent les paraphrases dans de nombreux aspects de la vie sans avoir à fournir d'efforts considérables (en exprimant une même idée par des formulations différentes afin, par exemple, de la simplifier et de la communiquer le plus clairement possible). Cependant, ce phénomène linguistique reste au cœur des préoccupations des théoriciens de la langue, d'une part (comme le dit Mel'cuk (1988), « *la linguistique théorique présuppose nécessairement une théorie de la paraphrase.* ») et des enjeux du TAL, d'autre part. La présence de la paraphrase complique considérablement l'ensemble des applications ayant pour objectif la modélisation, la compréhension et la production de langue à l'aide de machines.

Pour étayer cet argument, l'utilité des paraphrases se retrouve dans le domaine de recherche de réponses à des questions, où l'extraction d'une réponse précise à partir d'un ensemble de documents, nécessite de ramener l'information cherchée (contenue dans la question) sous une forme proche des différentes formes qu'elle pourrait prendre dans plusieurs passages de texte. Par exemple, la réponse à la question *Quand le projet de Génome Humain a-t-il abouti ?* apparaît sous plusieurs formulations dans les passages suivants :

> $\mathbf{P}_1$: *L'analyse complète du génome humain a été terminée en 2005.*
> $\mathbf{P}_2$: *Le Projet Génome Humain dont la mission était d'aboutir au séquençage complet de l'ADN du génome humain, a été entrepris en 1990.*

Si la sélection de la bonne réponse se fonde uniquement sur une similarité de surface entre question et réponse, il serait évidemment possible de proposer des réponses incorrectes et de ne pas trouver la réponse attendue. Ici la phrase $\mathbf{P}_2$ contient tous les mots clés de la question, c'est pourtant la phrase $\mathbf{P}_1$ qui contient la bonne réponse. Si on avait la connaissance que *a abouti* est équivalente à *a été terminée*, la réponse *en 2005* pourrait être identifiée.

Cependant, identifier automatiquement des équivalences sémantiques est une tâche complexe. Ces dernières années ont été marquées par un intérêt croissant porté à l'acquisition et à l'utilisation de paraphrases (Madnani et Dorr, 2010), même si la définition de cette notion reste difficile à cerner, comme en témoigne la vaste panoplie des définitions présentes dans la littérature (Fuchs, 1994).

La majorité des travaux adopte un point de vue utilitaire qui ne s'intéresse pas explicitement à comprendre la paraphrase dans ses différents niveaux de granularité mais se sert de ses caractéristiques pour améliorer les systèmes automatiques.

Par exemple, en évaluation de la traduction automatique, les hypothèses produites par un système sont évaluées en mesurant leur similarité à des traductions de référence créées par des humains. Ces mesures de similarité se fondent essentiellement sur le nombre de groupes de mots communs dans les deux phrases. Cependant, il est impossible d'identifier les différentes formulations d'un même contenu sémantique avec une seule traduction de référence. Cela peut pénaliser les hypothèses de traduction véhiculant le même sens mais utilisant des expressions différentes de celles présentes dans la référence. Considérons les deux hypothèses de traduction $\mathbf{H}_1$ et $\mathbf{H}_2$ (produites par deux systèmes différents) et la phrase de référence $\mathbf{R}$. Selon ces mesures de similarité, $\mathbf{H}_1$ partageant plus de mots avec $\mathbf{R}$, aura un meilleur score que $\mathbf{H}_2$, pourtant plus proche sémantiquement.

$\mathbf{H}_1$: *Mme Young a l'air heureuse.*
$\mathbf{H}_2$: *Mme Young semblait contente.*
$\mathbf{R}$: *Mme Young avait l'air heureuse.*

Une première solution est d'utiliser, par exemple, des références multiples (Papineni *et coll.*, 2002). Toutefois, constituer plusieurs références manuellement est une tâche très coûteuse. Une autre solution, adoptée dans plusieurs travaux, consiste à prendre en compte les paraphrases lors de l'évaluation afin d'autoriser des formulations différentes. Cela permet de considérer les parties d'une traduction candidate qui sont sémantiquement équivalentes mais lexicalement différentes (*semblait contente* $\leftrightarrow$ *avait l'air heureuse*, par exemple) comme des traductions correctes.

En génération automatique de texte, la paraphrase peut être également exploitée pour proposer à un rédacteur des reformulations pour de courtes unités textuelles afin de rendre son texte plus compréhensible et plus adapté à son contexte d'écriture. Par exemple, la connaissance d'une équivalence entre l'acronyme *BCE* et la forme étendue *Banque Centrale Européenne* permet d'abréger le texte, si ce segment a déjà été présenté sous sa forme étendue.

Généralement, dans ces travaux, la définition de la paraphrase dépend fortement de l'application. D'un point de vue purement applicatif, l'utilité des paraphrases se mesure uniquement par la capacité des systèmes à bien les exploiter dans un contexte spécifique.

Loin des cadres applicatifs, la problématique que nous abordons dans cette thèse se concentre sur la compréhension des paraphrases sous-phrastiques. Nous effectuons une exploration de cet objet linguistique en commençant par acquérir des phrases liées par une relation d'équivalence sémantique et allant jusqu'à l'analyse des paraphrases défiant les méthodes automatiques d'acquisition. Le point de départ de cette étude consiste à répondre à des questions importantes relatives à l'acquisition des paraphrases.

La première question concerne l'identification des types de ressources contenant des paraphrases sous-phrastiques, et des connaissances pour les identifier. Cela passe tout d'abord par l'étude des différentes caractéristiques des corpus existants. Dans ce document, nous montrons par une analyse quantitative et qualitative, suivant différentes stratégies et étudiant des corpus de différents degrés de parallélisme, que les corpus monolingues parallèles contenant des paires de phrases obtenues, par exemple, en traduisant deux fois la même phrase vers la même langue, constituent, malgré leur rareté et la difficulté de leur construction, les corpus les plus appropriés pour l'étude de la paraphrase sous-phrastique. Étant issues de la volonté d'exprimer la même idée, les paires de phrases, contenues dans ces corpus, contiennent une grande variétés d'équivalences beaucoup plus fiables que celles extraites indirectement par le biais des textes comparables. En outre, le contexte de ces équivalences peut être extrait de façon directe, ce qui est particulièrement important pour caractériser les conditions de leur validité.

En se fondant sur ces analyses, nous abordons la tâche d'acquisition des paraphrases en considérant diverses techniques automatiques mettant en jeu des connaissances à différents niveaux. Ces techniques ont été choisies pour leur complémentarité éventuelle ainsi que les différentes ressources mises en jeu. Nous présentons également différents scénarios de combinaison de leurs résultats.

Nous répondons en outre à des questions concernant l'impact de la langue et de la comparabilité des paires de phrases utilisées

sur les performances des techniques évaluées. Nous donnons également une description détaillée des paraphrases qui défient actuellement les approches automatiques.

Puisque les corpus monolingues parallèles que nous utilisons pour notre étude principale ne sont pas disponibles en grandes quantités, ils ne sont pas nécessairement représentatifs de tous les types de corpus que l'on peut réellement exploiter. Il est intéressant d'examiner la possibilité d'utiliser d'autres types de corpus contenant des paires de phrases liées sémantiquement et ayant différents degrés de parallélisme.

Dans le but de généraliser nos résultats, nous avons en outre mené toutes nos expériences en français et en anglais.

CONTRIBUTIONS PRINCIPALES

Le lecteur trouvera ici un premier bilan des principales contributions apportées par ce travail :

- Nous présentons une analyse comparative détaillée des différentes sources d'acquisition de paraphrases phrastiques et justifions l'utilité des corpus monolingues parallèles pour l'étude de la paraphrase sous-phrastique.
- Nous détaillons un cadre original d'acquisition de paraphrases en utilisant cinq techniques exploitant différentes propriétés de la langue pour extraire des paraphrases sous-phrastiques à partir de corpus monolingues parallèles.
- Nous présentons des approches originales fondées sur la combinaison de systèmes et sur l'apprentissage automatique des caractéristiques de paraphrases pour l'acquisition et la validation des paraphrases sous-phrastiques.
- Nous étudions l'impact du niveau de comparabilité des paires d'énoncés sur le nombre et la qualité des paraphrases acquises.
- Nous étudions les limites des techniques actuelles d'acquisition de paraphrases et définissons une typologie des paraphrases difficiles à identifier automatiquement.
- Nous évaluons les techniques d'acquisition étudiées sur différents types de corpus ayant différents degré de comparabilité.

PLAN DU MANUSCRIT

Ce mémoire est organisé de la manière suivante : Dans le chapitre 1, nous passons en revue les différents courants linguistiques traitant le phénomène paraphrastique et nous donnons quelques-unes des définitions ainsi que les différentes typologies associées à la paraphrase dans la littérature.

Dans le chapitre 2, nous passons en revue les travaux existants concernant l'acquisition de paraphrases de différents niveaux de granularité. Bien que le manque de mesures standard, comme celles utilisées en traduction automatique, soit considérée comme une des limitations au développement des systèmes automatiques d'acquisition et de production de paraphrases, nous donnons dans ce chapitre un aperçu des différentes méthodes d'évaluation de paraphrases en discutant leurs avantages et leurs inconvénients.

Par la suite, nous décrivons les expériences et les analyses des paraphrases menées dans le cadre de cette thèse. Le Chapitre 3 a pour objet la description des sources potentielles de paraphrases. Nous présentons dans ce chapitre différentes stratégies de construction de corpus de paraphrases en les définissant sur la base de l'origine du signal du contenu sémantique des paires de phrases contenues : un texte dans différentes langues (Texte), de la parole transcrite dans une autre langue (Parole), la description d'une scène visualisée (Scène), et une courte description (un titre d'article) d'un événement (Événement). Un examen approfondi de l'annotation manuelle de chaque corpus est également réalisé, permettant de définir un guide pour le choix de la source la plus appropriée relativement aux objectifs d'une étude particulière sur la paraphrase sous-phrastique.

Le chapitre 4 est dédié à une étude détaillée de la tâche d'acquisition de paraphrases sous-phrastiques à partir de paires de phrases parallèles. Nous abordons ce problème par l'intermédiaire de différentes techniques opérant sur des niveaux différents et exploitant diverses ressources :

1. des modèles statistiques d'alignements de mots ;
2. des métarègles de variation de termes ;
3. une similarité de structures syntaxiques ;
4. un taux d'édition sur des séquences de mots ;
5. des équivalences de traduction.

Nous faisons également le point sur l'annotation manuelle de paraphrases de référence permettant d'évaluer les performances des systèmes d'acquisition testés. Nous détaillons et analysons par la suite les résultats de l'évaluation de chaque technique individuellement. Puis, nous décrivons des analyses portant sur l'impact du degré de comparabilité des corpus sur les performances des techniques d'acquisition.

Dans le chapitre 5, nous étudions la complémentarité des différentes approches testées et présentons des méthodes hybrides d'acquisition de paraphrases sous-phrastiques par la combinaison de paraphrases candidates produites par les différentes techniques. Ces méthodes visent à améliorer les performances de l'ensemble des techniques en exploitant efficacement leurs différentes caractéristiques. Nous abordons également la tâche de validation des

paraphrases obtenues en présentant un système de combinaison exploitant différents modèles linguistiques et statistiques. Puis, nous présentons une typologie des paraphrases « difficiles à acquérir », où chaque catégorie est illustrée par des exemples représentatifs et quantifiée dans les deux langues étudiées.

Le chapitre 6 détaille une étude de l'acquisition de paraphrases depuis quatre corpus de différents types (décrits dans le chapitre 3) et nous présentons ensuite la performance du système de combinaison présenté dans le chapitre 5 sur chacun des types de corpus. Une analyse des paraphrases pouvant être acquises et étant effectivement acquises pour chaque corpus sera également proposée.

La dernière partie de cette thèse est consacrée à une discussion de nos contributions ainsi que la présentation des perspectives liées à l'étude de la paraphrase (Chapitre 7). Nous y présentons en outre des expériences exploratoires concernant des travaux futurs sur l'acquisition et la validation de paraphrases (Chapitre 8).

Première partie

CONTEXTE DE NOTRE ÉTUDE

1

LA PARAPHRASE

La paraphrase est un phénomène se trouvant au cœur des préoccupations de nombreuses théories linguistiques, qui la considèrent à la fois comme partie intégrante de la compétence linguistique d'un locuteur, étant donné qu'elle survient souvent dans les langues humaines, et un outil primordial pour la modélisation du langage : « *La linguistique théorique présuppose nécessairement une théorie de la paraphrase* » (Mel'cuk, 1988). Katz et Fodor (1963) insistent sur le fait qu'un critère important d'évaluation de la portée d'une théorie sémantique est sa capacité à traiter les phénomènes paraphrastiques. Certains travaux linguistiques ont porté sur l'étude de la paraphrase et plus particulièrement sur sa définition. Les linguistes s'accordent sur le fait que la paraphrase est une opération de reformulation de texte qui *conserve le sens* (Harris, 1957; Martin, 1976; Mel'cuk, 1988; Fuchs, 1982; Duclaye, 2003; Milićević, 2007). Paraphraser consiste à produire un *texte cible* à partir d'un *texte source* afin de clarifier, expliciter ou développer certains aspects. Dans certaines approches, la paraphrase a été définie comme un acte de *traduction intralingue* par lequel on peut remplacer un texte par un autre sans que cela entraîne des modifications notables de sens. Elle consiste en une « *interprétation des signes d'une langue par d'autres signes de la même langue* » (Jakobson, 1963). D'autres approches, se sont basées sur la notion de *synonymie* (Milićević, 2007) et sur des théories complètes (Mel'cuk, 1988) pour donner une définition assez détaillée de la paraphrase.

Dans ce chapitre, nous aborderons les questions suivantes : Pourquoi paraphrase-t-on ? En quoi consiste la paraphrase en linguistique ? Quels sont les différents types de paraphrases ? Et, jusqu'à quel niveau de granularité peut-on paraphraser ? Pour cela, nous passons en revue les principales définitions pertinentes attribuées à ce concept dans la littérature.

Un exposé chronologique détaillé des travaux liés à la paraphrase, se trouve par exemple dans (Martin, 1976; Fuchs, 1982; Mel'cuk, 1988; Cristea, 2001). Nous nous limitons ici à un aperçu des travaux les plus caractéristiques. Nous décrivons tout d'abord l'utilité de la paraphrase dans la pratique langagière (section 1.1). Puis, nous classons les approches par domaine d'utilisation de l'objet paraphrase : la linguistique théorique (section 1.2) et le traitement automatique des langues (section 1.3). Nous présentons en-

suite une description de la relation classique liant la paraphrase au contexte (section 1.4). Nous terminons par une description détaillée des différentes approches permettant de distinguer des catégories de paraphrases (section 1.5).

1.1 POURQUOI PARAPHRASE-T-ON ?

Dans leur comportement langagier ordinaire, les locuteurs d'une langue emploient la paraphrase pour des objectifs divers. Ils cherchent, par exemple, des synonymes pour varier leurs discours, éviter des répétitions, exposer de façon plus appropriée leurs idées ou en abréger d'autres. Lorsque l'on souhaite répondre à une question spécifique on utilise des informations générales ou provenant d'une source de référence que l'on reformule afin de les adapter au discours courant. En outre, les locuteurs recourent, spontanément, à la paraphrase pour changer de style. Utiliser des paraphrases dans ce cas permet de clarifier une pensée et simplifier sa compréhension. La paraphrase est donc nécessaire afin de trouver la meilleure expression pour un contenu sémantique donné dans une situation de communication donnée.

Fuchs (1994) distingue la **reformulation explicative** de celle **à visée imitative**. La première consiste à interpréter un texte T dont on cherche à expliciter le sens par l'intermédiaire d'un autre texte T′. Par différence avec la reformulation à visée explicative, la deuxième est tournée vers la production de T′ dont on cherche à construire les formes d'expression à partir du sens de T.

Ces deux types de reformulation trouvent leurs origines dans l'exégèse biblique et la rhétorique. En rhétorique classique, la paraphrase était un exercice préparatoire pour les futurs orateurs qui consistait à effectuer un ample développement d'un texte d'auteur ou d'une sentence. Un élève doit pouvoir reformuler des textes en remplaçant les mots par des équivalents ou en procédant à une paraphrase plus libre. On se servait de la paraphrase dans l'exégèse biblique afin d'expliciter le sens littéral d'un texte donné en reformulant les tournures linguistiques peu claires afin de dévoiler les significations cachées derrière le sens littéral apparent. Ceci permettait aux fidèles de comprendre les textes sacrés et d'en maîtriser le sens.

1.2 LA PARAPHRASE DANS LES THÉORIES LINGUISTIQUES

Dans un sens général, paraphraser signifie reformuler. Il s'agit d'une opération qui consiste à changer les mots (la forme, la syntaxe) d'un texte, tout en préservant sa signification (la sémantique). En effet, il est possible d'exprimer une idée de différentes manières en fonction de l'objectif du locuteur. Le terme *paraphrase* [1] est apparu en 1525. Il dérive du latin *paraphrasis* emprunté au grec *paraphrazein*. Il est composé de *para* (*à côté de*), et de *phrasis* (*discours*).

1. Cette définition est indiquée dans le TLFI : http://atilf.atilf.fr/dendien/scripts/tlfiv5/advanced.exe?8;s=3091673805;

Dans cette section, nous passons en revue les différentes définitions associées à la paraphrase dans le domaine de la linguistique.

Plusieurs travaux ont été menés pour étudier la paraphrase et tenter de lui associer une définition assez précise (Martin, 1976; Fuchs, 1994; Cristea, 2001). Des définitions linguistiques variées de la paraphrase ont été proposées qui s'accordent sur le fait que paraphraser signifie reformuler ou réécrire.

Dès les années 1950, Harris introduit, dans le cadre de la grammaire transformationnelle, la notion de *transformation* (Harris, 1957; Chomsky, 1957), comprise comme une relation d'équivalence entre phrases conservant le sens et générant, par conséquent, des paraphrases. Dans ce cadre, la paraphrase a été définie au niveau de schémas des phrases comme une relation d'équivalence à base de règles transformationnelles liant deux schémas de phrases. L'exemple 1 ci-dessous illustre une règle permettant de transformer une phrase active (telle que *He saw the man*) à la voix passive (*The man was seen by him*).

$$N_1 \text{ t } V \ N_2 \rightarrow N_2 \text{ t be } Ven \text{ by } N \tag{1}$$

Dans ces définitions, la paraphrase est considérée comme un système complet, *un système paraphrastique* faisant intervenir des transformations qui consistent à convertir une séquence de mots en une autre dans la limite de certaines contraintes spécifiées par la grammaire. Le principal inconvénient de la théorie de la paraphrase fondée sur des règles transformationnelles est qu'il traite la paraphrase comme un phénomène purement syntaxique en ignorant son aspect sémantique. En effet, un grand nombre de paraphrases sont de nature lexicale, par exemple, celles qui sont obtenues par le remplacement d'un mot dans une phrase par un synonyme. Cependant, des règles transformationnelles peuvent également exprimer des contraintes lexicales. C'est avec l'apparition de la *sémantique générative* (McCawley, 1968) que le champ d'étude de la paraphrase a été étendu en introduisant la notion d'*équivalence sémantique* par des règles de transformations lexicales impliquant la syntaxe et la sémantique des phrases.

Une approche différente de la paraphrase a été introduite dans la *Théorie Sens-Texte* (TST) élaborée dans les années 1960 par Mel'čuk. Cette approche sert de cadre linguistique théorique pour la construction de modèles des langues naturelles. Elle repose sur des principes généraux applicables à toutes les langues, ce qui la rend universelle (Mel'cuk, 1988). La TST décrit un modèle à sept niveaux linguistiques, de la morphologie/phonétique à la sémantique avec une même structure de représentation. Des transformations entre ces niveaux de représentation permettent de naviguer d'un texte vers son sens, et réciproquement. Mel'čuk (2003) donne la description suivante :

> *La langue naturelle est un système de correspondances entre les sens, modélisés par la représentation sémantique, et les textes, modélisés, eux, par la représentation phonétique.*

Le modèle linguistique défini par Mel'cuk met essentiellement l'accent sur la sémantique et accorde de fait une forte importance au lexique.

La TST rend compte de l'association que tout locuteur d'une langue est capable de faire entre un sens donné dans cette langue et l'ensemble des énoncés paraphrastiques exprimant ce sens par l'intermédiaire de descriptions lexicales détaillées et structurées dans le *Dictionnaire Explicatif et Combinatoire* : DEC (Mel'čuk et Arbatchewsky-Jumarie, 1992). Ce dictionnaire comporte une liste exhaustive de tous les sens possibles des termes (ou *lexèmes*) ainsi que leurs usages syntaxiques et sémantiques. Ce dictionnaire permet d'accéder aux dérivés sémantiques et aux collocations pour produire un texte à partir d'un sens, c'est-à-dire exprimer linguistiquement ses idées. Il permet aussi d'explorer le potentiel de paraphrasage, car « *la langue est [...] plus qu'un outil pour exprimer nos pensées : c'est un outil pour exprimer nos pensées de multiples façons [...]*(Mel'čuk et Polguère, 2007). Les règles de paraphrasage « *mettent en jeu des procédés se situant au confluent du lexique et de la grammaire* » (Mel'čuk et Polguère, 2007).

La théorie postule, en effet, que les langues sont définies par la façon dont leurs unités lexicales sont combinées par l'application des *fonctions lexicales*. Ces dernières expriment différents types de relations pouvant lier des unités lexicales dans une langue : **analogie sémantique** (*synonymie*, *antonymie*) ; **dérivations** (*nominalisation*, *adjectivisation*) ; **phénomènes de cooccurrence lexicale** déterminant le lexème précis à utiliser pour exprimer un sens donné (*intensification*.

La TST a défini un ensemble de 67 fonctions lexicales de paraphrasage réalisant des transformations sur des représentations syntaxiques profondes et permettant de faire correspondre des descriptions proches du langage avec une représentation sémantique. Il est généralement reconnu que cette théorie a apporté une contribution pertinente des notions utiles pour l'étude linguistique de la paraphrase de référence. La langue a été, généralement, décrite dans la TST comme un *système de paraphrasage*.

Alors que la TST donne un aperçu de la paraphrase, plusieurs facteurs ont limité son utilité. Tout d'abord, tout en donnant à la paraphrase un rôle important dans la théorie de la langue, la TST néglige dans sa description un aspect important du phénomène en ignorant les différences sémantiques et le rôle du contexte dans l'assimilation sémantique à l'origine de l'effet de paraphrase (Amghar, 1996). Ensuite, les sept niveaux linguistiques définissant la base de cette théorie rendent l'exploitation effective de ce modèle difficile.

Contrairement aux théories décrites jusqu'alors, Honeck (1971) présente une description à haut niveau des paraphrases. Il distingue entre la paraphrase : (1) *transformationnelle*, où la structure de surface de la phrase ou du segment d'origine est inchangée, et où seuls les mots utilisés sont modifiés ; (2) *lexicale*, où la structure de surface de la phrase ou du segment d'origine est conservée mais certains mots pleins sont remplacés par un synonyme ; (3) *formalexicale*, où la structure de surface ainsi que les mots pleins de la phrase ou le segment de base sont modifiés. Par exemple, la phrase (b) ci-dessous est une paraphrase transformationnelle de (a), alors que (c) en est une paraphrase lexicale. Enfin, (d) est considérée comme une paraphrase formalexicale.

(a) Le chat a mangé la souris.
(b) La souris a été mangée par le chat.
(c) Le félin a dévoré la souris.
(d) Le rongeur a été dévoré par le minou.

Martin (1976) a mis en évidence les difficultés auxquelles on se heurte lorsqu'on entreprend de formuler une définition de la paraphrase et a, par la suite, présenté une définition de la paraphrase comme une *relation d'équivalence* vérifiant un ensemble de caractéristiques logiques :

> *Deux phrases* P_i *et* P_j *seront dites en relation de paraphrase si, pour tout locuteur et en toute situation,* P_i *est logiquement équivalente à* P_j.

La notion d'*équivalence logique* fait référence aux trois propriétés classiques en mathématiques : symétrie, transitivité et réflexivité. Cependant, Fuchs (1994) affirme que « *... la reformulation en discours est une relation qui ne vérifie aucune des trois propriétés suivantes : transitivité, symétrie et réflexivité ...* ». Dans cette lignée, des données expérimentales présentées par Levrat et Amghar (1995) ont confirmé les limites de la définition de Martin. En effet, si la symétrie est souvent admise (Fuchs, 1994), il n'est pas rare que l'on considère la paraphrase comme un contenu étendu de la phrase initiale. De plus, de proche en proche, la succession des paraphrases induit des déformations sémantiques, ce qui fait perdre la transitivité des phrases. Enfin, une phrase est rarement considérée comme paraphrase d'elle-même (donc pas de réflexivité).

Le point de départ de cette théorie logico-sémantique de la paraphrase est la distinction entre différentes catégories de paraphrases qui seront décrites dans la section 1.5.

Un peu plus tard et dans le cadre de la théorie de l'*énonciation* (Culioli, 1976), Fuchs (1982) a démontré qu'une approche de la paraphrase ne pouvait se limiter à une analyse exclusivement linguistique qui ferait abstraction de la dimension énonciative. Elle la présente comme une activité exercée par un locuteur ou un interlocuteur :

Paraphraser c'est se livrer à une activité de reformulation, par laquelle on restitue le sens d'un discours (énoncé ou texte) déjà produit.

Plus récemment, Milićević (2007) tente de répondre à des questions fondamentales sur la modélisation de la paraphrase en se basant sur la TST. Pour elle, la paraphrase est définie selon deux aspects : (1) une relation de (quasi-)synonymie entre phrases, dans son aspect statique et (2) une opération permettant de produire des phrases (quasi-)synonymes, dans son aspect dynamique. Selon cette définition, la paraphrase est réduite à une manifestation de la synonymie portant sur des phrases complètes : une relation unissant des phrases *synonymes*.

L'auteur considère que la modélisation de la capacité paraphrastique d'un locuteur revient à proposer des règles linguistiques décrivant formellement les liens paraphrastiques que ceux-ci peuvent établir entre expressions, par exemple :

- X est sûr de Y
- X ne doute pas de Y
- Y est un fait certain pour X
- Y ne soulève chez X aucun doute

En conclusion, les théories que nous avons évoquées dans cette section attribuent à la paraphrase des propriétés qui en font une relation d'équivalence sémantique. Elle est fréquemment associée à la synonymie et opère généralement au niveau lexical ou syntagmatique. Plus globalement, ces théories décrivent la paraphrase dans un cadre linguistique à l'aide de modèles de différents niveaux de complexité, ou en proposant une définition à un niveau très abstrait. Les modèles définis en linguistique théorique sont difficiles à appréhender et ne peuvent être mis en œuvre que si l'on utilise des représentations complexes. Les définitions plus abstraites sont elles difficiles à employer en pratique.

1.3 LA PARAPHRASE TELLE QUE DÉFINIE DANS LE TAL

La majorité des systèmes de traitement automatique des langues sont d'une façon ou d'une autre confrontés au phénomène de paraphrase (Madnani et Dorr, 2010). Cependant, il y en a relativement peu qui se sont intéressés à la nature linguistique de ce phénomène. Ils s'appuient, en effet, sur des définitions simples et informelles, de nature quantitative telles que des nombres de mots ou de *n*-grammes communs.

En TAL, des définitions variées ont été associées à la paraphrase, qui ont en commun de se fonder sur le principe de l'équivalence sémantique. Il s'agit d'*un moyen alternatif* exprimant, dans une même langue, *le même contenu sémantique*, la *même information* ou la *même idée* que la forme originale (Barzilay et McKeown, 2001; Fujita, 2005;

Callison-Burch, 2007; Bhagat, 2009; Zhao *et coll.*, 2009; Madnani, 2010). Selon toutes les défnitions rencontrées, il existe une relation paraphrastique entre deux énoncés lorsqu'ils sont formulés pour véhiculer ou exprimer une même signification ou idée.

La notion d'équivalence sémantique a été modélisée selon plusieurs points de vue. En suivant, par exemple, les règles de la logique, Dras (1999) propose une définition de la paraphrase fondée sur un modèle de la sémantique vériconditionnelle : deux unités textuelles sont interchangeables si, pour les propositions A et B qu'elles contiennent, l'ensemble de vérité de B est un sous-ensemble de l'ensemble de vérité de A.

D'autres considèrent la paraphrase comme une forme de traduction avec une même langue source et cible (Quirk *et coll.*, 2004; Zhao *et coll.*, 2008b). Ainsi, les approches développées en traduction automatique pour extraire des équivalences sémantiques entre expressions dans différentes langues peuvent être directement utilisées pour apprendre des paraphrases sous-phrastiques (Ohtake et Yamamoto, 2003). D'autres approches considèrent la paraphrase comme un phénomène d'implication textuelle bidirectionnelle (Malakasiotis, 2011) et l'utilisent plus particulièrement pour améliorer la performance de leurs systèmes (Bosma et Callison-Burch, 2006). L'inconvénient majeur de ces définitions réside dans le fait qu'elles transfèrent à d'autres domaines le problème d'équivalence sémantique entre expressions : définition d'un ensemble de vérité des propositions et reconnaissance des implications textuelles.

Contrairement aux définitions basées sur des notions de logique qui sont très restrictives, Bhagat (2009) affirme que, même si certaines paraphrases potentielles ne sont pas équivalentes au sens logique, elles doivent être considérées comme paraphrases ou "quasi-paraphrases" pour des raisons purement pratiques. Par conséquent, les auteurs sur la paraphrase en TAL font souvent l'hypothèse que les paraphrases sont des constructions approximativement sémantiquement équivalentes.

Ces approches seront davantage décrites dans le chapitre suivant (Chapitre 2).

1.4 PARAPHRASE ET CONTEXTE

La plupart des travaux sur la paraphrase développent des modèles permettant de remplacer des unités textuelles par d'autres dans un énoncé afin de produire une paraphrase tout en s'assurant que le sens de l'énoncé d'origine est conservé. La relation d'équivalence, considérée comme l'élément clé dans la paraphrase, dépend d'une notion importante et familière aux linguistes : **le contexte** :

> *La paraphrase ne se comprend qu'en situation de communication réelle faisant intervenir un contexte commun et identifié.* (Fuchs, 1982)

Le terme « contexte » est souvent associé à deux concepts clés : « situation » et « voisinage ». Le contexte d'un évènement est défini en incluant les *circonstances* et *conditions* qui l'entourent. Par conséquent, le contexte d'une unité textuelle donnée inclut au moins les mots qui l'entourent.

L'existence des phénomènes d'ambiguïté dans ses formes lexicale et syntaxique, où l'on attribue à une même forme plusieurs interprétations (*avocat : le fruit* et *avocat : le métier*) rend la tâche de production de paraphrases assez difficile à effectuer. En effet, le choix des unités lexicales et des transformations syntaxiques à employer sont dépendantes de l'ensemble des mots l'entourant. Ce contexte spécifie le cadre/domaine dans lequel l'expression (à paraphraser) a été présentée et permet par conséquent de choisir la reformulation la plus appropriée. Délimiter le contexte dans lequel une paraphrase peut remplacer une expression donnée, revient à effectuer un test de *substituabilité* de l'une par l'autre, un critère intervenant dans la définition de la paraphrase elle-même (Dras, 1999). D'après Mel'cuk (1988), si deux unités textuelles sont *synonymes*, elles doivent être substituables dans au moins quelques contextes, sans changer la signification du texte.

Dans ses travaux de thèse, Duclaye (2003) distingue entre *contexte linguistique* et *contexte extra-linguistique*. Le premier représente le choix d'unités lexicales, de constructions syntaxiques spécifiques, etc. Elle appelle contexte extra-linguistique la situation concrète dont il est question. Cette distinction renvoie en quelque sorte à l'origine de la distinction entre paraphrase linguistique (portant sur les *phrases*) et paraphrase non-linguistique (impliquant les *énoncés*).

Une hypothèse largement suivie dans les travaux de TAL et qui rend compte de l'importance du contexte lors du traitement de la paraphrase est l'*hypothèse distributionnelle* de (Harris, 1954) provenant de la linguistique. Cette hypothèse veut que les mots qui apparaissent dans les mêmes contextes tendent à avoir des significations semblables. L'idée principale de cette hypothèse semble assez claire : il existe, souvent, une corrélation entre la similarité distributionnelle et la similarité du sens ce qui nous permet d'utiliser la première pour estimer la seconde. Ceci permet de décrire une langue par des relations entre ses éléments (mots, morphèmes, phonèmes, etc.) et ce en fonction de leur apparition avec d'autres éléments.

1.5 CATÉGORISATION DES PARAPHRASES

1.5.1 *Généralités*

En linguistique, quelques typologies de paraphrases ont été introduites, généralement au sein d'une théorie abordant le phénomène paraphrastique (Martin, 1976; Mel'cuk, 1988). Par conséquent, ces typologies sont strictement liées à un cadre particulier et sont difficilement exploitables dans d'autres contextes, car elles s'appuient sur des critères disparates et diffèrent d'une étude à l'autre.

Dans la TST de Mel'čuk et Arbatchewsky-Jumarie (1992), par exemple, une typologie prend la forme de listes de règles de construction de paraphrases lexicales et syntaxiques exprimées respectivement par l'intermédiaire de fonctions lexicales et d'un ensemble d'arbres de dépendances indiquant la restructuration syntaxique nécessaire à appliquer pour chaque règle lexicale. Milićević (2003) propose de nouvelles règles de paraphrase opérant au niveau sémantique de représentation, nécessaires pour rendre compte de certaines **paraphrases approximatives** ne possédant pas nécessairement un sens absolument identique pour qu'elles soient considérées comme des paraphrases.

Il existe quelques typologies de paraphrases présentées dans des travaux en TAL. La majorité sont décrites sous forme de listes de catégories de paraphrases utiles pour le fonctionnement d'un système ou une application spécifique (Marton *et coll.*, 2009; Max, 2010). Ces typologies ne présentent qu'une liste non exhaustives des paraphrases trouvées dans des corpus spécifiques utilisés pour une tâche particulière : production de certains types de paraphrases (synonymes, paraphrases syntaxiques, etc.) à partir de prédicats (Kozlowski *et coll.*, 2003) ou encore dans le cadre de la présentation d'un système d'annotation de corpus parallèle (dérivation morphologique, comparatifs vs superlatifs, etc.) (Dorr *et coll.*, 2004).

D'autres travaux distinguent d'une façon assez grossière quelques types de paraphrases (Barzilay, 2003; Shimohata, 2004). Dans sa thèse, Barzilay (2003) répartit les paraphrases dans deux classes : paraphrases atomiques ou paraphrases composites. Les atomiques sont des paraphrases liant des unités lexicales non décomposables (mots ou petits segments). Les paraphrases composites sont celles contenant des constructions qui peuvent être décomposées en de plus petites unités (phrases et segments complexes). Les paraphrases atomiques sont ensuite subdivisées en se basant sur des critères de taille des segments en relation de paraphrase. En outre, une catégorisation plus fine des paraphrases composites est proposée en considérant les modifications de haut niveau affectant la phrase d'origine. Ces modifications incluent les opérations d'édi-

tion et les transformations syntaxiques et lexicales. Il est à noter que cette définition a été suivie, par la suite, dans plusieurs travaux en TAL (Cohn *et coll.*, 2008).

Notons que d'autres typologies plus couvrantes et détaillées comme celles de Culicover (1968) et Dras (1999) ou plus récemment Fujita (2005) et Vila *et coll.* (2011) ont été proposées. Culicover a proposé le premier système traitant automatiquement la paraphrase dans le but de récupérer des passages de textes en réponse à des requêtes en langue "naturelle". Il a pour cela défini un ensemble cohérent regroupant les types linguistiques possibles pour la paraphrase et a également proposé une tentative de formalisation par la définition de certaines conditions structurelles et sémantiques nécessaires.

Les travaux plus récents de Dras (1999) développent une étude précise et approfondie des paraphrases syntaxiques permettant de les représenter plus formellement, par la suite, en utilisant des grammaires d'arbres adjoints synchrones (TAG). La catégorisation de paraphrases suivante, selon cinq axes, a été introduite [2] :

(1) CHANGEMENT DE PERSPECTIVE : changement dans la façon dont les éléments de texte sont représentés, tel que le remplacement d'un verbe par un adjectif dans une phrase.
Exemple :
Ce manuscrit peut être lu. ↔ *Ce manuscrit est lisible.*

(2) CHANGEMENT D'EMPHASE : changement de la structure syntaxique d'une phrase en modifiant son focus par exemple, passage de la voix active à la voix passive (exemple 2-a), ou conversion de la phrase sous forme de phrase pseudo-clivée (exemple 2-b).
Exemples :
(2-a) *Maman a servi le repas.* ↔ *Le repas a été servi par maman.*
(2-b) *Le chat a bu tout le lait.* ↔ *Celui qui a bu tout le lait, c'est le chat.*

(3) CHANGEMENT DE RELATION : changement de connexion entre les propositions des phrases.
Exemple :
Une étude conduite par « Gallup Poll » a indiqué qu'un Américain sur quatre croit aux fantômes. ↔ *Une étude a été conduite par le « Gallup Poll ». Elle a indiqué qu'un Américain sur quatre croit aux fantômes.*

(4) SUPPRESSION : suppression d'éléments périphériques de la phrase d'origine. Il est important de souligner que pour cette catégorie les paraphrases sont unidirectionnelles.
Exemple :
La situation de guerre a affecté de nombreuses personnes. → *La guerre a affecté de nombreuses personnes.*

2. Les exemples accompagnant chaque catégorie ont été adaptés en français.

(5) DÉPLACEMENT DE PROPOSITION : changement de la position de quelques expressions dans la phrase.
Exemple : *L'étudiant a copié les schémas importants avant de rendre le livre.* ↔ *Avant de rendre le livre, l'étudiant a copié les schémas importants*

Plus récemment, Fujita (2005, 2010) a proposé une typologie plus générale et exhaustive, linguistiquement fondée et axée sur la dénotation. Elle comporte également cinq classes de paraphrases : *extra-phrastiques*, *extra-clausales*, *syntaxiques*, *morpho-syntaxiques* et *lexicales*. La distinction entre classes repose sur la portée sémantique et sur le degré de généralité des paraphrases. Finalement, Vila *et coll.* (2011) font une analyse critique des principales typologies de paraphrases proposées jusqu'alors en TAL et affirment qu'il n'existe aucune caractérisation des paraphrases qui soit complète, linguistiquement fondée et automatiquement interprétable en même temps. Les auteurs proposent à leur tour une typologie contenant 9 classes de paraphrases phrastiques, combinant connaissances lexicales, syntaxiques, sémantiques et pragmatiques.

Comme nous l'avons vu, il existe différentes typologies de paraphrases dans la littérature en TAL, portant sur différents niveaux de la langue. Ces catégories peuvent être résumées dans une typologie fondée sur 2 axes principaux : le niveau de granularité du texte et les différents niveaux de la langue.

1.5.2 *Niveau de granularité*

La paraphrase peut se produire à plusieurs niveaux de granularité textuelle. Des éléments lexicaux individuels ayant le même sens sont généralement appelés *paraphrases lexicales* ou plus simplement *synonymes*. À une échelle plus grossière, le terme *paraphrase sous-phrastique* renvoie à des unités textuelles (segments ou fragments de texte) partageant le même contenu sémantique. Deux phrases ou énoncés représentant un même contenu sémantique sont considérées comme *paraphrases phrastiques*. Ces trois classes de paraphrases sont détaillées ci-dessous.

PARAPHRASE LEXICALE Des unités lexicales individuelles ayant la même signification sont généralement appelés *paraphrases lexicales* ou, plus couramment, des *synonymes* (par exemple, *manger* ↔ *consommer* et *bouquin* ↔ *livre*). Toutefois, la paraphrase lexicale ne se limite pas strictement à la notion de synonymie. Elle se manifeste sous plusieurs formes telle que l'*hyperonymie*, où un mot donné est en relation sémantique hiérarchique avec un autre. Cette relation peut être considérée comme une relation paraphrastique dans laquelle l'un des mots est plus spécifique/général que l'autre (par exemple, *bâtiment* ↔ *maison*, *chien* ↔ *animal*). Employer

des paraphrases lexicales consiste, généralement, en une substitution synonymique qui laisse intact le cadre syntaxique de la phrase source.

PARAPHRASE SOUS-PHRASTIQUE Le terme paraphrase *sous-phrastique* recouvre aussi bien une paire de mots qu'une paire de groupe de mots (syntagmes ou fragments textuels quelconques) dont la taille peut être aussi grande que nécessaire dans la limite d'une phrase et qui sont en relation d'équivalence sémantique dans un contexte donné (par exemple, « *envisage-t-elle* » ↔ « *a-t-elle l'intention* »). Ces unités textuelles peuvent, dans certains cas, se présenter sous forme de patrons liant des éléments textuels variables tels que « *X ne doute pas de Y* » ↔ « *X est sûr de Y* » .

PARAPHRASE PHRASTIQUE Deux phrases véhiculant le même contenu sémantique sont appelées *paraphrases d'énoncé* ou *paraphrases phrastiques* (par exemple « *Elle a grondé son enfant* » ↔ « *Elle s'est fâchée contre son enfant* »). Bien qu'il soit assez simple pour un humain de produire des paraphrases phrastiques en remplaçant un mot ou un groupe de mots dans la phrase d'origine avec leurs équivalents sémantiques respectifs, il est beaucoup plus difficile de produire des paraphrases telles que : « *Vous n'êtes même pas en mesure de me donner ce renseignement* » ↔ « *Tu n'es même pas fichu de me passer ce tuyau* » . Dans ce genre de situation, le locuteur doit également maîtriser les aspects stylistiques de son intervention.

1.5.3 *Niveau d'analyse de la langue*

La catégorisation des paraphrases par niveau d'analyse de la langue a été introduite par Martin (1976). La distinction entre *paraphrase sémantique* ou *linguistique* et *paraphrase pragmatique* ou *situationnelle* a été le point de départ de la théorie sémantique véri-conditionnelle. Selon cette théorie, comprendre une phrase, c'est connaître les conditions dans lesquelles elle est vraie. La question n'est donc pas de savoir si une phrase donnée est vraie ou fausse, mais de préciser quelles conditions doivent être remplies pour que la phrase soit vraie[3].

Il est à noter que la distinction entre ces deux catégories de paraphrases, en linguistique, passe par la compréhension de la différence entre deux concepts importants : *phrase* et *énoncé*. Une phrase est un groupe **stable** et **constant** de constituants structurés pour exprimer une idée ou fournir une signification. Une phrase est construite selon des règles structurales de la syntaxe et selon des critères de grammaticalité bien définis. Lorsqu'une phrase est

3. Cette définition est donnée dans le dictionnaire SÉMANTICLOPÉDIE : http:www.semantique-gdr.net/dico/index.php/V%C3%A9riconditionnel

prononcée dans un certain contexte (circonstances, lieu, moment) et dans un certain co-contexte (son entourage linguistique), elle devient un énoncé unique. L'énoncé est un phénomène **variable** lié à l'activité langagière. Il est relié à un contexte et il fournit le sens en fonction de la compréhension et de l'interprétation de ce dernier.

PARAPHRASE SÉMANTIQUE OU LINGUISTIQUE :

On appelle paraphrase linguistique (ou sémantique) celle qui est indépendante de la situation. Plus formellement, « *Deux phrases* P *et* Q *sont en relation de paraphrase linguistique si leur sens est le même et si elles ne s'écartent que par leurs topicalisations et leurs connotations.* » (Martin, 1976). Ces paraphrases sont employées par les locuteurs dans leur comportement langagier ordinaire pour s'exprimer de façon plus appropriée, changer de style ou encore pour clarifier leurs pensées (Mel'cuk, 1988).

