TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Paraphrases et modifications locales dans l'historique des révisions de Wikipédia

Camille Dutrey[1] Houda Bouamor[2,3] Delphine Bernhard[2] Aurélien Max[2,3]
(1) INALCO, Paris, France
(2) LIMSI-CNRS, Orsay, France
(3) Univ. Paris-Sud, Orsay, France
camille@dutrey.fr {prénom.nom}@limsi.fr

Résumé. Dans cet article, nous analysons les modifications locales disponibles dans l'historique des révisions de la version française de Wikipédia. Nous définissons tout d'abord une typologie des modifications fondée sur une étude détaillée d'un large corpus de modifications. Puis, nous détaillons l'annotation manuelle d'une partie de ce corpus afin d'évaluer le degré de complexité de la tâche d'identification automatique de paraphrases dans ce genre de corpus. Enfin, nous évaluons un outil d'identification de paraphrases à base de règles sur un sous-ensemble de notre corpus.

Abstract. In this article, we analyse the modifications available in the French Wikipédia revision history. We first define a typology of modifications based on a detailed study of a large corpus of modifications. Moreover, we detail a manual annotation study of a subpart of the corpus aimed at assessing the difficulty of automatic paraphrase identification in such a corpus. Finally, we assess a rule-based paraphrase identification tool on a subset of our corpus.

Mots-clés : Wikipédia, révisions, identification de paraphrases.

Keywords: Wikipedia, revisions, paraphrase identification.

1 Introduction

Wikipédia ne cesse de croître et est actuellement l'encyclopédie libre la plus volumineuse et la plus fréquentée au monde. Ses articles sont écrits et maintenus de manière collaborative et bénévole. Les énormes quantités de données présentes dans cette encyclopédie ont motivé de nombreux travaux sur l'acquisition automatique de ressources comme par exemple l'acquisition des connaissances lexico-sémantiques (Zesch *et al.*, 2008). Cependant, la majorité de ces études n'utilisent que la version la plus récente des articles de l'encyclopédie. Wikipédia met également à disposition l'historique des révisions de chacun de ses articles qui sont itérativement modifiés et affinés par de multiples utilisateurs du Web. Ces révisions rendent possible l'extraction de certains types de modifications locales reflétant l'évolution, la maturation et la correction de la forme linguistique des articles, et constituent donc une importante source de connaissances encore peu exploitée à ce jour.

Dans cet article, nous détaillons une typologie des modifications locales présentes dans le corpus WiCoPaCo [1]

1. Librement téléchargeable sur `http://wicopaco.limsi.fr`

de révisions extraites automatiquement de la version française de Wikipédia. Notre étude met l'accent sur le phénomène de *paraphrases locales*, qui sont de plus en plus utilisées pour améliorer les performances de plusieurs applications de TAL comme les systèmes de traduction automatique (Max, 2010) ou de question-réponse (Duclaye *et al.*, 2003), ainsi qu'en génération, pour aider des auteurs à trouver des formulations plus adaptées (Max, 2008).

Cet article est organisé comme suit : dans la section 2, nous passons tout d'abord en revue les principaux travaux portant sur l'utilisation de l'historique des révisions de Wikipédia, puis nous décrivons le corpus WiCoPaCo utilisé dans cette étude dans la section 3. Nous présentons la typologie des modifications locales que nous proposons dans la section 4. Dans la section 5 nous exposons nos premières expériences sur l'identification automatique de paraphrases dans ce corpus, et enfin nous présentons nos observations et décrivons nos travaux futurs dans la section 6.

2 Exploitation des révisions de Wikipédia : état de l'art

Les révisions de Wikipédia ont déjà été exploitées pour différentes tâches et applications. Nelken & Yamangil (2008) exploitent l'historique des révisions de Wikipédia pour acquérir une grande quantité de données d'apprentissage en comparant les versions adjacentes d'un même article pour trois tâches à différents niveaux de granularité linguistique : collecte des fautes d'orthographe et de leur correction (niveau du mot), données d'apprentissage pour les algorithmes de compression de phrases (niveau phrase) et d'amorçage pour les systèmes de résumé automatique (niveau document). Yatskar *et al.* (2010) utilisent l'historique des modifications dans la version anglaise simplifiée de Wikipédia pour en extraire des simplifications lexicales.

Max & Wisniewski (2010) décrivent WICOPACO (Wikipédia Correction and Paraphrase Corpus), une ressource construite en explorant automatiquement l'historique des révisions de Wikipédia et en extrayant les modifications locales effectuées par les contributeurs. Ce corpus comprend différents types de corrections et de réécritures. Le travail de Wisniewski *et al.* (2010) montre par exemple comment cette ressource peut être utilisée pour améliorer la performance d'un système de correction orthographique automatique. Zanzotto & Pennacchiotti (2010) exploitent quant à eux l'historique des révisions de Wikipédia pour extraire un grand nombre de paires d'unités textuelles en relation d'implication et appliquent des méthodes d'apprentissage semi-supervisé pour rendre l'ensemble des données extraites cohérentes par rapport aux données existantes.

Dans la mesure où l'utilisation et la popularité des wikis ainsi que d'autres systèmes collaboratifs s'accroît, les questions concernant la fiabilité de ces informations gagnent en importance. Hu *et al.* (2007) ont par exemple développé un modèle de confiance basé sur l'historique des éditions des articles de Wikipédia afin de calculer et contrôler leur fiabilité. D'autres travaux ont proposé des visualisations originales des révisions de Wikipédia, tels que History Flow (Viégas *et al.*, 2004) ou WikiDashboard (Suh *et al.*, 2008). WikipediaViz (F. Chevalier, S. Huot et J-D. Fekete, 2010) propose un ensemble de visualisations basé sur un mécanisme de collecte et d'aggrégation de données d'éditions de Wikipédia pour aider le lecteur à appréhender la maturité d'un article.

Comme nous l'avons montré, la plupart des travaux de recherche antérieurs sur l'historique des révisions de Wikipédia se concentrent sur des aspects spécifiques de la ressource et se fixent pour objectif des applications bien définies telles que la simplification de texte, la compression de phrases ou encore la visualisation des informations. À notre connaissance, il n'y a pas de vision globale des phénomènes de modifications locales disponibles dans les révisions de Wikipedia, bien qu'il existe une grande variété de types de modifications qui sont d'intérêt pour de nombreuses applications de traitement automatique des langues, et en particulier les paraphrases locales.

3 WICOPACO, un corpus de modifications locales de Wikipédia

L'acquisition de paires de segments textuels ayant le même sens (*paraphrases locales*) a été à l'origine d'un nombre important de travaux sur l'exploitation automatique de corpus de textes (voir par exemple (Madnani & Dorr, 2010)). Les corpus utilisés peuvent être organisés par le degré de correspondance entre deux unités de texte : des paires de paraphrases phrastiques, obtenues par exemple par traduction multiple (*corpus parallèles monolingues*) ; les paires de phrases ayant le même sens, obtenues à partir de corpus composés de textes dans la même langue partageant une partie du vocabulaire employé, ce qui implique généralement que les textes parlent d'un même sujet durant la même période (*corpus monolingues comparables*) ; les paires de phrases partageant des traductions dans d'autres langues (*corpus parallèles multilingues*). Les corpus monolingues parallèlles sont les corpus les plus appropriés pour observer et acquérir automatiquement des segments de texte en relation de paraphrase locale de haute qualité (Bouamor *et al.*, 2010). Or ce type de corpus existe en très faible quantité, et leur construction est une tâche compliquée et coûteuse. *A contrario*, un des principaux défauts des autres types de corpus est que les paraphrases potentielles n'y sont observées qu'*indirectement*, par exemple par l'intermédiaire d'une traduction commune ou d'un contexte jugé similaire.

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 étant destinées à ne pas modifier le sens du texte, mais à améliorer sa qualité, le rendre plus cohérent, ou limiter sa redondance. Des brouillons d'écrivains ont été notamment utilisés dans les critiques génétiques textuelles qui étudient les processus de création de textes (Bourdaillet & Ganascia, 2007). Ces documents annotés sont malheureusement disponibles en petites quantités et sont de plus difficiles à encoder en format électronique. En outre, ces projets contiennent souvent 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'historique des révisions de cette ressources constitue donc une source importante de phénomènes de réécriture *naturelle*, y compris des paraphrases locales dans leur contexte.