> *Elle ne pouvait se passer de sucre.* ↔ *Elle ne pouvait pas se priver de sucre.*
> *Pierre a ôté son manteau.* ↔ *Pierre a enlevé son manteau.*

Généralement, un tel système de paraphrase comporte des règles de deux types : syntaxiques et lexicales.

PARAPHRASE SYNTAXIQUE

Les phrases d'une paire paraphrastique peuvent être reliées par une relation fondée sur une règle qui spécifie les conditions du passage d'une phrase à l'autre. Ces paraphrases portent sur le niveau de la syntaxe des phrases, elles sont donc appelées des **paraphrases syntaxiques**. Les règles de transformation syntaxiques dans ce cadre portent sur la nominalisation, une transformation qui consiste à convertir un adjectif en syntagme nominal (a), l'épithétisation, une opération transformant une proposition relative en adjectif épithète (b) et la topicalisation par clivage (c-1) ou passivation (c-2) (Cristea, 2001).

(a) *L'eau est limpide, cela permet de voir les algues.* ↔ *La limpidité de l'eau permet de voir les algues.*

(b) *l'industrie du coton* ↔ *l'industrie cotonnière*

(c_1) *Un jeune écrivain a écrit ce roman* ↔ *C'est un jeune écrivain qui a écrit ce roman.*

(c_2) *Un jeune écrivain a écrit ce roman* ↔ *Ce roman a été écrit par un jeune écrivain.*

PARAPHRASE LEXICO-SYNTAXIQUE

La majorité des paraphrases opèrent sur plus d'un niveau de langue. C'est le cas, par exemple, des paraphrases lexico-syntaxiques. La production de telles paraphrases se fait non seulement par la modification de la structure de phrase d'origine (niveau syntaxique) mais aussi le change-

ment des unités lexicales employées (niveau lexical). Ce type de paraphrases peut être obtenu par exemple par redistribution des arguments sur des actants différents dans une phrase (a) ou encore par un mécanisme de double négation (b) [4].

(a) *Pierre a prêté des disques à Jean.* ↔ *Jean a emprunté des disques à Pierre.*

(b) *Il est probable que Pierre viendra.* ↔ *Il est improbable que Pierre ne vienne pas.*

PARAPHRASE NON-LINGUISTIQUE :

PARAPHRASE PRAGMATIQUE Telle que définie par Martin (1976), la paraphrase pragmatique peut être perçue comme l'étude des conditions et des circonstances entre les interlocuteurs et le rapport existant entre l'information contenue dans l'énoncé et étant des connaissances du destinataire. En effet, P_j est une paraphrase pragmatique de P_i si, dans une situation donnée, P_j se réfère à la même intention que P_i. Ainsi, « *Il y a un courant d'air* » peut renvoyer à « *Je veux que l'on ferme la fenêtre* », aussi bien que la phrase « *Fermez la fenêtre, s.v.p* ». La relation entre le courant d'air et la fermeture de la fenêtre est une donnée d'expérience, indépendante de la langue, mais commune à un grand nombre de personnes, de telle sorte que deux phrases de sens distinct peuvent s'interpréter, grâce à une expérience d'univers commune, comme des paraphrases pragmatiques produites en conséquence de l'énoncé initial. La paraphrase pragmatique relève de cette forme d'implicite qui laisse inexprimé ce que la situation permet de restituer en ce que le locuteur juge superflue.

PARAPHRASE RÉFÉRENTIELLE Contrairement au cas où la relation de paraphrase peut être établie sur la seule base du sens linguistique, il existe d'autres paraphrases où il est nécessaire de connaître la référence de certains termes pour déterminer s'il existe une relation de paraphrase. Fuchs (1982) présente trois types de paraphrase référentielle : (1) la paraphrase référentielle anaphorique, obtenue par exemple en remplaçant *Les femmes, les enfants, les hommes* par *Ils* dans *Les femmes, les enfants, les hommes veulent venir*. Il faut donc connaître la référence du terme anaphorique (*Ils* dans cet exemple) pour établir la relation de la paraphrase ; (2) la paraphrase référentielle des termes déictiques, où chaque fois qu'une phrase contient un terme déictique, il est nécessaire de connaître la référence de ce terme pour décider s'il y a ou non relation de paraphrase avec une phrase correspondante quand on a une expression

4. Ces exemples sont donnés par Cristea (2001).

descriptive au lieu d'une déictique[5] (*Il est allé te voir là-bas le mois dernier.* ↔ *Jack est allé te voir à la maison le mois dernier.*) ; (3) la paraphrase référentielle de description définie, où il s'agit de l'emploi d'une périphrase qui vise à désigner un certain objet ou une certaine personne de la réalité en le décrivant à l'aide d'une propriété ou lieu de la nommer directement. Pour pouvoir établir la relation de paraphrase entre une phrase contenant le nom de l'objet, il faut reconnaître la référence de la description de l'objet en question (par exemple, *La Cité des Doges* ↔ *Venise*).

5. Dans un énoncé oral ou écrit, les déictiques sont des mots ou expressions qui déterminent les conditions particulières de l'énonciation, liées à une situation de communication donnée (tiré de Wikipédia).

CONCLUSION DU CHAPITRE

Nous nous sommes intéressée, dans ce chapitre, au problème de la définition qui attire depuis longtemps l'attention de la communauté du TAL. Nous avons passé en revue les définitions issues de la linguistique les plus marquantes ainsi que celles suivies dans des travaux en TAL. Les diverses prises de position sur la notion de paraphrase lui attribuent des propriétés qui en font une relation d'équivalence sémantique obtenue par divers mécanismes et qui implique un ensemble de phénomènes linguistiques. Il est important de souligner que les définitions données à la paraphrase sont souvent subjectives et donnent une idée de la grande diversité du phénomène. Nous avons, par la suite, présenté les typologies de paraphrase les plus importantes dans la littérature. Ces typologies mettent à nouveau en évidence la complexité de la distinction entre différents types de paraphrases.

Dans le chapitre 2, nous allons reprendre les principales approches mises en œuvre pour acquérir et évaluer automatiquement des paraphrases à différents niveaux de granularité.

2

TRAITEMENT AUTOMATIQUE DE LA PARAPHRASE : ÉTAT DE L'ART

La variation paraphrastique est l'une des principales sources de complexité pour les processus de traitement automatique des langues. Les thésaurus encodés manuellement sont par nature incomplets et ne sont pas disponibles pour toutes les langues. De plus, ils ne comprennent souvent pas d'expressions composées de plusieurs mots qui sont nécessaires pour produire ou reconnaître automatiquement des paraphrases plus complexes. Le besoin d'acquérir automatiquement des paraphrases à partir de corpus de textes a ainsi été à l'origine de nombreux travaux (Madnani et Dorr, 2010). De récents événements et publications scientifiques dédiés à la paraphrase révèlent l'importance actuelle de ce phénomène dans le domaine du TAL.

La nécessité d'évaluer la qualité des paraphrases candidates produites par une technique d'acquisition particulière constitue un problème central pour les systèmes d'acquisition automatique. Ceci mène, cependant, à un problème de définition circulaire de la paraphrase puisque l'extraction de la paraphrase nécessite la capacité de déterminer si deux unités de texte répondent bien aux critères attendus.

Dans ce chapitre, nous présentons un aperçu global des différentes approches proposées pour l'acquisition et l'évaluation des paraphrases phrastiques et sous-phrastiques. La section 2.1 présente des travaux visant la construction de corpus de paraphrases d'énoncés *via* des techniques automatiques, semi-automatiques et manuelles. Dans la section 2.2, nous donnons une description détaillée de techniques d'acquisition de paraphrases sous-phrastiques exploitant diverses sources de connaissances allant des ressources sémantiques à différents types de corpus dans une ou plusieurs langues. Un aperçu des méthodes proposées pour l'évaluation des systèmes d'acquisition de paraphrases est donné dans la section 2.3. Certains travaux exploitant la paraphrase pour améliorer la performance des systèmes qu'ils présentent sont décrits dans la section 2.4. Puis, la tâche d'implication textuelle, considérée comme une tâche étroitement liée à la paraphrase, est discutée dans la section 2.5.

Contrairement à la traduction, pour laquelle il existe de grandes quantités d'exemples produits par des humains, les activités humaines ne produisent pas *explicitement* de quantités importantes de paraphrases qui pourraient servir de données d'apprentissage. Par conséquent, les techniques automatiques se sont principalement appuyées sur l'observation *indirecte* d'unités de texte plus ou moins en relation d'équivalence, ce qui soulève un certain nombre de questions.

Plusieurs méthodes de construction de corpus de paraphrases d'énoncés dans différentes langues ont été proposées. Barzilay et McKeown (2001) ont distingué trois méthodes pour collecter des paraphrases : (1) l'utilisation de ressources linguistiques existantes ; (2) l'extraction de mots ou d'expressions similaires à partir de corpus ; (3) l'acquisition manuelle de paraphrases.

Les deux premières méthodes incluent les méthodes d'acquisition automatique de paraphrases. Par exemple, Langkilde et Knight (1998) se sont basés sur les connaissances sémantiques fournies par la base lexico-sémantique WordNet (Miller, 1995) pour exploiter les relations de synonymie entre termes. Ces ressources linguistiques ne sont, cependant, pas disponibles dans toutes les langues. C'est pour cela que de nombreux travaux d'acquisition de paraphrases se sont appuyés sur d'autres sources de connaissances disponibles en quantité plus importante.

Barzilay et Lee (2003) introduisent une technique d'alignement multi-séquence à partir de phrases extraites à partir d'articles journalistiques relatant les mêmes informations et provenant de différentes sources : elles factorisent sous forme de treillis celles ayant la même structure syntaxique. Dans la même direction, Dolan *et coll.* (2004) présentent une méthode non supervisée pour l'acquisition de paires d'énoncés sémantiquement similaires d'articles de presse décrivant des événements similaires. Les paraphrases qu'ils ont obtenues ont été évaluées par des juges humains et sont souvent plus *comparables* que *parallèles*, dans le sens où il n'est pas toujours possible d'associer les éléments d'une phrase à ceux de l'autre phrase. Ganitkevitch *et coll.* (2011) ont utilisé des corpus multilingues de débats parlementaires pour la production de paraphrases phrastiques.

Ces méthodes ont le défaut d'être fondées sur des algorithmes nécessitant de grandes masses de données annotées de façon fiable. Pour pallier ces problèmes, une autre manière de procéder consiste à demander à des contributeurs de paraphraser directement un énoncé dans la langue souhaitée. C'est ce qui a été fait, par exemple, pour la construction des fichiers de développement du corpus BTEC (Takezawa *et coll.*, 2002), où des phrases en japon-

ais sont tout d'abord traduites en anglais, puis paraphrasées par 15 locuteurs monolingues, créant ainsi un corpus monolingue parallèle où chaque énoncé est lié à 16 paraphrases, dont l'une peut être considérée comme « phrase de référence ».

Une autre solution est d'exprimer ce problème d'acquisition comme une tâche de traduction multiple d'un même corpus de phrases. Cette méthode a par exemple été suivie pour constituer le MTC (Multiple-Translation Chinese corpus) [1]. Ce corpus a été développé pour la traduction automatique, afin de permettre l'utilisation de plusieurs traductions de référence en traduction vers l'anglais. Il contient 105 articles (993 phrases) extraits de 3 journaux écrits en mandarin. Les phrases sources ont été traduites indépendamment en anglais par 11 traducteurs. Chaque groupe de phrases traduites comporte donc 11 traductions sémantiquement équivalentes, qui peuvent être considérées comme étant des paraphrases d'énoncés. Ce corpus constitue une source riche pour l'apprentissage de paraphrases lexicales et structurelles, et a en particulier été utilisé dans l'étude de Pang *et coll.* (2003). Barzilay et McKeown (2001), quant à elles, ont exploité un corpus de paraphrases en anglais construit via des traductions vers l'anglais de 5 romans pour la production automatique de paraphrases d'énoncés. Cohn *et coll.* (2008) utilisent également des traductions multiples afin de construire un corpus monolingue parallèle pour faire de l'annotation manuelle de paraphrases.

Il existe d'autres activités humaines qui produisent des paires d'unités textuelles qui sont parfois exploitées pour constituer des corpus contenant des paires d'énoncés plus ou moins parallèles. Ainsi, dans le cadre de l'amélioration des systèmes de réponses aux questions, plusieurs travaux ont porté sur la construction de corpus de paraphrases de questions. Bernhard et Gurevych (2008) ont construit un corpus de paraphrases de questions en exploitant des réseaux sociaux sur le Web. Dans cette expérience, elles ont recueilli un corpus constitué manuellement de questions et de leurs paraphrases à partir du site WikiAnswers [2]. Lorsqu'un utilisateur entre une question qui ne fait pas déjà partie des questions référencées sur WikiAnswers, le site web affiche une liste de questions semblables à celle que vient de poser l'utilisateur. L'utilisateur choisit la question qui éventuellement paraphrase sa propre question. Ces reformulations de questions sont stockées et peuvent être récupérées de manière automatique pour constituer un corpus de paraphrases de questions.

L'historique des révisions des ressources collaboratives telles que Wikipédia rend possible l'extraction de certains types de modifications locales reflétant l'évolution, la maturation et la correction de la

1. Linguistic Data Consortium (LDC) Catalog Number LDC2002T01, ISBN 1-58563-217-1

2. http://wiki.answers.com

forme linguistique des articles, et constituent donc une importante source de connaissances de plus en plus exploitée à ce jour (Nelken et Yamangil, 2008; Max et Wisniewski, 2010). Nous avons réalisé une étude détaillée des types de modifications dans les révisions de Wikipédia (Dutrey *et coll.*, 2011a), qui a montré qu'une quantité importante de ces modifications peuvent être considérées comme des paraphrases, bien que, pour certaines, les reconnaitre automatiquement défie les techniques d'identification de paraphrases actuelles (section 8.2 du chapitre 8).

Il est aussi possible d'acquérir directement des paraphrases ciblées en formulant la tâche d'acquisition de paraphrases sous la forme d'un jeu en ligne. Par exemple, dans ses travaux, Chklovski (2005) [3], des joueurs doivent déterminer des paraphrases partiellement masquées pour une expression donnée. Chaque joueur a la possibilité de faire plusieurs essais pour trouver la paraphrase de « référence ». Une autre manière de procéder consiste à collecter des paraphrases par le biais des plateformes de *crowdsourcing*. Amazon's Mechanical Turk (Barr et Cabrera, 2006) est le service de *crowdsourcing* qui a gagné le plus en popularité, et qui a attiré de plus en plus l'attention des chercheurs dans tous les domaines (Bernstein *et coll.*, 2010). *Via* cette plateforme, Resnik *et coll.* (2010) collectent des paraphrases pour des unités de texte difficiles à traduire automatiquement, permettant de simplifier, par la suite, le processus de traduction. Denkowski *et coll.* (2010) ont également étudié l'impact de l'exploitation des paraphrases en traduction automatique en utilisant Mechanical Turk pour filtrer des paraphrases obtenues automatiquement. Les résultats de ces travaux sont des petits corpus contenant entre 1000 et 2500 paraphrases, facilitant le contrôle manuel de leur qualité.

En formulant la tâche d'acquisition de paraphrases d'énoncés comme une tâche de description de vidéos sur Mechanical Turk, Chen et Dolan (2011) ont construit à la fois des corpus multilingues et monolingues parallèles, une ressource très intéressante pour diverses applications de TAL. Plus récemment, et dans le cadre de la détection de plagiat, Burrows *et coll.* (2012) ont présenté Webis Crowd Paraphrase Corpus 2011, un corpus contenant 4067 paraphrases de passages de texte obtenues par *crowdsourcing*. Dans une prise de position concernant la plate-forme Amazon Mechanical Turk, Sagot *et coll.* (2011), ont démontré que ce genre de plateforme de travail en ligne est loin d'être idéale et soulève de nombreux problèmes concernant l'éthique, le prix et la situation des travailleurs ainsi que la qualité des ressources linguistiques.

Enfin, diverses techniques ont été développées pour produire automatiquement des paraphrases. Celles-ci ont notamment été utilisées pour enrichir des corpus d'apprentissage de systèmes

3. 1001 Paraphrases (http://ai-games.org/paraphrase.html)

de traduction automatique (Nakov, 2008), pour optimiser leurs paramètres (Madnani *et coll.*, 2008) et d'autres types d'applications (Quirk *et coll.*, 2004; Zhao *et coll.*, 2008a; Max, 2009; Zhao *et coll.*, 2010). Cependant, outre le fait que les performances actuelles de ces techniques sont limitées, il est à noter que leur utilisation pour produire des corpus de paraphrases pose un problème car elles sont tributaires de la disponibilité des ressources d'apprentissage. Les travaux portant sur la production automatique de paraphrases sont décrits en détails dans les travaux de Madnani et Dorr (2010) et Androutsopoulos et Malakasiotis (2010).

Certains types de corpus de paraphrases d'énoncés représentatifs décrits ci-dessus sont illustrés dans la figure 1.

2.2 Acquisition de paraphrases sous-phrastiques

Les approches présentées dans la section précédente portent sur l'acquisition de paraphrases d'énoncés. Or il est également utile d'extraire des reformulations pour des unités de texte plus fines : de telles paraphrases peuvent être exploitées d'une manière plus générale par les techniques d'identification et de production de paraphrases. Dans cette section, nous décrivons les principales approches employées pour la détection et l'acquisition de paraphrases sous-phrastiques ; un panorama plus large de ces techniques se trouve là encore dans (Madnani et Dorr, 2010; Androutsopoulos et Malakasiotis, 2010).

2.2.1 *Exploitation de ressources linguistiques*

L'acquisition de paraphrases peut être réalisée à l'aide de diverses méthodologies et en exploitant diverses ressources fournissant des connaissances sémantiques. Dans les méthodes exploitant des thésaurus, par exemple, pour un segment source, une paraphrase est obtenue en remplaçant certains mots par leurs synonymes (Langkilde et Knight, 1998; Bolshakov et Gelbukh, 2004; Kauchak et Barzilay, 2006). Ces méthodes comportent généralement deux phases : extraction et validation de paraphrases candidates. Dans la première phase, des synonymes pour les termes à substituer sont extraits à partir d'un réseau sémantique tel que WordNet (Miller, 1995). Par la suite, le synonyme le plus adapté au contexte d'apparition de chaque terme est retenu. WordNet a été utilisé comme ressource principale dans plusieurs travaux entrant dans cette catégorie (Langkilde et Knight, 1998; Jing, 1998) ou a été combiné avec d'autres ressources telles que les encyclopédies (Hassan *et coll.*, 2007). Ces méthodes sont assez simples à développer, mais dépendent de la disponibilité des ressources nécessaires. Or, ces ressources, qui sont d'une manière générale de bonne qualité,

corpus comparables (articles de journaux)
(Barzilay et Lee, 2003)

> *Prague is a centuries-old city with a wealth of historic landmarks.*
> *The physical attractions and landmarks of Prague are many.*

corpus comparables (articles de journaux)
(Dolan *et coll.*, 2004)

> *The Hartford Courant reported %%day%% that Tony Bryant said two friends were the killers.*
> *A lawyer for Skakel says there is a claim that the murder was carried out by two friends of one of Skakel's school classmates, Tony Bryan.*

corpus monolingues parallèles (traductions multiples de romans)
(Barzilay et McKeown, 2001)

> *Emma burst into tears and he tried to comfort her, saying things to make her smile.*
> *Emma cried, and he tried to console her, adorning his words with puns.*

corpus parallèles alignés manuellement (questions extraites de sites communautaires)
(Bernhard et Gurevych, 2008)

> *How many ounces are there in a pound ?*
> *What's the number of ounces per pound ?*

modifications extraites de l'historique de révisions (Wikipédia)
(Dutrey *et coll.*, 2011a)

> *Ce vers de Nuit rhénane d'Apollinaire qui paraît presque sans structure rythmique*
> *Ce vers de Nuit rhénane d'Apollinaire dont la césure est comme masquée*

paraphrases générées automatiquement (corpus parallèles bilingues)
(Madnani *et coll.*, 2008)

> *(hong kong, macau and taiwan) macau passed legalization to avoid double tax.*
> *macao adopted bills to avoidance of double taxation (hong kong, macao and taiwan)*

paraphrases générées automatiquement (systèmes de traduction automatiques multiples)
(Zhao *et coll.*, 2010)

> *he said there will be major cuts in the salaries of high-level civil servants .*
> *he said there are significant cuts in the salaries of high-level officials .*

FIGURE 1: Exemples extraits de sources et techniques représentatives pour la collecte de paraphrases d'énoncés.

demandent un travail de construction important, ce qui les rend rares et particulièrement limitées pour certaines langues. En outre, les méthodes basées sur ces connaissances ne couvrent qu'un seul phénomène de la paraphrase : la synonymie.

Outre les réseaux sémantiques, des dictionnaires généralistes (Wallis, 1993; Fujita *et coll.*, 2005) et plus spécifiques tels que les dictionnaires de structures lexicales conceptuelles (Baldwin *et coll.*, 2001; Fujita *et coll.*, 2004) ont été utilisées afin d'extraire des équivalences sémantiques entre unités textuelles.

Une autre approche consiste à mettre en place des bases de règles d'identification de paraphrases locales construites manuellement (McKeown, 1983; Dras, 1999). Ces règles permettent, en effet, de décrire explicitement des patrons de segments en relation d'équivalence. Les travaux de Jacquemin (1999) sur le repérage de variantes de termes exploitant des règles de réécriture et des connaissances *a priori* entrent dans ce cadre. L'approche proposée repose sur des métarègles de réécritures morphosyntaxiques écrites manuellement ainsi que des ressources énumérant des variantes morphologiques (mots de même lemme ou liés par dérivation morphologique) et sémantiques (synonymes).

Il est important de signaler que les règles d'identification de paraphrases ne couvrent pas tous les phénomènes paraphrastiques, sont coûteuses à construire et ne produisent pas nécessairement des équivalences substituables dans tout contexte. Ces limites sont à la source des travaux qui acquièrent des connaissances depuis des corpus.

2.2.2 *Approches fondées sur des corpus*

L'acquisition de *paraphrases* repose la plupart du temps sur l'appariement préalable d'unités de plus grande taille (typiquement des paires de phrases ou des documents comparables). Ces unités peuvent être obtenues directement par un processus supervisé, tel que la traduction humaine multiple, ou l'appariement automatique fondé sur la similarité entre textes (Mihalcea *et coll.*, 2006). On observe que les techniques pour l'acquisition de paraphrases sont généralement très dépendantes des types de corpus sur lesquels elles ont été développées (Madnani et Dorr, 2010). Les principaux types de corpus peuvent être grossièrement définis par :

1. **corpus monolingue** : un corpus constitué de documents similaires, par exemple issus du Web (Pasça et Dienes, 2005).
2. **corpus multilingues parallèles** : des paires de phrases disponibles dans deux langues ou plus (Bannard et Callison-Burch, 2005) (comme les transcriptions des débats parlementaires européens).

3. **corpus monolingues comparables** : des paires de textes associés en fonction d'une mesure de similarité textuelle en suivant éventuellement certaines heuristiques (tels que les articles de journaux publiés dans un même intervalle de temps (Sekine, 2001; Dolan *et coll.*, 2004)).
4. **corpus monolingues parallèles** : des paires d'énoncés de sens équivalents alignées de façon supervisée (comme les traductions multiples de livres (Barzilay et McKeown, 2001) ou les groupes de questions ayant la même réponse (Bernhard et Gurevych, 2008)).

Dans la suite de cette section, nous décrivons les principales approches d'acquisition de paraphrases à partir des différents types de corpus évoqués.

2.2.2.1 *Acquisition depuis des corpus monolingues*

Une hypothèse largement suivie pour extraire des segments textuels équivalents est que deux mots, et par extension deux segments, apparaissant dans des contextes similaires peuvent être interchangeables. Cette hypothèse de *distributionnalité* (Harris, 1954), déjà évoquée dans le chapitre 1, a notamment été appliquée au cas de chemins de dépendances syntaxiques pour extraire des règles d'inférence (pouvant être considérées comme règles de paraphrases ici) par Lin et Pantel (2001). Dans leur algorithme (DIRT), chaque chemin dans un arbre syntaxique représente une relation binaire entre deux noms. Si deux chemins relient un même ensemble de mots, DIRT émet alors l'hypothèse que les patrons correspondants sont équivalents. Les résultats obtenus prennent donc la forme de patrons d'équivalence à deux arguments tels que : *X asks for Y*, *X requests Y*, *X's request for Y*, *X wants Y*, *Y is requested by X*, etc. Lin et Pantel ont utilisé dans cette étude un corpus monolingue constitué d'articles de journaux.

Bien que les chemins d'arbres syntaxiques servent de représentations riches et utiles, le coût associé à leur construction les rend difficilement exploitables sur de très larges collections. Afin de limiter ce coût, Pasça et Dienes (2005) ont présenté une approche mesurant la similarité entre unités textuelles fondée sur les *n*-grammes du contexte. Pour cela, les auteurs ont collecté une grande collection d'articles de journaux à partir du Web[4] pour acquérir des équivalences entre segments apparaissant dans un même contexte, représenté par un ensemble d'ancres.

Bhagat et Ravichandran (2008) ont proposé une approche similaire visant à trouver des variations locales en mesurant la similarité

4. Ces documents sont extraits à partir de l'index du moteur de recherche Google

distributionnelle dans un corpus monolingue à large échelle (contenant 25 milliards de mots), extrait de l'agrégateur de d'articles de presse Google News. Pour cela, un algorithme de *Locality Sensitive Hashing* [5] afin de réduire les dimensions des vecteurs encodant les contextes.

Outre le Web, une autre source potentielle de paraphrases locales réside dans les nombreuses modifications que les rédacteurs font lors de la révision d'un texte, certaines d'entre elles n'étant pas destinées à modifier le sens du texte, mais à améliorer sa qualité, le rendre plus cohérent, ou limiter sa redondance. Des manuscrits d'écrivains ont été notamment utilisés dans le domaine de la critique génétique textuelle dont le but est d'étudier les processus de création de textes (Bourdaillet et Ganascia, 2007). Ces documents annotés sont malheureusement disponibles en petites quantités et sont, de plus, difficiles à encoder en format électronique. En outre, ces réécritures consistent souvent en des réorganisations textuelles importantes qui sont très difficiles à exploiter pour l'acquisition de paraphrases.

L'émergence et l'adoption des *wikis* a fait de l'écriture collaborative une pratique très courante. L'encyclopédie en ligne Wikipédia, en particulier, attire de nombreuses contributions sur un large éventail de sujets et dans de nombreuses langues. Bien que certaines contributions consistent en des changements importants (par exemple la création d'un article, la suppression d'une section, la réécriture complète d'un paragraphe), une proportion importante des modifications textuelles sont effectuées sur des textes courts pour corriger, améliorer ou enrichir le contenu de l'encyclopédie. L'étude effectuée par Dutrey *et coll.* (2011a) a montré que les révisions de Wikipédia contiennent de nombreuses réécritures *naturelles* à faible variation sémantique, incluant de nombreuses paraphrases locales dans leurs contexte (voir section 8.2 du chapitre 8). Les auteurs présentent une tentative de reconnaissance automatique de ces paraphrases en exploitant un moteur de reconnaissance de variantes de termes par règles.

2.2.2.2 *Acquisition de paraphrases depuis des corpus monolingues comparables*

Des corpus comparables, dans lesquels un même contenu est probablement décrit sous plusieurs formes, permettent de guider la mise en correspondance d'équivalences locales. Les techniques développées sur ces corpus se fondent là aussi sur l'hypothèse que des unités linguistiques apparaissant de nombreuses fois dans des

5. Méthode de recherche approximative dans des espaces de grande dimension. L'idée principale est d'utiliser une famille de fonctions de hachage choisies telles que des points proches dans l'espace d'origine aient une forte probabilité d'avoir la même valeur de hachage.

contextes similaires peuvent avoir la même signification, pour extraire des correspondances entre fragments de texte. Restreindre les corpus utilisés à des textes comparables, sélectionnés sur la base d'un même genre ou de thèmes communs, permet d'augmenter la probabilité que les correspondances obtenues soient effectivement valides.

Sekine (2001) a par exemple utilisé des corpus comparables constitués d'articles journalistiques publiés le même jour, relatant les mêmes informations mais provenant de différentes sources pour acquérir des expressions synonymes. Avant de procéder à l'alignement de phrases, il effectue une opération préalable de détection automatique d'articles traitant des mêmes sujets. En reprenant cette méthode, Shinyama *et coll.* (2002) ont réalisé une expérience permettant d'acquérir des patrons de paraphrases généralisées à partir de corpus d'articles de journaux. L'intuition suggérant cette approche est qu'une même actualité parue à une époque précise peut être décrite dans divers articles de presse par des journalistes utilisant des formulations différentes. Leur algorithme permet d'extraire automatiquement des articles de presse décrivant un même événement dans un domaine donné (« affaires personnelles », « assassinat », etc.) et se fonde sur l'hypothèse suggérant que les entités nommées sont généralement conservées dans une paire de paraphrases. Dans leur méthode, une mesure de similarité est calculée sur le base du nombre d'entités nommées partagées entre deux énoncés. Plus ce nombre est grand, plus leur probabilité d'équivalence est importante. Le résultat de l'algorithme est une liste de patrons de paraphrases généralisées avec des types d'entité nommées, représentés par des variables.

Avoir recours aux entités nommées et les utiliser comme ancres pour détecter des équivalences sémantiques entre unités textuelles a été également exploré dans (Shinyama et Sekine, 2003; Marius, 2005).

D'autres méthodes ont été mises en œuvre pour tenter d'identifier directement les correspondances dans deux corpus comparables monolingues en s'appuyant sur des algorithmes d'alignement combinant plusieurs heuristiques de similarité. Barzilay et Lee (2003) ont introduit une technique d'alignement multi-séquences factorisant des phrases structurellement similaires de corpus comparables sous forme de treillis. Plus récemment, Shen *et coll.* (2006) ont proposé une approche étendant l'algorithme d'alignement de Barzilay et Lee, en lui ajoutant des contraintes syntaxiques. Cette approche vise à limiter les erreurs d'alignement pouvant conduire à l'acquisition de phrases grammaticalement incompatibles [6].

6. Leur algorithme est semblable à celui introduit par Pang *et coll.* (2003) (décrit dans la section 2.2.2.4), mais en remplaçant le corpus parallèle par un corpus comparable.

Plus récemment et pour des unités textuelles de différents niveaux de granularité, Wang et Callison-Burch (2011) présentent une approche semi-supervisée visant à extraire des documents, puis des phrases et enfin des segments en relation d'équivalence sémantique à partir d'un corpus comparable contenant plusieurs articles de journaux. Dans leurs expériences, les auteurs ont adapté une méthode utilisée en traduction automatique pour l'extraction d'équivalences de traduction et évaluent la qualité des paraphrases acquises manuellement par l'intermédiaire de Mechanical Turk.

Outre les articles journalistiques, dans certains travaux sont exploités des corpus monolingues comparables constitués dans d'autres domaines. Dans le domaine médical, Elhadad et Sutaria (2007) et Deléger et Zweigenbaum (2009) travaillent sur un corpus composé d'articles scientifiques et de leur version « grand public », pour extraire des paraphrases entre deux types de discours, grand public contre spécialisé.

Il est important de signaler qu'en dépit de leur grande disponibilité, les corpus monolingues comparables demeurent difficiles à exploiter pour la tâche d'acquisition de paraphrases sous-phrastiques. En effet, le parallélisme entre les phrases est limité à des correspondances sémantiques et thématiques effectuées en se fondant sur des événements (des ancres) décrivant un même sujet au niveau des documents. Ceci rend la tâche d'extraction de segments en relation d'équivalence plus difficile. De plus, l'absence d'informations précises sur l'emplacement des paraphrases rend leur acquisition automatique plus complexe. Il est également à noter, que les paraphrases extraites à partir de ces corpus sont éventuellement spécifiques au domaine pour lequel les corpus sont constitués, ce qui peut limiter leur utilisation ultérieure.

2.2.2.3 *Approches exploitant des corpus bilingues parallèles*

Outre les corpus monolingues, des corpus bilingues parallèles ont été exploités pour l'acquisition de paraphrases en se fondant sur l'hypothèse que des segments partageant des traductions dans une autre langue peuvent être des paraphrases dans certains contextes. Ces corpus sont constitués de textes en deux langues, alignés phrase à phrase. Leur utilisation pour la paraphrase s'explique, essentiellement, par le fait que l'opération de traduction est censée conserver le sens de la phrase d'origine. Ce type de ressource est relativement facile à constituer et d'importants corpus utilisés en traduction automatique statistique existent déjà (Patry et Langlais, 2011).

Bannard et Callison-Burch (2005) ont décrit une approche dite « par pivot » exploitant un corpus parallèle bilingue. L'intuition derrière cette approche est que les termes et les expressions d'une

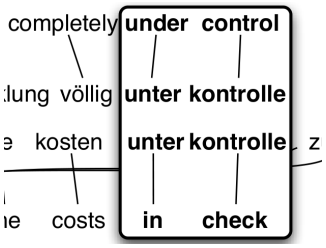


FIGURE 2: Exemple de paraphrases (*under control* ↔ *in check*) extraites par pivot à partir d'un corpus anglais-allemand

langue différente peuvent être utilisés comme *pivots* pour extraire des paraphrases : les segments alignés avec les mêmes termes dans la langue pivot sont considérés comme des paraphrases potentielles. Cela est illustré dans la figure 2 où la paraphrase *in check* ↔ *under control* est obtenue en utilisant l'allemand comme langue pivot.

Zhao *et coll.* (2008b) ont proposé une extension de cette approche permettant d'extraire des patrons à partir de tels corpus multilingues (voir figure 3). Il a été montré dans des travaux ultérieurs que l'acquisition de paraphrases par l'utilisation de plusieurs langues pivot diminue le bruit (Bannard et Callison-Burch, 2005). Plus récemment, Callison-Burch (2008) et Max (2009) ont utilisé des traductions de segments en pivot pour l'acquisition de paraphrases en prenant en compte une modélisation du contexte de la phrase d'origine.

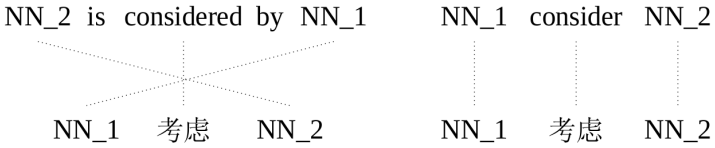


FIGURE 3: Exemple de patrons de paraphrases (NN_2 *is considered by* NN_1 ↔ NN_1 *consider* NN_2) extraites par pivot à partir d'un corpus anglais-chinois

De telles approches permettent d'identifier une grande variété de phénomènes paraphrastiques dans les paires de paraphrases produites, mais elles sont limitées par les techniques d'alignement utilisées et surtout par la disponibilité des corpus parallèles appropriés dans les langues souhaitées. L'acquisition de paraphrases par pivot fait partie des approches que nous avons étudiées dans notre travail et sera décrite plus en détail dans le chapitre 5.

2.2.2.4 *Acquistion de paraphrases depuis des corpus monolingues parallèles*

Les approches présentées ci-dessus sont limitées par le fait que de nombreux contextes ne correspondront pas à des emplois équivalents. Si des énoncés en relation de paraphrase sont disponibles (Cohn *et coll.*, 2008), alors l'extraction de segments en relation d'équivalence peut être formulée sous la forme d'une tâche d'alignement des mots et segments de la paire d'énoncés. Le plus grand avantage des corpus monolingues parallèles est que ces paires sont, par nature, sémantiquement équivalentes.

Dans ce cadre, les travaux précédents ont principalement exploité des traductions multiples produites indépendamment par différents traducteurs, qui font des choix lexicaux différents tout en préservant le sens du texte dans la langue d'origine. L'extraction de paraphrases sous-phrastiques à partir de ces corpus correspond bien alors à une mise en correspondance des segments sémantiquement similaires dans chaque paire d'énoncés. Ces dernières représentent un contexte dans lequel les paraphrases extraites sont valides.

Barzilay et McKeown (2001) ont présenté une approche d'acquisition d'unités textuelles sémantiquement similaires à partir de corpus monolingues parallèles. Elles présentent une méthode d'apprentissage non supervisé de paraphrases lexicales et syntaxiques à partir de paires de traductions proposées indépendamment. Leur méthode, exploitant des informations sur le contexte d'apparition des paraphrases candidates, est fortement inspirée de l'hypothèse de distributionnalité : les expressions des phrases alignées qui apparaissent dans des contexte similaires sont des paraphrases. Ainsi, des paraphrases telles que « *burst into tears* ↔ *cried* » et « *comfort* ↔ *console* » sont extraites à partir de la paire de paraphrases d'énoncé suivante :

> (1) *Emma burst into tears and he tried to comfort her, saying things to make her smile.*
> (2) *Emma cried, and he tried to console her, adorning his words with puns.*

En s'inspirant des approches de (Barzilay et McKeown, 2001) et (Lin et Pantel, 2001), Ibrahim *et coll.* (2003) proposent une approche non supervisée d'acquisition de paraphrases « structurelles », ou de fragments d'arbres syntaxiques sémantiquement équivalents, à partir de corpus monolingues alignés. L'équivalence sémantique au sein d'une paire est assurée ici par le partage de contextes syntaxiques (des noms ou des pronoms identiques ou ayant la même catégorie syntaxique). Ils appliquent une version modifiée de l'algorithme de similarité distributionnelle entre arbres de dépendances de Lin et Pantel (2001), conçue pour améliorer d'une part

les travaux de (Barzilay et McKeown, 2001) où les équivalences lexicales et sous-phrastiques devraient être contiguës en raison de l'absence de traitement syntaxique important et ceux de (Lin et Pantel, 2001) qui permettent d'acquérir des patrons de paraphrases sous-phrastiques pouvant parfois être antonymiques.

Il est possible d'acquérir des correspondances entre deux phrases parallèles en utilisant une approche plus directe que l'approche classique de distributionnalité. Pang *et coll.* (2003) ont présenté un algorithme exploitant la structure syntaxique d'un ensemble de traductions pour identifier des équivalences locales. Des automates à états finis encodant plusieurs paraphrases sont construits. Puis, des équivalences entre des unités textuelles de différents niveaux de granularité sont extraits. Les auteurs ont utilisé le corpus MTC comportant 11 traductions en anglais de 900 phrases issues d'articles de journaux en chinois.

Malgré leur rareté, plusieurs auteurs ont défendu le fait que les corpus parallèles monolingues sont les candidats les plus naturels pour l'acquisition de la paraphrase sous-phrastique. En effet, les phrases parallèles étant issues de la volonté d'exprimer les mêmes idées, les équivalences apprises seront par nature beaucoup plus fortes qu'en utilisant des ressources davantage « comparables ». De plus, le contexte de ces équivalences peut être extrait de façon directe, ce qui est particulièrement important pour pouvoir caractériser par la suite les conditions de leur substituabilité. Nous montrerons dans nos expériences que ces corpus contiennent une grande variété de phénomènes paraphrastiques, y compris un certain nombre qui défient les techniques automatiques actuelles (Partie ii).

Dans le tableau 1, nous présentons les travaux les plus représentatifs de l'acquisition de paraphrases sous-phrastiques classés en fonction du type de la ressource principale exploitée et de la connaissance considérée dans l'approche implémentée.

2.3 Évaluation des systèmes de paraphrase

Contrairement à la traduction automatique et malgré les efforts entrepris, le domaine de l'acquisition et de production des paraphrases manque d'une méthodologie d'évaluation « standard » complètement automatique. En effet, malgré la similarité entre les tâches de paraphrasage et de traduction automatique, plusieurs différences ne permettent pas de réutiliser les mesures établies pour l'évaluation des systèmes de traduction automatique. La plupart des travaux sur la paraphrase ont recours à des évaluations manuelles où des juges humains sont invités à répondre à des questions telle que : *Est ce que le sens est préservé ?* ou encore *Est ce que la phrase est grammaticalement correcte ?* et la performance du

Ressources exploitées	**Approches ou connaissances considérées**	
	Patrons de paraphrases	Paraphrases
Corpus monolingues	– Lin et Pantel (2001) – Bhagat et Ravichandran (2008)	– Bourdaillet et Ganascia (2007) – Dutrey *et coll.* (2011a)
Corpus monolingues comparables	– Shinyama *et coll.* (2002) – Barzilay et Lee (2003)	– Sekine (2001) – Shen *et coll.* (2006) – Deléger et Zweigenbaum (2009) – Wang et Callison-Burch (2011)
Corpus bilingues parallèles	– Zhao *et coll.* (2008b)	– Bannard et Callison-Burch (2005) – Zhou *et coll.* (2006) – Callison-Burch (2008) – Max (2009)
Corpus monolingues parallèles	– Ibrahim *et coll.* (2003)	– Barzilay et McKeown (2001) – Pang *et coll.* (2003)
Autres	[Langkilde et Knight (1998), Jacquemin (1999), Dras (1999) Fujita *et coll.* (2005), Hassan *et coll.* (2007)]	

Tableau 1: Synthèse des travaux sur l'acquisition de paraphrases en fonction des ressources utilisées et du type de connaissances considéré

système est mesurée en calculant le pourcentage des réponses positives obtenues. Cependant, ce type d'évaluation est coûteux et difficilement reproductible, ce qui empêche la comparaison des performances des techniques proposées. En outre, la complexité inhérente à la tâche rend cette évaluation encore plus difficile à en juger par la difficulté à obtenir de forts accords inter-annotateurs sur la tâche la plus élémentaire d'identification de paraphrases sous-phrastiques dans des paires d'énoncés en relation paraphrastique (Cohn *et coll.*, 2008).

En outre, l'évaluation des systèmes de paraphrase est limitée par une contrainte imposant le fait qu'une paraphrase *utile* doit être,

généralement, lexicalement différente de l'énoncé d'origine, tout en préservant évidemment son sens. C'est pour cette raison qu'une mesure comme BLEU (Papineni *et coll.*, 2002), qui compare une traduction candidate et une traduction de référence uniquement au niveau lexical, ne peut réellement permettre de distinguer les « bonnes » paraphrases des « mauvaises ».

Callison-Burch *et coll.* (2008) ont proposé PARAMETRIC, une mesure automatique spécifique à l'évaluation des systèmes d'acquisition de paraphrases. Dans PARAMETRIC, un ensemble de paraphrases acquises *via* une technique automatique est comparé à des alignements de référence réalisés manuellement pour le même ensemble de phrases (ou corpus d'évaluation). Des mesures classiques de *précision* et de *rappel* sont calculées en comptant le nombre de paires communes aux alignements de référence et aux alignements des paraphrases proposées. La limite principale de cette approche réside dans le fait qu'elle dépend beaucoup de la couverture du corpus d'évaluation et des choix des annotateurs.

En génération de paraphrases, Liu *et coll.* (2010), ont introduit la mesure PEM (Paraphrase Evaluation Metric), utilisant une langue pivot afin d'évaluer en plus de l'équivalence lexicale de surface, une équivalence sémantique entre des paraphrases candidates produites par un système donné. Ils proposent 3 critères définissant ce qu'ils appellent de *bonnes* paraphrases : la corrélation sémantique, le caractère naturel et la dissimilarité lexicale. Des scores représentant chaque critère sont associés aux paraphrases candidates. Les trois scores sont par la suite combinés en utilisant un classifieur SVM utilisant comme données d'entraînement des évaluations fournies par des juges humains. Bien que PEM ait montré une corrélation acceptable avec les jugements humains, cette mesure présente quelques limites : elle ne modélise que des paraphrases au niveau sous-phrastique ; elle nécessite un corpus bilingue parallèle approprié, ce qui présente un biais favorable aux techniques d'acquisition de paraphrases par pivot, par exemple.