WICOPACO (Max & Wisniewski, 2010) est un corpus de modifications locales, extrait à partir de l'historique des révisions des articles de Wikipédia, et disponible pour le moment pour le français. Ce corpus a été construit en 4 étapes :

1. Sélection de paires de versions d'articles.
2. Normalisation du texte (segmentation, suppression du « wikitexte », etc.).
3. Alignement des modifications par une recherche de plus longues sous-séquences communes.
4. Filtrage des modifications retenues (séquences initiales de 7 mots ou plus) et extraction du contexte avant et après modification (paragraphe englobant).

Le corpus que nous avons utilisé contient 408 816 entrées uniques. Celui-ci est disponible sous forme d'un fichier au format XML associant un élément à chaque modification. Une modification y est décrite par un contexte avant modification et un contexte après modification, ainsi que par un ensemble de métadonnées donnant des informations sur le contributeur de la révision et permettant de localiser le texte extrait dans la ressource Wikipédia d'origine. Un exemple d'un tel élément, correspondant à une simplification lexicale, est donné à la figure 1.

```
<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 1 – Exemple d'une modification dans le corpus WICOPACO.

4 Typologie des modifications locales

Nous avons analysé le corpus WICOPACO afin de développer une typologie détaillée des modifications locales [2] dans les révisions de Wikipédia. Cette typologie permet de représenter tous les phénomènes observables dans WICOPACO et d'indiquer le degré de variation sémantique entre les segments correspondant à des paires de modifications locales. Elle se compose de deux catégories distinguant deux grandes classes de variation sémantique : la classe des *faibles variations sémantiques* et la classe des *corrections factuelles et vandalismes*. Ces deux classes peuvent contenir des modifications locales pour lesquelles il n'existe pas de relation sémantique stricte entre le segment avant la modification et le segment après modification, ce qui est par exemple le cas pour les modifications typographiques [3].

4.1 Modifications à faible variation sémantique

La classe des modifications à faible variation sémantique comporte les *corrections* et les *reformulations* (voir Table 1).

Les corrections de surface font référence aux changements de surface qui visent à améliorer le texte afin qu'il soit conforme aux normes linguistiques, et se décomposent de la manière suivante :

- *Corrections typographiques* : changement de la disposition et du format du texte, par exemple, ajout ou suppression d'espaces ou de signes de ponctuation, changement de casse d'un caractère, modification du format d'une date ou d'une heure, écriture d'un nombre en toutes lettres ou en chiffres, etc.
- *Corrections orthographiques* : corrections affectant les fautes d'orthographe. Elles se réfèrent à la transformation d'un mot inexistant en un mot attesté dans le lexique, comme la modification de diacritiques ou le remplacement d'un ou plusieurs caractères.
- *Corrections grammaticales* : résolution des fautes d'orthographe qui ne peuvent être détectées et corrigées que par la prise en compte du contexte.

2. Ici nous héritons de la définition suivie dans la ressource utilisée pour la localité des modifications observées : il s'agit de segments d'au plus 7 mots (ponctuations et autres signes non inclus).

3. La typologie complète est disponible sous forme de document technique du LIMSI (Dutrey *et al.*, 2011) et est accessible à l'adresse suivante : `http://wicopaco.limsi.fr/pub/typologie-modifications-wikipedia.pdf`

CORRECTIONS DE SURFACE
Corrections typographiques
⇒ ex. une espace remplacée par un trait d'union pour corriger une erreur typographique : *Le triceps brachial est un muscle extenseur de l' **[avant bras → avant-bras]** sur le bras.*
Corrections orthographiques
⇒ ex. un caractère alphabétique supprimé pour transformer un *non-mot* en un mot attesté dans le lexique : *Ces trois parties se **[rejoingnent → rejoignent]** pour former une épaisse masse.* ⇒ ex. un diacritique remplacé par un autre pour transformer un *non-mot* en un mot attesté dans le lexique : *L' **[ëglise → église]** gothique Sainte-Marie...*