Une autre tentative pour l'évaluation des paraphrases d'énoncés générées automatiquement a été présentée par (Chen et Dolan, 2011). Dans la mesure qu'ils proposent, la conservation de sens est évaluée en utilisant BLEU avec plusieurs références. Pour mesurer la dissimilarité lexicale, les auteurs ont introduit PINC, un score mesurant le nombre de n-grammes différents entre deux paraphrases candidates. PINC consiste essentiellement en un score BLEU inversé, ce qui permet de réduire le nombre de n-grammes communs pour les deux phrases. Selon cette mesure, une *bonne* paraphrase doit partager un petit nombre de n-grammes avec la phrase source (garantissant une valeur élevée de dissimilarité lexicale), mais de nombreux n-grammes avec les phrases de référence (indiquant la conservation du sens d'origine). Ils construisent et

utilisent pour cela un corpus de descriptions multiples de vidéos. Leur mesure est très dépendante de l'existence de cette ressource.

Tout comme dans PARAMETRIC, (Metzler *et coll.*, 2011) ont présenté une mesure d'évaluation de paraphrases phrastiques obtenues *via* 3 techniques d'acquisition opérant à des niveaux différents, en les comparant à une référence construite par des annotateurs humains sur MECHANICAL TURK. Cette mesure combine le *rappel*, défini par le pourcentage de phrases pour lesquelles le système a renvoyé au moins une paraphrase, et la *précision* correspondant au nombre de paraphrases correctes parmi les paraphrases renvoyées.

En conclusion, le domaine d'acquisition de paraphrases manque d'une méthodologie d'évaluation automatique largement reconnue et utilisée. Le jugement humain des transformations dans une phrase entière est très complexe et de bons accords inter- et intra-annotateurs sont difficiles à atteindre avec des hypothèses ayant des qualités comparables. Utiliser des paraphrases d'énoncés pour appuyer une tâche donnée (comme en traduction automatique où l'on fournit par exemple aux systèmes de traduction des références alternatives afin d'optimiser leurs performances (Madnani *et coll.*, 2008)) peut être considéré comme une solution pour l'évaluation extrinsèque de la qualité des paraphrases. Mais les résultats publiés n'indiquent pas clairement que les améliorations sur une tâche sont corrélées avec la qualité des paraphrases utilisées.

2.4 APPLICATIONS DU TAL EXPLOITANT DES PARAPHRASES

Compte tenu des difficultés autour de l'évaluation automatique des méthodes d'acquisition de paraphrases et leur forte dépendance envers des annotations faites par des humains, plusieurs auteurs ont eu recours à l'évaluation extrinsèque de la qualité des paraphrases acquises en les intégrant dans une application afin d'améliorer ses performances (Duclaye, 2003; Marton *et coll.*, 2009).

Chevelu (2011) a distingué trois classes d'applications exploitant les paraphrases. La première classe regroupe les applications visant à produire des paraphrases pour se substituer à la phrase source. Ainsi, en résumé automatique, l'identification des paraphrases permet de condenser les informations contenues dans plusieurs documents (McKeown *et coll.*, 2002; Barzilay, 2003) et d'améliorer la qualité des résumés automatiques (Hirao *et coll.*, 2004). Produire une paraphrase plus courte qu'une phrase d'origine permet de condenser un texte (Knight et Marcu, 2000), une étape primordiale en résumé automatique. Dans cette classe figure aussi les applications de normalisation de texte visant à utiliser un vocabulaire contrôlé (Nasr, 1996) et celles ayant pour but d'aider des auteurs à trouver des formulations plus adaptées (Max, 2008).

La deuxième classe d'applications consiste à améliorer les performances de divers systèmes de TAL. Une des applications les plus fréquentes de la paraphrase est la génération automatique de différentes formulations de requêtes (Duclaye, 2003) ou de patrons (Ravichandran et Hovy, 2002) à soumettre à des systèmes de réponses à des questions. Les paraphrases permettent également d'améliorer le repérage d'une réponse à une question (McKeown, 1979; Pasça, 2003; Duboue et Chu-Carroll, 2006), d'une part et d'empêcher la pénalisation d'un passage contenant une formulation équivalente à une expression clé mais pas l'expression elle-même (Pasça et Dienes, 2005). Les paraphrases sont également utiles dans la tâche de catégorisation et d'alignement des textes ayant des caractéristiques similaires, en extraction d'information, ce qui réduit la disparité dans les modèles d'extraction (Shinyama et Sekine, 2003, 2005).

Un système de reconnaissance de paraphrases peut également permettre de détecter des plagiats (White et Joy, 2004; Uzuner *et coll.*, 2005; Burrows *et coll.*, 2012), de produire des phrases plus adaptées pour la synthèse vocale (Chevelu, 2011), d'améliorer les systèmes de traduction automatique (Callison-Burch *et coll.*, 2006; Marton *et coll.*, 2009; Max, 2010), simplifier des textes (Marsi et Krahmer, 2005; Zhu *et coll.*, 2010; Coster et Kauchak, 2011), ou encore l'adapter au lectorat (Deléger et Zweigenbaum, 2009).

La dernière classe mentionnée par Chevelu comporte les applications où les paraphrases sont utilisées pour améliorer l'évaluation des systèmes de TAL. La paraphrase est utilisée, par exemple, en traduction automatique pour autoriser des formulations différentes (Papineni *et coll.*, 2002; Lepage et Denoual, 2005; Kauchak et Barzilay, 2006; Snover *et coll.*, 2006; Nakov, 2008). En évaluation des systèmes de résumé automatique des documents, les résumés générés automatiquement sont souvent évalués en se fondant sur des résumés de référence proposés par des humains. Zhou *et coll.* (2006) ont proposé la mesure ParaEval qui exploite une base de paraphrases sous-phrastiques lors du calcul des n-grammes communs entre les résumés de référence et ceux proposés par un système.

L'évaluation de la paraphrase dans ces travaux est guidée par la tâche s'en servant. Ces travaux ne s'intéressent pas explicitement au phénomène paraphrastique et mesurent plutôt la capacité des systèmes automatiques à exploiter ses caractéristiques pour améliorer leurs performances.

2.5 Paraphrase et implication textuelle

Un problème étroitement lié à celui de l'identification de paraphrases est celui de la reconnaissance d'implications textuelles (do-

maine de la déduction sur textes) (Androutsopoulos et Malakasiotis, 2010). Une expression T est en relation d'implication textuelle avec une hypothèse H si un humain ayant maîtrise raisonnable d'une langue et des connaissances standard sur le monde peut déduire de l'expression T que l'hypothèse H est vraie. Elle consiste donc à déterminer si un texte en *implique* un autre. La reconnaissance d'inférences permet, par exemple, d'aider à déterminer si une phrase est une réponse à une question (Bernard, 2011), ou encore de détecter la redondance dans un résumé. Il existe de nos jours une campagne annuelle d'évaluation des systèmes d'inférence textuelle (Dagan *et coll.*, 2006).

L'implication textuelle et la paraphrase portent, en effet, sur des aspects pertinents de la sémantique en langue. Les deux phénomènes partagent une propriété : pour pouvoir établir une de ces relations entre deux expressions, il faut avoir une bonne compréhension du contexte dans lequel elles peuvent être vérifiées. Pour les distinguer, on se base sur la direction de validité logique de la relation dans une paire de textes donnée. L'inférence est une relation *directionnelle* entre deux expressions dans laquelle l'une implique l'autre, alors que la paraphrase est une relation dans laquelle deux unités textuelles expriment le même sens. La tâche de reconnaissance de paraphrases peut donc être formulée comme une reconnaissance d'*implication textuelle bidirectionnelle*, et la reconnaissance d'implication textuelle peut être abordée comme un problème de reconnaissance de *quasi-paraphrases*. La distinction entre ces deux phénomènes dépend fortement du contexte.

Dans ce chapitre, nous avons tout d'abord présenté les méthodes existantes de construction de paraphrases d'énoncés. Puis, nous avons présenté un aperçu général des approches les plus représentatives développées pour l'acquisition de paraphrases sous-phrastiques exploitant différents types de ressources. Il est à noter que ces approches ont en commun d'être fortement liées aux types de ressources auxquelles elles s'appliquent. En fait, dans la plupart des travaux mentionnés, le type de corpus utilisé a un impact direct sur les performances, mais aucune modélisation plus générale de ce que sont les paraphrases sous-phrastiques ne semble avoir clairement émergé, les liens entre ces différents travaux étant souvent difficiles à établir.

Nous verrons dans la partie suivante que nous adoptons certaines approches évoquées dans ce chapitre, principalement celle basée sur des corpus monolingues parallèles telles que la méthode proposée par Pang *et coll.* (2003), pour l'étude de la paraphrase sous-phrastique. Ces corpus sont considérées par plusieurs comme le *matériau naturel* où des segments sémantiquement équivalents peuvent exister (Barzilay et McKeown, 2001; Pang *et coll.*, 2003; Madnani et Dorr, 2010).

Avant de détailler nos choix de techniques et de ressources utiles pour l'étude des paraphrases sous-phrastiques, faisant l'objet de cette thèse, nous présentons dans le chapitre suivant une analyse contrastive des différentes sources d'acquisition de paraphrases d'énoncés, ressource essentielle pour étudier finement le phénomène paraphrastique.

Deuxième partie

ACQUISITION DE PARAPHRASES SOUS-PHRASTIQUES DEPUIS DES PAIRES DE PHRASES

3

CONSTRUCTION DE CORPUS DE PARAPHRASES D'ÉNONCÉS

Nous avons vu, à travers les travaux de l'état de l'art cités dans le chapitre précédent, que les méthodes d'acquisition automatique de paraphrases sous-phrastiques sont étroitement liées aux types de ressources exploitées. L'acquisition de ces connaissances est souvent réalisée en effectuant tout d'abord une étape de collecte de paires d'énoncés supposés avoir des sens liés. Les activités humaines ne produisent pas *explicitement* de quantités importantes de paraphrases phrastiques, ce qui donne une importance particulière aux travaux visant la constitution de corpus monolingues parallèles. Cependant, la nature des paires de paraphrases obtenues a généralement un effet considérable sur la quantité et la qualité des paraphrases sous-phrastiques que ces corpus contiennent.

Ce chapitre est consacré à une étude quantitative et qualitative de l'impact des sources à partir desquelles des paires de paraphrases d'énoncés en français sont acquises sur leur degré de parallélisme. Cette étude s'avère nécessaire pour guider le choix de la source la plus appropriée relativement aux objectifs d'une étude particulière sur la paraphrase sous-phrastique.

Nous décrivons dans un premier temps cinq stratégies d'acquisition de paraphrases d'énoncés (section 3.1). Puis, nous présentons une expérience d'annotation manuelle effectuée au niveau sous-phrastique que nous avons menée sur des échantillons de 50 paires d'énoncés. Ces paires sont issues de corpus construits selon les stratégies choisies (section 3.2). Enfin, une typologie détaillée des paraphrases sous-phrastiques trouvées manuellement dans chaque corpus est proposée (section 3.3).

3.1 CORPUS DE PAIRES D'ÉNONCÉS POUR L'ACQUISITION DE PARAPHRASES

Nous détaillons dans cette partie la méthode de constitution de différents types de corpus servant de matériau à la tâche d'acquisition de paraphrases à partir de paires d'énoncés en relation. Un corpus pour chaque type a été construit et comporte 50 paires d'énoncés. Les cinq stratégies choisies pour la collecte des paires de paraphrases représentent divers types de signaux de contenus sémantiques pouvant être exploités pour l'étude de la paraphrase. Comme les analyses présentées dans ce chapitre le montrent, ces stratégies nous ont permis d'obtenir des corpus de paires d'énoncés à différents niveaux de parallélisme sur lesquelles des techniques d'acquisition de paraphrases auront vraisemblablement des performances différentes.

1. **Traductions multiples à partir d'une seule langue (Texte$_{en \rightarrow fr}$)** : des paires d'énoncés obtenus par traductions multiples d'un même texte à partir d'une seule langue source vers une même langue cible.
2. **Traductions multiples à partir de plusieurs langues (Texte$_{xx \rightarrow fr}$)** : des paires d'énoncés résultant de traductions multiples d'un même texte à partir de plusieurs langues vers une même langue cible.
3. **Traductions multiples de dialogues oraux (Parole)** : des paires d'extraits de textes résultant de traductions multiples de mêmes extraits de dialogues oraux de films dans une même langue cible.
4. **Descriptions multiples de vidéos (Scène)** : des paires d'énoncés résultant de descriptions multiples d'une même scène visuelle.
5. **Titres d'articles de journaux (Événement)** : des paires d'énoncés résultant de descriptions multiples d'un même événement ou de deux événements proches, dans une même langue.

Des exemples de chacun de ces corpus sont donnés dans le tableau 2.

3.1.1 *Traductions multiples*

Une façon d'obtenir des énoncés en relation de paraphrases est de traduire plusieurs fois indépendamment un même énoncé. En effet, deux traductions d'un même énoncé doivent conserver le sens d'origine, ceci indépendamment des langues source et cible. Néanmoins, chaque traducteur pourra faire des choix de traduction différents, qui mèneront donc dans la majorité des cas à des traductions différentes.

Source	Paraphrases
TEXTE$_{en \rightarrow fr}$	Plusieurs orateurs ont considéré que ceci avait trop tardé. Plusieurs locuteurs ont jugé cela nécéssaire depuis longtemps.
TEXTE$_{xx \rightarrow fr}$	Plusieurs intervenants l'ont considéré comme une chose indispensable. Le retard avec lequel s'accomplie cette étape a été souligné dans de nombreuses interventions.
PAROLE	On ne voudrait pas qu'ils imaginaient qu'on n'est pas heureux Personne ne doit douter de notre bonheur conjugal
SCÈNE	Superman déplace des rochers. Superman dégage l'entrée d'une grotte bloquée par des rochers.
ÉVÉNEMENT	Algues vertes : un décret favorisera leur prolifération. Algues vertes : parution d'un décret controversé sur l'épandage.

Tableau 2: Exemples de paraphrases acquises à partir de sources différentes

Nous avons construit le corpus MULTITRAD (Bouamor, 2010) selon ce principe, en obtenant des traductions multiples par des humains vers le français.

L'expérience a consisté à soumettre un ensemble de 500 phrases, exprimées chacune dans 10 langues européennes (anglais, allemand, italien, portugais, espagnol, néerlandais, danois, suédois, finnois et grec) à un groupe de participants francophones en leur demandant de traduire chacune des phrases tout en conservant autant que possible le sens d'origine [1]. Les 500 phrases proposées sont extraites aléatoirement du corpus parallèle multilingue Europarl (Koehn, 2005), composé de transcriptions des débats parlementaires européens. Nous avons choisi Europarl car celui-ci comporte un ensemble de phrases disponibles en 11 langues [2] (les 10 langues d'origine des phrases à traduire et le français), ce qui est une caractéristique rare pour un corpus parallèle.

Nous avons mesuré l'impact de la langue source de traductions multiples sur le degré de comparabilité des paraphrases obtenues. Pour cela, nous avons mesuré un degré de similarité lexicale entre les paraphrases obtenues à partir des différentes langues pour des paires de langues contenant au moins 20 paires de traductions. Le tableau 3 regroupe les moyennes des similarités obtenues entre différentes paires de langues d'origine en utilisant le coefficient de chevauchement (CC) qui représente le pourcentage de chevauche-

1. MULTITRAD contient, en plus des paraphrases correctes, des paraphrases comportant des erreurs causées par des fautes d'orthographe (*nécéssaire, peche, apprôché*), des erreurs de traductions (*M. Le Président* au lieu de *Mme La Présidente*), des phrases non grammaticales (telle que celle le premier exemple de du tableau 2) ou encore la présence de phrases incomplètes.

2. La version d'Europarl que nous avons utilisée ne nous permet pas de connaître la langue d'origine pour chaque phrase.

ment lexical entre les vocabulaires P_1 et P_2 de deux phrases, défini par :

$$CC = \frac{|P_1 \cap P_2|}{min(|P_1|, |P_2|)} \tag{2}$$

Par exemple, nous observons que les 172 [3] paraphrases obtenues à partir de l'anglais comportent 90% de formes communs en moyenne. En revanche, les paraphrases provenant de deux langues différentes comportent entre 36% et 42% [4] de formes différents en moyenne. Ces valeurs montrent que nous obtenons davantage de variations lexicales en traduisant à partir de langues différentes.

	Toutes formes				
	en	**es**	**de**	**it**	**pt**
en	0,90 172	0,64 69	0,59 89	0,63 84	0,62 58
es	*	-	0,62 57	0,63 57	0,64 51
de	*	*	-	0,58 67	0,61 53
it	*	*	*	-	0,65 50
pt	*	*	*	*	-
	Lemmes des mots pleins				
en	0,90 172	0,65 69	0,61 89	0,66 84	0,64 58
es	*	-	0,57 57	0,68 57	0,68 51
de	*	*	-	0,59 67	0,62 53
it	*	*	*	-	0,66 50
pt	*	*	*	*	-

Tableau 3: Valeurs de similarité lexicale entre groupes d'au moins 20 paires de paraphrases pour tous types de formes (partie gauche) et uniquement pour les lemmes de mots pleins (partie droite)

Nous avons refait ces mesures, mais en ne considérant cette fois que les mots pleins pour la mesure de chevauchement, ce qui élimine de fait les nombreuses *formes* correspondant à des signes de ponctuation et à des mots grammaticaux. Les résultats sont donnés dans la partie droite du tableau 3. Nous observons dans ce cas

3. C'est le nombre indiqué en indice dans le tableau 3 représentant le nombre de traductions communes obtenues à partir de deux langues.

4. Ces valeurs représentent les complémentaires de celles figurant dans la partie gauche du tableau 3.

qu'en moyenne le degré de similarité entre les paires d'énoncés est plus faible[5].

Nous avons extrait deux sous-corpus particuliers constitués de groupes de paraphrases pour les 50 mêmes énoncés à partir de MULTITRAD. Le premier (TEXTE$_{en \rightarrow fr}$) est constitué de paraphrases obtenues par traductions indépendantes depuis un même énoncé en anglais (en) ; un second corpus (TEXTE$_{xx \rightarrow fr}$) est constitué de paraphrases obtenues par traduction depuis l'allemand (de), l'espagnol (es), l'italien (it) et le portugais (pt). L'objectif poursuivi ici est de pouvoir comparer les paraphrases obtenues lorsqu'elles résultent de la traduction depuis la même langue ou depuis des langues diverses, correspondant donc à plusieurs traductions. Un exemple pour un même énoncé d'origine dans les deux corpus est donné figure 4.

Langue	**Phrase source dans plusieurs langues et traductions proposées en français**
en	There should be absolutely no ambiguity about our message.
fr	Notre message devrait être parfaitement clair.
es	No debe haber la menor ambigüedad en nuestro mensaje.
fr	Il ne doit pas y avoir la moindre ambiguïté dans notre message.
de	Unsere Botschaft muß eindeutig sein.
fr	Notre message doit être clair.
it	Non può esserci ambiguità alcuna nel nostro messaggio.
fr	Il ne peut y avoir aucune ambiguïté dans notre message.
pt	Não deve haver qualquer ambiguidade na nossa mensagem.
fr	Il ne devrait y avoir aucune ambiguïté dans notre message.

FIGURE 4: Exemples de traductions obtenues à partir de plusieurs langues sources pour l'énoncé « *Il ne doit y avoir aucune ambiguïté dans notre message.* »

Nous disposons donc de deux corpus constitués chacun de 50 groupes de 4 paraphrases. Nous considérons une paraphrase de chaque groupe comme une phrase de référence, qui devra être alignée avec chacune des 3 autres paraphrases de son groupe. Nous avons alors constitué 300 paires de paraphrases (2 corpus * 50 groupes * 3 alignements). Pour chaque groupe de quatre paraphrases, la paraphrase la plus similaire en moyenne aux autres paraphrases a été identifiée et associée aux trois autres. Cette similarité est calculée par la moyenne des taux d'édition entre énoncés telles que mesurée par la mesure TER (Translation Edit Rate) (Snover *et coll.*, 2006). Les 50 paires les plus similaires ont été retenues pour cette étude.

5. Pour effectuer les opérations de lemmatisation et d'analyse morphosyntaxique, nous avons utilisé l'outil TREETAGGER (Schmid, 1994).

3.1.2 *Traductions multiples de sous-titres*

Sur le Web, les sous-titres de films et de séries télévisées sont généralement proposées par des contributeurs volontaires dans différentes langues. Deux versions de sous-titres issues de la traduction d'un même dialogue vers une même langue cible, peuvent contenir des paires d'énoncés sémantiquement équivalents.

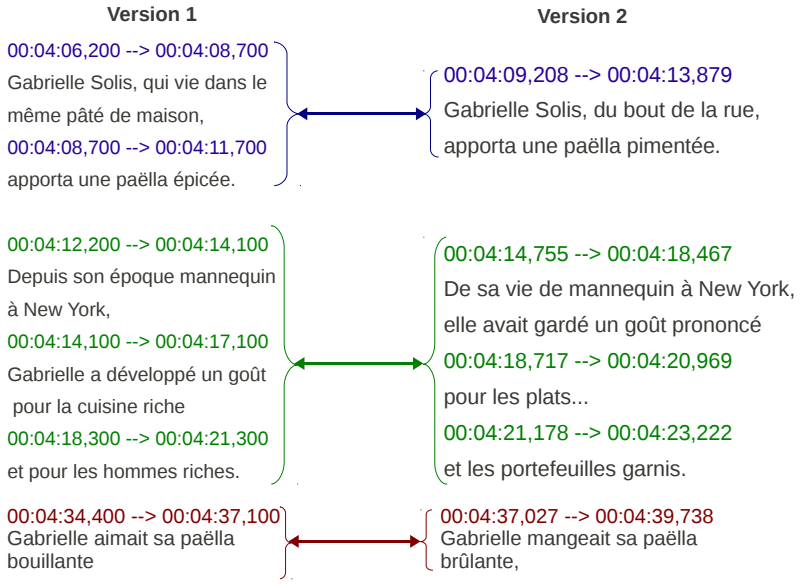


FIGURE 5: Alignement de segments extraits de deux versions de sous-titres en français de la série « *Desperate Housewives* »

C'est cette hypothèse que nous avons suivie pour construire un corpus, appelé PAROLE dans la suite de ce document, contenant 50 paires d'énoncés à partir de deux versions de sous-titres. Nous avons utilisé la série télévisée tournée en anglais américain *Desperate Housewives*[6]. Nous avons tout d'abord aligné ce corpus en utilisant l'algorithme décrit dans (Tiedemann, 2007), basé sur des indices de durée et développé pour l'alignement de sous-titres multilingues. Pour pouvoir utiliser cet algorithme, nous avons effectué un ensemble d'opérations de pré-traitement permettant de nettoyer la base de données initiale et de convertir les fichiers de sous-titres en des fichiers XML[7]. Un exemple d'alignement produit par cet algorithme est donné dans la figure 5[8], où les intervalles de

6. Ces sous-titres sont librement disponibles sur `http://www.opensubtitles.org`

7. les script utilisés sont accessibles sur : `http://opus.lingfil.uu.se/tools`

8. Les exemples sont données tels que produits, ce qui explique les quelques fautes (« vie » au lieu de « vit »).

temps utilisés pour l'alignement sont en couleur. Dans cette figure, sont illustrés des alignements impliquant une seule phrase telle que *Gabrielle aimait sa paëlla bouillante* ↔ *Gabrielle mangeait sa paëlla brûlante* et ceux impliquant un ensemble de phrases telle que la paire apparaissant en bleu dans l'exemple.

Nous avons ensuite retenu des paires d'énoncés en dessous d'un seuil minimal de taux d'édition afin de garantir un minimum de similarité pour les paires d'énoncés extraits et filtré manuellement les erreurs manifestes de l'algorithme d'alignement, nous permettant d'obtenir des paires telles que celles présentées dans la figure 2.

3.1.3 *Descriptions multiples de vidéos*

Outre le texte et la parole transcrite, d'autres types de médias peuvent servir de support pour acquérir des paraphrases, notamment les scènes représentées dans des vidéos. Chen et Dolan (2011) ont ainsi formulé la tâche d'acquisition de paraphrases comme une tâche de description de vidéos par le biais du *crowdsourcing* pour construire un corpus contenant des groupes de descriptions multiples de courtes vidéos dans plusieurs langues.

Nous avons construit Scène, un corpus de 50 paires de descriptions en français extrait à partir du corpus de Chen et Dolan (2011). Un point important est que les descriptions en français dans cette ressource sont disponibles dans de faibles quantités. Ce corpus ne comporte que 1197 descriptions de vidéos faites en français, dont seulement 80 vidéos ont été décrites plus d'une fois. En outre, aucune des phrases proposées n'a le statut « vérifiée » indiquant que leur contributeur a été sélectionné de façon supervisée . Un exemple d'une telle paire de descriptions pour une vidéo impliquant un super héros est donné dans la figure 2.

3.1.4 *Titres d'articles de journaux sur le même sujet*

La collecte des titres des articles journalistiques relatant les mêmes informations et provenant de différentes sources peut être considérée comme une source intéressante pour l'acquisition de paraphrases phrastiques. En effet, un même événement peut être décrit par différents journalistes et agences de presse de diverses manières, mettant en jeu différentes expressions. Un tel corpus peut être facilement collecté à partir du Web, par exemple *via* des agrégateurs d'informations tels que Google News[9], qui mettent à disposition des quantités importantes de pointeurs vers des articles de presse regroupés par sujet.

Wubben *et coll.* (2009) ont ainsi constitué un corpus contenant des paraphrases d'énoncés potentielles. Nous avons de la même

9. http://news.google.com

manière extrait des titres pour 100 ensembles d'articles d'actualité provenant de Google contenant au total 1462 titres. Nous avons, ensuite, affiné l'algorithme de regroupement en ne retenant que les paires de titres correspondant à des articles dont les dates de publication n'étaient pas séparées de plus d'un jour. Puis, pour chaque groupe de phrases, nous avons retenu les paires ayant le plus petit taux d'édition (tel que donné par la mesure TER) au-dessus d'un seuil fixé empiriquement. Dans chaque groupe, une première itération permet d'extraire les paires d'énoncés les plus similaires. Puis, afin d'assurer une couverture maximale de l'ensemble des groupes, le reste des paires ont été extraites à partir des groupes déjà utilisés. Nous avons ainsi obtenu ÉVÉNEMENT, un corpus comportant 50 paires de titres appariés par la méthode décrite. Un exemple d'une paire de titres d'articles obtenue pour un sujet d'actualité français est illustré dans la figure 2.

3.2 EXPÉRIENCE D'ANNOTATION MANUELLE ET ANALYSE DES RÉSULTATS

Afin de comparer les paires d'énoncés obtenues par chaque approche, nous avons calculé différentes mesures de similarité, incluant des mesures utilisées dans l'évaluation des systèmes de traduction automatique :

- *coefficient de chevauchement (CC)*, représentant le pourcentage de chevauchement lexical entre les vocabulaires de deux phrases (calculé ici sur les lemmes des mots pleins) ;
- score BLEU (Papineni *et coll.*, 2002), basé sur la précision *n*-grammes ;
- score TER (Snover *et coll.*, 2006), basé sur un taux d'édition [10] ;
- score METEOR (Lavie et Agarwal, 2007), basé sur la moyenne harmonique de la précision et du rappel *n*-grammes [11].

	Statistiques			Similarité			
	# formes	# formes différentes	# formes$_{/\text{phrase}}$	CC	TER	BLEU	METEOR
TEXTE$_{\text{en}\rightarrow\text{fr}}$	2 138	320	21,4	73,3	56,7	28,2	55,1
TEXTE$_{\text{xx}\rightarrow\text{fr}}$	2 594	494	25,5	71,9	70,6	19,3	42,7
PAROLE	1 426	520	14,3	66,3	89,2	14,0	31,1
SCÈNE	676	70	6,8	72,2	54,7	20,1	39,3
ÉVÉNEMENT	925	441	9,2	67,2	52,6	31,8	52,2

Tableau 4: Propriétés de tous les corpus, avec une moyenne des similarités des 50 paires d'énoncés.

10. La similarité entre les phrases d'une paire est inversement proportionnelle à la valeur renvoyée par cette mesure.

11. La synonymie, considérée lors de la troisième passe du calcul de l'alignement entre les deux phrases, n'a pas été prise en compte dans le calcul de ce score puisque Wordnet n'est pas disponible en français.

Le tableau 4 indique différentes statistiques pour les corpus collectés. La première observation est que le corpus construit par traductions multiples contient des phrases significativement plus longues que les autres types de corpus en termes de nombre de formes différentes, plus de deux fois plus longues que celles de celui contenant des sous-titres. La description de vidéos produit en revanche des phrases très courtes (moyenne de 6,8 mots par phrase).

Nous avons réalisé une annotation manuelle des paraphrases sous-phrastiques dans les corpus décrits ci-dessus, en suivant l'essentiel des consignes décrites dans (Cohn *et coll.*, 2008) [12] à l'aide de l'outil d'annotation en ligne YAWAT (Germann, 2008).

50 paires de paraphrases ont été annotées indépendamment par deux annotateurs francophones. Les principales consignes étaient que les paraphrases *sûres* et *possibles* devaient être distinguées, que les alignements les plus petits devaient être privilégiés sans décourager néanmoins les alignements groupe-à-groupe ($n - m$), et que les phrases devaient être alignées autant que possible. Nous ne considérerons dans la suite, pour toutes les statistiques et les expériences, que les paraphrases qui ne sont pas des paires identiques (telles que (*petit pont de bois* ↔ *petit pont de bois*)), car on peut les considérer comme triviales au regard de la tâche d'acquisition.

	Accord inter-annotateurs		**Stat. sur les formes dans les paraphrases** (sans les paraphrases identiques)			
			para. sûres		para. possible	
	para. sûres	para. possible	% formes [13]	# formes	% formes	# formes
TEXTE$_{en \to fr}$	65,3	15,9	40,46	865	17,26	369
TEXTE$_{xx \to fr}$	61,2	13,4	19,01	493	28,03	727
PAROLE	82,7	20,8	23,50	335	35,13	501
SCÈNE	42,8	9,3	8,14	55	4,73	32
ÉVÉNEMENT	67,8	3,8	10,59	98	13,62	126

Tableau 5: Statistiques sur l'annotation manuelle des paraphrases sous-phrastiques extraites à partir des paires d'énoncés dans chaque type de corpus divisées en paraphrases *sûres* et *possibles*.

Le tableau 5 montre les valeurs d'accords inter-annotateurs où, pour chaque type de paraphrase, nous calculons la moyenne des valeurs de rappel obtenues en prenant chaque annotateur comme

12. Voir http://staffwww.dcs.shef.ac.uk/people/T.Cohn/paraphrase_guidelines.pdf

13. Le pourcentage de formes des paraphrases sûres par rapport à toutes les paraphrases obtenues.

référence. Les valeurs obtenues pour les paraphrases sûres sont acceptables (accords de 65,3 pour TEXTE$_{en \rightarrow fr}$ et 82,7 pour PAROLE), mais celles obtenues pour les paraphrases possibles sont assez faibles (accord de 3,8 pour ÉVÉNEMENT). Ce dernier résultat était relativement prévisible, étant donné les difficultés d'interprétation pour les paraphrases pouvant entrer dans cette catégorie.

Le tableau 5 indique également les pourcentages et les proportions de paraphrases pour chaque niveau de certitude pour chacun des corpus. Nous obtenons approximativement le même nombre de paraphrases sûres et possibles (resp. 1846 sûres et 1755 possibles) pour l'union de tous les corpus.

Les autres résultats remarquables concernent les deux corpus obtenus par traductions multiples. Ces corpus ont à peu près le même nombre de formes participant à des paraphrases (1234 contre 1220), mais les proportions des paraphrases *sûres* et *possibles* sont inversées : les paraphrases trouvées dans les traductions proposées à partir d'une seule langue semblent être plus certaines pour nos annotateurs que celles obtenues à partir de plusieurs langues où nous obtenons en outre beaucoup plus de variation lexicale (cf. Tableau 4). TEXTE$_{en \rightarrow fr}$ contient beaucoup plus de paraphrases que les autres corpus. Cependant, si l'on considère uniquement les paraphrases possibles, PAROLE est le corpus contenant le plus grand nombre de paraphrases sous-phrastiques comparé à SCÈNE qui comporte nettement le plus petit nombre de paraphrases.

Nous remarquons donc que les méthodes basées sur la traduction facilitent la présence et l'identification des paraphrases, surtout quand elles sont obtenues à partir d'une même langue. Cependant, de tels corpus sont d'une disponibilité limitée puisque les traductions multiples ne sont pas disponibles à large échelle. En outre, elles ne sont pas disponibles pour tous les domaines, ne couvrent pas plusieurs langues et leur construction est une tâche coûteuse nécessitant des compétences en traduction. Les titres d'articles de journaux sont eux, en comparaison, plus faciles à acquérir mais appartiennent à un genre et à des domaines limités, ce qui implique que les paraphrases extraites auront peut être des applications plus limitées.

3.3 TYPOLOGIE DES PARAPHRASES SOUS-PHRASTIQUES PAR CORPUS

Après l'étude comparant paraphrases *sûres* et *possibles* présentée dans la section précédente, nous examinons à présent la nature des paraphrases sous-phrastiques obtenues dans chaque corpus. Nous commençons par décrire les différents types de paraphrases que nous distinguons, puis nous rendons compte des résultats obtenus pour nos différents types de corpus.

Les différentes classes définies pour la catégorisation des paraphrases sont les suivantes :

- **synonymie** (Syno.) : équivalences lexicales ou au niveau des segments non décomposables (par exemple, *Vous allez adorer ça* ↔ *ça va vous plaire*) ;
- **variations pragmatiques** (Pragma.) : paraphrases très particulières au contexte d'acquisition (par exemple, *au Bangladesh* ↔ *dans ce pays*) ;
- **variations typographiques** (Typo.) : comporte toutes les variations des nombres (lettres ou chiffres), ainsi que l'utilisation des acronymes (par exemple, *UE* ↔ *Union Européenne*) ;
- **inclusion** : comporte les cas où l'un des segments constituant une paire de paraphrases est plus précis que l'autre (par exemple, *droits qu'ils ont acquis* ↔ *droits acquis*) ;
- **variations morphologiques** (Morpho.) : variations de la formation des mot (par exemple, *en Chine* ↔ *chinois*) ;
- **variations syntaxiques** (Synt.) : où les paraphrases correspondent à différentes constructions syntaxiques (par exemple, *Souvenons-nous* ↔ *On se rappelle*) ;
- **variations d'accord** (Accord) : cette classe contient les variations de nombre et genre dans les segments nominaux et adjectivaux ainsi que la variation de temps pour les verbes (par exemple, *souhaites* ↔ *souhaitais*).

Dans cette typologie, nous reprenons quelques classes définies dans les typologies existantes (détaillées dans le chapitre 2) telles que les variations syntaxiques, les inclusions, etc. Nous avons choisi de définir les types de paraphrases en nous basant sur des exemples observés dans nos corpus d'étude. De plus, les typologies présentées alors concernent en grande partie des paraphrases au niveau phrastiques, qui sortent de notre étude.

	Accord	**Inclusion**	**Typo.**	**Morpho.**	**Synt.**	**Syno.**	**Pragma.**
Texte$_{en \rightarrow fr}$	28,5	2,1	9,0	3,0	6,6	46,9	3,6
Texte$_{xx \rightarrow fr}$	20,7	20,7	9,3	0,7	4,3	43,6	0,7
Parole	33,8	8,9	5,3	0,0	5,3	46,4	0,0
Scène	14,2	8,0	14,2	3,5	11,6	45,5	2,6
Événement	12,2	16,0	19,7	7,4	8,6	28,3	7,4

Tableau 6: Distribution (en pourcentages) des catégories de paraphrases dans les 50 paires de paraphrases par type de corpus étudié.

Le tableau 6 indique les résultats obtenus pour chaque catégorie de paraphrases. Nous observons que la *synonymie* (lexicale ou sous-phrastique) est la classe la plus représentée pour tous les corpus.

Cependant, cette classe ne représente que 28,3% pour Événement, une valeur nettement plus faible que celles des autres corpus. Nous remarquons ensuite que le corpus Événement contient des proportions équilibrées des autres types de paraphrases, alors que, par exemple, la *synonymie* et l'*accord* correspondent aux 3/4 de toutes les paraphrases dans le corpus $\textsc{Texte}_{\texttt{en}\rightarrow\texttt{fr}}$. Le nombre de variations typographiques dans le corpus Événement est probablement dû à la quantité importante des acronymes utilisés dans les paraphrases. Enfin, en comparant les deux corpus obtenus par traductions multiples, nous observons que la principale différence réside dans le nombre d'inclusions qui est plus important dans les traductions effectuées à partir de plusieurs langues.

CONCLUSION DU CHAPITRE

Nous avons proposé dans ce chapitre une analyse des paraphrases sous-phrastiques présentes dans différents types de corpus constitués de paires d'énoncés sémantiquement liés en français.

Tout d'abord, nous avons remarqué que les paires d'énoncés les plus parallèles dans notre étude sont celles obtenues par traductions multiples à partir d'une même langue source et que les paires les plus dissimilaires sont extraites à partir des descriptions de vidéos. Il n'est en revanche pas surprenant que les paires d'énoncés les plus similaires contiennent le plus grand nombre de paraphrases sous-phrastiques et plus particulièrement de paraphrases *sûres*. Cependant, ce type de corpus est difficile à construire. Néanmoins, les corpus contenant deux versions de sous-titres et les titres d'articles de journaux sont plus faciles à collecter. Ces corpus peuvent contenir des paires d'énoncés assez similaires et, par conséquent, un nombre intéressant de paraphrases sous-phrastiques même s'ils comportent plus de paraphrases *possibles* que de paraphrases *sûres*. Nous avons aussi pu remarquer que les titres d'articles contiennent des paraphrases variées.

4

ACQUISITION DE PARAPHRASES SOUS-PHRASTIQUES DEPUIS DES PARAPHRASES D'ÉNONCÉS

Après avoir décrit et comparé différentes méthodes de construction de corpus de paraphrases d'énoncés, nous présentons, dans ce chapitre, une étude détaillée de la tâche d'acquisition de paraphrases sous-phrastiques. Nous justifions le choix de corpus monolingues parallèles, constitués de paires d'énoncés sémantiquement équivalents obtenues par traductions multiples. Si ces ressources sont évidemment rares, nous défendons le fait qu'elles sont les candidates les plus naturelles pour l'observation de la paraphrase sous-phrastique. En outre, les énoncés parallèles étant issues de la volonté d'exprimer les mêmes idées, les équivalences apprises seront par nature beaucoup plus fortes qu'en utilisant des ressources que l'on peut qualifier de « comparables » plutôt que de « parallèles ». De plus, le contexte de ces équivalences peut être extrait de façon directe, ce qui est particulièrement important pour pouvoir caractériser par la suite les conditions de leur substituabilité. Nous montrerons que ces corpus contiennent une grande variété de phénomènes paraphrastiques qui défient les techniques automatiques actuelles. Nous avons suivi les principes généraux de l'approche décrite par Cohn *et coll.* (2008), dans laquelle des paires d'énoncés en relation de paraphrase sont alignées manuellement au niveau des mots, et des techniques d'acquisition sont comparées sur leur capacité à trouver les paires de paraphrases sous-phrastiques de la référence et sur la qualité des paires qu'elles prédisent.

Le but de cette étude approfondie est de répondre à des questions sur l'acquisition de paraphrases sous-phrastiques : De quel type de connaissances a-t-on besoin ? Quelles techniques doit-on implémenter ? Quel est l'impact de la langue et de la comparabilité des paires d'énoncés utilisées sur les paraphrases acquises ?

Nous présentons, dans un premier temps, une étude détaillée de la tâche d'acquisition de paraphrases sous-phrastiques avec des expériences menées sur deux langues, l'anglais et le français, avec cinq techniques d'acquisition développées originellement pour des besoins divers, mettant en œuvre différents principes. Ces techniques ont notamment été sélectionnées pour le type de traitements ou de ressources qu'elles mettent en jeu, ainsi que leur complé-

mentarité potentielle. La première est fondée sur l'apprentissage statistique d'alignements entre mots ; la deuxième est fondée sur l'-expression symbolique de la variation entre termes ; la troisième est fondée sur l'alignement de structures syntaxiques ; la quatrième est fondée sur un taux d'édition entre séquences de mots ; la cinquième est fondée sur des traductions communes dans une langue pivot ;

Ce chapitre est organisé de la façon suivante : nous allons tout d'abord présenter notre cadre expérimental (section 4.1), puis les différentes techniques étudiées (section 4.2). Nous détaillerons et analyserons les résultats obtenus (section 4.3). Enfin, nous décrivons des analyses portant sur l'impact du degré de comparabilité des corpus sur les performances des techniques d'acquisition seront décrites (section 4.4).

4.1 EXPÉRIENCES EN ACQUISITION DE PARAPHRASES SOUS-PHRASTIQUES

L'étude présentée dans ce chapitre a pour objectif initial d'étudier les caractéristiques des paraphrases sous-phrastiques difficiles à obtenir par des techniques d'acquisition représentatives des approches proposées, afin de mettre en évidence les types de connaissances requises. Pour cela, il est nécessaire d'établir une distinction claire entre les paraphrases et les autres phénomènes de reformulation. Nous commençons par décrire ici les données et la méthodologie d'évaluation choisies pour effectuer nos expériences.

4.1.1 *Langues et corpus*

Les expériences que nous détaillerons dans la suite de ce document ont été menées sur deux langues, l'anglais et le français. Pour chaque langue, nous avons constitué un corpus contenant 125 paires d'énoncés utilisées pour le développement et le réglage des paramètres, et 500 paires d'énoncés pour l'évaluation. Nous avons choisi d'utiliser des corpus monolingues parallèles, la ressource la plus directe pour observer et acquérir des relations d'équivalence entre segments tel que nous l'avons montré dans le chapitre 3. En effet, en travaillant sur de tels corpus, naturellement denses en paraphrases sous-phrastiques, nous espérons extraire des paraphrases *précises*.

Pour l'anglais, le corpus de traductions multiples MTC du LDC [1] a été développé afin de soutenir le développement des systèmes automatiques d'évaluation de la traduction. 11 traducteurs humains ont été sollicités pour traduire en anglais des articles journalistiques écrits en chinois mandarin. Cohn *et coll.* (2008) présentent une description détaillée de ce corpus et l'utilisent, avec d'autres ressources, pour créer un corpus monolingue parallèle annoté au niveau sous-phrastique. Un exemple d'énoncés parallèles obtenu manuellement par traduction multiple est donné dans la figure 6.

Pour le français, nous avons utilisé le corpus CESTA [2]. Celui-ci a été construit dans le cadre d'une campagne d'évaluation des systèmes de traduction automatique. Il comporte des paires de paraphrases obtenues par traductions multiples en français à partir de l'anglais. La figure 7 illustre un exemple de traductions extraites de ce corpus.

1. disponible sur : http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2002T01
2. http://www.elda.org/article125.html

Favorable to Hong Kong's long-term economic prosperity and stability.
Good for the long-term prosperity and stability of Hong Kong's economy.
Help ensure long-term prosperity and stability in Hong Kong's economy.
It is helpful for Hong Kong to have a long-term prosperous and steady economy.
Beneficial for the stability and prosperity of Hong Kong in the long run.
Long-term. stable prosperity advantageous to Hong Kong's economy.
Beneficial to the long-term prosperity and stability of HK economy.
It's advantageous to the long-term prosperity and stability of Hong Kong.
It helps to maintain Hongkong's economy's long-term prosperity and stability.
Bringing lasting prosperity and steady economic growth to Hong Kong.
It will assist in stabilizing long-term prosperity for Hong Kong's economy.

FIGURE 6: Exemple de 11 traductions d'un même énoncé extraites du MTC.