Corrections grammaticales
⇒ ex. un diacritique remplacé par un autre pour corriger une erreur portant sur un mot attesté dans le lexique : *L'anathème pour le **[pêcheur → pécheur]** : ce dernier est privé de sépulture chrétienne.* ⇒ ex. un mot remplacé par un autre pour corriger une erreur portant sur un mot attesté dans le lexique : *Il chante avec une **[voie → voix]** de troubadour.*
REFORMULATIONS
Reformulations lexicales
⇒ ex. un emprunt remplacé par la forme correspondante en français standard : ***[L'implémentation → La mise en œuvre]** de l'algorithme...*
Reformulations syntaxiques
⇒ ex. une permutation entre deux segments sur l'axe syntagmatique : *Source : **[l'Invention de l'Europe d'Emmanuel Todd → Emmanuel Todd, L'Invention de l'Europe]**.* ⇒ ex. une proposition circonstancielle transformée en une proposition relative : *Un infomercial pseudo-scientifique **[en exposant → qui expose]** grossièrement...*
Reformulations sémantiques
⇒ ex. un mot remplacé par un autre appartenant au même champ lexical (hyponymie) : *Il fonde le **[journal → quotidien]** francophone « Le Tunisien »en 1907.* ⇒ ex. une paraphrase servant différents propos (infra précision de sens) : *Ce vers de Nuit rhénane d'Apollinaire **[qui paraît presque sans structure rythmique → dont la césure est comme masquée]**...*

TABLE 1 – Types de modifications à faible variation sémantique

Les reformulations correspondent à des changements plus importants qui modifient les choix lexicaux et syntaxiques faits par le contributeur précédent sans modifier profondément la signification du texte :

- *Reformulations lexicales* consistant, par exemple, à remplacer un acronyme avec son nom complet, traduire un mot étranger ou un emprunt, remplacer une variante régionale par sa version standard, etc.
- *Reformulations syntaxiques* permettant, par exemple, de modifier l'ordre des propositions, de transformer une phrase à la voix active ou passive ou de changer le type de proposition.
- *Reformulations sémantiques* comme l'utilisation d'hyperonymes ou d'hyponymes, la normalisation encyclopédique, l'utilisation de synonymes ou l'ajout d'informations additionnelles peu significatives.

4.2 Corrections factuelles et vandalismes

CORRECTIONS FACTUELLES
⇒ ex. un mot remplacé par son antonyme : *Un catalyseur solide (phase **[liquide → solide]**) avec de l'hydrogène (phase gazeuse).* ⇒ ex. un segment remplacé par un autre n'ayant aucun lien sémantique avec le premier : *représente pour eux **[l'Occident chrétien → la supériorité de la race celto-germanique]**.*
VANDALISMES
⇒ ex. une chaîne insérée produisant un *non-mot* (vandalisme manifeste) : *L'Autriche a été occupée **[par → psh ! ! ar]** les Romains.* ⇒ ex. un mot remplacé par un autre qui ne produit aucun sens compte tenu du contexte (vandalisme subtil) : *Devant la Cour de **[Cassation → Castration]**…*

TABLE 2 – Corrections factuelles et vandalismes

Cette classe se décompose en deux sous-types (voir la Table 2), les *corrections factuelles* et les *vandalismes*, pour lesquels le sens du texte est fortement affecté et peut être totalement changé.

Les corrections factuelles correspondent soit à une modification qui induit une forte variation de sens, soit à une modification ne présentant aucun lien sémantique avec le texte initial. Elles consistent, par exemple, à remplacer un mot par un antonyme ou à changer le temps d'un verbe de sorte à ce que le sens de la phrase soit modifié. Ce genre de modifications vise à améliorer le contenu de Wikipédia.

Le vandalisme fait référence aux modifications qui, délibérément, modifient ou détruisent le contenu afin de nuire à la qualité de Wikipédia. Le vandalisme manifeste se caractérise par l'insertion de non-mots ou d'insultes tandis que le vandalisme subtil se caractérise par des reformulations grammaticales mais dont l'interprétation est en complète contradiction avec le sens initial. Il est particulièrement important de détecter ce dernier type, puisque dans certains cas un lecteur peu attentif ou crédule pourra considérer à tort l'information décrite comme fiable.

4.3 Annotation manuelle