Suspicion de détournement de crédits communautaires d'aide au développement dans l'île de Saint-Martin.
Le présumé détournement des Fonds de Développement CE à St Martin.
Détournements supposés de Fonds de Développement de la CE à St-Martin.
Détournement allégué des fonds de développement de la CE dans la rue Martin.

FIGURE 7: Exemple de 4 traductions d'un même énoncé extraites de CESTA.

4.1.1.1 *Annotation et analyse*

Deux annotateurs ont réalisé une annotation des paraphrases dans les corpus décrits ci-dessus, en suivant les consignes décrites dans la section 3.2 du chapitre 3. Le but de cette tâche d'annotation est de distinguer au sein des paires d'énoncés les segments en relation de paraphrase en distinguant les paraphrases *sûres* et *possibles*.

La table 7 donne différentes statistiques décrivant les corpus anglais et français utilisés ainsi que les valeurs d'accords inter-annotateurs calculés sur des sous-ensembles de 50 paires d'énoncés annotées indépendamment par deux annotateurs. Les deux corpus sont de taille comparable, avec des énoncés légèrement plus longs pour le français (24 mots par énoncé en moyenne contre 21 pour l'anglais). Un détail intéressant concerne les valeurs d'accords inter-annotateurs proches obtenues pour les deux langues sur les paraphrases sûres (66,1 pour l'anglais et 64,6 pour le français). Ces

Statistiques du corpus 500 paires d'énoncés		**Accord inter-annotateurs** 50 paires d'énoncés		**Stat. sur les formes dans les paraphrases** (sans les paraphrases identiques)			
				para. sûres		para. possibles	
# formes	# formes /énoncé	para. sûres	para. possibles	% formes	# formes	% formes	# formes
anglais							
21 473	21,0	66,1	20,4	18,6	4 004	12,3	2 651
français							
24 641	24.0	64,6	16,6	29,2	7 218	6,2	1 527

Tableau 7: Description des sous-corpus extraits de MTC (pour l'anglais) et de CESTA (pour le français) et des annotations de référence pour les paraphrases obtenues.

valeurs sont généralement considérées comme acceptables pour ce genre de tâche (Cohn *et coll.*, 2008), mais celles obtenues pour les paraphrases possibles sont faibles (20,4 et 16,6 pour l'anglais et le français, respectivement). Ce dernier résultat était relativement prévisible, étant donné le nombre d'interprétations pour les paraphrases entrant probablement dans cette catégorie. Ce défaut d'accord ne constituera toutefois pas un problème dans la suite pour nos évaluations par rapport à ce qui constitue notre référence : comme nous le verrons dans la section 4.3, notre mesure d'évaluation ne les considèrera pas comme des solutions attendues, et se limitera à ne pas les considérer comme fausses lorsqu'elles apparaitront parmi les hypothèses d'un système.

Le tableau 7 montre enfin combien de paraphrases (en pourcentage et en nombre) sont obtenues pour chaque niveau de certitude (*sûres* ou *possibles*) et pour chacun des corpus. Nous obtenons des ensembles de paires d'énoncés de tailles comparables en nombre d'occurences de formes pour l'anglais (21 473) et le français (24 641). Le corpus anglais contient quasiment deux fois plus de paraphrases *sûres* (4 004) que de paraphrases annotées comme *possibles* (2 651). En français, on trouve encore davantage de paraphrases *sûres* (7 218) qu'en anglais. Ceci peut s'expliquer par deux faits principaux : premièrement, le corpus français a été obtenu par traduction à partir d'une langue source plus proche du français (l'anglais), que ne l'est le chinois de l'anglais. Il n'est donc pas surprenant qu'il soit plus facile d'aligner des énoncés qui sont plus similaires. Deuxièmement, les annotateurs, ayant travaillé indépendamment et sur une seule langue à la fois, ont peut-être interprété différemment la distinction entre les deux types de paraphrases.

Les exemples donnés dans les figures 8 et 9 illustrent la complexité de la tâche d'alignement au niveau des mots entre deux énoncés en relation de paraphrase en anglais et en français. Les alignements présentés ont été obtenus par annotation manuelle par des annota-

teurs. Bien que les accords inter-annotateurs obtenus soient acceptables pour cette tâche, de nombreux défauts apparaissent. En particulier, dans la figure 8, nous observons l'impact direct du choix typographique ainsi que de l'absence de segmentation en occurences de forme pour le segment *US$9*, ce qui impose l'alignement du segment *US$9 billion* avec le segment *9 billion US dollars*.

4.1.2 *Méthodologie d'évaluation*

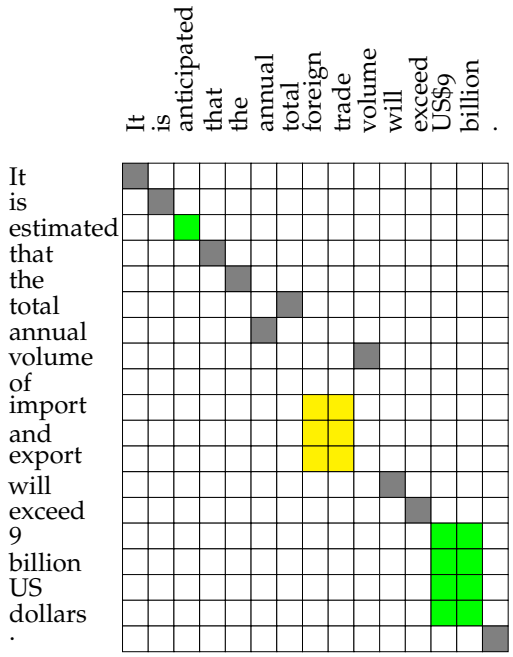
Nous adoptons la méthodologie PARAMETRIC de Callison-Burch *et coll.* (2008) pour évaluer la performance des systèmes sur la tâche d'acquisition de paraphrases depuis des paires d'énoncés. Dans cette méthode, un ensemble de paraphrases candidates extraites d'une paire d'énoncés est comparé à un ensemble de paraphrases de référence, obtenues par annotation manuelle décrivant les alignements entre mots.

La performance d'une technique se décompose en des valeurs usuelles de *précision* (P) et de *rappel* (R). La première valeur correspond à la proportion de paires d'hypothèses de paraphrases produites par un système, ensemble noté H, qui sont correctes par rapport à l'ensemble de référence composite contenant les paraphrases *sûres* et *possibles*, noté E_{tout}. Les paires de paraphrases composites de référence sont obtenues en joignant des paraphrases atomiques ou composites adjacentes. Par exemple, à partir de la matrice donnée dans la figure 8, la paire de paraphrases composites *foreign trade volume* ↔ *volume of import and export* est, par exemple, impossible à cause de la présence de la préposition *of* qui n'est alignée avec aucun mot de le second énoncé. Cela n'est pas le cas pour l'exemple en français de la figure 9, où, en plus de la paire *qui devront l'être* ↔ *à effectuer*, il est possible de constituer la paire *ou qui devront l' être* ↔ *ou à effectuer*.

Le rappel est obtenu en calculant la proportion de l'ensemble de référence de paraphrases *sûres*, noté $E_{sûr}$, qui sont trouvées par un système. Nous calculons également une valeur de *F-mesure* (F_1) combinant précision et rappel en leur donnant une même importance. Ces mesures d'évaluation sont donc définies de la manière suivante :

$$P = \frac{|H \cap E_{tout}|}{|H|} \quad R = \frac{|H \cap E_{sûr}|}{|E_{sûr}|} \quad F_1 = \frac{2PR}{P+R}$$

Il est important de noter sur ces définitions que la façon dont les ensembles E_{tout} et $E_{sûr}$ de paires de paraphrases de référence sont définis garantit que les hypothèses de paraphrases incluant les paraphrases de référence annotées comme *possibles* ne pénaliseront pas la précision, sans toutefois augmenter le rappel.

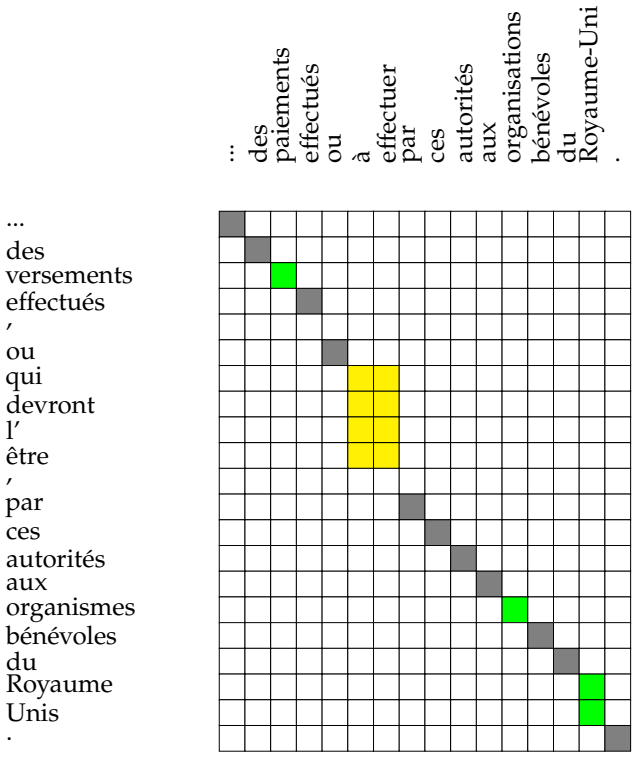


Paraphrases composites sûres		
It is estimated	↔	It is anticipated
is estimated	↔	is anticipated
estimated	↔	anticipated
estimated that	↔	anticipated that
estimated that the	↔	anticipated that the
estimated that the total annual	↔	anticipated that the annual total
total annual	↔	annual total
will exceed 9 billion US dollars	↔	will exceed US$9 billion
exceed 9 billion US dollars	↔	exceed US$9 billion
9 billion US dollars	↔	US$9 billion
9 billion US dollars .	↔	US$9 billion .
Paraphrases composites possibles		
import and export	↔	foreign trade

FIGURE 8: Alignements de référence sûrs (en vert) et possibles (en jaune) pour une paire de traductions en anglais extraite à partir du corpus de référence MTC et liste des paraphrases composites obtenues à partir de ces alignements. L'identité est marquée en gris.

4.2 TECHNIQUES INDIVIDUELLES POUR L'ACQUISITION DE PARAPHRASES SOUS-PHRASTIQUES

Dans le but d'obtenir des alignements entre paraphrases d'énoncés, nous avons implémenté cinq techniques proposées dans la lit-



Paraphrases composites sûres		
des versements	↔	des paiements
des versements effectués	↔	des paiements effectués
versements	↔	paiements
versements effectués	↔	paiements effectués
	⋮	
organismes bénévoles	↔	organisations bénévoles
organismes	↔	organisations
Royaume Unis	↔	Royaume-Uni
Paraphrases composites possibles		
qui devront l' être	↔	à effectuer
ou qui devront l' être	↔	à effectuer ou

FIGURE 9: Alignements de référence sûrs (en vert) et possibles (en jaune) pour un extrait de paire de traductions en français extraite à partir du corpus de référence CESTA, et liste des paraphrases composites obtenues à partir de ces alignements. L'identité est indiquée en gris

térature et développées initialement pour des besoins différents. Ces techniques ont été choisies parce qu'elles opèrent à des niveaux

différents, reposent sur des principes différents et exploitent des ressources distinctes, ce qui notamment devrait permettre de tirer parti de leur complémentarité potentielle.

La première est fondée sur l'apprentissage statistique d'alignements entre mots (Mot), et requiert donc des données d'apprentissage en quantité relativement importante. La seconde exploite des règles de description de variantes de termes et des connaissances *a priori* sur la variation lexicale (Terme). La troisième utilise la structure syntaxique des énoncés pour mettre en correspondance des segments (Synt), et requiert par conséquent un analyseur syntaxique. La quatrième calcule une transformation au niveau des mots pour passer d'une séquence de mots à une autre en mettant en jeu des opérations de transformation dont le coût est appris automatiquement (Edit). La cinquième, enfin, exploite des équivalences de traduction obtenues via une langue pivot (Pivot). L'évaluation des performances de chacune de ces techniques sera présentée dans la section 4.3.

4.2.1 *Apprentissage d'alignements entre mots (*Mot*)*

En traduction automatique statistique fondée sur les segments, où des corpus parallèles bilingues en quantités importantes sont disponibles, des correspondances entre segments dans les deux langues servent de base à l'apprentissage de modèles de traduction. L'approche la plus utilisée consiste à apprendre des alignements entre mots dans chaque direction de traduction (Och et Ney, 2003), à symétriser les résultats puis à utiliser une heuristique d'extraction de segments alignés. Les modèles d'alignement alors appliqués pour aligner les mots de deux énoncés parallèles résultent de l'apprentissage sur l'ensemble des bitextes disponibles. Ainsi, plus la quantité de données est importante et plus les données sont "parallèles" (c'est-à-dire que les traductions sont relativement littérales, par opposition à des traductions telles que des traductions d'idiomes), et plus les correspondances au niveau de chaque énoncé sont précises.

L'application au cas monolingue a déjà été essayée, par exemple par Quirk *et coll.* (2004) qui ont traité le problème de la réécriture de phrases comme un problème de traduction automatique monolingue. Pour cela, ces auteurs ont construit des systèmes complets de traduction statistique sur la base d'un corpus monolingue constitué de paires d'énoncés extraites depuis des corpus comparables à l'aide d'heuristiques "raisonnables" [3].

Les techniques d'alignement purement statistiques nécessitent typiquement des quantités de données importantes pour produire

3. À notre connaissance, aucune tentative d'utilisation de corpus parallèles monolingues n'a jamais été décrite dans aucune application, vraisemblablement du fait des faibles quantités de données disponibles.

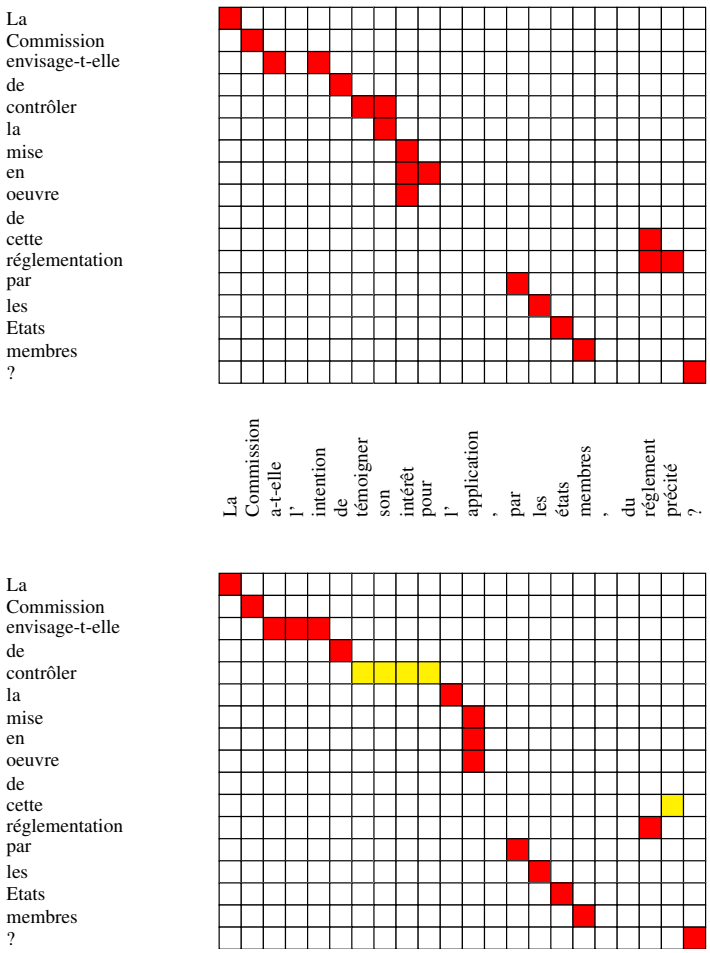


FIGURE 10: Matrice d'alignement pour une paire d'énoncés en relation de paraphrase produite par la technique MOT (partie supérieure), et matrice correspondante dans la base de référence (partie inférieure).

des alignements fiables. Afin d'augmenter la quantité de données d'apprentissage et d'améliorer ses capacités d'alignement, nous avons mis à disposition de notre système MOT un corpus parallèle contenant toutes les paires de paraphrases possibles (obtenues en mettant en correspondance des paires de paraphrases appartenant à des groupes constitués de plusieurs paraphrases pour un même énoncé) obtenues par des traductions multiples. Ceci constitue donc un avantage pour cette technique, car les autres techniques n'exploiteront pas l'information provenant d'autres paires d'énoncés pour construire leurs alignements. Nous utilisons le système MOSES (Koehn *et coll.*, 2007) pour réaliser l'extraction des bi-segments, lequel utilise GIZA++ (Och et Ney, 2003) pour l'alignement dans chaque direction de traduction puis des heuristiques d'extraction de bi-segments.

L'aligneur GIZA++ est un programme implémentant des algorithmes d'alignement au niveau des mots sur des corpus parallèles pour calculer des modèles de traduction. Cet outil[4] repose sur une combinaison de modèles génératifs, les modèles IBM (Brown *et coll.*, 1993), qui sont des modèles appris de manière non supervisée à partir des grands corpus alignés phrase à phrase. Ces modèles sont décrits dans plusieurs travaux en traduction automatique statistique (Allauzen et Yvon, 2011). Bien qu'ils soient conçus à l'origine pour la tâche d'alignement bilingue entre mots pour la traduction automatique statistique, rien n'empêche leur utilisation dans un cadre monolingue.

Giza	*(La Commission a-t-elle l' intention de ↔ La Commission envisage-t-elle de), (de contrôler la ↔ de témoigner son), (a-t-elle l' intention ↔ envisage-t-elle) (de contrôler la mise en œuvre ↔ de témoigner son intérêt pour), (contrôler la ↔ témoigner son), (contrôler la mise en œuvre ↔ témoigner son intérêt pour), (mise en œuvre ↔ intérêt pour), (cette réglementation ↔ règlement précité), (par les Etats ↔ par les États), (par les Etats membres ↔ par les États membres), (les Etats ↔ les États), (les Etats membres ↔ les États membres), (Etats ↔ États), (Etats membres ↔ États membres)*
Référence Sûre	*(La Commission a-t-elle l' intention ↔ La Commission envisage-t-elle), (La Commission a-t-elle l' intention de ↔ La Commission envisage-t-elle de), (Commission a-t-elle l' intention ↔ Commission envisage-t-elle), (Commission a-t-elle l' intention de ↔ Commission envisage-t-elle de), (a-t-elle l' intention ↔ envisage-t-elle), (a-t-elle l' intention de ↔ envisage-t-elle de), (l' ↔ la), (l' application ↔ la mise en œuvre), (application ↔ mise en œuvre), (par les États ↔ par les Etats), (par les États membres ↔ par les Etats membres), (les États ↔ les Etats), (les États membres ↔ les Etats membres), (États ↔ Etats), (États membres ↔ Etats membres), (règlement ↔ réglementation)*
Référence Possible	*(témoigner son intérêt pour ↔ contrôler), (précité ↔ cette)*

Tableau 8: Paraphrases extraites à partir des matrices d'alignement données dans la figure 10

Sur la matrice d'alignement de mots obtenue, nous appliquons les règles suivantes pour l'extraction de bi-segments correspondant à des paraphrases locales candidates : considérant un segment du côté *source* s_i^{i+m-1} commençant au mot d'indice i et de taille m, et un segment du côté *cible* c_j^{j+n-1}, les deux segments forment un

4. Librement disponible sur : http://code.google.com/p/giza-pp

bi-segment si tous les mots source (resp. cible) dont l'indice est compris entre i et $i + m - 1$ (resp. j et $j + n - 1$) sont alignés avec au moins un mot du segment cible (resp. source) et ne sont alignés qu'avec des mots de ce segment [5].

La figure 10 présente un exemple de matrice d'alignement produite par MOT : dans cet exemple, 12 paraphrases différentes sont trouvées parmi les 21 paraphrases de la référence (voir le tableau 8.).

4.2.2 *Expression symbolique de la variation (*TERME*)*

Outre des mots communs, deux énoncés en relation de paraphrase peuvent utiliser des mots en relation d'équivalence (des synonymes en contexte) ou plus généralement des groupes de mots en relation d'équivalence (définissant la notion que nous utilisons de paraphrases locales). Pour chaque paire de tels groupes, et sous certaines hypothèses, il est possible d'exprimer des règles régissant les variations syntagmatiques et paradigmatiques acceptables. Dans le domaine de l'acquisition terminologique, par exemple, les *termes* d'un domaine peuvent connaître des variations linguistiques importantes, dont le repérage automatique a été l'objet de nombreux travaux (Bourigault et Jacquemin, 1999; Castellví *et coll.*, 2001). Ceci offre donc une solution assez directe au problème de mise en correspondance étudié ici.

Le système FASTR (Jacquemin, 1999) [6], un outil dédié au repérage de variantes terminologiques de nature morpho-syntaxique, définit par un système de métarègles (écrites manuellement) s'appliquant à des règles de termes les variations acceptables. Ces métarègles permettent d'exprimer les réécritures morphosyntaxiques possibles, ainsi que les relations (au niveau des lemmes) d'ordre morphologique ou sémantique contenues dans des ressources préexistantes (familles morphologiques et sémantiques). Elles permettent, par exemple, de repérer une relation de variation entre « *les groupes de développement du squelette* » et « *développement squelettique* » (dérivation morphologique), « *fixation de l'azote* » et « *fixation biologique de l'azote* » (variation syntaxique de type insertion).

Un exemple de métarègle, illustré dans la figure 11, nommée ici **NAtoVASyn**, porte sur un segment formé par un **N**om suivi d'un **A**djectif, qui peut être réécrit en un segment formé au minimum d'un **V**erbe suivi d'un **N**om et d'un **A**djectif (N1 A1). La réecriture n'est autorisée que si le nom et le verbe appartiennent à la même famille morphologique (attribut `root`) et que les deux adjectifs sont

5. Contrairement à ce qui est fréquemment suivi en traduction statistique fondée sur les segments (Koehn *et coll.*, 2003), nous n'ajoutons pas les mots non alignés préents aux frontières des bi-segments extraits.

6. Disponible pour le français et l'anglais sur : `http://perso.limsi.fr/Individu/jacquemi/FASTR`

connus comme synonymes (attribut `syn`). Une telle métarègle permet par exemple de reconnaître le segment *protéger de façon permanente* comme une variante du terme *protection constante*. Le schéma de la règle initiale est donné par la partie gauche de la métarègle, le schéma de la variation par la partie droite.

```
Metarule NAtoVAsyn
  ( X1 -> N1 A1) = X1 -> V1 <ART? | PRON? | PREP?> N A2:
    <N1 root> = <V1 root>
    <A1 syn> = <A2 syn>
    <X1 metaLabel> = 'XX'.
```

FIGURE 11: Exemple de métarègle de `Fastr` permettant de reconnaitre *protéger de façon permanente* comme variante de *protection constante*.

L'opération d'indexation contrôlée de FASTR consiste à repérer dans un corpus l'ensemble des variantes correspondant à une liste de termes fournie en entrée. Nous exploitons ce mécanisme de la manière suivante : considérant une paire de paraphrases d'énoncés, nous recherchons avec TERME dans la première phrase (notre corpus) des variantes pour chacun des segments possibles de l'autre énoncé (à concurrence d'une certaine taille), puis nous inversons la recherche et retenons l'intersection des résultats. L'usage que nous faisons du moteur de détection de variantes de termes semble *a priori* favorable à l'obtention d'une bonne précision. À l'inverse, les règles définies pour le repérage de variantes de termes ne sont pas nécessairement les mieux adaptées pour assurer une bonne couverture des phénomènes paraphrastiques entre segments de nature quelconque, comme démontré dans l'étude réalisée dans (Dutrey *et coll.*, 2011a).

4.2.3 *Alignement de structures syntaxiques (SYNT)*

L'exploitation du caractère parallèle de deux énoncés peut être poussée encore plus loin : si ces énoncés partagent une même structure syntaxique de haut niveau, il est possible de réaliser un alignement guidé par des représentations syntaxiques permettant de faire apparaître des correspondances sous-phrastiques. C'est cette idée qui est mise en avant dans l'approche de fusion syntaxique proposée par Pang *et coll.* (2003), illustrée sur la figure 12. Cet algorithme a été initialement proposé afin de produire de nouvelles paraphrases d'énoncés. Il permet de construire des graphes de mots encodant plusieurs paraphrases, en se basant sur leurs indices syntaxiques. Il est alors possible d'extraire de chaque graphe des

équivalences de différents niveaux de granularité : lexicale, sous-phrastique et phrastique. Même si la fusion est guidée par les constituants, les résultats peuvent être de nature quelconque.

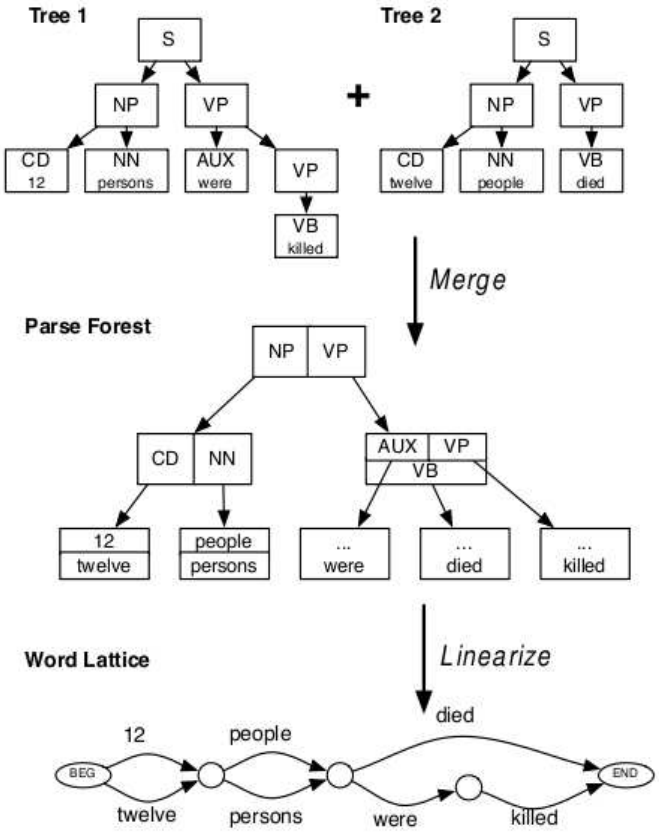


FIGURE 12: Processus de fusion de deux arbres syntaxiques et création de graphe de mots de Pang *et coll.* (2003)

Deux énoncés sont tout d'abord analysés syntaxiquement (cf. partie supérieure de la figure 12). Une étape de fusion est alors effectuée de la façon suivante : deux sous-arbres sont récursivement fusionnés si leur catégorie racine et la liste de leurs catégories filles sont les mêmes. Lorsque les listes de filles diffèrent, une liste d'alternatives est créée. Par ailleurs, un mécanisme de blocage lexical empêche toute fusion si un mot plein, présent dans la descendance d'une catégorie fille du premier arbre fusionné, se retrouve dans la descendance d'une autre catégorie fille du second arbre, pour éviter de fusionner à tort deux nœuds qui ne seraient pas équivalents.

La forêt d'analyse ainsi construite (partie centrale de la figure 12) est finalement linéarisée pour construire un automate à états finis (en bas de la figure 12), lequel peut être minimisé pour fusionner

tous les arcs correspondant à des hypothèses partageant de mêmes préfixes ou suffixes. Les sous-chemins partageant un même nœud de départ et un même nœud d'arrivée représentent donc des segments en relation d'équivalence. Il est alors possible d'extraire de façon efficace d'un tel automate l'ensemble des bi-segments correspondant à des paraphrases candidates.

Notre réimplémentation de cet algorithme a mis en évidence plusieurs limitations que nous avons tenté de corriger en essayant plus particulièrement d'améliorer sa robustesse.

Tout d'abord, l'algorithme décrit par Pang *et coll.* (2003) arrête toute fusion si un blocage lexical est actif et n'utilise pas le nouvel énoncé dont la fusion n'a pu être réalisée[7]. Nous avons donc implémenté un mode de fusion flexible, où les parties de l'énoncé non concernées par le blocage sont tout de même fusionnées. Ce mode minimise le risque de perdre des paraphrases candidates correctes apparaissant ailleurs entre les syntagmes.

Par ailleurs, l'algorithme est très dépendant de la qualité des analyses syntaxiques effectuées, qui, en cas d'erreurs, peuvent par exemple bloquer des fusions qui seraient tout à fait légitimes, ce que nous avons observé à de nombreuses reprises sur nos données de développement. Pour pallier ce problème, nous avons implémenté un mode dans lequel les k meilleures analyses produites par un analyseur probabiliste sont utilisées, et où la combinaison retenue entre la i-ème analyse du premier énoncé et la j-ème analyse du second parmi les k^2 combinaisons possibles est celle minimisant le nombre de nœuds du graphe obtenu avant *réduction* (fusion d'arcs correspondant à des préfixes ou à des suffixes communs). L'intuition suivie est ici que plus un automate sera compact avant réduction et mieux les deux énoncés seront alignés, et que l'opération de réduction doit supprimer le moins possible d'arcs qui ne résulteraient pas d'une fusion autorisée préalablement par la syntaxe commune des deux énoncés. Pour notre implémentation, nous avons utilisé l'analyseur syntaxique probabiliste de Berkeley (Petrov *et coll.*, 2006) et sa transposition au français (Candito *et coll.*, 2010) pour produire les 5 meilleures analyses pour chaque énoncé, et nous avons effectué une recherche exhaustive de la meilleure fusion pour chaque paire d'énoncés.

Un exemple d'automate obtenu par application de Synt est donné dans la figure 13 : ont été fusionnées les 3 énoncés commençant par *La BCE veut conserver l'inflation sous la barre des ...*, *La BCE veut garder l'inflation sous la barre des ...* et *La Banque Centrale Européenne veut maintenir l'inflation sous la barres des ....* Un parcours des chemins possibles dans l'automate obtenu permet d'extraire les paraphrases suivantes pour l'extrait donné : *BCE* ↔ *Banque Centrale*

7. Notons qu'en limitant le blocage lexical aux cas où les mots rencontrés sont strictement identiques, on autorise des fusions erronées lorsque des synonymes sont utilisés.

Européenne, *maintenir* ↔ *garder* ↔ *conserver*, *La BCE* ↔ *La Banque Centrale Européenne*, *BCE veut* ↔ *Banque Centrale Européenne veut*, *veut maintenir* ↔ *veut garder* ↔ *veut conserver*, etc.

Tout comme TERME, cette technique semble *a priori* plus adaptée à l'extraction précise de paraphrases, mais contrairement à TERME il est attendu qu'elle ne parvienne pas à extraire de correspondances lorsque les structures syntaxiques de haut niveau des paraphrases d'énoncés ne sont pas compatibles.

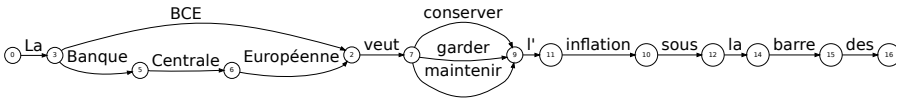


FIGURE 13: Exemple d'une partie d'automate obtenu par application de SYNT sur 3 énoncés en relation de paraphrase.

4.2.4 *Taux d'édition sur des séquences de mots (*EDIT*)*

Une relation entre deux paraphrases peut également s'exprimer sous forme de la séquence d'éditions la plus directe permettant de transformer l'une en l'autre. Une telle séquence d'éditions sur les mots est, par exemple, implémentée dans la technique TER_p (Translation Edit Rate plus) (Snover *et coll.*, 2009), originellement développée pour le calcul d'un taux d'édition servant de mesure en traduction automatique pour évaluer une hypothèse de traduction relativement à une traduction de référence. Ce calcul met en jeu des opérations de transformation de chaîne incluant l'insertion, la suppression et la substitution de mots, ainsi que le déplacement et la substitution de segments. Le coût de chaque opération est déterminé par au minimum un poids pouvant être optimisé [8].

Bien que TER_p ne soit pas conçu explicitement pour l'alignement monolingue, cette mesure produit des alignements entre mots ou séquences de mots, ce qui correspond bien à la définition des paires de paraphrases. Utiliser TER_p pour évaluer une sortie d'un système de traduction automatique donné, revient entre autre, à calculer l'effort nécessaire pour obtenir la phrase de référence à partir d'une traduction hypothèse, indépendamment des alignements produits. Toutefois, si nous remplaçons les traductions par des paraphrases d'énoncés, cette mesure peut être utilisée comme aligneur automatique, fournissant des paraphrases sous-phrastiques.

8. Snover *et coll.* (2010) ont utilisé l'algorithme de *hill climbing*, technique d'optimisation implémentant un algorithme itératif commençant par une solution arbitraire à un problème donné, puis tente de trouver la meilleure solution en changeant incrémentalement un seul élément de la solution.

Pour son calcul, TER_p produit donc un alignement au niveau des mots entre deux énoncés. Pour nos besoins, nous avons implémenté une méthode EDIT qui extrait l'ensemble des bi-segments (à concurrence d'une taille maximale) dérivables des alignements produits par TER_p. Nous avons exploité la possibilité d'optimiser TER_p pour nos besoins, en optimisant ses paramètres par *hill climbing*, en utilisant une première itération à partir de poids uniformes puis 100 itérations à partir de poids aléatoires. Les substitutions de mots ou segments, qui sont optionnelles, peuvent exploiter des listes fournies à l'algorithme, et les substitutions de segments ont une probabilité associée [9].

Pour nos expériences, nous avons exploité la tâche d'alignement sous-phrastique de deux énoncés sous forme de la séquence d'éditions la moins coûteuse permettant de transformer l'une en l'autre. Par la suite, nous dénoterons $\text{EDIT}_{\rightarrow P}$, $\text{EDIT}_{\rightarrow R}$ et $\text{EDIT}_{\rightarrow F_1}$, les variantes de EDIT correspondant à des optimisations réalisées sur un corpus de développement maximisant respectivement la précision, le rappel ou la F-mesure (voir la section 4.1) pour des annotations de référence. Un exemple de résultat d'alignement avec EDIT est donné dans la figure 14. Sont obtenues une substitution de segments (P), *ce dégrèvement* ↔ *cet allongement*, une substitution lexicale (S), *équivaut* ↔ *revient* et des paraphrases composites, *ce dégrèvement fiscal* ↔ *cet allongement fiscal*.

Reference	ce	dégrèvement	fiscal	équivaut
	P	P		S
Hyp After Shifts	cet	allègement	fiscal	revient

FIGURE 14: Exemple d'un alignement obtenu par EDIT entre deux extraits de paraphrases d'énoncés.

4.2.5 *Traductions communes par langue pivot* (PIVOT)

Des travaux précédents (Bannard et Callison-Burch, 2005) ont montré que les équivalences de traduction peuvent être exploitées pour déterminer si deux unités textuelles constituent des paraphrases sous-phrastiques. Au lieu de s'appuyer sur des corpus monolingues parallèles *rares*, Bannard et Callison-Burch ont proposé une approche exploitant la grande disponibilité de corpus

9. La version standard de TER_p implémente des techniques de racinisation et met en jeu des ressources de synonymies et de paraphrases, mais pour l'anglais uniquement : nous ne les avons donc pas utilisées.

parallèles bilingues pour produire des reformulations locales de courtes unités textuelles par traduction dans une langue puis par rétro-traduction et sélection dans la langue d'origine. Un segment source seg_1 est tout d'abord traduit dans une langue pivot avant d'être traduit à nouveau dans la langue d'origine pour obtenir la paraphrase candidate seg_2. L'exemple de la figure 15 illustre ce processus de construction de paraphrases pour le segment *ce n'est pas le moment de*, en utilisant l'anglais comme langue pivot.

Dans ce cadre, une *probabilité de paraphrasage* entre deux segments seg_1 et seg_2 exploitant l'existence d'un alignement commun avec un segment $pivot$ dans une autre langue, est introduite :

$$P_{para}(seg_1, seg_2) = \sum_{Pivot} P_{trad}(pivot|seg_1) P_{trad}(seg_2|pivot) \tag{3}$$

Pour cette approche, nous avons utilisé le corpus parallèle multilingue des débats parlementaires européens EUROPARL (Koehn, 2005) en anglais et français, constitué d'environ 1,7 million d'énoncés parallèles : ceci nous permet d'utiliser la même ressource pour construire les paraphrases dans les deux langues, en utilisant chaque langue comme pivot pour l'autre langue. Comme pour MOT, nous utilisons le système de traduction automatique statistique MOSES (Koehn *et coll.*, 2007) qui a recours à l'outil GIZA++ (Och et Ney, 2003) pour aligner les énoncés au niveau des mots.

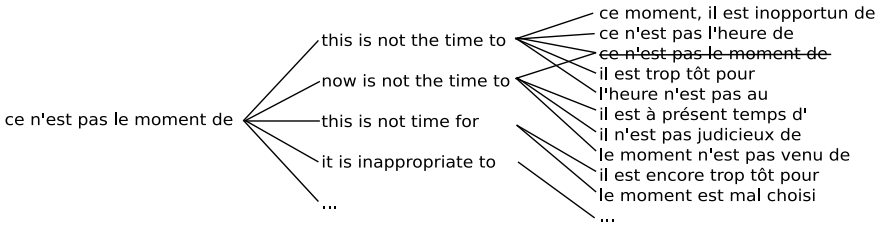


FIGURE 15: Exemple de paraphrases obtenues par pivot pour le segment français "*ce n'est pas le moment de*" par la technique du pivot bilingue (tirée de (Max, 2008))

Pour chaque segment d'une paire d'énoncés, nous construisons l'ensemble de ses paraphrases potentielles et récupérons la paraphrase présente dans l'autre énoncé ayant la plus forte probabilité. Nous répétons ce processus dans les deux directions et retenons pour chaque segment la paraphrase, de n'importe quelle direction, ayant la plus forte probabilité.

Nous avons évalué chacune des méthodes présentées ci-dessus sur nos données d'évaluation décrites dans la section 4.1. Nous détaillons dans le tableau 9 les résultats obtenus pour les 5 techniques décrites précédemment.

Le résultat le plus marquant concerne les différences de performance des techniques sur les deux langues, tant au niveau de la précision que du rappel. Nous constatons que toutes les techniques, à l'exception de TERME, obtiennent de meilleures performances (en F-mesure) sur le corpus français. Cela peut s'expliquer par le fait que ce dernier, obtenu par des traductions multiples plus littérales (à partir de l'anglais), comparées aux traductions en anglais à partir du chinois, est plus facilement alignable. Ceci est clairement illustré dans les résultats de la méthode d'alignement statistique MOT, qui obtient un avantage de 2,46 en F-mesure pour le français relativement à l'anglais.

	MOT	PIVOT	TERME	SYNT	EDIT		
					$\rightarrow$ P	$\rightarrow$ R	$\rightarrow$ F_1
	anglais						
P	48,3	**73,4**	63,1	63,6	60,5	36,2	41,2
R	59,0	25,8	5,9	6,3	9,2	**67,8**	66,4
F_1	**53,1**	38,2	10,7	11,5	16,0	47,2	50,9
	français						
P	52,5	**64,5**	56,9	57,9	61,5	43,1	46,4
R	58,9	30,3	4,9	7,3	3,1	61,3	**61,4**
F_1	**55,5**	41,2	9,1	12,9	5,9	50,6	52,8

Tableau 9: Résultats obtenus pour chaque technique d'acquisition de paraphrases individuelle sur l'anglais (partie supérieure) et le français (partie inférieure). Les meilleurs scores de chaque ligne sont **en gras.**

En revanche, la différence de degré de parallélisme des énoncés n'affecte pas TERME, qui obtient des résultats proches pour les deux langues (10,7 et 9,1 sur l'anglais et le français, respectivement). Les deux techniques exploitant des informations linguistiques, TERME et SYNT se distinguent par une précision relativement forte (63,1 pour TERME et 63,6 pour SYNT sur l'anglais) mais une valeur de rappel beaucoup moins importante que les autres techniques(5,9 pour TERME et 6,3 pour SYNT sur l'anglais). Une explication possible

pour la différence de performance de Synt entre les deux langues (meilleure précision sur le corpus anglais) est que l'analyseur syntaxique a été entraîné sur une plus grande quantité de données pour l'anglais, ce qui permet d'obtenir des analyses plus précises. Les faibles valeurs de rappel obtenues avec ces deux techniques ne sont pas contradictoires : les métarègles de Terme sont très focalisées sur l'extraction de variantes de termes et ne peuvent couvrir tous les phénomènes paraphrastiques. Synt nécessite quant à elle deux arbres syntaxiques fortement comparables pour pouvoir les fusionner et extraire par la suite des correspondances fines entre nœuds. Lorsque ces conditions sont vérifiées, ces deux techniques obtiennent de bons résultats en termes de précision.

Les résultats de la technique statistique d'alignement entre mots, Mot, et celle fondée sur un calcul de taux d'édition, Edit, sont très comparables et obtiennent les meilleures valeurs de F-mesure. Mot obtient une valeur de précision plus élevée que Edit, qui obtient un meilleur rappel, Mot obtenant des valeurs de F-mesure meilleures (respectivement +2,2 et +2,7 pour l'anglais et le français).

Pivot obtient les meilleures valeurs de précision dans les deux langues avec un avantage sur Terme de respectivement +10,3 et +7,6 pour l'anglais et le français. Cependant, cette méthode ne trouve que peu de paraphrases sûres, avec une différence relativement à Edit de respectivement -42,0 et -31,1 en rappel sur l'anglais et le français. Ceci peut être dû en partie au corpus bilingue utilisé : Pivot extrait, en effet, ses paraphrases candidates d'un corpus de débats parlementaires, alors que notre ensemble d'évaluation provient du domaine journalistique. On peut donc notamment observer le résultat de différences relatives aux domaines et donc la présence de nombreux bi-segments "hors-vocabulaire", en particulier pour les entités nommées qui sont largement utilisées dans le domaine journalistique.

La meilleure valeur de rappel obtenue est de 67,8 pour l'anglais ($\text{Edit}_{\rightarrow R}$) et de 61,4 sur le français ($\text{Edit}_{\rightarrow F_1}$), ce qui montre que Edit identifie le plus grand nombre de paraphrases sûres pour les deux langues. Le fait que les résultats sont ici meilleurs pour l'anglais n'est pas contradictoire avec nos remarques précédentes : le taux d'édition implémenté est en mesure d'aligner des mots et des segments assez distants indépendamment de la syntaxe et d'identifier des correspondances entre les mots proches restants.

Enfin, le fait que le rappel ne soit globalement pas très élevé vient confirmer la complexité de notre tâche d'identification de paraphrases sous-phrastiques.

Rappelons toutefois que, dans nos mesures, les paraphrases identités ne sont pas considérées. De plus, les valeurs d'accords inter-annotateurs données dans le tableau 7 montrent que la constitution du corpus de référence est un processus d'annotation difficile. Nos

résultats sont également influencés par le fait que les mesures adoptées comptent comme fausses de petites variations par rapport à la référence.

4.4 PERFORMANCE EN FONCTION DU DEGRÉ DE COMPARABILITÉ DES ÉNONCÉS

Les paires de paraphrases d'énoncés que nous étudions dans ce travail posent des problèmes de difficulté variable. En effet, ces paraphrases peuvent être très proches et ne différer que par quelques mots. Certaines paraphrases peuvent en revanche avoir des structures syntaxiques très différentes et/ou peuvent être lexicalement très distantes. Il est ainsi instructif de considérer la performance des différentes méthodes testées en fonction de ces difficultés. Celle-ci pourrait se mesurer par un accord inter-annotateurs au niveau de chaque énoncé, mais nous avons choisi d'utiliser une mesure automatique fondée sur un taux d'édition sur les formes, $(1 - \mathrm{TER}(\mathit{énoncé}_1, \mathit{énoncé}_2))$, qui est donc d'autant plus grande que les énoncés sont proches. Les résultats obtenus pour l'ensemble des techniques étudiées sont présentées dans la figure 16. Dans cette figure, la valeur de chaque barre dans les intervalles discrétisés est une moyenne des éléments de cet intervalle, et ne rend pas compte du nombre de ces éléments. Pour la précision, une valeur de 0 peut indiquer soit l'absence de proposition pour les énoncés de cet intervalle, soit des propositions toutes incorrectes. Le nombre des énoncés dans chaque intervalle est également donné dans la figure 16.