Nous avons conçu un schéma d'annotation basé sur la typologie décrite précédemment. L'objectif de cette annotation est d'évaluer le degré de complexité de l'identification manuelle des paraphrases dans les modifications locales. L'annotation est guidée par notre application cible qui est l'identification automatique des paraphrases dans WICOPACO. Dans notre typologie, les paraphrases correspondent aux reformulations présentes dans la classe des réécritures à faible variation sémantique. Elles doivent être distinguées des corrections de surface et des reformulations qui induisent un changement sémantique majeur. Cette annotation a donc deux buts principaux : repérer des phénomènes liés à une réécriture à faible variation sémantique et repérer les vandalismes, notamment dans une optique ultérieure d'apprentissage supervisé.

Afin de faciliter cette tâche, nous avons élaboré un schéma d'annotation composé de quatre classes principales :
– *Les corrections de surface*, qui englobent toutes les modifications visant à rendre le texte conforme aux normes de la langue.

– *Les reformulations*, qui correspondent aux différents types de paraphrases, y compris les précisions et les simplifications.
– *Les corrections factuelles et les vandalismes*
– *Les défauts d'alignement* qui correspondent aux cas où les modifications locales identifiées présentent un défaut dans leur alignement (voir Figure 2). Cependant, même avec un défaut d'alignement un segment peut contenir une modification locale.

Henri IV	**fut assassiné par Ravaillac en 1610**.
A partir de la conversion d' Henri IV	**la fidélité au roi l' a emporté sur l' appartenance religieuse**.

FIGURE 2 – Exemple d'un défaut d'alignement dans le corpus WICOPACO.

Une annotation couvre l'ensemble du segment identifié comme une modification locale (notée par une balise XML `m` dans le corpus WICOPACO, comme illustré dans la Figure 1) : l'objectif est de déterminer le type de la modification de partir d'une paire de segments, mais pas de réaligner les mots dans ces segments. En outre, il était possible d'attribuer plusieurs étiquettes à la même modification.

Pour réaliser cette annotation, nous avons utilisé l'outil d'alignement Yawat (Germann, 2008) conçu à l'origine pour l'alignement de textes parallèles bilingues au niveau du mot. Nous avons adapté le schéma d'annotation de cet outil pour notre annotation multi-niveaux. L'annotation a été réalisée par quatre annotateurs [4] sur 200 paires de segments tirées d'une version filtrée du corpus WICOPACO. Comme les modifications de ponctuation sont fréquentes, seules les modifications d'une distance d'édition (Levenshtein) d'au moins 4 ont été considérées pour l'annotation.

4.4 Résultats de l'annotation

La Table 3 décrit l'accord inter-annotateur de notre annotation, calculé à l'aide de la mesure du Kappa (κ) [5]. L'accord inter-annotateur varie de modéré à fort, en fonction de la classe. Globalement, les valeurs du κ sont proches des valeurs déclarées par Dolan & Brockett (2005) pour l'identification de paraphrases (κ de 0,62) et par Glickman *et al.* (2005) (κ de 0,6) pour l'implication textuelle.

Type	**κ moy**	**Interprétation**	**κ maximum**	**κ minimum**
Corrections factuelles et vandalismes	0,65	Accord fort	0,71	0,61
Reformulation	0,60	Accord modéré	0,71	0,51
Correction	0,54	Accord modéré	0,81	0,40
Défaut d'alignement	0,48	Accord modéré	0,62	0,28

TABLE 3 – Accord inter-annotateur pour l'annotation des révisions de Wikipédia.

Nous indiquons également le nombre d'annotations identiques attribuées par 1 à 4 annotateurs (voir la Table 4) ainsi que les annotations uniques, attribuées par un seul annotateur. Cela permet de quantifier approximativement les phénomènes présents dans le corpus. Les paraphrases ont le plus grand nombre d'occurrences, suivies par les corrections factuelles et vandalismes. Ceci montre que les révisions de Wikipédia constituent un corpus bien

4. Co-auteurs du présent article.
5. Nous avons utilisé le calculateur κ en ligne pour annotateurs et classes multiples diponible sur `http://cosmion.net/jeroen/software/kappa/`.

adapté pour l'acquisition automatique de paraphrases [6]. En outre, les défauts d'alignement sont assez rares, ce qui montre que la méthode d'alignement utilisée pour la construction de WICOPACO est suffisamment précise pour fournir des modifications utiles.

	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

TABLE 4 – Nombre d'annotations identiques attribués par 1, 2, 3 ou 4 annotateurs.

L'étude des annotations a souligné certains problèmes potentiels pour l'identification automatique des classes décrites dans notre typologie. Tout d'abord, plusieurs phénomènes peuvent se produire simultanément, par exemple une transformation de diathèse (voix grammaticale) peut inclure une correction d'un non-mot (erreur). Dans ce cas, un classifieur automatique devrait être en mesure d'assigner plusieurs classes à une modification. Deuxièmement le contexte phrastique fourni par le corpus WICOPACO n'est parfois pas suffisant pour prendre une décision sur un type de modification spécifique. Un contexte plus large pourrait être utile aux classifieurs automatiques. Troisièmement, le typage correct d'une modification nécessite parfois une certaine connaissance des intentions du contributeur. Ce type d'information est parfois disponible dans les commentaires associés à une révision, mais peut être difficile à interpréter de façon automatique.

5 Identification de paraphrases : une méthode à base de règles

Nous avons mis en œuvre une méthode automatique destinée à distinguer les paraphrases des autres modifications dans WICOPACO, en adaptant l'outil de reconnaissance de variantes de termes `Fastr` (Christian Jacquemin, 1994). L'opération d'*indexation contrôlée* de ce système définit les variations acceptables par un système de métarègles s'appliquant à des règles de termes. Elles permettent d'exprimer les réécritures morphosyntaxiques possibles, ainsi que les relations d'ordre morphologique ou sémantique contenues dans des ressources préexistantes.

Nous avons dû créer un nouvel ensemble de métarègles pour la reconnaissance de paraphrases car le jeu de métarègles original s'est révélé innaproprié pour notre étude, cet ensemble ayant été développé avec l'objectif de reconnaissance de variantes de termes qui recouvrent une définition beaucoup plus permissive de la paraphrase. Nous avons cependant pu réutiliser les familles morphologiques et les familles sémantiques fournies par `Fastr`, pour exprimer des contraintes sémantiques et morphologiques. L'utilisation de `Fastr` nous permet d'évaluer si un système à base de règles est adapté pour l'identification des paraphrases dans un corpus tel que le nôtre, présentant une riche variété de phénomènes.

Nous avons utilisé un autre type de corpus pour le développement des nouvelles métarègles, afin de vérifier si les règles sont suffisamment générales pour être appliquées sur un corpus de type différent. Ce corpus est extrait de MULTITRAD (Bouamor, 2010), un corpus construit par collecte de paraphrases d'énoncés par traduction multiple multilingue, et annoté au niveau des mots. Ce corpus de développement a permis l'extraction de patrons

6. Cet article ne présente pas d'usage concret de paraphrases extraites de la ressource utilisée, mais il est évident que différentes paraphrases ne seront pas nécessairement adaptées pour les mêmes usages.

de reformulations exprimés sous forme de séquences de catégories morphosyntaxiques. La Table 5 montre par exemple les principaux patrons de reformulations observés pour le patron initial NOM VER VER.

Réécriture	Fréquence	Exemple
NOM VER VER	7	orateurs ont estimé → locuteurs ont jugé
NOM VER VER PRP NOM CONJ	2	Parlement a demandé → Parlement a fait des demandes pour
NOM VER PRP DET NOM	1	lois sont écrites→ lois restent dans les livres

TABLE 5 – Principaux patrons de réécriture associés à la séquence NOM VER VER dans MULTITRAD

La Table 6 illustre un exemple de nouvelle métarègle pour FASTR. Cette métarègle a pour nom **NAtoVASyn** car elle porte sur un segment source dont la structure est un **N**om suivi d'un **A**djectif réécrit (**to**) en un segment dont la structure est au minimum un **V**erbe suivi d'un **N**om suivi d'**A**djectif. La métarègle intègre également certaines contraintes morphologiques et sémantiques qui précisent que (i) le nom du segment source et le verbe du segment cible ont une racine morphologique commune et (ii) les adjectifs des segments source et cible sont synonymes. Il est à noter que l'utilisation d'un moteur de détection de variante est comparable aux travaux de Deléger & Zweigenbaum (2009) sur l'extraction de paraphrases de vulgarisation en langue de spécialité.