Pour la précision, on constate tout d'abord que MOT est très sensible à la difficulté telle que nous la définissons, et que les alignements que cette technique produit sont d'autant moins bons que les énoncés sont différentes. De façon un peu plus surprenante, SYNT et EDIT ne semblent pas trop affectés par cette difficulté. Cependant, ceci est dû au fait que les valeurs des barres, pour chaque intervalle discrétisé, sont une moyenne qui ne rend pas compte du nombre d'éléments. Il est possible que SYNT extraie peu de paraphrases sur des paires d'énoncés difficiles, mais que, lorsqu'elle parvient à trouver des structures syntaxiques compatibles, celles-ci permettent un alignement précis. Enfin, TERME est insensible à cette difficulté, ce qui était attendu puisque cette technique fonctionne avec des patrons morphosyntaxiques pouvant impliquer des mots et des structures syntaxiques de haut niveau différents. Nous déduisons donc de ces remarques que ces différentes techniques peuvent être utilisées à bon escient pour différents niveaux de parallélisme des corpus d'acquisition.

Le rappel fait apparaître une tendance beaucoup plus marquée : MOT, EDIT et SYNT extraient d'autant moins de paraphrases de

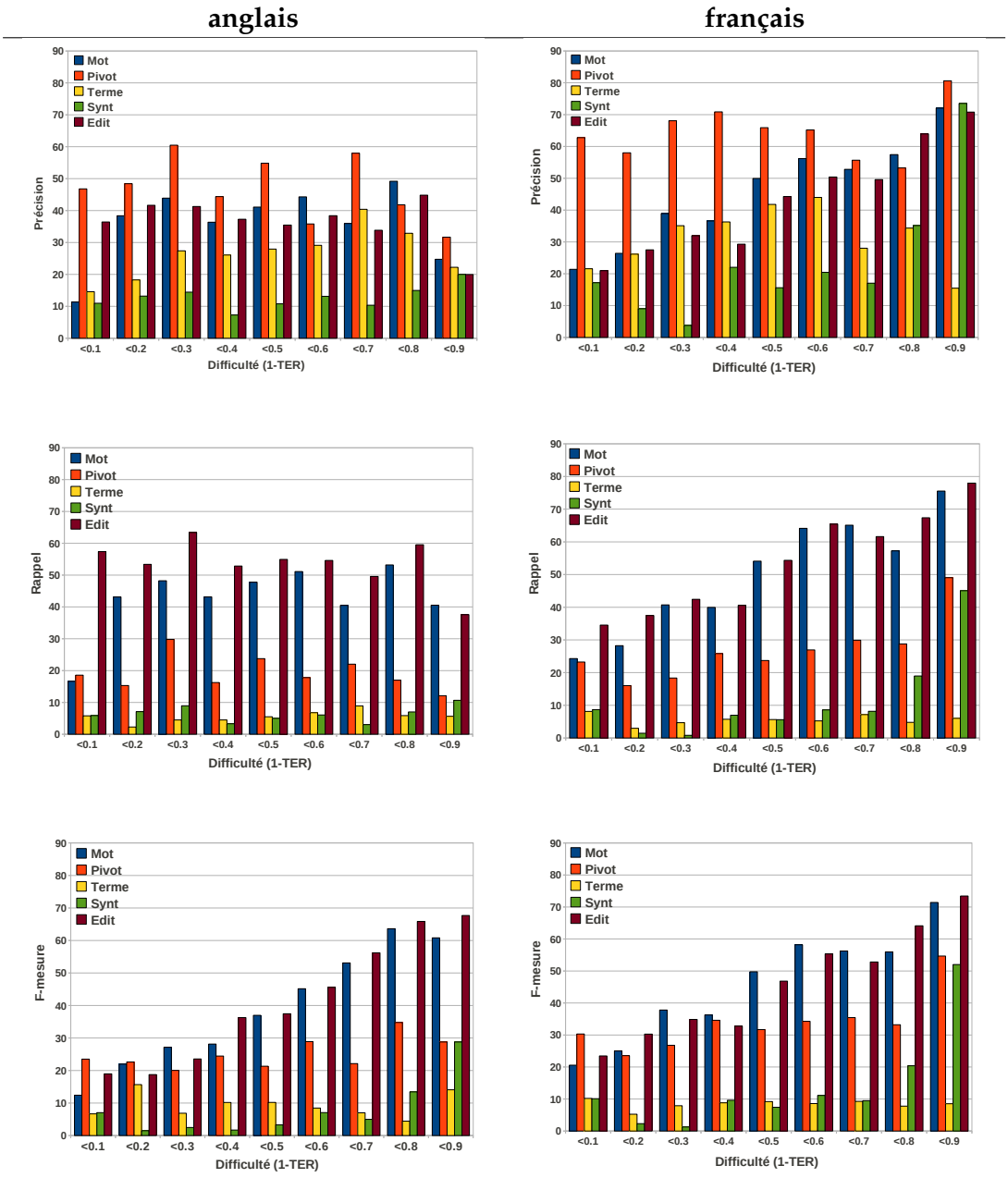
la référence que les énoncés sont difficiles. À nouveau, TERME y semble insensible. PIVOT présente un comportement différent du reste des techniques : sa précision ne semble pas être très affectée par le degré de parallélisme des énoncés, alors qu'elle présente un meilleur rappel sur les paraphrases d'énoncés les plus parallèles.

Cette analyse confirme l'hypothèse qu'il est préférable d'avoir des paraphrases d'énoncés les plus « parallèles » possibles pour obtenir une bonne performance en acquisition car plus les textes à aligner sont différents et plus il apparait difficile d'identifier des paraphrases correctes. Un autre enseignement intéressant concerne les techniques PIVOT et TERME : celles-ci sont particulièrement utiles pour extraire des paraphrases sous-phrastiques précises dans des paraphrases d'énoncés de formes très différentes, ce qui reflète assez bien leurs usages originels.

CONCLUSION DU CHAPITRE

Dans ce chapitre, nous avons décrit et situé la tâche d'acquisition de paraphrases sous-phrastiques à partir d'énoncés en relation de paraphrase, ressources difficiles à obtenir mais qui permettent de se concentrer sur un cadre naturel d'étude des phénomènes de paraphrase. Les différentes techniques que nous avons mises en œuvre, à l'origine développées pour des besoins différents, se sont révélées relativement complémentaires, et permettent sous certaines conditions d'obtenir des résultats acceptables en termes de précision et de rappel relativement à un ensemble de référence.

Nous avons présenté cinq méthodes d'acquisition de paraphrases sous-phrastiques à partir de corpus monolingues parallèles. Ces méthodes reposent sur des caractéristiques linguistiques différentes : MOT sur l'apprentissage statistique, TERME sur un approche symbolique de la variation de termes, SYNT sur des proximités syntaxiques, PIVOT sur les variations de traductions et enfin EDIT sur un taux d'édition. En évaluant ces méthodes, nous avons constaté qu'effectivement leurs résultats semblent complémentaires, ce qui nous a mené à un second objectif, l'hybridation de ces méthodes.



Intervalles										
$< 0,1$	$< 0,2$	$< 0,3$	$< 0,4$	$< 0,5$	$< 0,6$	$< 0,7$	$< 0,8$	$< 0,9$	< 1	total
anglais										
18	24	66	68	96	94	77	43	12	2	500
français										
38	27	63	80	100	71	65	41	12	3	500

FIGURE 16: Performance en précision, rappel et F-mesure des techniques étudiées pour l'anglais (à gauche) et le français (à droite) en fonction de la difficulté d'alignement des paires d'énoncés mesurée par la valeur de (1-TER). La distribution des énoncés dans les intervalles est fournie dans la table (en bas).

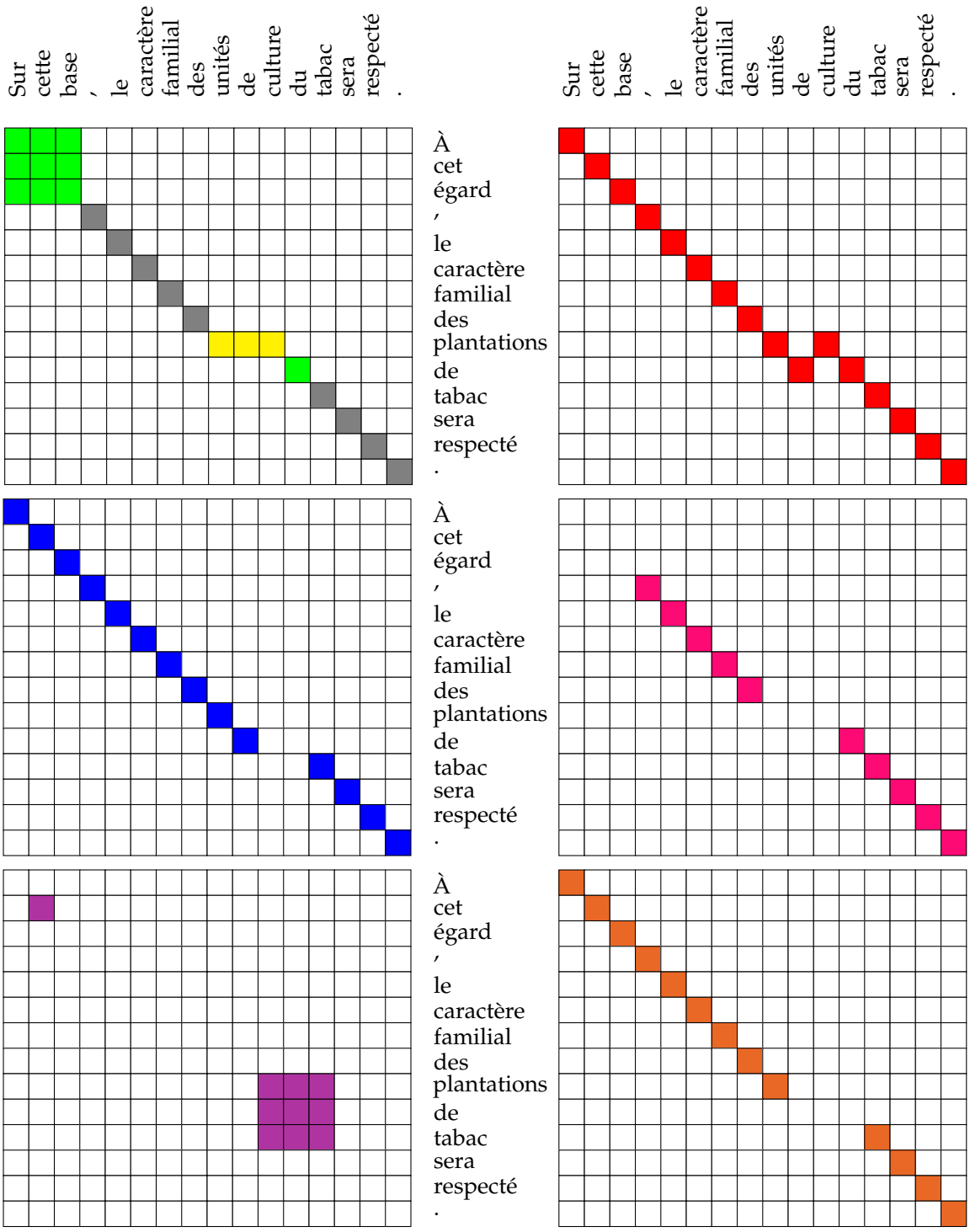


FIGURE 17: Exemples de matrices d'alignement pour la paire d'énoncés : *À cet égard, le caractère familial des plantations de tabac sera respecté.* ↔ *Sur cette base, le caractère familial des unités de culture du tabac sera respecté.*
La matrice de référence est donnée en haut, à gauche. La matrice rouge illustre les alignements produits par la technique MOT, la matrice en bleu, représente les alignements fournis par EDIT, celle en violet indique les paraphrases produites par TERME. Les alignements produits par PIVOT sont donnés dans la matrice rose. Enfin les paraphrases trouvées par SYNT sont indiquées sur la matrice orange.

5

COMBINAISON D'INFORMATIONS POUR L'ACQUISITION DE PARAPHRASES SOUS-PHRASTIQUES

Dans le chapitre 4, nous avons montré que différentes approches peuvent être utilisées, avec des performances variables, pour faire l'acquisition de paraphrases sous-phrastiques depuis des corpus monolingues parallèles. Outre l'amélioration individuelle de ces approches, il est possible de parvenir à une amélioration des performances obtenues en exploitant utilement les résultats de chacune.

Nous décrivons dans ce chapitre des approches hybrides d'extraction de paraphrases sous-phrastiques par la combinaison de paraphrases candidates produites par les différentes techniques décrites dans le chapitre précédent, avec pour objectif d'améliorer leur performance en tirant profit de leurs différentes caractéristiques. Dans la section 5.1, nous examinons leur complémentarité potentielle. Puis, nous introduisons deux approches de combinaison de leurs résultats. Dans la première, les résultats produits par chaque technique sont combinés *a posteriori* (section 5.2). Dans la deuxième, une technique est *adaptée* pour tenir compte des résultats produits par les autres techniques (section 5.3). Ensuite, nous introduisons une méthode basée sur la classification automatique des paraphrases permettant de valider des paraphrases candidates par apprentissage de différentes caractéristiques aussi bien contextuelles que syntaxiques (section 5.4). Nous terminons ce chapitre par une analyse du problème de l'identification d'équivalence de sens entre segments textuels et nous proposons finalement une typologie des paraphrases difficiles à acquérir automatiquement (section 5.5).

Les techniques présentées dans le chapitre précédent, opèrent à des niveaux d'appréhension du texte différents et exploitent des ressources distinctes. Avant d'examiner une éventuelle combinaison de leurs sorties, il est intéressant d'étudier s'il existe entre elles une certaine *complémentarité*. Pour cela, nous examinerons l'apport d'une technique dans une combinaison d'autres techniques en termes de nombre de paraphrases correctes.

5.1.1 *Étude de la complémentarité des techniques*

Nous avons estimé la complémentarité entre deux techniques à l'aide de la différence entre le rappel de leur union et le meilleur de leurs rappels individuels. Nous avons donc utilisé la formule suivante entre un ensemble de candidats t_i extraits en utilisant une technique i, et l'ensemble t_j des paraphrases proposées par une autre technique j :

$$C(t_i, t_j) = \text{rappel}(t_i \cup t_j) - \max(\text{rappel}(t_i), \text{rappel}(t_j)) \quad (4)$$

Dans le tableau 10, nous détaillons les complémentarités ainsi mesurées sur les corpus TEXTE comportant 500 paires d'énoncés pour l'ensemble des techniques décrites dans le chapitre précédent. Les valeurs de complémentarité sont calculées pour chaque paire de techniques individuelles, et pour chaque technique individuelle relativement à l'ensemble des autres techniques [1].

Il apparaît que plusieurs paires de techniques sont assez fortement complémentaires. La plus forte valeur de complémentarité est obtenue, pour les deux langues, en combinant MOT et EDIT, avec des gains de 9,76 points en rappel sur l'anglais et 12,55 sur le français. Selon ces résultats, PIVOT arrive à identifier des paraphrases qui sont plus similaires à celles de EDIT que de MOT. SYNT et TERME présentent elles aussi une forte complémentarité en français malgré leur faible nombre de paraphrases. L'apport de chaque technique relativement à l'ensemble des autres techniques montre que EDIT apporte le plus grand nombre de paraphrases inédites pour l'anglais et que EDIT et MOT sont les plus utiles pour le français.

Ces expériences, qui révèlent que les différentes techniques étudiées sont relativement complémentaires entre elles, confirment tout d'abord l'intérêt de la sélection que nous avons effectuée. Elles permettent en outre de confirmer le potentiel d'une combinaison de leurs résultats.

1. Le tableau 10 est symétrique, par exemple la complémentarité entre PIVOT et SYNT est bien sûr identique à celle entre SYNT et PIVOT.

	Mot	Pivot	Terme	Synt	Edit$_{\rightarrow F_1}$	*Toutes les autres*
	anglais					
Mot	-	7,21	2,57	1,02	**9,76**	7,22
Pivot	**7,21**	-	3,26	3,53	4,63	2,49
Terme	2,57	3,26	-	**5,62**	1,62	0,94
Synt	1,02	3,53	**5,62**	-	0,80	0,27
Edit$_{\rightarrow F_1}$	**9,76**	4,63	1,62	0,80	-	12,89
	français					
Mot	-	8,00	2,17	1,49	**12,55**	9,20
Pivot	**8,00**	-	3,06	3,86	6,14	3,03
Terme	2,17	3,06	-	**4,53**	1,50	0,86
Synt	1,49	3,86	**4,53**	-	1,45	0,35
Edit$_{\rightarrow F_1}$	**12,55**	6,14	1,50	1,45	-	9,26

Tableau 10: Valeurs de complémentarité pour un ensemble d'évaluation dans les deux langues telles que mesurées par l'équation 4. Les valeurs données en gras indiquent les valeurs les plus élevées pour chaque technique.

Nous faisons ici une synthèse des points forts et des limitations de chacune de ces techniques de façon à guider la recherche d'un mode de combinaison plus efficace :

- Mot : très sensible à la fréquence d'observation de mots et de cooccurrences entre mots, cette technique peut être informée par la connaissance d'associations *a priori*, qui peuvent par exemple être transmises sous forme de données d'apprentissage additionnelles.
- Terme : cette technique est spécialisée dans l'extraction d'un type de bi-segments contraints par des règles de réécriture et de variation lexicale. Les métarègles, qui ont été développées manuellement, sont assez précises mais ne peuvent couvrir tous les phénomènes de paraphrase. Leur apprentissage automatique pourrait améliorer la couverture, mais éventuellement au détriment de la précision. L'enrichissement automatique des familles morphologiques et sémantiques devrait également permettre d'augmenter le rappel.
- Synt : cette technique est très sensible au degré de parallélisme des énoncés qui décide de la fusion de constituants syntaxiques. Nous avons déjà pris en compte le problème de la correction des analyses syntaxiques en autorisant la fusion à opérer sur les k-meilleures analyses syntaxiques. Le blocage lexical empêche une fusion lorsqu'un mot présent dans le constituant d'un énoncé est présent dans un constituant non aligné de l'autre énoncé. Ce blocage pourrait être amélioré par la con-

naissance *a priori* de paraphrases locales, ce qui, néanmoins, ne pourrait bénéficier qu'à la précision.
- Pivot : cette technique exploite la connaissance de traductions communes dans une autre langue. Il est donc possible d'améliorer ses performances en utilisant plusieurs langues pivots simultanément, et en utilisant des corpus parallèles thématiquement adaptés.
- Edit : cette technique transforme une séquence de mots en une autre en un coût minimal, en utilisant des pondérations optimisées pour les différentes opérations utilisées. L'algorithme manipule des segments qui n'ont pas nécessairement de motivation linguistique, ce qui peut mener à des transformations aberrantes. En outre, des opérations d'insertion et de suppression peuvent être utilisées à tort lorsque des correspondances au niveau des mots ou des segments ne sont pas connues. Ainsi, si de telles correspondances peuvent être fournies à Edit, il est possible de diminuer le nombre d'opérations de transformation aberrantes et ainsi d'augmenter la performance.

5.1.2 *Approches pour la combinaison des techniques*

Dans la section précédente nous avons montré qu'il existe plusieurs voies pour améliorer la performance de l'alignement monolingue auquel nous nous intéressons à partir des techniques décrites. Sans considérer davantage, à ce stade, l'amélioration individuelle de chacune des techniques, nous pouvons décrire les deux grandes familles d'approches possibles pour l'hybridation de plusieurs techniques de la manière suivante : 1) les résultats produits indépendamment par chaque technique sont combinés *a posteriori*, ou 2) une technique est *adaptée* en lui apportant la connaissance des résultats produits par les autres techniques. Ces deux possibilités sont examinées ci-dessous.

5.2 combinaison par simple union

Nous considérons tout d'abord une combinaison naïve obtenue par l'union *a posteriori* des résultats de toutes les techniques tel qu'illustré en haut de la figure 18. Les résultats de l'évaluation de cette approche élémentaire sont donnés sur la dernière colonne du tableau 11, dans laquelle les premières colonnes rappellent les résultats des techniques individuelles présentées dans le chapitre 4. Ces résultats sont encourageants : dans les deux langues, environ 8 paraphrases sur 10 de la référence sont trouvées par au moins une technique. Ce résultat vient valider la discussion précédente sur l'apport de la combinaison des différentes techniques à la tâche d'alignement sous-phrastique. La précision a cependant baissé de

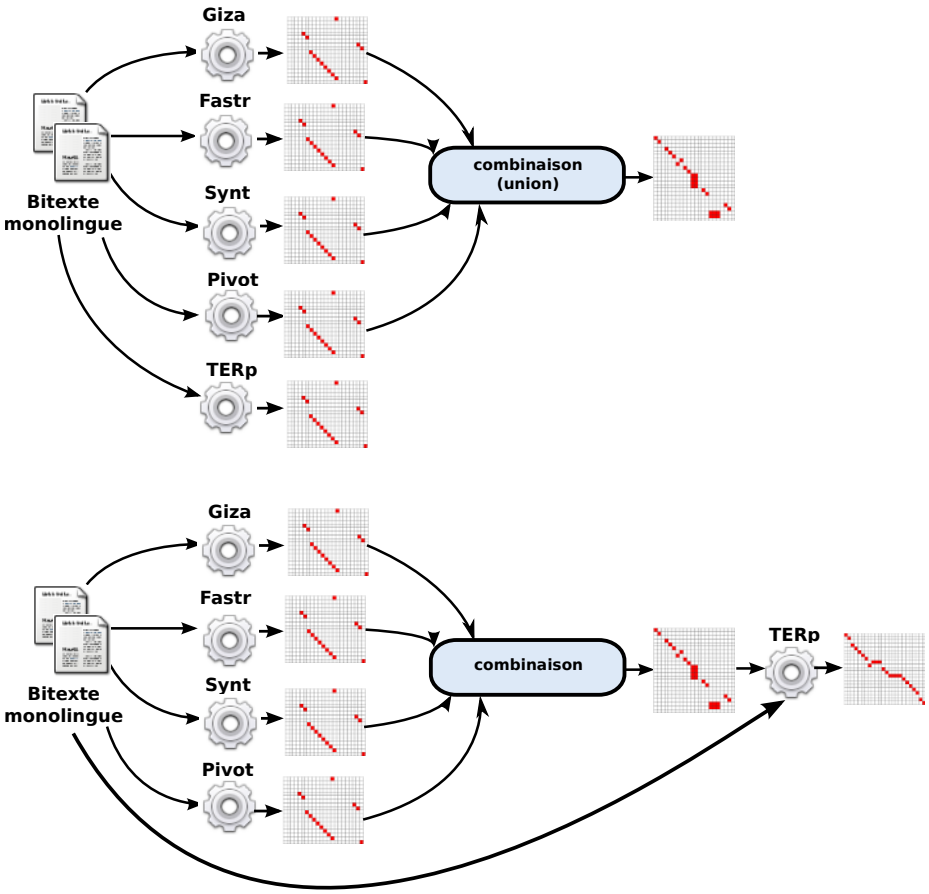


FIGURE 18: Approches implémentées dans ce travail pour combiner des informations pour l'alignement monolingue.

façon significative (3,5 et 4 paraphrases correctes sur 10 pour respectivement l'anglais et le français), et les valeurs de F-mesure sont plus faibles que celles de EDIT et MOT, et à peine plus élevées que celles de PIVOT. Ceci montre clairement que l'union des techniques n'est intéressante que pour des applications pour lesquelles le rappel est primordial.

Dans la section 5.4, nous allons revenir sur cette union en étudiant comment apprendre automatiquement à distinguer les paraphrases correctes dans ce sur-ensemble.

5.3 COMBINAISON PAR ADAPTATION

Nous considérons à présent une seconde approche, illustrée en bas de la figure 18, qui consiste à informer la technique EDIT, obtenue en optimisant TER_p sur la F-mesure, avec des paires de paraphrases candidates extraites par les autres techniques. D'après nos observations, EDIT est un candidat assez naturel pour l'adaptation. En effet, la connaissance d'alignements non triviaux et cor-

	Mot	Pivot	Terme	Synt	Edit → P	Edit → R	Edit → F_1	**Union**
	anglais							
P	48,3	**73,4**	63,1	63,6	60,5	36,2	41,2	35,75
R	59,0	25,8	5,9	6,3	9,2	**67,8**	66,4	**81,06**
F_1	**53,1**	38,2	10,7	11,5	16,0	47,2	50,9	49,62
	français							
P	52,5	**64,5**	56,9	57,9	61,5	43,1	46,4	40,84
R	58,9	30,3	4,9	7,3	3,1	61,3	**61,4**	**78,43**
F_1	**55,5**	41,2	9,1	12,9	5,9	50,6	52,8	53,71

Tableau 11: Résultats obtenus pour chaque technique d'acquisition de paraphrases individuelle, ainsi que pour l'union naïve (à droite) sur l'anglais (partie supérieure) et le français (partie inférieure). Les meilleurs scores de chaque ligne sont **en gras.**

rects au niveau des mots ou des segments peut diminuer le nombre d'opérations effectuées à tort. Il s'agit précisément de la motivation majeure pour l'évolution de Ter à Ter_p (Snover *et coll.*, 2010), liée à la possibilité d'utiliser une base de paraphrases locales connues *a priori*. Ainsi le système est plus robuste face aux hypothèses de traduction qu'il accepte lorsque leurs segments ne correspondent pas exactement à ceux d'une traduction de référence. Ter_p exploite ainsi une table de paraphrases, dans laquelle chaque paraphrase est associée à une probabilité de paraphrasage utilisée dans la définition du coût de l'opération de substitution (de paraphrase).

Intuitivement, plus deux énoncés sont parallèles, plus les correspondances entre leurs segments sont faciles à repérer, ce qui facilite la mise en correspondance des parties restantes par Ter_p. Mais, en généralisant cette approche, les résultats peuvent être améliorés en utilisant des paraphrases de segments.

Contrairement à ce qui est fait dans Ter_p, nous n'utiliserons pas, dans nos expériences, une base de connaissances externe [2], mais nous construisons dynamiquement la base de paraphrases utilisées en combinant les paraphrases candidates extraites par les autres techniques [3]. Ces paraphrases sont nombreuses et assez précises pour Mot, et peu nombreuses mais plus précises pour Pivot, Terme et Synt. Comme nous l'avons montré dans la section 5.1.1, ces techniques peuvent être complémentaires quant aux types de paraphrases qu'elles permettent d'identifier, ce qui suggère égale-

2. Afin d'assurer la comparabilité des résultats entre les deux langues étudiées, nous n'avons pas exploité les synonymes issus de Wordnet qui ne sont pas disponibles pour d'autres langues que l'anglais.

3. La base de paraphrases est utilisée de façon globale pour toutes les paires testées.

ment une utilisation conjointe de toutes les techniques pour construire la table de paraphrases utilisée pour un tel système hybride.

Un problème important à considérer concerne la manière dont la table de paraphrases utilisée par TER$_p$ est construite à partir des hypothèses produites par les différentes techniques. À ce stade de nos travaux, nous ne disposons pas de *mesures de confiance* données par chaque technique pour chacune de ces hypothèses, la solution la plus simple est de les considérer initialement comme équiprobables. Nous réalisons donc une combinaison simple par calcul d'union : chaque hypothèse apparaissant au moins une fois parmi les hypothèses des différents systèmes est retenue et est associée à un poids constant uniforme.

Un autre aspect important concerne là encore la pondération associée à chacune des paires de paraphrases *a priori* fournies à TER$_p$. Considérons le cas où deux paraphrases sont fournies à TER$_p$ et où l'une est un sous-segment de l'autre : par exemple, (*ce dégrèvement* ↔ *cet allègement*) inclut (*dégrèvement* ↔ *allègement*). Si ces deux paraphrases sont fournies avec le même score à TER$_p$, celui-ci préfèrera, dans de nombreux cas, utiliser la plus couvrante des deux, car cela minimisera souvent la quantité d'opérations de transformation restant à faire, et donc le coût global de transformation (comme illustré en haut de la figure 19). Cela peut ne pas être un défaut en soi, car l'identification des plus longues sous-unités paraphrastiques peut être utile. Cependant, nos mesures d'évaluation (définies dans la section 4.1.2 du chapitre 4) se fondent sur l'ensemble des bi-segments pouvant être extraits par composition de bi-segments plus petits. Ainsi, si dans l'exemple précédent l'alignement de référence inclut à la fois les paraphrases *ce* ↔ *cet*, *dégrèvement* ↔ *allègement* et *ce dégrèvement* ↔ *cet allègement*, l'utilisation de cette dernière ne permettrait pas de reconstruire facilement l'alignement le plus fin, ce qui pourrait pénaliser le rappel mesuré.

Plusieurs solutions sont envisageables pour pallier ce problème. La pondération des paraphrases pourrait prendre en compte le nombre de mots couverts en favorisant les courts segments. Ne disposant néanmoins pas de solutions génériques applicables à toutes les techniques ni de moyen d'intégrer des scores de confiance motivés, nous préférons nous en remettre à une solution initiale plus simple. Ainsi, ne seront gardés pour construire la table de paraphrases utilisée par TER$_p$ que les bi-segments n'étant inclus dans aucun autre bi-segment proposé, que nous appelons *bi-segments minimaux*.

5.3.1 *Expériences et résultats*

Nous avons testé 4 configurations d'hybridation en informant à chaque fois EDIT par les résultats de chacune des autres tech-

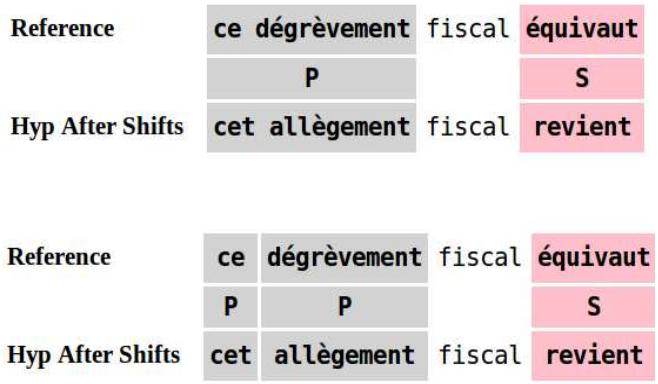


FIGURE 19: Exemple de deux alignements résultats de TER_p, en utilisant l'ensemble des bi-segments non filtrés (en haut), et un ensemble de bi-segments minimaux (en bas) et des opérations de substitution simple S et de substitution de paraphrases P.

niques individuelles. Les résultats, présentés dans le tableau 12, sont obtenus en suivant la méthodologie d'évaluation décrite dans le chapitre précédent (section 4.1).

À l'exception de deux configurations, toute hybridation impliquant une technique individuelle améliore à la fois la F-mesure de EDIT et celle de la technique utilisée, dans les deux langues. Les deux exceptions concernent l'utilisation de MOT et SYNT en français. Ces résultats négatifs peuvent peut-être provenir de l'incapacité de la technique d'optimisation utilisée à échapper à des optimaux locaux.

Les meilleures performances sont obtenues en utilisant PIVOT avec des gains respectifs de +8,76 et +21,48 relativement à EDIT et PIVOT (rappelés dans le tableau 11) sur l'anglais et de +3,61 et +15,21 sur le français. Nous remarquons par ailleurs que, contrairement aux systèmes individuels, nos systèmes hybrides semblent plus performants pour l'anglais que pour le français. Les différences tiennent essentiellement aux gains en rappel, qui s'avère être déjà nettement plus fort pour la technique EDIT sans hybridation en anglais qu'en français (66,47 contre 61,45 respectivement).

Le tableau 12 donne également la performance de EDIT lorsqu'elle est informée par l'ensemble complet des paraphrases produites par toutes les autres techniques. Cette fois-ci, nous ne notons aucune amélioration des performances relativement aux meilleurs résultats obtenus précédemment. Il est probable que la quantité de bruit combiné ne lui permette pas d'utiliser de façon efficace les paraphrases candidates correctes.

	Union	Systèmes hybrides (Edit$_{+X \rightarrow F_1}$)				
	TOUT	+MOT	+PIVOT	+TERME	+SYNT	+TOUT
	anglais					
P	35,75	47,07	57,29	51,64	47,25	47,97
R	**81,06**	66,34	62,30	66,44	62,65	64,46
F_1	49,62	55,06	**59,69**	58,11	53,87	55,01
	français					
P	40,84	49,54	57,89	49,66	48,54	50,37
R	**78,43**	57,97	55,18	58,98	54,03	62,12
F_1	53,71	53,42	**56,50**	53,92	51,14	51,15

Tableau 12: Résultats obtenus pour les deux langues par l'union des 5 techniques individuelles (UNION/TOUT), pour les systèmes hybrides EDIT utilisant chaque technique individuelle (+X) et de l'union des résultats de l'ensemble des techniques (+TOUT).

5.3.2 *Étude oracle*

Les résultats de EDIT peuvent être améliorés si on y intègre un ensemble de paraphrases précises qu'il exploite efficacement (tel que l'utilisation des résultats précis de PIVOT). Nous avons mené une expérience *oracle* dans laquelle notre technique hybride EDIT est informé par le sous-ensemble de la référence comportant uniquement les paraphrases non identifiées par nos techniques automatiques. Le but de cette expérience est double : d'une part, elle permet d'évaluer les résultats de cette combinaison intégrant des connaissances précises défiant les techniques d'acquisition automatique. D'autre part, elle permet de confirmer l'utilité de l'acquisition de ce genre de connaissances pour de tel mode efficace de combinaison d'information. Nous donnons dans le tableau 13 les résultats obtenus et nous rappelons les résultats de l'union naïve (UNION) ainsi que ceux de EDIT optimisée sur la F-mesure.

Ces résultats montrent que des gains importants peuvent être obtenus en fournissant à EDIT la liste des paraphrases qu'aucune des techniques individuelles n'avait su proposer. Ces gains atteignent +20,2 et +18,8 en F-mesure (pour l'anglais et le français, respectivement) relativement à la technique EDIT non informée, et environ 10 points de rappel, qui n'atteint pas, cependant, le rappel de l'union élémentaire (autour de 80, pour les deux langues).

Ce résultat vient confirmer le fait que fournir à EDIT des paraphrases précises permet de mieux guider son processus d'alignement. Par conséquent, il serait intéressant d'étudier les résultats d'une expérience plus réaliste, où la technique EDIT serait informée

	UNION	EDIT$_{+\text{TOUT}\rightarrow F_1}$	ORACLE$_{\rightarrow F_1}$
	anglais		
P	35,75	41,2	**66,1**
R	**81,06**	66,4	77,1
F_1	49,62	50,9	**71,1**
	français		
P	40,8	46,4	**69,4**
R	**78,4**	61,4	73,9
F_1	53,7	52,8	**71,6**

Tableau 13: Résultats des expériences oracle menées sur des ensembles d'évaluation pour l'anglais et le français.

par une technique filtrant efficacement les résultats contenus dans l'union des techniques individuelles.

5.4 VALIDATION DE PARAPHRASES PAR CLASSIFICATION AUTOMATIQUE

Dans les sections précédentes, nous avons décrit deux types de combinaisons possibles des techniques d'acquisition individuelles. Dans la première, les résultats produits indépendamment par chaque technique sont combinés *a posteriori* par une union naïve. Dans la seconde, une technique particulière, EDIT, est *adaptée* en lui apportant la connaissance des résultats des autres techniques. Nous avons évalué ces deux méthodes et nous avons obtenu des résultats encourageants, principalement en rappel. Cependant, la prise en compte d'un grand nombre de paraphrases candidates augmente mécaniquement la quantité de bruit en entrée. Pour tenter de diminuer ce bruit, nous avons mis en place un processus de *validation* des paraphrases, formulé comme une classification binaire (« *paraphrase* » ou « *non paraphrase* ») des paraphrases candidates applicable à l'ensemble des techniques individuelles considérées.

5.4.1 *Classification par Maximum d'Entropie : traits utilisés*

Nous avons abordé ce problème avec une classification discriminante à maximum d'entropie MAXENT (Berger *et coll.*, 1996) [4]. Un tel classifieur cherche à maximiser la probabilité conditionnelle $P(y|x)$ en faisant l'hypothèse qu'elle suit une loi exponentielle.

Cette tâche de classification permet d'inclure des traits qui n'étaient pas nécessairement pris en compte ou possibles à considérer.

4. Nous avons utilisé l'implémentation disponible sur : http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html

Plus généralement, ceci permet de tenter d'apprendre une caractérisation plus générique des paraphrases, qui pourrait s'adapter trivialement à un nombre quelconque par les techniques individuelles en entrée. Les traits que nous utilisons, calculés pour toutes les paires de paraphrases candidates quelles que soient la ou les techniques l'ayant proposée, sont les suivants :

Traits dérivés des paires de segments

- *Distance d'édition entre les paraphrases* : Les variantes morphologiques représentent souvent des formes de surface proches. Plus généralement, dans une paire de paraphrases, la similarité de surface entre ses formes peut être un bon indicateur de leur équivalence sémantique. Nous avons alors calculé un trait discrétisé représentant la distance d'édition de Levenstein entre les deux paraphrases candidates dans une paire donnée.
- *Similarité entre racines* : Dans une paire de paraphrases, les racines communes des variantes flexionnelles et dérivationnelles peuvent indiquer des formes sémantiquement liées. Nous avons défini un trait binaire indiquant si les racines des paraphrases d'une paire donnée sont identiques[5].
- *Identité d'ensemble de formes* : Une réorganisation syntaxique au sein d'une paire de paraphrases peut impliquer le même ensemble de mots, dans différents ordres. Nous avons, alors défini un trait binaire indiquant si les deux ensembles de formes représentant la paire de paraphrases comportent les mêmes formes introduits dans un ordre différent.
- *Longueur des segments* : Nous avons utilisé un trait discrétisé indiquant le rapport de longueur des segments dans chaque paire de paraphrases candidate.

Traits dérivés des paires d'énoncés

- *Similarité entre paires d'énoncés* : un ensemble de traits représentant la similarité entre les paires de d'énoncés d'origine. Nous avons mesuré la similarité en utilisant différents scores définis en traduction automatique : BLEU, TER et METEOR, ainsi que la mesure de similarité classique : cosinus.
- *Position relative des paraphrases* : En fonction de la langue dans laquelle des énoncés parallèles sont analysé, il est possible que des paraphrases sous-phrastiques apparaissent dans des positions proches dans leurs énoncés d'origine. Afin de prendre en compte cette caractéristique, nous avons

5. Nous avons utilisé une implémentation de l'outil de racinisation Snowball pour les deux langues, disponible sur : http://snowball.tartarus.org

utilisé un trait indiquant la position relative des deux segments dans leurs énoncés d'origine.

- *Présence de formes communes aux frontières des paraphrases* : Trait indiquant si les mots entourant les paraphrases sont identiques.
- *Présence d'une autre paire de paraphrases de chaque système aux frontières de la paraphrase* : Trait indiquant si les mots entourant les paraphrases sont elles-même paraphrases.
- *Présence d'une paraphrase à un autre endroit dans l'autre énoncé* : La présence d'un segment d'une paire de paraphrase sous-phrastique dans un autre endroit dans l'autre énoncé d'origine peut être un bon indicateur de la non équivalence entre les segments constituant de la paire. Nous avons donc défini un trait binaire représentant cet hypothèse.

Traits distributionnels

Traits indiquant le degré de similarité entre les contextes dans lesquels les paraphrases d'une paire donnée apparaissent. Pour cela, nous avons utilisé l'ensemble complet du corpus bilingue français-anglais fourni pour la dernière version de l'atelier de traduction automatique WMT [6]. Ce corpus comporte environ 30 millions de phrases parallèles : cela permet à nouveau de garantir que les mêmes ressources ont été utilisées pour les expériences menées sur les deux langues.

Nous avons collecté toutes les occurrences des segments apparaissant dans une paire de paraphrases, puis nous avons construit des vecteurs de mots pleins en ne gardant que les mots du voisinage distant de moins de 10 tokens du segment considéré. Nous calculons enfin le cosinus entre les vecteurs représentant les deux paraphrases, modélisant ainsi l'hypothèse classique de distributionnalité.

Traits dérivés des systèmes

Traits indiquant quelle technique ou combinaison de techniques a proposé une paire de paraphrases candidate donnée.

5.4.2 *Expériences et résultats*

Nos exemples *positifs* sont constitués de l'ensemble des paraphrases proposées par au moins l'une des 5 techniques d'acquisition étudiées et qui appartiennent à l'ensemble des paraphrases *sûres* de la référence. Nous avons extrait un nombre correspondant d'exemples *négatifs* à partir des paires de paraphrases proposées ne

6. http://www.statmt.org/wmt11/translation-task.html

figurant pas dans la totalité de la référence[7]. Nous avons choisi de ne pas considérer les paraphrases *possibles* car, étant proposées par les annotateurs n'ayant pas de certitude concernant leur classe de rattachement, elles ne nous permettent pas d'avoir le classifieur le plus discriminant dont on a besoin ici pour distinguer convenablement les paraphrases des autres phénomènes.

Nous détaillons dans le tableau 14 les résultats[8] obtenus pour cette validation[9].

	Mot	Pivot	Terme	Synt	Edit → P	Edit → R	Edit → F_1	**Union**	**Validation**
	anglais								
P	48,3	**73,4**	63,1	63,6	60,5	36,2	41,2	35,8	**68,8**
R	59,0	25,8	5,9	6,3	9,2	67,8	66,4	**81,1**	63,3
F_1	53,1	38,2	10,7	11,5	16,0	47,2	50,9	49,6	**65,9**
	français								
P	52,5	64,5	56,9	57,9	61,5	43,1	46,4	40,8	**74,8**
R	58,9	30,3	4,9	7,3	3,1	61,3	61,4	**78,4**	61,1
F_1	55,5	41,2	9,1	12,9	5,9	50,6	52,8	53,7	**67,2**

Tableau 14: Résultats obtenus pour les techniques individuelles ainsi que pour leur union et leur validation sur l'anglais (partie supérieure) et le français (partie inférieure).

Nous obtenons le meilleur résultat de cette étude en termes de F-mesure à l'aide de notre processus de validation. Ce résultat dépasse tous ceux obtenus pour les autres configurations de combinaison. Pour l'anglais, un gain de +12,9 en F-mesure est obtenu relativement au résultat du meilleur système individuel (Mot) et de +16,3 par rapport à l'union naïve. Pour le français, cette validation apporte des améliorations en F-mesure de +11,7 relativement au résultat de la meilleure technique individuelle (Mot) et de +13,5 relativement à l'union de toutes les techniques individuelles. Sans surprise, les valeurs maximales pour le rappel demeurent celles

7. Cela est rendu possible par le fait que la précision de l'union est inférieure à 50 sur les deux langues ce qui nous fournit donc un nombre plus élevé d'exemples négatifs.

8. Les valeurs de performance sont obtenues en effectuant une validation croisée avec 10 sous-corpus, et correspondent donc aux valeurs moyennes en considérant tour à tour chacun de ces sous-corpus comme données d'évaluation et le complément comme données d'apprentissage.