Metarule NAtoVAsyn(X1 → N1 A1) = X1 → V1 {ART ? | PRON ? | PREP ?} N A2 : <N1 root> = <V1 root> <A1 syn> = <A2 syn> <X1 metaLabel> = 'XX'.
protection constante → protéger de façon permanente

TABLE 6 – Exemple d'une métarègle de `Fastr`

Un ensemble de 83 métarègles a été développé pour la reconnaissance de paraphrases. Nous avons d'abord évalué la couverture des règles construites manuellement à l'aide de 206 paires de paraphrases d'énoncés issues du corpus MULTITRAD n'ayant pas été utilisés pour le développement des métarègles. `Fastr` a été en mesure d'identifier 185 paraphrases candidates, dont certaines sont illustrées dans la table 7.

MultiTrad	**WiCoPaCo**
décrit dans la proposition ↔ proposé	décéda ↔ mourut
objectif ultime ↔ but ultime	abritant ↔ qui abrite
reste ↔ demeure	standardisation ↔ normalisation

TABLE 7 – Exemples de paires de paraphrases locales identifiées par `Fastr` avec le jeu de métarègles développé.

Afin d'évaluer les règles sur les révisions de Wikipédia, nous avons construit manuellement un corpus de 200 paraphrases positives (paires de modifications en relation de paraphrases) et 200 négatives (paires de modifications sans lien de paraphrase) à partir du corpus WICOPACO. `Fastr` a identifié 31 paires de paraphrases candidates dans le corpus positif. Parmi elles 22 (70%) sont correctes (la modification est identifiée entièrement comme paraphrase), 7 (22,5%) correspondent à une sous-partie de la modification et 2 (6%) n'existent pas dans la référence (c'est-à-dire qu'elles couvrent une autre partie du contexte). Dans le corpus négatif, seulement 4 paraphrases candidates ont été trouvées, parmi lesquelles une seulement se trouve dans le corpus de référence.

Ces résultats préliminaires montrent que les patrons de réécriture morphosyntaxiques peuvent atteindre une bonne précision pour identifier des paraphrases locales dans les révisions de Wikipédia, mais que parfois des informations

plus fines sur le contexte syntaxique et sémantique sont nécessaires. La couverture obtenue est très limitée du fait de la grande variété de phénomènes concernés, en outre difficilement capturés par les règles développées sur un corpus de développement différent du corpus sur lequel a porté notre évaluation. Par ailleurs, l'étude de plusieurs exemples a révélé que les ressources morphologiques et sémantiques utilisées par `Fastr` pourraient être enrichies afin d'assurer une meilleure couverture pour notre tâche.

6 Conclusions et perspectives

Dans cet article, nous avons décrit une typologie des modifications locales présentes dans les révisions des articles de Wikipédia. Cette typologie pourra servir de repère utile pour des travaux ultérieurs sur cet ensemble de données. Si nous ne l'avons pas formellement démontré, nous pensons que la structure de haut niveau de notre typologie s'applique assez directement quelle que soit la langue étudiée. Nous travaillerons prochainement sur une version anglaise du corpus WICOPACO et testerons alors cette hypothèse.

Nous avons également effectué une annotation manuelle d'un sous-ensemble du corpus étudié. Cette étude a montré qu'une quantité importante de modifications correspondent à des reformulations avec de faibles variations sémantiques. Ceci constitue donc un résultat encourageant dans l'optique d'exploiter les révisions d'une ressource importante et dynamique telle que Wikipédia pour l'acquisition de paraphrases. Nos premières expériences exploitant un moteur de reconnaissance de variantes de termes adapté à nos besoins ont révélé des résultats encourageants, permettant avant tout d'obtenir une bonne précision sur les paraphrases identifiées. Les différentes limites de l'outil utilisé que nous avons identifiées nous ont permis de spécifier un nouveau moteur d'identification par règles plus adapté à nos besoins, permettant l'expression de règles avec une combinaison quelconque de contraintes portant sur le lexique, les constituants, ou les dépendances syntaxiques.