9. Nous avons également décidé de recopier les résultats des techniques individuelles ainsi que ceux obtenus par l'union naïve sur laquelle s'opère la validation, afin de faciliter l'analyse.

obtenues par l'union. On constate que l'ensemble des techniques individuelles étudiées permet de trouver des quantités comparables de paraphrases dans les deux langues. Bien que ces résultats soient satisfaisants étant donné la complexité de notre tâche, une analyse plus fine des faux positifs et des faux négatifs pourra éventuellement nous permettre d'améliorer les performances obtenues par l'ajout de traits pertinents.

5.4.3 *Étude d'ablation de techniques*

	Union	**Validation**					
	Tout	\Mot	\Pivot	\Terme	\Synt	\Terp$_{\rightarrow F}$	Tout
	anglais						
P	35,75	68,79	64,53	**68,96**	68,49	67,48	68,83
R	**81,06**	57,94	59,66	62,92	62,84	58,66	63,26
F_1	49,62	62,90	62,00	65,80	65,54	62,76	**65,93**
	français						
P	40,84	**79,56**	69,45	74,62	74,70	77,32	74,75
R	**78,43**	55,37	56,96	60,55	61,06	57,13	61,10
F_1	53,71	65,30	62,59	66,85	67,19	65,72	**67,24**

Tableau 15: Résultats obtenus en retirant à tour de rôle une technique individuelle de l'expérience de validation des paraphrases.

Les techniques individuelles employées dans ce travail ont été choisies principalement pour leurs caractéristiques et pour leur complémentarité potentielle. Il est donc intéressant d'analyser les résultats obtenus lorsque les paraphrases produites par chacune des techniques ne sont pas prises en compte à tour de rôle dans le processus de validation. Ainsi, si une technique particulière met en jeu des ressources spécifiques (par exemple des corpus bilingues dans Pivot), la retirer peut nous fournir des indications sur sa contribution par rapport aux autres techniques utilisées.

Le tableau 15 indique les résultats obtenus pour ces expériences. Il n'est pas surprenant que la suppression de Terme ou de Synt de l'ensemble des techniques n'entraîne pas de baisse sensible des résultats compte tenu du très faible nombre de paraphrases proposées par ces techniques. Le retrait de toutes les autres techniques provoque une baisse notable de la F-mesure. Mot et Edit ne produisent pas de résultats très différents, tandis que la contribution de Pivot apparaît comme nettement plus importante que toutes les autres techniques bien qu'elle ne présente pas la meilleure valeur

de complémentarité avec l'ensemble des autres techniques (voir le tableau 10). Ce résultat souligne la contribution positive de l'utilisation des corpus parallèles bilingues sur l'acquisition de paraphrases sous-phrastiques.

5.4.4 *Étude d'ablation de traits et de la variation de la taille du corpus d'apprentissage*

Afin d'étudier l'apport de chaque classe de modèles fournis dans notre classifieur, d'une part, et l'impact du nombre d'exemples dans notre corpus d'apprentissage, nous avons réalisé une étude d'ablation de modèles variant en fonction de la taille du corpus d'apprentissage. La principale motivation de cette analyse est de vérifier que la méthode de classification que nous proposons est capable de se concentrer sur les traits les plus fiables dans conditions expérimentales diverses.

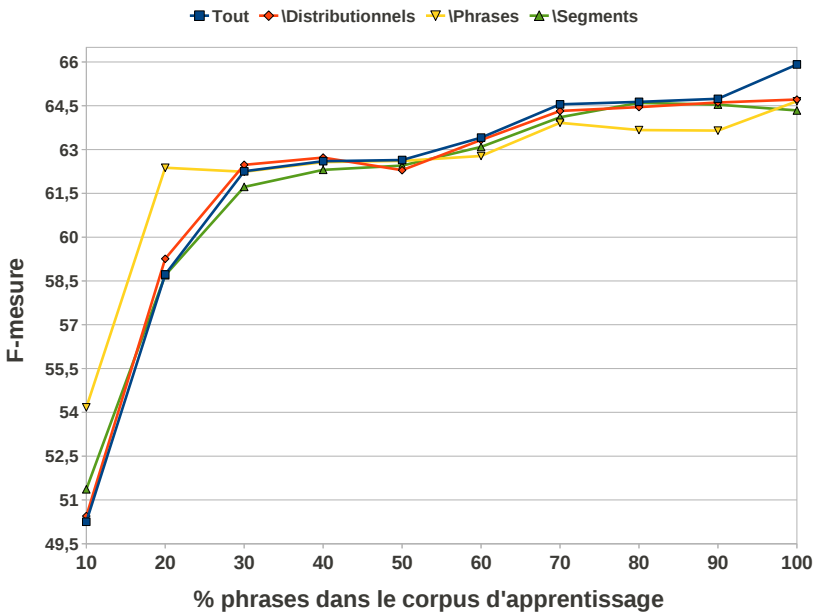


FIGURE 20: Courbes d'apprentissage obtenues par les suppression des classes de traits individuellement, pour l'anglais. La classe de traits des "systèmes" n'apparaît volontairement pas ici : son retrait mène à une F-mesure variant de 48,1 (10%) à 56,3 (100%), soit des valeurs bien inférieures aux autres présentées ici.

La figure 20 montre comment les performances en termes de F-mesure varient en fonction du nombre d'exemples considérés dans

le corpus d'apprentissage, pour l'anglais. Nous ne montrons pas les résultats obtenus en supprimant les traits dérivés des systèmes afin de garantir une certaine visibilité des courbes. La suppression de ces traits provoquent une chute des valeurs de F-mesure pour toutes les configurations. Les valeurs varient entre 48,1 (quand seuls 10% des exemples sont considérés) et 56,3 (lorsque tout le corpus d'apprentissage est pris en compte). Ce résultat n'est pas surprenant dans la mesure où le calcul de ces traits est basé sur le nombre de systèmes proposant chaque exemple de paraphrase. On observe en effet que les paraphrases proposées par tous les systèmes sont parfaites (la précision de l'intersection est de 100% pour un très faible rappel de 0.18), ce qui exploite l'hypothèse bien connue que les résultats consensuels sont souvent corrects.

Cependant, en supprimant la classe des traits dérivés des paires d'énoncés, et en n'exploitant qu'un petit nombre d'exemples (entre 10 et 20% d'exemples), la performance de notre classifieur semble être meilleure que la combinaison de tous les traits. Ceci est probablement dû aux exemples positifs et négatifs partageant les mêmes énoncés d'origine et donc les mêmes mesures de similarité (BLEU, COSINUS, TER, METEOR), ce qui n'aide pas le classifieur à distinguer les paraphrases des autres phénomènes pour les rattacher à leurs bonne classe d'appartenance.

Les courbes correspondant à la suppression du reste des classes de traits se comportent globalement d'une manière cohérente. Les performances du classifieur varient de façon presque linéaire avec le nombre d'exemples considérés dans le corpus d'apprentissage.

En combinant toutes les classes de traits et en utilisant la totalité des exemples du corpus d'apprentissage, nous obtenons le meilleur résultat (65,93 de F-mesure). Ce résultat confirme l'utilité des traits sélectionnés pour caractériser les paraphrases candidates et laisser entrevoir de possibles gains par ajout de nouvelles données d'apprentissage.

Bien que ces résultats soient déjà satisfaisants, étant donnée la complexité de cette tâche de classification, une étude plus fine des faux positifs et négatifs pourrait nous aider à détecter les cas problématiques et développer d'autres modèles afin d'obtenir de meilleures performances de classification.

5.4.5 *Performance en fonction du degré de comparabilité des énoncés*

En utilisant le même calcul de la difficulté d'alignement d'une paire d'énoncés $(1 - \mathrm{TER}(\mathrm{énoncé}_1, \mathrm{énoncé}_2))$ présentées dans la section 4.4, nous avons mesuré l'impact du degré de comparabilité des paraphrases d'énoncés sur les résultats de la validation.

Les résultats obtenus pour l'ensemble des techniques étudiées ainsi que pour la technique de validation de leurs résultats par

classification sont présentées dans la figure 21. Le résultat le plus intéressant ici concerne la validation dont les performances dépassent ceux de toutes les autres techniques pour tous les niveaux de difficulté d'alignement, pour les deux langues. Cependant, le degré de comparabilité des énoncés alignés a un impact clair sur le nombre de paraphrases correctes retenues après validation : plus les deux énoncés sont similaires (parallèles), plus le nombre de paraphrases sous-phrastiques correctement reconnues est important.

Ces observations complémentaires viennent confirmer l'apport de la validation par classification pour la tâche d'acquisition de paraphrases sous-phrastiques.

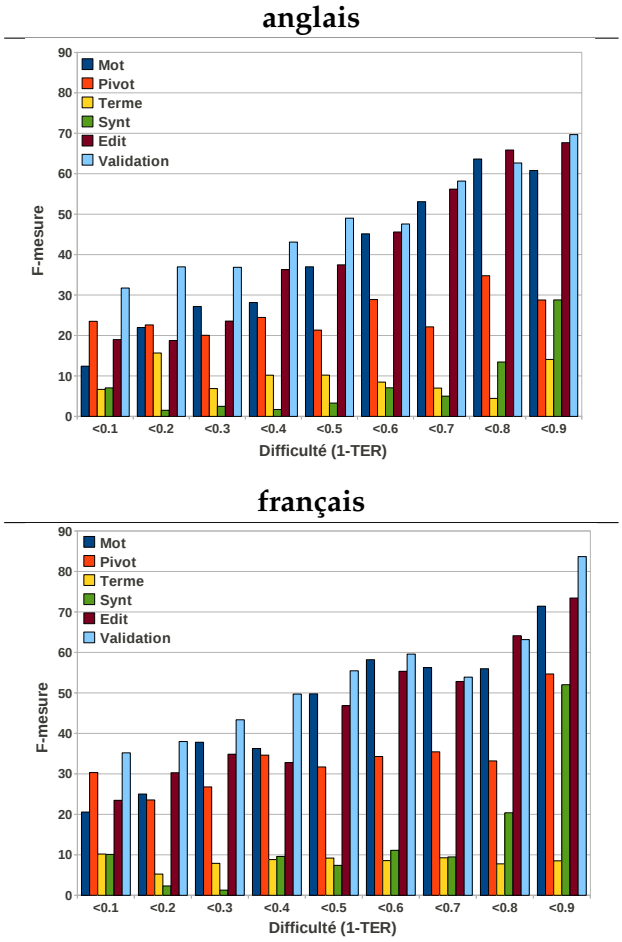
5.5 TYPOLOGIE DES PARAPHRASES DIFFICILES À ACQUÉRIR

Un objectif important de cette étude consiste à caractériser les paraphrases qui sont « difficiles » à acquérir automatiquement par les techniques étudiées. Les rappels de l'ordre de 80% obtenus par l'union naïve sur les deux langues nous semblent à la fois souligner la complexité de la tâche visée et indiquer la bonne performance relative obtenue par les techniques choisies. Ce dernier point confirme donc l'intérêt qui existe à étudier les paraphrases de la référence défiant les techniques automatiques.

Nous avons décidé de nous concentrer sur la configuration où l'on considère les résultats de l'union de chacune des techniques, puisque, naturellement, c'est ainsi qu'est obtenu le meilleur rappel (voir le tableau 14). Nous avons extrait le sous-ensemble du corpus de référence contenant des paraphrases ne faisant pas partie des paraphrases obtenues par l'union des techniques. Ces paraphrases constituent le sous-ensemble de paraphrases de référence (*sûres* et *possibles*) qu'aucune technique individuelle n'a pu identifier.

Nous avons défini de grandes classes de paraphrases par annotation itérative effectuée par deux annotateurs. Nous réutilisons certaines classes utilisées dans la définition de la typologie des paraphrases apparaissant dans les corpus étudiés dans le chapitre 3. La classification obtenue est décrite dans le tableau 16 : chaque classe y est illustrée par des exemples représentatifs dans les deux langues, et les classes ont été ordonnées par ordre décroissant de leur fréquence dans le corpus anglais.

Tout d'abord, nous constatons qu'il existe une corrélation claire entre les deux langues, à quelques exceptions près. Par exemple, il existe plus d'*équivalences lexicales et sous-phrastiques* dans notre corpus français, mais moins de *variations typographiques* et de *variations morphologiques*. L'*équivalence lexicale et sous-phrastique* est de loin la catégorie la plus représentée avec un peu moins d'1/3 de toutes les paraphrases difficiles à repérer dans le corpus anglais et 4/10 de ces paraphrases dans le corpus français. En se référant à la ty-



Intervalles										
<0,1	<0,2	<0,3	<0,4	<0,5	<0,6	<0,7	<0,8	<0,9	<1	total
anglais										
18	24	66	68	96	94	77	43	12	2	500
français										
38	27	63	80	100	71	65	41	12	3	500

FIGURE 21: Performance en F-mesure de nos 5 techniques étudiées et de notre système de validation pour l'anglais (en haut) et le français (au centre) en fonction de la difficulté d'alignement des paires d'énoncés mesurée par la valeur de (1-TER). La distribution des énoncés dans les intervalles est fournie dans le tableau (en bas).

pologie des paraphrases trouvées dans l'échantillon de 50 paires d'énoncés extraites des corpus obtenus par traductions multiples (voir section 3.3), nous remarquons que pour le français, par exemple, 46,9% des paraphrases ont été classées dans cette catégorie, ce qui explique la forte proportion des synonymes (lexicaux ou sous-phrastiques) difficiles à acquérir.

Catégorie	Exemples	# en	% en	# fr	% fr
Équivalence lexicales et sous – phrastiques	businesses ↔ entreprises at a rapid rate ↔ fast conference ↔ general assembly basic installations ↔ infrastructure bonne ↔ appropriée maintenant ↔ à présent Même si ↔ En dépit des	329	31,76	420	48,72
Variations pragmatiques	to south korea ↔ home are taking emergency actions ↔ roll up their sleeves investigation bureau ↔ department endémiques locaux ↔ sédentaires de la communauté ↔ membre	244	23,55	176	20,42
Variations typographiques	hong kong ↔ hongkong 11 ↔ eleven 20 billion us dollars ↔ us$4,1 billion february ↔ feb, voice of america ↔ voa programme-cadre ↔ programme cadre UPU ↔ Union Postale Universelle	171	16,51	33	3,83
Inclusions	the hopewell group ↔ hopewell pfizer now is ↔ pfizer is now ↔ right now south korea ↔ korea centrale ↔ centrale thermique d'ordre interne ↔ interne	95	9,17	58	6,73
Variations morphologiques	to resign ↔ resigning iraqi ↔ iraq british ↔ by Great Britain hong kong people ↔ honkongnese postaux ↔ de poste environnementales ↔ relatives à l'environnement	116	11,2	82	6,61
Variations syntaxiques	temperature on the surface ↔ surface temperature it is an urgent task ↔ has become urgent pour quel montant ↔ quel était le montant qui lui ont soumis ↔ lui ayant été soumis	48	4,63	53	9,51
Nombre	assertion ↔ assertions industries ↔ industry	21	2,03	23	2,67
Anaphore	Pinochet ↔ he east timor ↔ it Somalie ↔ pays	12	1,16	17	1,97
	Total	1,036	100	862	100

Tableau 16: Classes et exemples de paraphrases sous-phrastiques « difficiles à acquérir » par les techniques automatiques étudiées. Les catégories ont été ordonnées par fréquence décroissante en anglais.

Quelques autres paires impliquant des segments assez longs tel que *en train d' être réalisé à grands pas* ↔ *en cours* sont difficiles à identifier pour toutes les techniques, à l'exception de Synt dans le cas où elles apparaissent dans des structures syntaxiques compatibles. L'approche d'alignement statistique (Mot) peut, en particulier, avoir des difficultés à identifier des équivalences entre des éléments rares dans une paire de paraphrases comme *Même si* ↔ *En dépit des*. Quelques autres exemples, tel que *bonne* ↔ *appropriée*, auraient pu

être capturés si l'on disposait de suffisamment de connaissances lexicales *a priori*. Ces connaissances pourraient provenir de dictionnaires ou de ressources lexico-sémantiques (tel que Wordnet pour l'anglais) et être intégrées dans la liste des synonymes exploitée par FASTR.

La classe des *variations pragmatiques*, représentant plus de 20% d'exemples dans les deux langues, correspond aux segments qui ne sont des paraphrases que dans certains contextes très spécifiques et qui sont, par conséquent, difficiles à identifier sans passer par une analyse complexe comme *de la communauté* $\leftrightarrow$ *membre*. Il n'est donc pas surprenant que ces phénomènes, résultant des différents choix faits par les traducteurs humains impliqués, soient difficiles à acquérir automatiquement, bien que des techniques comme EDIT ou SYNT soient capables de les identifier dans certaines conditions. Néanmoins, il peut s'avérer difficile de réutiliser ces paraphrases dans des applications concrètes, ce qui peut limiter par conséquent leur intérêt dans le cadre de l'étude des phénomènes de paraphrase. Il est également à noter que certains cas correspondent à des choix difficiles, et parfois incorrects ou discutables, faits par les annotateurs humains lors de la construction du corpus de référence, ou de leur incapacité à suivre correctement les directives de la tâche complexe d'annotation.

CONCLUSION DU CHAPITRE

En évaluant les cinq méthodes d'acquisition de paraphrases décrites dans le chapitre 4, nous avons constaté que leurs résultats étaient relativement complémentaires, ce qui nous a mené à essayer de combiner leurs résultats. Outre la combinaison des résultats des techniques individuelles *a posteriori* (union naïve), nous avons choisi d'utiliser les résultats de certaines méthodes comme données d'entrée d'une autre. Les résultats de cette approche ont confirmé notre hypothèse de complémentarité en montrant que l'hybridation de ces techniques permet souvent des gains significatifs en acquisition de paraphrases. Nous avons ensuite proposé une méthode de validation de paraphrases en formulant ce problème sous la forme d'une classification automatique exploitant différents traits pour l'identification des paraphrases. Ceci nous a permis d'obtenir des gains significatifs relativement aux techniques individuelles : nous obtenons une amélioration relative de +27% environ en F-mesure par rapport à la meilleure technique individuelle sur les deux langues. Un résultat important de notre étude est également l'identification, la description et la quantification de paraphrases qui défient les techniques étudiées.

Bien que nous ayons identifié le manque de ressources appropriées et suffisantes comme un important frein pour les recherches abordant les phénomènes paraphrastiques, nous croyons que la disponibilité d'un grand nombre de corpus monolingues parallèles permet déjà de poursuivre ces travaux dans plusieurs directions pour améliorer les résultats obtenus. En particulier, il semble intéressant d'étudier l'acquisition de paraphrases à partir de corpus de différents degrés de comparabilité, ce qui est l'objectif du chapitre suivant.

6

ACQUISITION DE PARAPHRASES SOUS-PHRASTIQUES : EXPLOITATION D'AUTRES TYPES DE CORPUS

Dans le chapitre 3, nous avons montré l'impact des sources à partir desquelles des paires de paraphrases d'énoncés sont acquises, sur leur degré de parallélisme, et sur la quantité et la qualité des paraphrases y figurant. Les expériences décrites dans le chapitre 4 ont confirmé ces résultats.

Malheureusement, comme nous l'avons déjà évoqué, les corpus monolingues parallèles que nous avons exploités dans le chapitre précédent sont peu nombreux et difficiles à construire. En outre, ils ne sont pas nécessairement représentatifs des types de corpus permettant d'acquérir des paraphrases.

Il est donc important de considérer à présent comment la technique décrite dans le chapitre 5 se comporte lorsque les paires d'énoncés présentent différents niveaux de parallélisme, problème déjà étudié dans la section 5.4.5 pour des paires d'énoncés de la même origine. Afin d'obtenir des paires d'énoncés à différents niveaux de similarité, tout en garantissant une proximité sémantique suffisante pour l'acquisition de paraphrases, nous avons considéré la possibilité d'utiliser des corpus de paires d'énoncés obtenus à partir de *signaux d'origine* différents et de les comparer. Les corpus que nous utiliserons dans ce chapitre seront ainsi les suivants :

- Texte [1] : des paires d'énoncés résultant de traductions multiples indépendantes d'un même texte (Barzilay et McKeown, 2001; Cohn *et coll.*, 2008) ;
- Parole : des paires d'extraits de textes résultant de traductions multiples indépendantes de mêmes extraits de parole (Tiedemann, 2007; Lavecchia *et coll.*, 2007) ;
- Scène : des paires d'énoncés résultant de descriptions multiples d'une même scène visuelle (Chen et Dolan, 2011) ;
- Événement : des paires d'énoncés résultant de descriptions multiples d'un même événement ou de deux événements proches (Dolan *et coll.*, 2004; Wubben *et coll.*, 2009).

Cette étude vient donc en complément de celle présentée dans le chapitre précédent, qui correspond à l'acquisition de paraphrases

1. Texte correspond en fait au corpus utilisé dans les chapitres 4 et 5

depuis le corpus que nous appellerons désormais TEXTE. Nous avons utilisé des corpus comportant un nombre identique de paires d'énoncés, toujours sur les deux mêmes langues d'étude, l'anglais et le français. Quatre des systèmes d'acquisition de paraphrases décrits dans le chapitre 4 ont été utilisés et leur résultats ont été combinés par la méthode détaillée dans le chapitre 5, qui s'est montrée la plus performante de notre étude.

Nous allons tout d'abord décrire la méthodologie de construction de ces corpus et leurs caractéristiques principales (section 6.1). Nous détaillons ensuite la performance du système de combinaison sur chacun des types de corpus (section 6.2), puis la performance de ce système lorsque des données d'apprentissage additionnelles provenant des autres types de corpus sont utilisées (section 6.3). Nous étudions finalement les paraphrases sous-phrastiques présentes dans chacun des corpus étudiés, en rendant compte des proportions trouvées pour chaque type par notre système (section 6.4).

Notre travail s'est concentré jusqu'ici sur l'acquisition de paraphrases sous-phrastiques depuis des paires d'énoncés en relation, et nous considérons à présent le fait d'utiliser des corpus dans lesquels les paires d'énoncés sont issues de différentes origines. Nous avons construit 8 corpus, un par signal d'origine et par langue. Ces corpus comportent tous 625 d'énoncés. Ces paires ont été obtenues en suivant la même méthodologie de construction que celle utilisée pour nos petits corpus de 50 paires utilisés dans le chapitre 3 et elles ont été annotées manuellement en suivant les mêmes consignes d'annotation. Nous décrivons ainsi brièvement ici les spécificités de ces nouveaux corpus.

6.1.1 *Corpus* TEXTE

Nous allons reprendre ici le corpus utilisé dans les chapitres 4 et 5, qui avait été choisi pour le développement de nos techniques d'acquisition car il s'est avéré être le plus dense en paraphrases sous-phrastiques lors de l'expérience d'annotation manuelle du chapitre 3. Des matrices illustrant des alignements de référence, pour les deux langues, sont données dans la figure 22. Les cases sur fond vert et gris (identité) correspondent aux alignements qualifiés de *sûrs* par l'annotateur, et celles sur fond jaune à des alignements qualifiés de *possibles*.

6.1.2 *Corpus* PAROLE

Pour l'anglais, nous avons utilisé des fichiers de sous-titres de films tournés en français [2], *Le Fabuleux Destin d'Amélie Poulain* et *Les Choristes*, et pour le français nous avons agrandi le corpus utilisé précédemment en extrayant un plus grand nombre de paires à partir des fichiers de la série américaine *Desperate Housewives*. Chaque corpus a été aligné à l'aide du même algorithme décrit dans le chapitre 3, puis nous avons extrait des paires d'énoncés en dessous d'un seuil minimal de taux d'édition. Des exemples de matrices de référence pour ce corpus sont donnés dans la figure 23.

6.1.3 *Corpus* SCÈNE

Nous avons utilisé le corpus *Multiple Video Description* (Chen et Dolan, 2011) obtenu à partir de descriptions multiples de courtes vidéos. De façon analogue à ce qui a été fait pour l'échantillon utilisé dans l'étude de l'impact des sources sur la nature des paraphrases acquises du chapitre 3, nous avons choisi des paires

2. Librement disponibles sur : http://www.opensubtitles.org

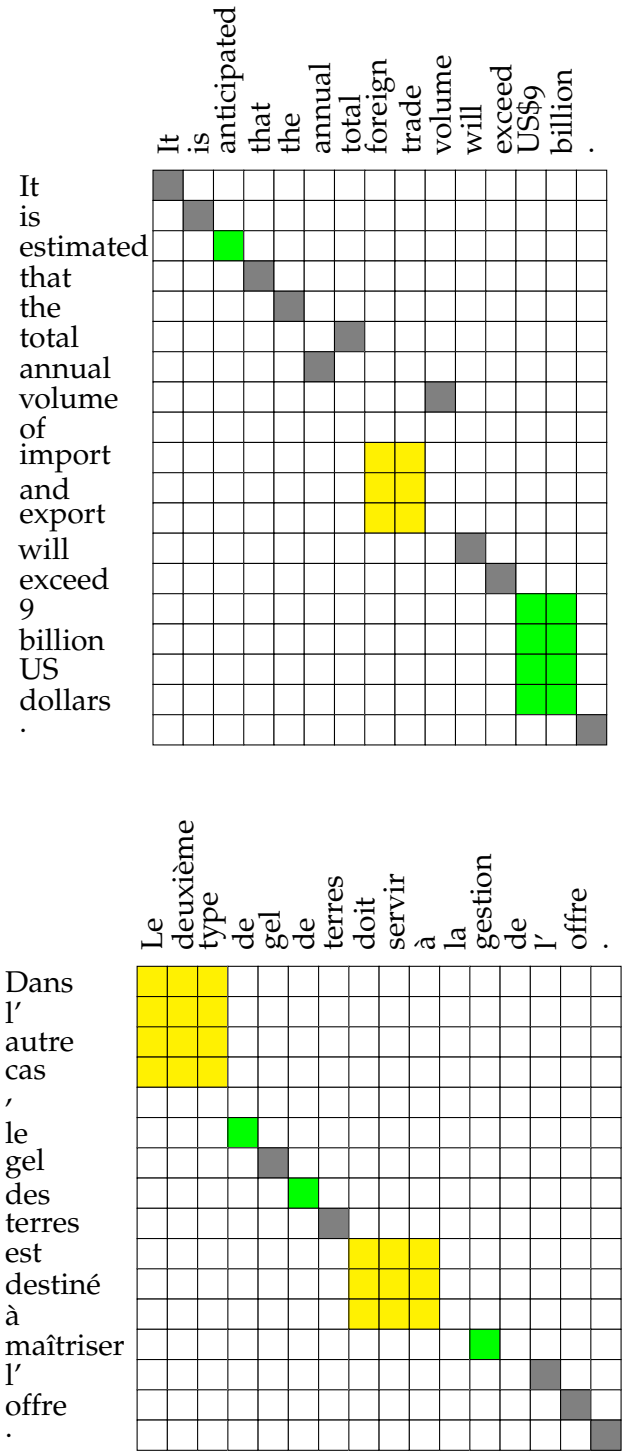


FIGURE 22: Exemples de matrices d'alignement de référence pour des paires d'énoncés extraits du corpus TEXTE pour l'anglais (en haut) et le français (en bas). Les alignements sur fond vert sont *sûrs*, ceux sur fond gris sont *sûrs* entre des formes identiques (*identité*), et ceux sur fond jaune sont *possibles*.

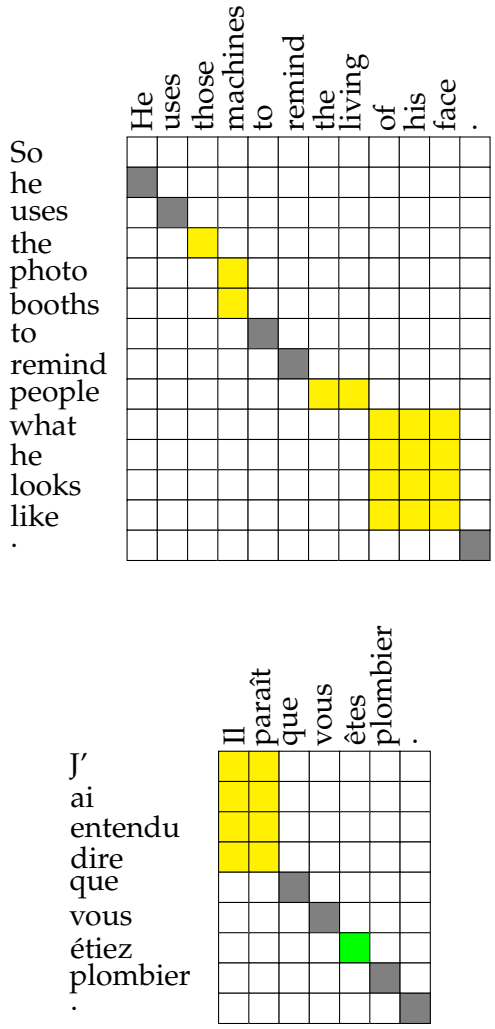


FIGURE 23: Exemples de matrices d'alignement de référence pour des paires d'énoncés extraits des corpus PAROLE pour l'anglais (en haut) et le français (en bas).

d'énoncés au sein de ces groupes en fonction d'un taux d'édition au-dessus d'un certain seuil. Un point important est que, pour l'anglais, nous avons pu utiliser des descriptions qualifiées de "vérifiées", signifiant que leurs contributeurs ont été sélectionnés de façon supervisée. Les descriptions en français dans cette ressource sont disponibles dans des quantités bien moins importantes, et en outre aucune n'a le statut de "vérifiée". Nous avons tout de même décidé d'utiliser ce corpus, mais en gardant à l'esprit que cette source est de nettement moins bonne qualité[3]. Des alignements de référence pour des paires de descriptions de vidéos sont donnés dans la figure 24.

3. Ce type de corpus sera par la suite désigné entre parenthèse pour le français ("(SCÈNE)") afin de rappeler son caractère particulier.

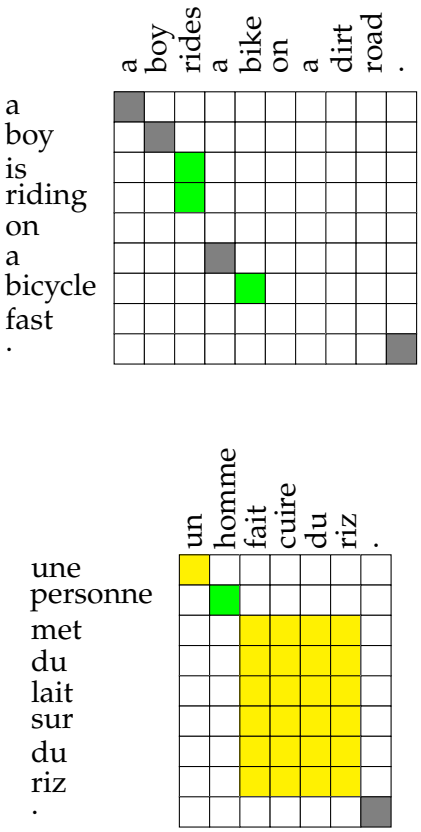


FIGURE 24: Exemples de matrices d'alignement de référence pour des paires d'énoncés extraits du corpus SCÈNE pour l'anglais (en haut) et le français (en bas).

6.1.4 *Corpus* ÉVÉNEMENT

Nous avons utilisé des titres de groupes d'articles d'actualité provenant du service d'agrégation Google News [4] dans ses versions anglophone et francophone. Nous avons téléchargé 534 groupes d'articles pour l'anglais et 295 groupes pour le français. Nous avons ensuite affiné l'algorithme de regroupement en ne retenant pour chaque groupe que les paires d'articles dont les dates de publication n'était pas espacées de plus d'un jour. Les paires retenues dans ce corpus ont été choisies en suivant le même protocole que pour l'échantillon utilisé dans l'étude préalable présentée dans le chapitre 3. Un exemple de paires d'énoncés alignées est donné, pour l'anglais et le français, dans la figure 25.

4. http://news.google.com

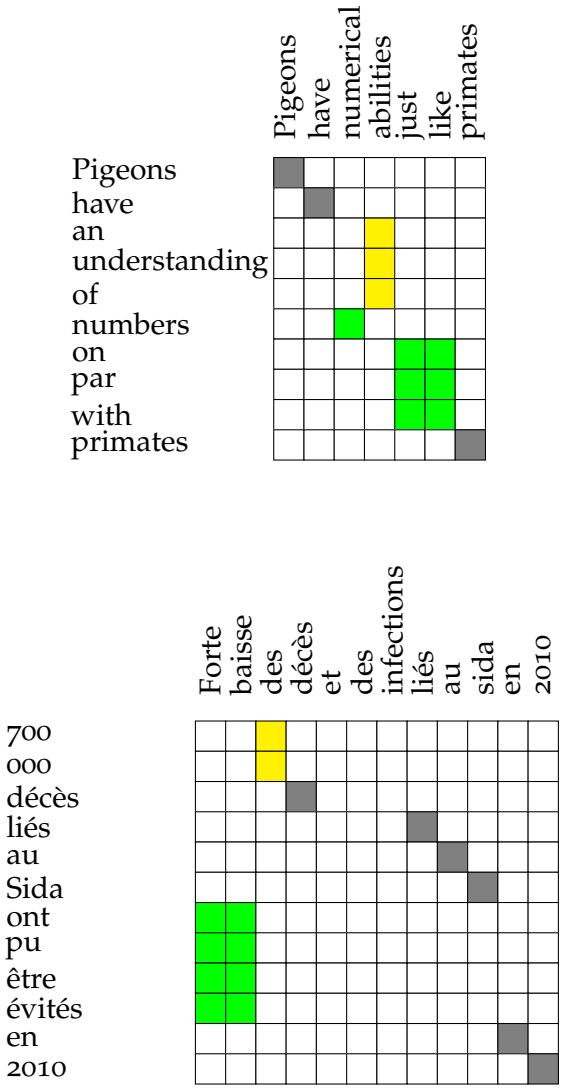


FIGURE 25: Exemples de matrices d'alignement de référence pour des paires d'énoncés extraits du corpus ÉVÉNEMENT pour l'anglais (en haut) et le français (en bas).

6.1.5 *Analyse des résultats d'annotation des corpus*

La figure 26 présente tout d'abord différentes mesures de similarité (définies dans le chapitre 3) fréquemment utilisées pour comparer des paires d'énoncés. Nous constatons que dans la majorité des cas, ces mesures, fondées sur des critères de similarité différents, classent les corpus de façon similaire. L'analyse comparative des différents corpus révèle que TEXTE contient les paires

d'énoncés les plus similaires selon toutes les mesures pour les deux langues, suivi de près par Événement. Scène contient des paires d'énoncés qui sont plus similaires que celles de Parole pour l'anglais, ce qui n'est pas le cas pour le français où les résultats sont plus contrastés selon les mesures. Ceci est probablement dû à la faible qualité du corpus Scène en français.

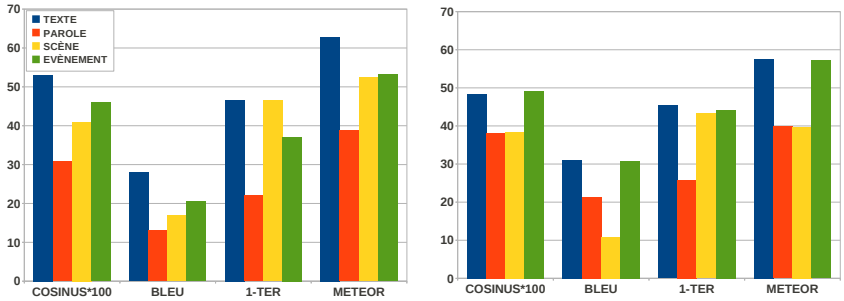


Figure 26: Moyenne de différentes mesures de similarité entre paires d'énoncés pour l'ensemble des corpus pour l'anglais (à gauche) et le français (à droite). Les mesures incluent : le cosinus des vecteurs de formes, BLEU (Papineni *et coll.*, 2002), TER (Snover *et coll.*, 2006) et METEOR (Lavie et Agarwal, 2007).

	Statistiques du corpus 500 paires d'énoncés		**Accords inter-annotateurs** 50 paires d'énoncés	
	# formes par énoncé	# formes	para. sûres	para. possibles
	anglais			
Texte	21 473	21,0	66,1	20,4
Parole	11 049	10,5	79,1	10,9
Scène	7 783	7,5	80,5	35,2
Événement	8 609	8,0	65,3	20,5
	français			
Texte	24 641	24,0	64,6	16,67
Parole	11 850	11,5	82,7	20,8
(Scène)	7 012	6,5	42,8	9,3
Événement	9 121	9,1	67,8	3,8

Tableau 17: Description de l'ensemble des corpus collectés et des annotations de référence pour les paraphrases en anglais et en français.

Les 4 corpus ont été annotés en paraphrases sous-phrastiques en suivant la procédure d'annotation utilisée dans le chapitre 3, toujours par deux annotateurs par langue. Le tableau 17 décrit les caractéristiques principales des corpus annotés ainsi que les accords inter-annotateurs obtenus.

Pour les deux langues, nous observons que l'augmentation du nombre de paires d'énoncés dans l'ensemble des corpus donnent des résultats qui restent cohérents avec ceux obtenus pour les petits corpus utilisés dans le chapitre 3. Le corpus Texte, obtenu par traduction multiple, contient des énoncés significativement plus longs que les autres corpus, par exemple plus de trois fois plus longs que ceux correspondant à des descriptions de vidéos (Scène).

Comme dans le chapitre 3, les valeurs d'accords inter-annotateurs ont été calculées sur des sous-ensembles de 50 paires d'énoncés annotées indépendamment par deux annotateurs et reportées dans le tableau 17. Pour l'anglais, tout comme pour le français, nous considérons les valeurs obtenues pour les paraphrases sûres comme acceptables mais les valeurs obtenues pour les paraphrases possibles sont très faibles (65,3 comparé à 20,5 pour Événement).

Le tableau 18 présente enfin les pourcentages et les nombres de paraphrases pour chaque niveau de certitude pour chacun des corpus. Nous obtenons approximativement le même nombre total de paraphrases pour l'anglais (16 799) et le français (17 001). Les corpus anglais dans leur totalité contiennent environ le même nombre de paraphrases des deux niveaux de certitude (8 303 *sûres* et 8 496 *possibles*), alors qu'en français on trouve davantage de paraphrases sûres (11 953 contre 5 048). Ceci peut s'expliquer par le fait que les annotateurs ont travaillé indépendamment, avec des interprétations souvent différentes de la tâche d'annotation, et que les corpus, aussi *comparables* soient-ils entre langue, sont différents par nature. Les autres faits remarquables sont que Texte contient beaucoup plus de paraphrases que les autres corpus, que Parole comporte proportionnellement plus de paraphrases *possibles* que les autres corpus, et que Scène contient nettement moins de paraphrases, en pourcentage et en nombre.

6.2 Acquisition de paraphrases sous-phrastiques par type de corpus

6.2.1 *Contexte expérimental*

Nous avons testé les performances des systèmes d'acquisition de paraphrases que nous avons développés sur nos divers corpus de paires d'énoncés. Nous avons retenu quatre des systèmes, déjà

	Stat. sur les formes dans les paraphrases sans les paraphrases identiques			
	para. sûres		para. possibles	
	% formes	# formes	% formes	# formes
	anglais			
Texte	18,6	4 004	12,3	2 651
Parole	17,5	1 942	31,6	3 500
Scène	10,9	851	14,0	1 094
Événement	17,5	1 506	14,5	1 251
	français			
Texte	29,2	7 218	6,2	1 527
Parole	22,5	2 667	16,7	1 981
(Scène)	3,9	275	9,4	664
Événement	19,6	1 793	9,6	876

Tableau 18: Description des formes contenus dans les paraphrases *sûres* et *possibles* extraites des corpus annotés manuellement en anglais (partie haute) et en français (partie basse). Ces mesures ne prennent pas en compte les paires de paraphrases constituées de segments identiques.

décrits dans le chapitre 4 : Mot, Terme, Edit et Pivot [5]. Nous avons également testé la méthode de validation des paraphrases issues de l'union des sorties des quatre systèmes considérés, en utilisant les ensembles de traits pour l'identification des paraphrases décrits dans le chapitre 5. Nous suivons finalement toujours la méthodologie d'évaluation décrite dans le chapitre 4 pour évaluer les performances des différents systèmes sur les différents corpus.

6.2.2 *Analyse des résultats*

Les résultats pour les systèmes appliqués individuellement, leur union et nos systèmes de combinaison entraînés sur chaque type de corpus (colonne "appr.=C") sont donnés dans le tableau 21 [6].

Nous constatons tout d'abord que tous les systèmes obtiennent de meilleurs résultats sur Texte, corpus pour lequel les plus grandes quantités de données d'apprentissage sont disponibles et

5. Le système Synt n'a pas été testé pour ces nouvelles expériences : cette technique obtenait des scores en rappel très faibles sur le corpus le plus parallèle, Texte (6,3 pour l'anglais et 7,3 pour le français), et les énoncés des corpus tels que Scène ou Événement se prêtent mal à une analyse syntaxique.

6. Nous avons recopié les résultats obtenus sur le corpus Texte dans la table afin de faciliter leur comparaison avec les nouveaux résultats. Les résultats donnés dans cette table correspondent à une moyenne de ceux obtenus en réalisant une validation croisée sur 10 ensembles de test pour chaque type de corpus.

dans lequel les équivalences sémantiques entre paires d'énoncés sont les plus probables.

En termes de performance en F-mesure par type de corpus, l'ordre des techniques testées en termes de performances est à peu près le même, à l'exception de Mot ayant une performance moins bonne que celle de Edit sur le corpus Événement. Ce résultat peut s'expliquer par le fait que Mot ne disposait pas de suffisamment de données d'apprentissage pour ce genre de corpus, contenant généralement de nombreux mots rares et de nombreuses entités nommées. Mot obtient de meilleurs résultats sur Texte et Parole, qui contiennent des énoncés longs, avec des répétitions probablement plus nombreuses au niveau du corpus, alors que Edit$_{\rightarrow F}$ a de meilleurs résultats sur Scène et Événement, où les équivalences qui sont rares au niveau du corpus sont plus fréquentes.

Dans toutes les configurations, notre combinaison de systèmes améliore de façon importante la F-mesure relativement au meilleur des systèmes individuels pour chaque type de corpus, ainsi que relativement à l'union des résultats de l'ensemble des systèmes. Les améliorations importantes précédemment observées sur Texte (respectivement +12,5 et +11,6 sur l'anglais et le français) existent également sur Parole (+11,7 et +11,1) et dans une moindre mesure sur Scène (+3,2 et +6,4) et sur Événement (+5,4 et +7,1). Nous avons constaté (voir le tableau 18) que Texte et Parole sont les deux types de corpus ayant le plus grand nombre d'exemples de paraphrases *sûres* pour les deux langues : nos résultats semblent indiquer que notre classifieur a été capable de les utiliser efficacement.

Le corpus Événement apparaît lui comme le type le plus difficile pour l'acquisition de paraphrases, ce qui pourrait être considéré comme un résultat décevant dans la mesure où il s'agit du type de corpus pour lequel de grandes quantités de données sont disponibles. Néanmoins, les valeurs de F-mesure de 46,0 pour l'anglais et 46,8 pour le français correspondent à des résultats provisoirement acceptables pour ce niveau de difficulté.

Nous notons finalement que les valeurs de rappel pour l'union sont assez fortes pour tous les types de corpus, allant de 71,4 (pour Parole) à 83,4 (pour Scène en anglais, et allant de 72,5 (pour Événement) à 86,4 (pour Scène) en français. Il y a, cependant, une nette baisse entre les valeurs de rappel entre les unions et nos systèmes de combinaison (colonne "appr.=C"), bien que ces dernières demeurent autour de 60, pour les deux langues. Nous pouvons également noter que la précision est en général meilleure pour un des systèmes individuels (Pivot) ce qui permet à la combinaison d'atteindre des valeurs intéressantes en particulier sur Texte, où nous disposons du plus grand nombre d'exemples (Précision de respectivement 68,4 et 74,6 pour l'anglais et le français).

	+Texte	+Parole	+Scène	+Événement	+Tous
anglais					
# *ex+*	7 342	2 296	1 784	1 171	12 593
Texte	65,5	**66,2**	65,1	**66,2**	65,1
Parole	56,0	53,5	52,8	54,8	**56,6**
Scène	49,7	**54,3**	53,7	53,8	42,7
Événement	51,1	45,3	42,5	46,0	**56,2**
français					
# *ex+*	12 961	3 340	966	2 160	19 427
Texte	**67,1**	**67,2**	66,7	67,0	66,6
Parole	57,6	**60,0**	56,4	59,6	57,9
(Scène)	23,7	22,0	**29,8**	23,9	21,1
Événement	45,2	45,6	44,3	46,8	**49,3**

Tableau 19: Résultats de l'évaluation (scores F_1) pour tous les types de corpus pour l'anglais (en haut) et le français (en bas) avec ajout des données d'apprentissage issues des autres types de corpus.

6.3 Exploitation des données d'apprentissage issues des autres types de corpus

Nous considérons à présent la possibilité d'améliorer la performance de notre système de combinaison par l'utilisation de données d'apprentissage supplémentaires provenant d'autres types de corpus. Pour cela, nous construisons des systèmes en utilisant tout d'abord les données additionnelles provenant d'un autre type de corpus, puis de l'ensemble des types de corpus disponibles. Les résultats obtenus sont donnés dans la table 19[7], où les valeurs sur fond grisé de la diagonale correspondent aux cas où aucune donnée supplémentaire n'est ajoutée. Les lignes "#*ex+*" indiquent le nombre d'exemples positifs de paraphrases apportés par chaque type de corpus supplémentaire sur le même nombre de paires d'énoncés.

Nous observons qu'il existe deux cas de figure. Dans le premier, la performance en F-mesure est améliorée pour l'anglais sur Texte (+0,7), Parole (+3,1) et Scène (+0,6) en utilisant soit un seul type de corpus supplémentaire, soit l'ensemble des corpus disponibles, alors que pour le français aucun ajout de données d'apprentissage n'améliore la performance pour ces types de corpus. Dans le second cas, Événement est amélioré à la fois pour l'anglais (+10,2) et pour le français (+2,5) en utilisant toutes les données d'apprentis-

7. Il est à noter que ces résultats sont toujours donnés en procédant à une validation croisée qui réalise une moyenne des résultats obtenus sur 10 ensembles de test pour chaque type de corpus testé.

sage supplémentaires disponibles. Hormis la condition où les données provenant de Texte sont ajoutées pour l'anglais, tous les ajouts d'autres types de corpus diminuent la performance quand ils sont ajoutés individuellement : on observe donc ici nettement une contribution collective attribuable à l'ajout d'au moins deux sources. La nature des exemples pertinents ainsi ajoutés retiendra notre attention pour de futurs travaux : la sélection plus fine d'exemples pourrait effectivement améliorer davantage la performance atteinte.

On peut encore noter que la performance sur Texte n'est pratiquement pas affectée par l'ajout de données supplémentaires, ce qui peut s'expliquer en partie par le fait que ce type de corpus contient à lui seul la moitié du nombre total d'exemples dans les deux langues. À l'opposé, Scène, qui a le plus petit nombre d'exemples d'entraînement, voit sa performance baisser sensiblement, assez fortement par exemple avec l'ajout des données provenant de Texte (respectivement -4,0 et -6,1 pour l'anglais et le français) et par tous les corpus ensemble (respectivement -11,0 et -8,7). Ceci souligne à nouveau la nature spécifique de ce type de corpus : des descriptions indépendantes de la même scène vidéo peuvent être verbalisées de façons très diverses, à différents niveaux. Finalement, il y a nettement plus d'exemples positifs en français (19 427) qu'en anglais (12 593) : ceci peut s'expliquer par le fait que les énoncés en français dans nos corpus contiennent plus de formes (cf. Tableau 17) et que les paraphrases en français contiennent plus de variantes morphologiques telles que différentes formes conjuguées des verbes.

En conclusion, les expériences menées sur les différents corpus confirment le fait qu'ils sont de natures relativement différentes, même si dans certains cas des gains plus ou moins marqués peuvent être obtenus par l'ajout d'exemples provenant d'autres types de corpus. Il nous reste bien entendu à mieux comprendre les caractéristiques des exemples les plus utiles.

6.4 Typologie des paraphrases acquises

Une analyse fine des différents types de paraphrases serait nécessaire pour servir de guide pour des travaux futurs afin de repousser les limites des systèmes actuels. Pour cela, nous avons établi une typologie détaillée des paraphrases trouvées dans 50 paires d'énoncés annotées (%réf) ainsi que de celles acquises par notre meilleur système (%sys), ceci pour les quatre types de corpus et les deux langues de notre étude. Les résultats apparaissent dans le tableau 20[8].

8. Il convient de signaler qu'à la différence des typologies de paraphrases décrites dans le chapitre 1, la nôtre ne porte que sur les paraphrases sous-phrastiques et présente, en plus des classes de paraphrases représentées, une

		Texte	Parole	Scène	Événement
		anglais			
Synonymie	%réf	51,2	39,8	50,0	36,9
	%sys	43,5	34,0	46,8	41,6
Typographie	%réf	7,6	25,6	1,3	15,0
	%sys	7,0	38,2	2,1	22,2
Accord	%réf	11,5	13,2	20,2	13,6
	%sys	17,6	19,0	19,1	13,8
Inclusion	%réf	12,1	12,3	21,6	19,1
	%sys	16,4	6,3	23,4	16,6
Pragmatique	%réf	0,6	1,7	0,0	**1,3**
	%sys	0,0	0,0	0,0	0,0
Syntaxe	%réf	4,4	3,5	1,3	6,8
	%sys	4,7	0,0	0,0	2,7
Morphologie	%réf	12,1	3,5	5,4	6,8
	%sys	10,5	2,1	8,5	2,7
		français			
Synonymie	%réf	46,9	45,5	46,4	28,3
	%sys	26,0	43,9	51,3	16,6
Typographie	%réf	9,0	14,2	5,3	19,7
	%sys	20,6	19,5	2,7	27,7
Accord	%réf	28,5	14,2	33,8	12,2
	%sys	47,7	24,3	40,5	38,8
Inclusion	%réf	2,1	8,0	8,9	16,0
	%sys	1,0	7,3	5,4	11,1
Pragmatique	%réf	3,6	2,6	0,0	**7,4**
	%sys	1,0	0,0	0,0	0,0
Syntaxe	%réf	6,6	11,6	5,3	8,6
	%sys	0,0	2,4	0,0	5,5
Morphologie	%réf	3,0	3,5	0,0	7,4
	%sys	3,2	2,4	0,0	0,0

Tableau 20: Distribution des catégories de paraphrases mesurée dans 50 paires d'énoncés annotées (%réf) et des hypothèses de paraphrases sur ces mêmes paires pour notre meilleur système (%sys) pour l'anglais (en haut) et le français (en bas).

Une première remarque porte sur la synonymie lexicale et sous-phrastique (comme *conduire* ↔ *déboucher*, *en complément* ↔ *faisant*

analyse quantitative des paraphrases dans les annotations de référence et dans les sorties du meilleur système de notre étude.

suite) qui est le phénomène le plus observé dans les deux langues, qui de plus représente le principal type d'hypothèses correctes proposées par nos systèmes (43,5% et 26% des paraphrases contenues dans les corpus TEXTE pour respectivement l'anglais et le français). Ceci peut s'expliquer par les choix lexicaux des traducteurs et l'impact des langues source de traduction. À l'exception de ÉVÉNEMENT en anglais, la référence annotée contient un nombre plus élevé d'équivalences synonymiques. En général, ce phénomène peut être capturé en utilisant suffisamment de connaissances lexicales *a priori* (par exemple, *via* l'enrichissement de la ressource sémantique portant uniquement sur les mots utilisée dans TERME).

Nous observons également que tous les corpus contiennent une proportion importante de variations d'accord couvrant les variations de *temps* (*moque* ↔ *moquait*) et de *nombre* (*particulière* ↔ *particulières*), qui sont clairement moins intéressantes en ce qui concerne les phénomènes paraphrastiques.

En terme de variation typographique, le corpus de sous-titres contient le plus grand nombre de telles réécritures dans les deux langues. ÉVÉNEMENT contient également une proportion importante de ce type (22,2% et 27,7% pour l'anglais et le français des paraphrases reconnues par notre système). Ce phénomène correspond, par exemple, à l'utilisation d'acronymes (comme *services postaux belges et de télécommunications* ↔ *PTT belges*) ou de majuscules (*communautaire* ↔ *Communautaire*).

L'inclusion, correspondant à des variations de précision entre deux segments (*BNP* ↔ *BNP Paribas*, *article sur la journée* ↔ *article sur l'étrange journée*) est un phénomène beaucoup plus fréquent dans les corpus anglais, plus particulièrement dans SCÈNE, regroupant des descriptions de vidéos faites par des contributeurs pouvant effectivement être plus ou moins précis dans leurs descriptions (*the boy* ↔ *the small boy*, *three cartoons* ↔ *three female cartoons*). Dans la version anglaise de ce corpus, 23,4% des paraphrases identifiées par notre système sont classées dans cette catégorie.

Nous remarquons la présence de quelques variations syntaxiques dans l'ensemble des corpus et dans les deux langues. Ces variations correspondent généralement à des segments relativement longs tels que *quelle en était la justification*↔ *pour quelles raisons*. Elles sont difficiles à identifier : seules 4,7% sont identifiées dans TEXTE en anglais, et aucune paraphrase de ce type n'a été trouvée par notre système dans le corpus SCÈNE dans les deux langues. Ce phénomène paraphrastique aurait pu être capturé par des techniques opérant au niveau syntaxique telles que SYNT dans le cas où elles apparaissent dans des structures syntaxiques compatibles, ce qui est souvent peu adapté pour les types de corpus étudiés.

Les paraphrases morphologiques, correspondant à des variations dérivationnelles portant sur la formation des mots (*refroidie* ↔

froide), sont absentes du corpus Scène en français, et très peu représentées en anglais (5,4% des paraphrases de la référence). Ceci peut être attribué à la qualité des paires d'énoncés contenues dans ce corpus. Texte contient, cependant, une proportion importante pour ce type de paraphrases en anglais, soit 10,5% des paraphrases identifiées par notre système, *Cambodian* ↔ *Cambodia 's*, *proposal* ↔ *proposed by*.

Il est finalement intéressant de noter que le corpus Événement contient des paraphrases de référence dans toutes les catégories. Ce corpus contient notamment des synonymes, représentant 41,6% des paraphrases identifiées par notre système pour l'anglais et 16,6% pour le français, ainsi que des paraphrases pragmatiques (7,4% de la référence en français) que notre système est incapable d'identifier. Ces paraphrases correspondent en effet à des segments dont la signification ne peut être comprise qu'en connaissant le contexte de leur emploi, obtenu par l'emploi d'anaphores, comme dans China enterprises ↔ *They*. Il n'est pas surprenant que Événement, en particulier, comporte ce genre de phénomènes étant donné le type de connaissances qui est généralement attendu de ses lecteurs (par exemple, *Coopération politique européenne* ↔ *Ministres*).

CONCLUSION DU CHAPITRE

Dans ce chapitre, nous nous sommes intéressée à l'acquisition de paraphrases depuis et entre différents types de corpus, en les définissant sur la base de l'origine du signal du contenu sémantique des paires d'énoncés contenues : un texte dans différentes langues (TEXTE), de la parole transcrite dans une autre langue (PAROLE), une scène visualisée (SCÈNE), et une courte description (un titre d'article) d'un événement (ÉVÉNEMENT).

Pour ce faire, nous avons construit un corpus annoté contenant 2 500 paires d'énoncés en anglais et en français. Nous avons ensuite évalué notre meilleur système de combinaison qui exploite les hypothèses de quatre de nos systèmes d'acquisition, ainsi que l'impact produit par l'utilisation des données d'apprentissage des autres types de corpus.

Nous avons également décrit une typologie détaillée des paraphrases pouvant être acquises (présentes dans la référence) et étant effectivement acquises (automatiquement) pour chaque corpus. Cette typologie regroupe plusieurs phénomènes paraphrastiques allant de la simple *synonymie* aux variations d'ordre *pragmatique*. Cette typologie pourra, dans le futur, servir de guide dans le choix du corpus approprié pour une utilisation particulière de la paraphrase.

Notre résultat le plus prometteur est certainement l'amélioration obtenue sur le type de corpus ÉVÉNEMENT en utilisant les données d'apprentissage de tous les corpus disponibles. Étant donné que les autres corpus sont beaucoup plus rares par nature (souvent de petite taille), ils apportent néanmoins des connaissances utiles pour améliorer la reconnaissance des paraphrases sur ce qui s'est avéré être le type de corpus le plus difficile, mais disponible en grande quantité (ÉVÉNEMENT). Un résultat de cette nature incite à appliquer et améliorer nos techniques pour l'acquisition de paraphrases à une plus large échelle (Pasça et Dienes, 2005; Bhagat et Ravichandran, 2008), comme sur le Web où les paires d'énoncés en relation peuvent être très nombreuses.

Corpus type (C)	Systèmes individuels												Combinaison de systèmes								
	Mot			Terme			Edit$_{\rightarrow F}$			Pivot			union			appr.=C			appr.=tout		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
anglais																					
Texte	48,2	58,9	53,0	63,1	5,9	10,7	41,2	66,4	50,9	73,4	25,8	38,2	20,8	**80,8**	33,1	68,4	62,8	**65,5**	**77,5**	56,1	65,1
Parole	39,7	44,2	41,8	27,1	3,5	6,3	25,0	50,3	33,4	**79,2**	15,3	25,7	25,5	**71,4**	37,6	51,0	56,3	53,5	67,7	48,7	**56,6**
Scène	44,8	57,7	50,5	47,4	5,2	9,5	40,1	67,9	50,4	**84,6**	14,6	25,0	36,2	**83,4**	50,5	44,9	66,8	**53,7**	33,2	59,7	42,7
Évén.	19,0	33,9	24,3	62,9	3,1	6,0	28,8	68,7	40,6	**97,4**	11,2	20,1	20,8	**75,5**	32,7	35,0	67,1	46,0	56,4	56,1	**56,2**
français																					
Texte	52,5	58,9	55,5	56,9	4,9	9,1	46,4	61,4	52,8	64,5	30,3	41,2	41,5	**77,9**	54,1	**74,7**	61,0	**67,1**	**74,5**	60,2	66,6
Parole	44,0	54,9	48,9	30,7	4,3	7,6	34,8	60,2	44,1	**75,5**	19,0	30,4	31,4	**76,2**	44,5	60,2	59,7	**60,0**	55,1	61,0	57,9
(Scène)	14,4	43,6	21,7	53,0	4,0	7,4	13,8	75,3	23,4	**94,6**	5,21	9,8	12,7	**86,4**	22,2	19,9	59,8	**29,8**	12,5	69,4	21,1
Évén.	28,7	44,2	34,8	34,4	2,3	4,3	29,9	58,9	39,7	**79,5**	15,0	25,2	25,2	**72,5**	37,4	40,0	56,3	46,8	62,4	40,7	**49,3**

Tableau 21: Résultats de l'évaluation pour chaque système individuel (à gauche) et les systèmes combinés (à droite) sur tous les types de corpus, pour l'anglais (en haut) et le français (en bas). Les valeurs en gras indiquent les meilleurs résultats pour une mesure donnée pour chaque type de corpus et chaque langue.

Troisième partie

CONCLUSION ET OUVERTURES

7

CONCLUSION DU MANUSCRIT

Nous nous sommes intéressée dans ce travail à l'étude des paraphrases sous-phrastiques, en nous concentrant sur leur acquisition depuis des corpus parallèles monolingues.

Notre point de départ a consisté à montrer que les corpus monolingues parallèles constituent un type de ressources approprié relativement à l'objectif de l'étude de la paraphrase sous-phrastique. Nous avons pour cela réalisé une analyse de différents types de corpus constitués de paires de phrases en relation de différents degrés de parallélisme. Cela a été rendu possible en considérant différents types de signaux d'origine pour le contenu sémantique des paires de phrases, ce qui constitue à notre connaissance une étude originale : des traductions multiples d'une phrase écrite (corpus Texte) ; des traductions multiples de dialogues oraux (corpus Parole) ; des descriptions multiples d'une scène visuelle (corpus Scène) ; et des descriptions multiples d'un même événement ou de deux événements proches (corpus Événement). Cette étude a permis de montrer que le type de corpus contenant les phrases les plus similaires, Texte, se révélait le plus intéressant pour l'observation de la paraphrase sous-phrastique, et ce en dépit de la difficulté à obtenir de telles phrases.

Munie de cette première observation, nous avons ensuite situé la tâche d'acquisition de paraphrases sous-phrastiques depuis des paires de phrases. Une méthodologie d'analyse existante a été décrite, qui consiste à comparer les paraphrases extraites automatiquement à celle d'un ensemble de référence obtenu manuellement. Nous avons initialement attaqué cette tâche en considérant une large gamme de techniques, développées originellement pour des besoins différents et opérant à différents niveaux. Ces techniques ont notamment été choisies pour le type d'algorithmes ou de ressources qu'elles mettent en œuvre, ainsi que leur complémentarité éventuelle. Nous avons ainsi retenu une approche fondée sur l'apprentissage statistique d'alignements entre mots (Mot) ; une approche fondée sur l'expression symbolique de la variation entre termes (Terme) ; une approche fondée sur l'alignement de structures syntaxiques (Synt) ; une approche fondée sur des traductions communes dans une langue pivot (Pivot) ; et enfin une approche fondée sur une mesure de taux d'édition entre séquences de mots (Edit).

Nous avons ensuite mené une évaluation de ces techniques individuelles sur deux langues, l'anglais et le français. L'alignement statistique de mots et l'alignement par calcul de taux d'édition ont obtenu de relativement bonnes performances sur cette tâche d'acquisition, ainsi que l'utilisation de traductions communes dans une moindre mesure ; les techniques d'identification de termes et de fusion d'arbres syntaxiques se sont elles révélées précises, mais trop restrictives pour permettre d'obtenir un bon rappel.

Nous avons ensuite cherché à améliorer les performances obtenues par le biais de différentes combinaisons des résultats produits par ces techniques individuelles, après qu'une mesure de complémentarité entre techniques se fut révélée encourageante. Nous avons mis en œuvre trois approches. La première approche construit une union naïve de l'ensemble des résultats produits par toutes les techniques. Si sa performance est peu intéressante, elle révèle que les techniques individuelles étudiées identifient 8/10 des paraphrases de référence, un chiffre dont nous avons tiré un double enseignement : 1) l'ensemble des techniques étudiées permet d'obtenir des résultats assez forts, signifiant que les connaissances et techniques mises en jeu ont été bien choisies, et 2) l'étude fine des paraphrases non reconnues doit être menée pour poursuivre l'amélioration des performances sur notre tâche d'acquisition.

La seconde approche adapte le fonctionnement de la technique de calcul de taux d'édition par la connaissance des paraphrases identifiées par les autres techniques. Ceci a, dans la majorité des cas, permis d'améliorer les performances en F-mesure des techniques individuelles mises en jeu. Cependant, les gains les plus marqués ont été obtenus lorsque l'adaptation se fait par la technique fondée sur des traductions communes, qui s'avère être assez précise.

Nous avons finalement étudié un cadre plus général, permettant de tirer à la fois profit des hypothèses produites par plusieurs systèmes d'acquisition, et d'intégrer des traits permettant de caractériser les paraphrases sous-phrastiques. Ceci a été formulé comme une tâche de validation de l'union des techniques individuelles par l'utilisation d'un classifieur à maximum d'entropie. Nous avons distingué deux classes : celle représentant des paraphrases sous-phrastiques, dont les exemples sont issus des résultats des techniques individuelles correspondant à des paraphrases de référence *sûres*, et celle représentant des segments n'étant pas des paraphrases, constituée des résultats ne correspondant ni à des paraphrases de référence *sûres* ni à des paraphrases de référence *possibles*. Les exemples ont été décrits par un ensemble de traits, qui outre les techniques ayant proposé l'exemple, incluent des caractéristiques au niveau des segments, des paires des phrases, et des profils de cooccurrences des segments. Nous avons de cette manière obtenu, sur les deux langues, des améliorations significatives des

meilleurs résultats obtenus jusqu'alors. Ce cadre de combinaison peut servir de socle à de futures études, car il permet d'intégrer aisément de nouvelles techniques individuelles ou de nouvelles tentatives de caractérisation des paraphrases.

Ces résultats performants nous ont permis de répondre à une question nouvelle et indispensable à nos yeux pour guider les travaux ultérieurs en acquisition de paraphrases : nous avons établi une typologie bilingue des paraphrases *difficiles à acquérir*, définies comme les paraphrases de référence non trouvées par l'union des techniques considérées. Cette analyse a révélé des résultats relativement comparables sur l'anglais et le français, et a notamment mis en évidence les importantes quantités de paraphrases correspondant aux équivalences lexicales et sous-phrastique qui ne sont pas extraites automatiquement (cf. Tableau 16). Ces paraphrases, qui ont la caractéristique d'être *énumérables*, par opposition à des variations linguistiques sur la morphologie ou la syntaxe, apportent une justification supplémentaire à ce type d'étude et à son application à l'acquisition de telles paraphrases sur de grandes masses de données textuelles.

Une fois ces résultats posés, nous avons considéré un nouvel axe d'analyse, en cherchant à évaluer l'impact de l'origine des paires de phrases sur la tâche d'acquisition. Nous avons repris les différents types de corpus de paires de phrases obtenues à partir de différents signaux étudiés initialement, et avons construit des corpus pour chacun dans nos deux langues, dont les paraphrases sous-phrastiques ont été annotées manuellement.

Outre le caractère fondé sur le signal d'origine des paires de phrases, ces différents corpus se distinguent par différents niveaux de complexité de construction : des traductions multiples d'un même texte sont par exemple peu disponibles et coûteuses à obtenir, mais elles se sont révélées précieuses pour notre étude ; des titres d'articles portant sur le même sujet sont au contraire faciles à collecter en continu et en grandes quantités, et leur étude ouvre la voie à l'acquisition à partir de corpus monolingues plus *comparables* que *parallèles*.

L'évaluation de l'acquisition de paraphrases sur l'ensemble de ces

corpus a mis en évidence une grande disparité de performance, la difficulté de la tâche pouvant être interprétée comme un niveau d'équivalence sémantique entre phrases. Ces expériences plus complètes ont confirmé la supériorité de notre approche de combinaison par classification, qui s'est avérée capable de faire progresser la F-mesure relativement au meilleur système individuel dans toutes les configurations de langue et de type de corpus. Nous avons en outre pu répondre à la question de la nature des paraphrases trouvées par ce système sur chacune des configurations, et nous avons

mis en vis-à-vis les proportions de paraphrases trouvées par type avec les proportions dans les annotations de référence. Cela pourra à nouveau s'avérer utile pour l'amélioration future des systèmes.

Notre dernière contribution a consisté à poser la question de l'impact de l'utilisation de données d'entraînement issues d'autres types de corpus, sans avoir recours à des techniques avancées de sélection d'exemples. Les résultats obtenus ont été contrastés, certaines configuration entraînant une diminution sensible des résultats relativement à la seule utilisation des données d'entraînement pour un type de corpus donné. Ceci met en exergue les différences de nature parfois importantes qui peuvent exister entre deux types de corpus obtenus à partir de signaux d'origine différents. Cependant, certaines configurations se sont trouvées améliorées : en particulier, le type de corpus qui était à la fois le plus difficile jusque-là et le plus facile à collecter, celui appariant des descriptions d'événements liés, s'est vu amélioré de façon significative par l'utilisation de l'ensemble de toutes les données d'entraînement disponibles. Ceci constitue un résultat encourageant pour des travaux futurs : s'agissant du type de corpus le plus facile à collecter, il est raisonnable d'avoir un coût fixe de construction de données annotées provenant de types de corpus plus difficiles à collecter. La nature de la contribution apportée par chaque type de corpus reste cependant un sujet d'étude que nous n'avons pas abordé ici.

Ce travail a ainsi apporté de nombreuses réponses permettant de mieux comprendre la nature des paraphrases sous-phrastiques, appuyées par de nombreux résultats empiriques. Nous pensons qu'une étude de cette ampleur, impliquant deux langues, cinq techniques individuelles, de nombreux traits pour caractériser les paraphrases, et quatre types de corpus, est une contribution utile et inédite au champ de recherche sur la paraphrase, se situant plus en amont qu'une majorité de travaux récents en TAL sur la paraphrase. Néanmoins, nous sommes consciente que de nombreuses questions mériteraient d'être abordées par la suite. Nous pensons en particulier aux questions suivantes :

- Quels seraient les résultats obtenus sur d'autres langues ? Si nous avons pu mener l'ensemble de nos expériences sur l'anglais et le français, langues que nous savons lire, il est intéressant, et pragmatiquement utile, de considérer le cas de langues d'autres familles.
- Quelles seraient les performances obtenues par d'autres techniques et d'autres connaissances *a priori* ? Il nous fut impossible de mener une étude exhaustive, et l'inclusion de certains outils ou ressources aurait été légitime.

– Quels traits caractérisant les paraphrases, indépendamment des techniques individuelles combinées, pourraient être mis en œuvre ? Seraient-ils utiles quel que soit le type de corpus, en particulier s'agissant de corpus davantage *comparables* que *parallèles* ?
– Finalement, la question de l'usage qui peut être fait des paraphrases acquises n'a jusqu'ici pas été abordée directement dans ce travail. Un problème important pour cela serait de caractériser les contextes d'équivalence pour les paires acquises, afin de pouvoir déterminer si une paire collectée depuis un ou des contextes précis peut être utilisée correctement dans un contexte applicatif particulier.

Nos réflexions autour de la notion de paraphrases sous-phrastiques nous ont par ailleurs entraînée vers différentes voies de réflexion et problèmes ayant fait l'objet de travaux en collaboration pendant notre thèse, que nous décrivons brièvement dans le dernier chapitre de ce manuscrit. Nous nous sommes en particulier penchée sur les problèmes suivants, qui constituent des fondations possibles pour la suite de notre étude principale :

1. Comment serait-il possible de faire, à grande échelle et en plusieurs langues, une acquisition *ciblée* de paraphrases, en particulier pour des segments difficiles à acquérir ?
2. Quelles ressources existantes riches en paraphrases auraient été jusqu'alors négligées, et que contiennent-elles ?
3. Comment utiliser utilement une *banque de paraphrases* existante dans une application apportant une aide concrète à un utilisateur ?

8 OUVERTURES

8.1 ATTAQUER LE PROBLÈME DE COUVERTURE : ACQUISITION MANUELLE CIBLÉE PAR LE JEU

Nos expériences ont révélé que certains types de paraphrases étaient plus difficiles à acquérir depuis nos corpus. S'il est de la responsabilité des chercheurs de poursuivre l'amélioration des techniques d'acquisition, certains contextes suggèrent la possibilité d'obtenir de manière *ciblée* des paraphrases pour certains segments. L'acquisition manuelle trouve donc ici un terrain tout à fait légitime. Dans le contexte de la traduction, Resnik *et coll.* (2010) ont par exemple recours au *crowdsourcing* pour obtenir des paraphrases sous-phrastiques correspondant à des segments de phrases difficiles à traduire. Le corpus décrit dans ce travail, que nous avons appelé SCÈNE, a été obtenu par le même biais (Chen et Dolan, 2011). Néanmoins, l'acquisition de données langagières par ce type d'approche peu rémunérée pose de nombreux problèmes, certains moraux, sur lesquels nous renvoyons le lecteur aux arguments de (Sagot *et coll.*, 2011).

Nous avons choisi de considérer la possibilité d'acquérir de telles données, en particulier donc pour des segments appartenant à nos paires de paraphrases difficiles à acquérir, par le jeu. Ce type d'acquisition peut effectivement rencontrer un certain succès, à en juger par exemple par des données acquises dans le cadre du site Web JEUXDEMOTS (Lafourcade et Joubert, 2008), portant sur l'acquisition de mots en relation. Des tentatives précédentes portant sur l'acquisition de paraphrases ont été publiées (Chklovski, 2005; España Bonet *et coll.*, 2009), sans qu'elle n'a mené à des collectes importantes de données. Nous proposons un cadre permettant une acquisition massive et multilingue de paraphrases, où l'acquisition des paraphrases est couplée à l'obtention d'une mesure de qualité.

Le jeu Web que nous avons conçu et développé [1] permet à un concepteur de surligner à la main dans une page Web des segments sur lesquels portera une partie particulière [2]. Les joueurs

1. Nous sommes reconnaissante ici pour le travail d'implémentation effectué par Léton Attanon, dans le cadre d'un stage qu'il a effectué au LIMSI.

2. Le choix finalement retenu d'utiliser des pages Web avec leur rendu d'affichage réel est de permettre à terme l'intégration de ce type de jeu dans les pratiques de lecture.

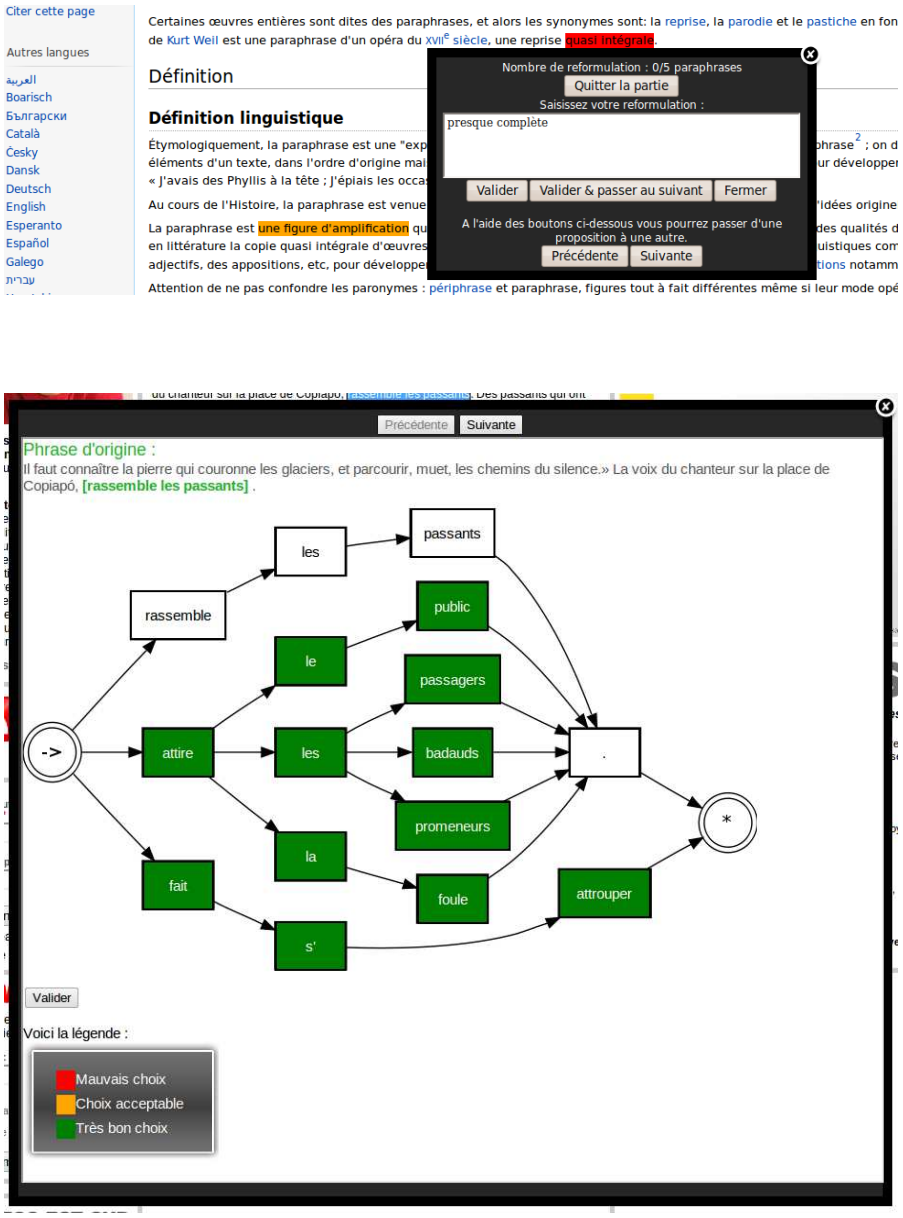


FIGURE 27: Interface de notre application de jeu sur le Web pour l'acquisition et l'évaluation de paraphrases sous-phrastiques.

cherchent ensuite des parties (par exemple, par langue, par thème, etc.) et peuvent participer à deux phases distinctes, pour des parties différentes. Dans la première, le joueur propose, pour chaque segment d'une partie, une reformulation. Celle-ci doit permettre de conserver la grammaticalité du texte modifié, et sera évaluée selon deux critères décrits ci-après. Cette première phase est illustrée dans la partie supérieure de la figure 27. La deuxième phase propose une vue compacte de l'ensemble des paraphrases proposées par des joueurs, telle qu'illustrée sur la partie inférieure de la figure 27. Pour minimiser le travail d'un évaluateur, les mots appa-

raissent initialement en vert, signifiant qu'ils participent à des paraphrases *correctes* : un clic les transforme en choix *acceptables* (sur fond orange), et deux en choix *incorrects* (sur fond rouge).

Le score d'une partie est déterminé d'une part par l'évaluation moyenne que reçoit la proposition d'un joueur, mais également par la *rareté* de cette proposition relativement à l'ensemble des propositions de tous les joueurs. Du point de vue de l'acquisition de données, cela permet donc de collecter préférentiellement des données variées, qui sont en outre associées à des scores d'évaluation obtenus auprès de nombreux autres joueurs.

Dans le but d'étudier la faisabilité de l'acquisition manuelle de paraphrases, nous avons mené une expérience à petite échelle visant à collecter des paraphrases auprès de 22 utilisateurs natifs du français pour une vingtaine de segments participant à des paraphrases difficiles à acquérir dans nos corpus. La première moitié des contributeurs ont proposé des reformulations dans le contexte d'une phrase du corpus, tandis que la deuxième moitié ont reformulé le segment aligné dans la référence dans le contexte de la deuxième phrase.

La figure 28 illustre un exemple de paraphrases obtenues. La majorité les paraphrases obtenues correspondent à des variations lexicales ou sous-phrastiques (50 paraphrases) et à des paraphrases syntaxiques (40 paraphrases). Des paraphrases qualifiées précédemment de "pragmatiques" sont également obtenues (4 paraphrases).

phrase #1	*La Commission est en étroite relation avec les autres Etats membres , et les presse de se conformer à cette demande* ***dès que possible*** *.*
paraphrases	*ASAP*, *au plus tôt* (x3), *au plus vite*, *aussi tôt qu'ils le pourront*, *dans les plus brefs délais*, *immédiatement*, *le plus tôt possible*, *rapidement* (x2)
phrase #2	*La Commission est en relation étroite avec les autres États membres , et les incite à se conformer au règlement* ***dans les meilleurs délais*** *.*
paraphrases	*à brève échéance*, *au plus tôt*, *au plus vite* (x3), *aussi rapidement que possible* (x2), *dans les plus brefs délais*, *le plus rapidement possible* (x2), *prestement*

FIGURE 28: Exemple d'une acquisition manuelle pour la paire de paraphrases *dès que possible* ↔ *dans les meilleurs délais* à partir d'une paire d'énoncés. Les paraphrases obtenues dans les deux directions sont soulignées.

Ces résultats préliminaires sont encourageants, et appuient selon nous l'intérêt de notre approche : celle-ci met à contribution des personnes volontaires souhaitant se distraire, et permet en théorie de produire de grandes quantités de données paraphrastiques pour un grand nombre de langues. Cependant, nous n'avons pas cru opportun, jusqu'à ce jour, de publier le jeu en ligne. Le mode de jeu actuel, qui requiert une inscription et l'utilisation d'un site dédié, est probablement trop lourd pour être largement adopté. Nous avons depuis redéfini le jeu pour qu'il soit intégré à de grands réseaux sociaux tels que Facebook, afin de pouvoir bénéficier de la souplesse d'intégration et du mécanisme d'invitation à prendre part à des parties, formulables ici sous forme de "défis". En cas de succès, nous comptons publier régulièrement l'intégralité des données collectées sur l'ensemble des langues pour la communauté de la recherche sur la paraphrase.

8.2 Utiliser des sources de paraphrases sous-exploitées : étude des révisions locales dans les traces d'édition de Wikipédia

Notre étude principale a porté sur l'acquisition de paraphrases depuis des paires d'énoncés. Nous avons notamment remarqué que, contrairement aux activités de traduction qui génèrent chaque jour de nombreuses traductions, il n'existe pas d'activités humaines produisant des données facilement exploitables pour l'acquisition de paraphrases[3]. Nous avons porté notre attention sur le cas des traces d'édition : il est vraisemblable que certaines correspondent à des reformulations sans changement de sens. Une étude approfondie[4] a été menée sur les traces d'édition de l'encyclopédie collaborative Wikipédia en français, pour déterminer la nature des reformulations de types paraphrastiques qui s'y trouvent. Cette étude s'est fondée sur le corpus WiCoPaCo (Max et Wisniewski, 2010), qui regroupe 408 816 descriptions de réécritures locales (portant sur au plus 7 mots) dans l'encyclopédie francophone. Il est important de noter qu'il s'agit là de reformulations particulièrement intéressantes, puisqu'elles découlent d'une activité *naturelle*[5]. Un exemple d'entrée du corpus, correspondant à une simplification lexicale, est donné dans la figure 29.

Une typologie a été construite (Dutrey *et coll.*, 2011a) pour décrire les différents phénomènes observables dans WiCoPaCo. Celle-ci se compose de deux grandes catégories : 1) *faibles variations séman-*

3. Bien entendu, les activités de description multiples et indépendantes, telles que représentées dans notre corpus Événement, peuvent entrer dans cette catégorie.

4. Réalisée par Camille Dutrey lors d'un stage au LIMSI, auquel j'ai participé.

5. Il est évidemment regrettable de constater que le vandalisme est une activité "naturelle" pour certains "contributeurs".

```
<modif id="407851" wp_page_id="1830844" wp_before_rev_id="20691183" wp_after_rev_id="20691225"
wp_user_id="287861" wp_user_num_modif="81" wp_comment="">
<before>Le genre Archaeopteris possède plus de caractéristiques communes avec les plantes à graines que toute autre
<m num_words="1">ptéridophyte</m> connue et les analyses cladistiques récentes le placent en groupe-frère des
plantes à graines .</before>
<after>Le genre Archaeopteris possède plus de caractéristiques communes avec les plantes à graines que toute autre
<m num_words="2">plante fossile</m> connue et les analyses cladistiques récentes le placent en groupe-frère des
plantes à graines .</after>
</modif>
```

FIGURE 29: Exemple d'entrée (au format XML) du corpus WICOPACO.

tiques et 2) *corrections factuelles et vandalismes*[6]. La première classe, qui nous intéresse plus précisément ici, se divise en des *corrections* et des *reformulations*. Ces dernières correspondent à des modifications lexicales ou sous-phrastiques apportées sans intention de modifier significativement le sens. Les principales sous-classes révélées par l'étude sont les reformulations lexicales (ex. ***[L'implémentation → La mise en œuvre]*** *de l'algorithme...*), syntaxiques (ex. *Un infomercial pseudo-scientifique* ***[en exposant → qui expose]*** *grossièrement...*) et sémantiques (*Il fonde le* ***[journal → quotidien]*** *francophone "Le Tunisien" en 1907.*).

La typologie établie a servi de référence pour une expérience d'annotation réalisée par 4 annotateurs sur 200 entrées du corpus, sélectionnées pour avoir une distance d'édition (Levenshtein) entre les segments avant et après réécriture supérieure ou égale à 4. L'accord inter-annotateur mesuré sur la classe des reformulations s'est avéré être modéré (la seule classe obtenant un accord fort étant sans surprise celles des corrections factuelles et du vandalisme). Le tableau 22 indique le nombre d'annotations identiques attribuées par 1 à 4 annotateurs, ce qui permet de quantifier approximativement les phénomènes présents dans le corpus. Les reformulations apparaissent comme les plus représentées, avec 30% de phrases concernées (60 occurrences reconnues par les 4 annotateurs sur les 200 phrases). L'identification à grande échelle de paraphrases depuis cette ressource s'avère donc très prometteuse.

	4 ann.	3 ann.	2 ann.	unique ann.	Total
Correction de surface	9	2	7	23	41
Reformulation	60	33	24	15	132
Corrections factuelles et vandalismes	47	15	13	32	107
Défaut d'alignement	2	4	8	6	20

Tableau 22: Nombre d'annotations identiques attribué par nombre d'annotateurs sur un corpus de 200 phrases tirées du corpus WICOPACO.

6. La typologie complète est disponible sous forme de rapport technique (Dutrey *et coll.*, 2011b).

8.3 UTILISER DES PARAPHRASES : APPLICATION À L'ASSISTANCE À LA RÉDACTION

Nos travaux ont décrit des méthodes et corpus permettant d'acquérir des paraphrases sous-phrastiques. Il est intéressant de considérer l'usage concret qui peut ensuite en être fait, dans des contextes par nature différents des contextes d'acquisition. Nous avons dans ce cadre considéré la tâche d'assistance à un rédacteur souhaitant réécrire un segment de texte particulier. Comme corpus d'entraînement et d'évaluation, nous avons naturellement choisi d'utiliser à nouveau le corpus WiCoPaCo, qui offre comme nous l'avons montré de nombreux exemples réels de reformulations à faible variation sémantique. Nous avons comparé différentes réécritures supposées appartenir à une *banque de paraphrases* existante, incluant notamment la réécriture d'origine extraite de WiCoPaCo. Afin d'obtenir d'autres réécritures candidates de différentes qualités, nous avons utilisé deux autres méthodes d'acquisition : *a*) des traductions automatiques par pivot (par l'espagnol, langue proche du français, et par le chinois, langue éloignée), et *b*) des paraphrases obtenues par notre jeu décrit dans la section 8.1. Un exemple de contexte de réécriture et de 5 propositions de réécriture dans notre interface d'annotation d'exemples est présenté dans la figure 30.

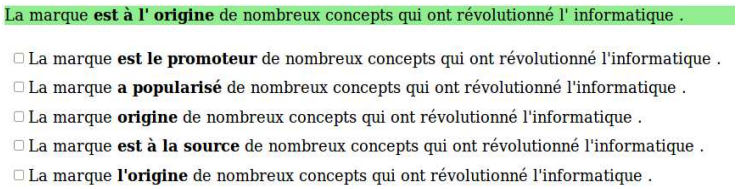


FIGURE 30: Exemple d'interface d'annotation pour une phrase d'origine (sur fond vert) et ses 5 paraphrases candidates (présentées dans un ordre aléatoire). Le segment en gras dans la phrase d'origine, *est à l'origine*, est ici paraphrasé par *est le promoteur* , *a popularisé* , *origine* , *est à la source* et *l'origine*.

Nous avons cherché à répondre à la question de l'acceptabilité d'une réécriture particulière dans chaque contexte d'évaluation. Nous avons abordé ce problème comme une tâche de classification binaire, de façon analogue aux travaux de (Brockett et Dolan, 2005), exploitant des modèles calculés à partir d'informations du Web. Le recours au Web semble ici indispensable : seule une telle échelle nous permet d'accéder à des exemples en nombres suffisants pour certains segments. En outre, il a été montré qu'un certain nombre d'applications de Traitement Automatique des Langues peuvent

être améliorées grâce à l'exploitation de fréquences de n-grammes sur le Web (Lapata et Keller, 2005)[7].

Nous avons extrait aléatoirement 150 exemples de réécritures à faible variation sémantique de WiCoPaCo. Un sous-ensemble de 100 phrases a été utilisé comme corpus d'apprentissage, les 50 phrases restant ayant servi pour l'évaluation. L'annotation des exemples en contexte, à l'aide de l'interface illustrée dans la figure 30, a été réalisée par 4 annotateurs, et a donné un taux d'accord inter-annotateur relativement fort de 0,65. Nous avons distingués trois sous-cas pour l'évaluation : 1) seuls les exemples annotés comme "paraphrases" par au moins l'un des juges sont utilisés (**Possibles**) ; 2) les exemples que les deux juges n'ont pas annotés comme "paraphrases" ou "non paraphrases" ne sont pas retenus (**Sûres**) ; et 3) seuls les exemples pour lesquels les deux juges proposent la même annotation sont retenus (**Sûres++**)[8].

Nous avons fait l'apprentissage d'un classifieur de type séparateur à vastes marges (SVM) (Cortes et Vapnik, 1995)[9] exploitant les traits suivants :

- **Score de modèle de langue** : probabilité donnée par un modèle de langue appris sur des données du Web[10].
- **Score de similarité thématique hors contexte** : similarité (calcul de cosinus) entre profils de mots cooccurrents dans un ensemble de documents sur le Web.
- **Score d'un modèle thématique contextuel** : similarité (calcul de cosinus) entre profils de mots cooccurrents dans la phrase d'origine.
- **Similarité des segments** : coût TER (Snover *et coll.*, 2010) de transformation entre les lemmes du segment d'origine et une paraphrase.

Outre notre classifieur, nous avons évalué les trois approches simples suivantes :

- ML_Web : considère une phrase comme paraphrase d'une phrase d'origine si son score de modèle de langue issu du Web est plus élevé que celui de la phrase d'origine.
- ML_Frontières : considère qu'une phrase est paraphrase d'une phrase d'origine si la fréquence sur le Web des bigrammes traversant les frontières gauche et droite après substitution est supérieure à 10.

7. Nous avons dans nos expériences utilisé le service Web Yahoo ! Search BOSS pour obtenir le nombre de documents du Web indexés contenant une expression littérale, alors disponible à :
http://developer.yahoo.com/search/boss

8. Il faut noter que ces différentes conditions correspondent de fait à des nombres d'exemples différents.

9. Nous avons utilisé l'implémentation libsvm (Chang et Lin, 2011).

10. Nous avons pour cela utilisé le Service Web N-gram de Microsoft (Wang *et coll.*, 2010).

– DepCont : considère qu'un segment est paraphrase d'un autre segment en contexte si les sous-ensembles des dépendances syntaxiques [11] entre les deux segments et leur contexte sont les mêmes.

Les résultats obtenus par les différentes techniques comparées et les différentes configurations sont donnés dans le tableau 23. L'observation la plus marquante est que cette tâche de classification de paraphrases s'avère sans surprise être difficile : la meilleure performance obtenue par l'un des systèmes est de 70,69 pour la condition Sûres. En outre, il existe une variation importante entre les différentes conditions testées avec un résultat faible pour notre classifieur de 57,67 dans la condition Possibles (cas de désaccord entre annotateurs, où un seul reconnaît le statut de paraphrase). D'une manière plus générale, la technique ML_Web et notre classifieur sont plus performants que les autres techniques de référence. ML_Frontières et DepCont ne modélisent que des contraintes grammaticales locales, ce qui fait qu'il n'est pas surprenant que ces informations ne permettent pas la reconnaissance de variations sémantiques licites entre paraphrases candidates. WebLM, qui se limite à la comparaison de scores de modèles de langue dérivés du Web, apparaît donc comme une technique relativement compétitive [12], mais sa performance est peu élevée (56,79) pour la condition Sure++. Puisque cette condition ne prend en compte que les annotations consensuelles pour l'apprentissage et l'évaluation, nous considérons cette condition comme la plus utile pour l'interprétation des résultats de ces travaux préliminaires. Ici, notre système obtient la meilleure performance, avec un avantage de 6,06 points par rapport à WebLM. Ceci montre que la seule utilisation d'un modèle de langue, aussi bien estimé soit-il, est trop limitée pour rendre compte correctement de l'ensemble des phénomènes de paraphrases présents dans notre corpus d'évaluation, ce qui confirme des résultats précédents où les modèles de langue n'étaient pas issus de comptes du Web (Bannard et Callison-Burch, 2005).

Finalement, le tableau 24 détaille les performances obtenues par chacune des méthodes d'acquisition de paraphrases pour chacune des 3 conditions. Il n'est tout d'abord pas surprenant que les reformulations extraites de WiCoPaCo soient largement identifiées comme de bonnes paraphrases en contexte, en particulier dans les conditions Possibles et Sûres++. Ces paraphrases sont le résultat de reformulations par des contributeurs de Wikipédia dans le contexte d'évaluation, et avaient déjà été reconnues comme telles par

11. Nous avons utilisé la version française (Candito *et coll.*, 2010) de l'analyseur probabiliste de Berkeley (Petrov et Klein, 2007).

12. Une explication peut résider dans le fait que nos méthodes d'acquisition de paraphrases utilisant Google Translate comme un traducteur automatique par pivot ont tendance à produire des segments ayant une forte valeur de probabilité dans le modèle de langue utilisé, qui est certainement assez comparable à celui utilisé dans nos expériences.

	ML_Web	LM_Frontières	DepCont	Classifieur
Possibles	**62,79**	54,88	48,53	57,67
Sûres	68,37	36,27	51,90	**70,69**
Sûres++	56,79	51,41	42,69	**62,85**

Tableau 23: Résultats de la performance de la classification (*accuracy*) pour les 3 techniques de référence et notre classifieur sur le corpus d'évaluation et les 3 conditions. Il convient de noter que la condition Sûres++ n'est pas directement comparable aux autres conditions puisque les tailles des corpus d'apprentissage et d'évaluation sont différentes de celles des deux autres conditions.

une première annotatrice. Les paraphrases obtenues par collecte manuelle sur des contextes issus du Web, donc d'un contexte possiblement différent de celui de l'évaluation, obtiennent une performance relativement acceptable. Les résultats confirment cependant le fait attendu que la substituabilité des paraphrases dépend fortement du contexte. Par exemple, la substitution du segment *de l'éditeur* par *publiée par les éditions* dans le contexte de la phrase "*Neopolis est une collection de bandes dessinées* ***de l'éditeur*** *Delcourt* permet de conserver le sens d'origine ainsi que la grammaticalité de la phrase. *A contrario*, la substitution par le segment *du logiciel* n'est pas adaptée à ce contexte. Sans surprise, les paraphrases obtenues automatiquement par traduction par pivot ne sont pas de bonne qualité, l'utilisation d'un pivot plus proche du français obtenant cependant de meilleurs résultats.

	WiCoPaCo	Humain	Pivot_{es}	Pivot_{zh}
Possibles	**89,33**	67,00	47,33	20,66
Sûres	**64,00**	44,50	31,33	10,66
Sûres++	**86,03**	57,34	37,71	12,60

Tableau 24: Performance (valeurs d'*accuracy*) de nos différentes méthodes d'acquisition pour nos trois conditions d'évaluation.

Ces expériences préliminaires ont montré que nous pouvions construire un classifieur obtenant de meilleurs résultats que plusieurs systèmes de référence simples lorsque l'on ne considère que les paraphrases obtenant des jugements consensuels dans la référence utilisée. Bien que ces premières expériences soient positives, leurs résultats pourraient être améliorés sur différents as-

pects. Tout d'abord, il est possible d'élargir l'exploration des différentes caractéristiques que nous mettons en œuvre dans le classifieur. Intégrer d'autres traits, dont des modèles mettant en jeu des dépendances syntaxiques calculées sur des données du Web, pourrait améliorer nettement les résultats. Une autre piste consiste à analyser plus finement nos résultats afin d'identifier les cas problématiques, dont certains ne peuvent pas être modélisés sans avoir recours à des connaissances sur le monde, ce qui suggérera notamment l'intégration de connaissances du domaine, éventuellement dérivées de méta-informations provenant des articles Wikipédia concernés. L'ensemble de ces expériences pourra être conduit en plusieurs langues, les données utilisées et les méthodes employées pouvant facilement être transposées. Finalement, il est intéressant de considérer l'approche décrite ici comme un cadre pour l'évaluation des systèmes d'acquisition de paraphrases.

BIBLIOGRAPHIE

Allauzen, A. et Yvon, F. (2011). Méthodes statistiques pour la traduction automatique. *Dans* Gaussier, E. et Yvon, F., éditeurs : *Mod'eles statistiques pour l'acc'es à l'information textuelle*, chapitre 7, pages 271–356. Herm'es, Paris. (Cité à la page 73.)

Amghar, T. (1996). *Une contribution à la modélisation informatique du phénomène de paraphrase : le jugement de paraphrase et la métonymie*. Thèse de doctorat, École Centrale de Nantes. (Cité à la page 13.)

Androutsopoulos, I. et Malakasiotis, P. (2010). A Survey of Paraphrasing and Textual Entailment Methods. *Journal of Artificial Intelligence Research*, 38:135–187. (Cité aux pages 31 and 45.)

Baldwin, T., Bond, F. et Ogura, K. (2001). Dictionary-driven analysis of japanese verbal alternations. *Dans Proceedings of the Seventh Annual Meeting of the Association of Natural Language Processing*, pages 281–284, Tokyo, Japon. (Cité à la page 33.)

Bannard, C. et Callison-Burch, C. (2005). Paraphrasing with Bilingual Parallel Corpora. *Dans Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 597–604, Ann Arbor, États-Unis. (Cité aux pages 33, 37, 38, 41, 79, and 142.)

Barr, J. et Cabrera, L. (2006). AI gets a brain. *Queue*, 4(4):24–29. (Cité à la page 30.)

Barzilay, R. (2003). *Information fusion for multidocument summarization : paraphrasing and generation*. Thèse de doctorat, Columbia University. (Cité aux pages 18 and 43.)

Barzilay, R. et Lee, L. (2003). Learning to paraphrase : an unsupervised approach using multiple-sequence alignment. *Dans Proceedings of NAACL-HLT*, pages 16–23, Edmonton, Canada. (Cité aux pages 28, 32, 36, and 41.)

Barzilay, R. et McKeown, K. R. (2001). Extracting Paraphrases from a Parallel Corpus. *Dans Proceedings of 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse, France. (Cité aux pages 15, 28, 29, 32, 34, 39, 40, 41, 46, and 109.)

Berger, A. L., Pietra, V. J. D. et Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71. (Cité à la page 96.)

Bernard, G. (2011). *Réordonnancement de candidats réponses pour un système de questions-réponses*. Thèse de doctorat, Université Paris Sud. (Cité à la page 45.)

Bernhard, D. et Gurevych, I. (2008). Answering Learners' Questions by Retrieving Question Paraphrases from Social Q&A Sites. *Dans Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 44–52, Columbus, États-Unis. (Cité aux pages 29, 32, and 34.)

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D. et Panovich, K. (2010). Soylent : a word processor with a crowd inside. *Dans Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 313–322. (Cité à la page 30.)

Bhagat, R. (2009). *Learning paraphrases from text*. Thèse de doctorat, University of Southern California. (Cité à la page 16.)

Bhagat, R. et Ravichandran, D. (2008). Large Scale Acquisition of Paraphrases for Learning Surface Patterns. *Dans Proceedings of ACL-08 : HLT*, pages 674–682, Columbus, États-Unis. (Cité aux pages 34, 41, and 125.)

Bolshakov, I. A. et Gelbukh, E. (2004). Synonymous paraphrasing using wordnet and internet. *Dans 9th International Conference on Applications of Natural Language to Information Systems, NLDB 2004*, pages 312–323, Salford, Royaume-Uni. (Cité à la page 31.)

Bosma, W. et Callison-Burch, C. (2006). Paraphrase substitution for recognizing textual entailment. *Dans Proceedings of the Cross-Language Evaluation Forum*, pages 502–509, Alicante, Espagne. (Cité à la page 16.)

Bouamor, H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multilingue multisource. *Dans Récital-TALN*, Montréal, Canada. (Cité à la page 51.)

Bourdaillet, J. et Ganascia, J.-G. (2007). Machine Assisted Study of Writers' Rewriting Processes. *Dans Proceedings of the International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2007)*, Madeire, Portugal. (Cité aux pages 35 and 41.)

Bourigault, D. et Jacquemin, C. (1999). Term extraction + Term Clustering : An integrated Platform for Computer-Aided Terminology. *Dans Proceedings of the 9th conference on European chapter of the Association for Computational Linguistics*, pages 15–22, Bergen, Norv'ege. (Cité à la page 74.)

Brockett, C. et Dolan, W. B. (2005). Support vector machines for paraphrase identification and corpus construction. *Dans Proceedings of The 3rd International Workshop on Paraphrasing IWP*, pages 1–8, Jeju Island, Corée du Sud. (Cité à la page 140.)

Brown, P., Pietra, V., Pietra, S. et Mercer, R. (1993). The mathematics of statistical machine translation : Parameter estimation. *Computational linguistics*, 19(2):263–311. (Cité à la page 73.)

Burrows, S., Potthast, M. et Stein, B. (2012). Paraphrase Acquisition via Crowdsourcing and Machine Learning. *Transactions on Intelligent Systems and Technology*, 5:22. (Cité aux pages 30 and 44.)

Callison-Burch, C. (2007). *Paraphrasing and Translation*. Thèse de doctorat, University of Édimbourg. (Cité à la page 16.)

Callison-Burch, C. (2008). Syntactic constraints on paraphrases extracted from parallel corpora. *Dans Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 196–205, Hawai, États-Unis. (Cité aux pages 38 and 41.)

Callison-Burch, C., Cohn, T. et Lapata, M. (2008). ParaMetric : An Automatic Evaluation Metric for Paraphrasing. *Dans Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 97–104, Manchester, Royaume-Uni. (Cité aux pages 42 and 68.)

Callison-Burch, C., Koehn, P. et Osborne, M. (2006). Improved statistical machine translation using paraphrases. *Dans Proceedings of the Human Language Technology Conference of the NAACL*, pages 17–24, New York, États-Unis. (Cité à la page 44.)

Candito, M., Crabbé, B. et Denis, P. (2010). Statistical French dependency parsing : treebank conversion and first results. *Dans Proceedings of LREC*, Valletta, Malte. (Cité aux pages 77 and 142.)

Castellví, M., Bagot, R. et Palatresi, J. (2001). Automatic term detection : A review of current systems. *Bourigault, D., Jacquemin, C., L'Homme, M.-C.(Eds.), Recent Advances in Computational Terminology. John Benjamins*, pages 53–88. (Cité à la page 74.)

Chang, C.-C. et Lin, C.-J. (2011). LIBSVM : A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27 :1–27 :27. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`. (Cité à la page 141.)

Chen, D. et Dolan, W. (2011). Collecting highly parallel data for paraphrase evaluation. *Dans Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, pages 190–200, Portland, USA. (Cité aux pages 30, 42, 55, 109, 111, and 135.)

Chevelu, J. (2011). *Production de paraphrases pour les systémes vocaux humain-machine*. Thèse de doctorat, Université de Caen. (Cité aux pages 43 and 44.)

Chklovski, T. (2005). Collecting paraphrase corpora from volunteer contributors. *Dans Proceedings of the 3rd international conference on Knowledge capture*, pages 115–120, Banff, Canada. (Cité aux pages 30 and 135.)

Chomsky, N. (1957). *Syntactic structures*. The Hague, Mouton, S.S. (Cité à la page 12.)

Cohn, T., Callison-Burch, C. et Lapata, M. (2008). Constructing corpora for development and evaluation of paraphrase systems. *Computational Lingustics*, 34(4):597–614. (Cité aux pages 19, 29, 39, 41, 57, 63, 65, 67, and 109.)

Cortes, C. et Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297. (Cité à la page 141.)

Coster, W. et Kauchak, D. (2011). Learning to simplify sentences using wikipedia. *Dans Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 1–9, Portland, États-Unis. (Cité à la page 44.)

Cristea, T. (2001). *Structures signifiantes et relations sémantiques*. Bucureşti, Editura FundaŃiei România de Mâine. (Cité aux pages 9, 12, 22, and 23.)

Culicover, P. (1968). Paraphrase generation and information retrieval from stored text. *Mechanical Translation and Computational Linguistics*, 11(1-2):78–88. (Cité à la page 19.)

Culioli, A. (1976). Transcription du seminaire de DEA :"recherche en linguistique : Théorie des opérations enonciatives". *Paris : Université de Paris VII*. (Cité à la page 14.)

Dagan, I., Glickman, O. et Magnini, B. (2006). The Pascal Recognising Textual Entailment Challenge. *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190. (Cité à la page 45.)

Deléger, L. et Zweigenbaum, P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. *Dans Proceedings of the 2nd Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, pages 2–10, Suntec, Singapour. (Cité aux pages 37, 41, and 44.)

Denkowski, M., Al-Haj, H. et Lavie, A. (2010). Turker-Assisted Paraphrasing for English-Arabic Machine Translation. *Dans Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 66–70, Los Angeles, États-Unis. (Cité à la page 30.)

Dolan, B., Quirk, C. et Brockett, C. (2004). Unsupervised construction of large paraphrase corpora : Exploiting massively parallel news sources. *Dans Proceedings of Coling 2004*, pages 350–356, Geneva, Suisse. (Cité aux pages 28, 32, 34, and 109.)

Dorr, B., Green, R., Levin, L., Rambow, O., Farwell, D., Habash, N., Helmreich, S., Hovy, E., Miller, K., Mitamura, T., Reeder, F. et Siddharthan, A. (2004). Semantic annotation and lexico-syntactic paraphrase. *Dans Actes de the 4th International Conference on Language Resources and Evaluation (LREC) Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbonne, Portugal. (Cité à la page 18.)

Dras, M. (1999). *Tree adjoining grammar and the reluctant paraphrasing of text*. Thèse de doctorat, Macquarie University, Sydney, Australie. (Cité aux pages 16, 17, 19, 33, and 41.)

Duboue, P. et Chu-Carroll, J. (2006). Answering the question you wish they had asked : The impact of paraphrasing for question answering. *Dans Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume : Short Papers*, pages 33–36, New York, États-Unis. (Cité à la page 44.)

Duclaye, F. (2003). *Apprentissage automatique de relations d'équivalence sémantique à partir du Web*. Thèse de doctorat, Télécom ParisTech. (Cité aux pages 9, 17, 43, and 44.)

Dutrey, C., Bouamor, H., Bernhard, D. et Max, A. (2011a). Paraphrases et modifications locales dans l'historique des révisions de wikipédia. *Dans Actes de TALN*, Montpellier, France. (Cité aux pages 30, 32, 35, 41, 75, and 138.)

Dutrey, C., Bouamor, H., Bernhard, D. et Max, A. (2011b). Typologie des modifications dans les révisions de Wikipédia. Notes et documents du LIMSI 2011-01, LIMSI-CNRS. (Cité à la page 139.)

Elhadad, N. et Sutaria, K. (2007). Mining a lexicon of technical terms and lay equivalents. *Dans Proceedings of the Workshop on BioNLP 2007 : Biological, Translational, and Clinical Language Processing*, Prague, République tch'eque. (Cité à la page 37.)

España Bonet, C., Vila, M., Marti, M. et Rodriguez, H. (2009). CoCo, a web interface for corpora compilation. *Procesamiento del lenguaje natural*, 43:367–368. (Cité à la page 135.)

Fuchs, C. (1982). *La paraphrase*. Presses Universitaires de France. (Cité aux pages 9, 14, 17, and 23.)

Fuchs, C. (1994). *Paraphrase et énonciation*. Ophrys, Paris. (Cité aux pages 2, 11, 12, and 14.)

Fujita, A. (2005). *Automatic Generation of Syntactically Well-formed and Semantically Appropriate Paraphrases*. Thèse de doctorat, Nara Institute of Science and Technology, Japon. (Cité aux pages 15, 19, and 20.)

Fujita, A. (2010). Typology of paraphrases and approaches to compute them. *Dans Workshop on Corpus-Based Approaches to Paraphrasing and Nominalization : Invited Talk*, Barcelone, Espagne. (Cité à la page 20.)

Fujita, A., Furihata, K., Inui, K., Matsumoto, Y. et Takeuchi, K. (2004). Paraphrasing of Japanese Light-verb Constructions Based on Lexical Conceptual Structure. *Dans Second ACL Workshop on Multiword Expressions : Integrating Processing*, pages 9–16, Barcelone, Espagne. (Cité à la page 33.)

Fujita, A., Inui, K. et Matsumoto, Y. (2005). Exploiting lexical conceptual structure for paraphrase generation. *Dans* Dale, R.,

Wong, K.-F., Su, J. et Kwong, O., éditeurs : *Natural Language Processing IJCNLP 2005*, volume 3651 de *Lecture Notes in Computer Science*. Springer Berlin / Heidelberg. (Cité aux pages 33 and 41.)

Ganitkevitch, J., Callison-Burch, C., Napoles, C. et Van Durme, B. (2011). Learning sentential paraphrases from bilingual parallel corpora for text-to-text generation. *Dans Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, Édimbourg, Royaume-Uni. (Cité à la page 28.)

Germann, U. (2008). Yawat : Yet Another Word Alignment Tool. *Dans Proceedings of the ACL-08 : HLT Demo Session*, pages 20–23, Columbus, États-Unis. (Cité à la page 57.)

Harris, Z. (1954). Distributional structure. *Word*. (Cité à la page 34.)

Harris, Z. (1957). Co-occurrence and transformation in linguistic structure. *Language*, 33(3):283–340. (Cité aux pages 9 and 12.)

Hassan, S., Csomai, A., Banea, C., Sinha, R. et Mihalcea, R. (2007). UNT : Subfinder : Combining knowledge sources for automatic lexical substitution. *Dans Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, République tch'eque. (Cité aux pages 31 and 41.)

Hirao, T., Fukusima, T., Okumura, M., Nobata, C. et Nanba, H. (2004). Corpus and evaluation measures for multiple document summarization with multiple sources. *Dans Proceedings of the 20th international conference on Computational Linguistics*, Genève, Suisse. (Cité à la page 43.)

Honeck, R. (1971). A study of paraphrases. *Journal of Verbal Learning and Verbal Behavior*, 10(4):367–381. (Cité à la page 14.)

Ibrahim, A., Katz, B. et Lin, J. (2003). Extracting Structural Paraphrases from Aligned Monolingual Corpora. *Dans Proceedings of the Second International Workshop on Paraphrasing*, pages 57–64, Sapporo, Japon. (Cité aux pages 39 and 41.)

Jacquemin, C. (1999). Syntagmatic and paradigmatic representations of term variation. *Dans Proceedings of ACL*, College Park, États-Unis. (Cité aux pages 33, 41, and 74.)

Jakobson, R. (1963). Essais de linguistique générale (1960). (Cité à la page 9.)

Jing, H. (1998). Usage of WordNet in natural language generation. *Dans Proceedings of The 17th International Conference on Computational Linguistics (COLING '98)*, pages 128–134, Montréal, Canada. (Cité à la page 31.)

Katz, J. et Fodor, J. (1963). The structure of a semantic theory. *Language*, 39(2):170–210. (Cité à la page 9.)

Kauchak, D. et Barzilay, R. (2006). Paraphrasing for Automatic Evaluation. *Dans Proceedings of the Human Language Technology Conference of the NAACL*, pages 455–462, New York, États-Unis. (Cité aux pages 31 and 44.)

Knight, K. et Marcu, D. (2000). Statistics-based summarization-step one : Sentence compression. *Dans Proceedings of the National Conference on Artificial Intelligence*. (Cité à la page 43.)

Koehn, P. (2005). Europarl : A parallel corpus for statistical machine translation. *Dans Proccedings of The Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailande. (Cité aux pages 51 and 80.)

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. et Herbst, E. (2007). Moses : Open Source Toolkit for Statistical Machine Translation. *Dans Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, République tch'eque. (Cité aux pages 72 and 80.)

Koehn, P., Och, F. J. et Marcu, D. (2003). Statistical Phrase-Based Translation. *Dans Proceedings of NAACL-HLT*, pages 48–54, Edmonton, Canada. (Cité à la page 74.)

Kozlowski, R., McCoy, K. et Vijay-Shanker, K. (2003). Generation of single-sentence paraphrases from predicate/argument structure using lexico-grammatical resources. *Dans Proceedings of the second international workshop on Paraphrasing*, Sapporo, Japon. (Cité à la page 18.)

Lafourcade, M. et Joubert, A. (2008). JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. *Dans Actes de JADT'08 : Journées internationales d'Analyse statistiques des Données Textuelles*, pages 657–666, Lyon, France. (Cité à la page 135.)

Langkilde, I. et Knight, K. (1998). Generation that exploits corpus-based statistical knowledge. *Dans Proceedings of COLING/ACL*, Montréal, Canada. (Cité aux pages 28, 31, and 41.)

Lapata, M. et Keller, F. (2005). Web-based Models for Natural Language Processing. *ACM Transactions on Speech and Language Processing*, 2(1):1–31. (Cité à la page 141.)

Lavecchia, C., Smaïli, K. et Langlois, D. (2007). Building Parallel Corpora from Movies. *Dans The 4th International Workshop on Natural Language Processing and Cognitive Science - NLPCS 2007*, Funchal, Portugal. (Cité à la page 109.)

Lavie, A. et Agarwal, A. (2007). METEOR : An automatic metric for MT evaluation with high levels of correlation with human

judgments. *Dans Proceedings of the ACL Workshop on Statistical Machine Translation*, Prague, République tch'eque. (Cité aux pages 12, 56, and 116.)

Lepage, Y. et Denoual, E. (2005). Automatic Generation of Paraphrases To Be Used as Translation References in Objective Evaluation Measures of Machine Translation. *Dans Proceedings of the IJCNLP Workshop on Paraphrasing*, pages 57–64, Jeju Island, Corée du Sud. (Cité à la page 44.)

Levrat, B. et Amghar, T. (1995). Paraphrase et reformulation : une présentation. Rapport technique, Université Paris-Nord. (Cité à la page 14.)

Lin, D. et Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, 7(4):343–360. (Cité aux pages 34, 39, 40, and 41.)

Liu, C., Dahlmeier, D. et Ng, H. T. (2010). PEM : A paraphrase evaluation metric exploiting parallel texts. *Dans Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 923–932, Cambridge, États-Unis. (Cité à la page 42.)

Madnani, N. (2010). *The Circle of Meaning : From Translation to Paraphrasing and Back*. Thèse de doctorat, University of Maryland, College Park, États-Unis. (Cité à la page 16.)

Madnani, N. et Dorr, B. J. (2010). Generating Phrasal and Sentential Paraphrases : A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3):341–387. (Cité aux pages 2, 15, 27, 31, 33, and 46.)

Madnani, N., Resnik, P., Dorr, B. et Schwartz, R. (2008). Are multiple reference translations necessary ? Investigating the value of paraphrased reference translations in parameter optimization. *Dans Proceedings of AMTA*, Hawai'i, États-Unis. (Cité aux pages 31, 32, and 43.)

Malakasiotis, P. (2011). *Paraphrase and Textual Entailment Recognition and Generation*. Thèse de doctorat, Department of Informatics, Athens University of Economics and Business, Gr'ece. (Cité à la page 16.)

Marius, P. (2005). Mining Paraphrases from Self-anchored Web Sentence Fragments. *Dans Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 193–204, Porto, Portugal. (Cité à la page 36.)

Marsi, E. et Krahmer, E. (2005). Explorations in sentence fusion. *Dans Proceedings of the European Workshop on Natural Language Generation*, AberdeenRoyaume-Uni. (Cité à la page 44.)

Martin, R. (1976). *Inférence, Antonymie et Paraphrase*. Librairie C. Klincksiec, Paris. (Cité aux pages 9, 12, 14, 18, 21, 22, and 23.)

Marton, Y., Callison-Burch, C. et Resnik, P. (2009). Improved Statistical Machine Translation Using Monolingually-Derived Paraphrases. *Dans Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 381–390, Suntec, Singapour. (Cité aux pages 18, 43, and 44.)

Max, A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. *Dans Actes de TALN*, Avignon, France. (Cité aux pages 11, 43, and 80.)

Max, A. (2009). Sub-sentencial Paraphrasing by Contextual Pivot Translation. *Dans Proceedings of the ACL Workshop on Applied Textual Inference*, pages 18–26, Suntec, Singapour. (Cité aux pages 31, 38, and 41.)

Max, A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. *Dans Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 656–666, Cambridge, États-Unis. (Cité aux pages 18 and 44.)

Max, A. et Wisniewski, G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. *Dans Proceedings of LREC*, Valetta, Malte. (Cité aux pages 30 and 138.)

McCawley, J. (1968). Lexical insertion in a transformational grammar without deep structure. (Cité à la page 12.)

McKeown, K. (1983). Paraphrasing questions using given and new information. *Computational Linguistics*, 9(1):1–10. (Cité à la page 33.)

McKeown, K., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J., Nenkova, A., Sable, C., Schiffman, B. et Sigelman, S. (2002). Tracking and summarizing news on a daily basis with Columbia's Newsblaster. *Dans Proceedings of the second international conference on Human Language Technology Research*, San Diego, États-Unis. (Cité à la page 43.)

McKeown, K. R. (1979). Paraphrasing using given and new information in a question-answer system. *Dans Proceedings of the 17th Annual Meeting of the Association for Computational Linguistics*, La Jolla, États-Unis. (Cité à la page 44.)

Mel'cuk, I. (1988). Paraphrase et lexique dans la théorie linguistique sens-texte in lexique et paraphrase. *Lexique*, (6):13–54. (Cité aux pages 1, 9, 12, 17, 18, and 22.)

Mel'čuk, I. (2003). Collocations dans le dictionnaire. *Les écarts culturels dans les dictionnaires bilingues. Paris, France : H. Champion*, pages 19–64. (Cité à la page 12.)

Mel'čuk, I. et Arbatchewsky-Jumarie, N. (1992). Dictionnaire explicatif et combinatoire du français contemporain : recherches

lexico-sémantiques. *Presses de l'Université de Montréal*, 3. (Cité aux pages 13 and 18.)

Mel'čuk, I. et Polguère, A. (2007). *Lexique actif du français : l'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*. De Boeck. (Cité à la page 13.)

Metzler, D., Hovy, E. et Zhang, C. (2011). An empirical evaluation of data-driven paraphrase generation techniques. *Dans Proceedings of ACL-HLT*, Portland, États-Unis. (Cité à la page 43.)

Mihalcea, R., Corley, C. et Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. *Dans Proceedings of AAAI*, Boston, États-Unis. (Cité à la page 33.)

Milićević, J. (2003). *Modélisation sémantique, syntaxique et lexicale de la paraphrase*. Thèse de doctorat, Université de Montréal, Canada. (Cité à la page 18.)

Milićević, J. (2007). *La paraphrase. Modélisation de la paraphrase langagi'ere*. Bern : Peter Lang. (Cité aux pages 9 and 15.)

Miller, G. A. (1995). WordNet : a lexical database For English. *Communications of the ACM*, 38(11):39–41. (Cité aux pages 28 and 31.)

Nakov, P. (2008). Improved Statistical Machine Translation Using Monolingual Paraphrases. *Dans Proceeding of the 18th European Conference on Artificial Intelligence (ECAI08)*, pages 338–342, Patras, Grèce. (Cité aux pages 31 and 44.)

Nasr, A. (1996). *Un modèle de reformulation automatique fondé sur la Théorie Sens-Texte–Application aux langues contrôlées*. Thèse de doctorat, Université Paris 7. (Cité à la page 43.)

Nelken, R. et Yamangil, E. (2008). Mining Wikipedia's Article Revision History for Training Computational Linguistics Algorithms. *Dans Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence : An Evolving Synergy*, Chicago, États-Unis. (Cité à la page 30.)

Och, F. J. et Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*. (Cité aux pages 71, 72, and 80.)

Ohtake, K. et Yamamoto, K. (2003). Applicability analysis of corpus-derived paraphrases toward example-based paraphrasing. *Dans Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, Sentosa, Singapour. (Cité à la page 16.)

Pang, B., Knight, K. et Marcu, D. (2003). Syntax-based alignement of multiple translations : Extracting paraphrases and generating new sentences. *Dans Proceedings of NAACL-HLT*, pages 102–109, Edmonton, Canada. (Cité aux pages 11, 29, 36, 40, 41, 46, 75, 76, and 77.)

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002). Bleu : a Method for Automatic Evaluation of Machine Translation. *Dans Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, États-Unis. (Cité aux pages 12, 2, 42, 44, 56, and 116.)

Pasça, M. (2003). Open-domain question answering from large text collections. *Computational Linguistics*, 29(4):665–667. (Cité à la page 44.)

Pasça, M. et Dienes, P. (2005). Aligning Needles in a Haystack : Paraphrase Acquisition Across the Web. *Dans Proceedings of IJCNLP*, Jeju Island, South Korea. (Cité aux pages 33, 34, 44, and 125.)

Patry, A. et Langlais, P. (2011). Identifying Parallel Documents from a Large Bilingual Collection of Texts : Application to Parallel Article Extraction in Wikipedia. *Dans 4th ACL/SIGWAC Workshop on Building and Using Comparable Corpora (BUCC'2011)*, Portland, États-Unis. (Cité à la page 37.)

Petrov, S., Barrett, L., Thibaux, R. et Klein, D. (2006). Learning accurate, compact, and interpretable tree annotation. *Dans Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australie. (Cité à la page 77.)

Petrov, S. et Klein, D. (2007). Improved inference for unlexicalized parsing. *Dans Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics ; Proceedings of the Main Conference*, pages 404–411, Rochester, États-Unis. (Cité à la page 142.)

Quirk, C., Brockett, C. et Dolan, W. (2004). Monolingual Machine Translation for Paraphrase Generation. *Dans Proceedings of EMNLP 2004*, pages 142–149, Barcelone, Espagne. (Cité aux pages 16, 31, and 71.)

Ravichandran, D. et Hovy, E. (2002). Learning surface text patterns for a Question Answering System. *Dans Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 41–47, Philadelphia, États-Unis. (Cité à la page 44.)

Resnik, P., Buzek, O., Hu, C., Kronrod, Y., Quinn, A. et Bederson, B. B. (2010). Improving translation via targeted paraphrasing. *Dans Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'10)*, pages 127–137, Cambridge, États-Unis. (Cité aux pages 30 and 135.)

Sagot, B., Fort, K., Adda, G., Mariani, J. et Lang, B. (2011). Un turc mécanique pour les ressources linguistiques : critique de la myriadisation du travail parcellisé. *Dans Actes de TALN*, Montpellier, France. (Cité aux pages 30 and 135.)

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Dans Proceedings of International Conference on New*

Methods in Language Processing, Manchester, Royaume-Uni. (Cité à la page 53.)

Sekine, S. (2001). Extracting synonymous expressions from multiple newspaper documents. *Dans Proceedings of the ANLP Workshop on Automatic Paraphrasing*, Pittsburgh, États-Unis. (Cité aux pages 34, 36, and 41.)

Shen, S., Radev, D. R., Patel, A. et Erkan, G. (2006). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. *Dans Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 747–754, Sydney, Australie. (Cité aux pages 36 and 41.)

Shimohata, M. (2004). *Acquiring Paraphrases from Corpora and its Application to Machine Translation*. Thèse de doctorat, Nara Institute of Science and Technology, Japon. (Cité à la page 18.)

Shinyama, Y. et Sekine, S. (2003). Paraphrase Acquisition for Information Extraction. *Dans Proceedings of the Second International Workshop on Paraphrasing*, pages 65–71, Sapporo, Japon. (Cité aux pages 36 and 44.)

Shinyama, Y. et Sekine, S. (2005). Using repeated patterns across comparable articles for paraphrase acquisition. Rapport technique, New York University. Proteus Technical Report. (Cité à la page 44.)

Shinyama, Y., Sekine, S. et Sudo, K. (2002). Automatic paraphrase acquisition from news articles. *Dans Proceedings of the second international conference on Human Language Technology Research*, pages 313–318, San Francisco,États-Unis. (Cité aux pages 36 and 41.)

Snover, M., Dorr, B. J., Schwartz, R., Micciulla, L. et Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Dans Proceedings of AMTA*, pages 223–231, Boston, États-Unis. (Cité aux pages 12, 44, 53, 56, and 116.)

Snover, M., Madnani, N., Dorr, B. et Schwartz, R. (2009). Fluency, adequacy, or HTER ? Exploring different human judgments with a tunable MT metric. *Dans Proceedings of the Fourth Workshop on Statistical Machine Translation*, Athènes, Grèce. (Cité à la page 78.)

Snover, M., Madnani, N., Dorr, B. J. et Schwartz, R. (2010). TER-Plus : paraphrase, semantic, and alignment enhancements to Translation Edit Rate. *Machine Translation*, 23(2-3). (Cité aux pages 78, 92, and 141.)

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H. et Yamamoto, S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. *Dans Proceedings of LREC 2002*, Las Palmas, Espagne. (Cité à la page 28.)

Tiedemann, J. (2007). Building a multilingual parallel subtitle corpus. *Dans CLIN17*, Louvain, Belgique. (Cité aux pages 54 and 109.)

Uzuner, O., Katz, B. et Nahnsen, T. (2005). Using Syntactic Information to Identify Plagiarism. *Dans Proceedings of the Second Workshop on Building Educational Applications Using NLP*, pages 37–44, Ann Arbor, États-Unis. (Cité à la page 44.)

Vila, M., Martí, M. et Rodríguez, H. (2011). Paraphrase concept and typology, a linguistically based and computationally oriented approach. *Procesamiento de Lenguaje Natural*, 46:83–90. (Cité aux pages 19 and 20.)

Wallis, P. (1993). Information retrieval based on paraphrase. *Dans Proceedings of PACLING 1993*, Vancouver, Canada. (Cité à la page 33.)

Wang, K., Thrasher, C., Viegas, E., Li, X. et Hsu, B.-j. P. (2010). An overview of microsoft web n-gram corpus and applications. *Dans Proceedings of the NAACL HLT 2010 Demonstration Session*, pages 45–48, Los Angeles, États-Unis. (Cité à la page 141.)

Wang, R. et Callison-Burch, C. (2011). Paraphrase fragment extraction from monolingual comparable corpora. *Dans Proceedings of the 4th Workshop on Building and Using Comparable Corpora : Comparable Corpora and the Web*, Portland, États-Unis. (Cité aux pages 37 and 41.)

White, D. et Joy, M. (2004). Sentence-based natural language plagiarism detection. *Journal on Educational Resources in Computing (JERIC)*, 4(4):2. (Cité à la page 44.)

Wubben, S., van den Bosch, A., Krahmer, E. et Marsi, E. (2009). Clustering and matching headlines for automatic paraphrase acquisition. *Dans EWNLG*, Athènes, Grèce. (Cité aux pages 55 and 109.)

Zhao, S., Lan, X., Liu, T. et Li, S. (2009). Application-driven Statistical Paraphrase Generation. *Dans Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Suntec, Singapour. (Cité à la page 16.)

Zhao, S., Niu, C., Zhou, M., Liu, T. et Li, S. (2008a). Combining Multiple Resources to Improve SMT-based Paraphrasing Model. *Dans Proceedings of ACL-08 : HLT*, pages 1021–1029, Columbus, États-Unis. (Cité à la page 31.)

Zhao, S., Wang, H., Lan, X. et Liu, T. (2010). Leveraging multiple mt engines for paraphrase generation. *Dans Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1326–1334, Pékin, Chine. (Cité aux pages 31 and 32.)

Zhao, S., Wang, H., Liu, T. et Li, S. (2008b). Pivot Approach for Extracting Paraphrase Patterns from Bilingual Corpora. *Dans Proceedings of ACL-08 : HLT*, pages 780–788, Columbus, États-Unis. Association for Computational Linguistics. (Cité aux pages 16, 38, and 41.)

Zhou, L., Lin, C.-Y., Munteanu, D. S. et Hovy, E. (2006). Paraeval : Using paraphrases to evaluate summaries automatically. *Dans Proceedings of the Human Language Technology Conference of the NAACL*, pages 447–454, New York, États-Unis. (Cité aux pages 41 and 44.)

Zhu, Z., Bernhard, D. et Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. *Dans Proceedings of The 23rd International Conference on Computational Linguistics (COLING 10)*, pages 1353–1361, Pekin, Chine. (Cité à la page 44.)

PUBLICATIONS

Les idées et les résultats présentés dans cette thèse ont déjà été publiés dans les articles suivants, disponibles sur http://perso.limsi.fr/hbouamor/publications :

Bouamor, H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multilingue multisource. *Dans Récital-TALN*, Montréal, Canada. (Cité à la page 51.)

Bouamor, H., Max, A., Illouz, G. et Vilnat, A. (2011a). Web-based validation for contextual targeted paraphrasing. *Dans Proceedings of the Workshop on Monolingual Text-To-Text Generation*, pages 10–19, Portland, États-Unis.

Bouamor, H., Max, A., Illouz, G. et Vilnat, A. (2012a). A contrastive review of paraphrase acquisition techniques. *Dans Proceedings of LREC*, Istanbul, Turquie.

Bouamor, H., Max, A., Illouz, G. et Vilnat, A. (2012b). Validation sur le web de reformulations locales : application à la wikipédia. *Dans Actes de TALN*, Grenoble, France.

Bouamor, H., Max, A. et Vilnat, A. (2009). Amener des utilisateurs à créer et évaluer des paraphrases par le jeu. *Dans Actes de TALN, session de démonstrations*, Senlis, France.

Bouamor, H., Max, A. et Vilnat, A. (2010a). Acquisition de paraphrases sous-phrastiques depuis des paraphrases d'énoncés. *Dans Actes de TALN 2010*, Montréal, Canada.

Bouamor, H., Max, A. et Vilnat, A. (2010b). Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases. *Dans Proceedings of IceTAL*, Reykjavik, Islande.

Bouamor, H., Max, A. et Vilnat, A. (2011b). Combinaison d'informations pour l'alignement monolingue. *Dans Actes de TALN*, Montpellier, France.

Bouamor, H., Max, A. et Vilnat, A. (2011c). Monolingual alignment by edit rate computation on sentential paraphrase pairs. *Dans Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies (ACL-HLT11)*, pages 395–400, Portland, États-Unis.

Bouamor, H., Max, A. et Vilnat, A. (2012c). Multi-technique paraphrase alignment : A contribution to pinpointing sub-sentential paraphrases. *ACM Transactions on Intelligent Systems and Technology, Special Issue On Paraphrasing*, 5:À paraitre.

Bouamor, H., Max, A. et Vilnat, A. (2012d). Validation of sub-sentential paraphrases acquired from parallel monolingual corpora. *Dans Proceedings of the 13th Conference of the European Chapter*

of the Association for Computational Linguistics (EACL12), pages 716–725, Avignon, France.

Bouamor, H., Max, A. et Vilnat, A. (2012e). Étude bilingue de l'acquisition et de la validation automatiques de paraphrases sous-phrastiques. *Traitement Automatique des Langues (TAL)*, 53(1):À paraitre.

Bouamor, H., Max, A. et Vilnat, A. (2012f). Une étude en 3d de la paraphrase : types de corpus, langues et techniques. *Dans Actes de TALN*, Grenoble, France.

Dutrey, C., Bouamor, H., Bernhard, D. et Max, A. (2011a). Local modifications and paraphrases in wikipedia's revision history. *SEPLN Journal*, (46).

Dutrey, C., Bouamor, H., Bernhard, D. et Max, A. (2011b). Paraphrases et modifications locales dans l'historique des révisions de wikipédia. *Dans Actes de TALN*, Montpellier, France. (Cité aux pages 30, 32, 35, 41, 75, and 138.)