Cependant, la richesse des types de réécriture possibles rendent difficile l'obtention d'un jeu de règles suffisamment couvrant. Nous comptons donc par la suite étudier différents types de classifieurs automatiques avec différents jeux de traits, portant sur les caractéristiques linguistiques des modifications mais également sur les métadonnées des révisions. Ce type d'apprentissage se fondera vraisemblablement sur une quantité de données importante dont nous ne disposions pas encore à ce stade de notre étude. Les données qui pourraient être prochainement disponible *via* un jeu en ligne développé dans notre laboratoire [7] seraient ici particulièrement utiles, puisqu'elles incluent des paraphrases candidates en contexte proposées par des joueurs et des évaluations chiffrées attribuées par d'autres joueurs. Il pourrait donc par exemple s'agir d'un cadre d'annotation original pour les données de WICOPACO.

Le présent travail a montré que les révisions de Wikipédia contiennent de nombreuses réécritures à faible variation sémantique, incluant de nombreuses paraphrases locales. Nos propositions pour l'identification de paraphrases ainsi que nos travaux en cours produiront des listes de paires de paraphrases candidates en contexte, possiblement associées à un score de confiance. Nous comptons par la suite évaluer les paraphrases extraites par la tâche, en bénéficiant des travaux en recherche d'information précise et en traduction automatique menés dans notre laboratoire, qui correspondent à des applications du TAL très sensibles à la variation en langue.

Un autre type d'application sur lequel nous comptons travailler porte sur l'exploitation des patrons de réécriture acquis (incluant les patrons de corrections orthographiques et grammaticaux obtenus par Wisniewski *et al.* (2010)) pour l'aide à la rédaction d'articles sous Wikipédia. Une interface efficace permettrait notamment d'anticiper certaines corrections ultérieures par d'autres contributeurs et ainsi de rendre plus efficace le travail d'un contributeur,

7. Voir (Bouamor *et al.*, 2009) pour une description initiale.

et de réduire globalement le nombre de révisions nécessaires à l'échelle de l'encyclopédie. Par exemple, une normalisation fréquemment apportée aux textes de l'encyclopédie pourrait être suggérée de façon interactive au contributeur qui l'introduirait dans une nouvelle contribution.

Remerciements

Les auteurs tiennent à remercier Julien Boulet, Martine Hurault-Plantet et Guillaume Wisniewski pour leur participation à la création du corpus WICOPACO utilisé dans ce travail, ainsi que les très nombreux contributeurs à la création du corpus MULTITRAD.

Références

BOUAMOR H. (2010). Construction d'un corpus de paraphrases d'énoncés par traduction multiple multilingue. In *Actes de RÉCITAL 2010*, Montréal, Canada.

BOUAMOR H., MAX A. & VILNAT A. (2009). Amener des utilisateurs à créer et évaluer des paraphrases par le jeu. In *Actes de TALN, session de démonstrations*, Senlis, France.

BOUAMOR H., MAX A. & VILNAT A. (2010). Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases. In *Proceedings of the 7th International Conference on NLP, IceTAL 2010*, volume 6233 of *Lecture Notes in Computer Science, Advances in Natural Language Processing* : Springer Berlin / Heidelberg.

BOURDAILLET J. & GANASCIA J.-G. (2007). Machine Assisted Study of Writers' Rewriting Processes. In *Proceedings of the International Workshop on Natural Language Processing and Cognitive Science (NLPCS 2007)*.

CHRISTIAN JACQUEMIN (1994). Recycling terms into a partial parser. In *Proceedings of the fourth conference on Applied natural language processing*, Stuttgart, Germany.

DELÉGER L. & ZWEIGENBAUM P. (2009). Extracting lay paraphrases of specialized expressions from monolingual comparable medical corpora. In *Proceedings of the Workshop on Building and Using Comparable Corpora : from Parallel to Non-parallel Corpora*, Singapore.

DOLAN W. B. & BROCKETT C. (2005). Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

DUCLAYE F., COLLIN O. & YVON F. (2003). Apprentissage automatique de paraphrases pour l'amélioration d'un système de questions-réponses. In *Actes de TALN*, Batz-sur-mer, France.

DUTREY C., BOUAMOR H., BERNHARD D. & MAX A. (2011). *Typologie des modifications dans les révisions de Wikipédia*. Notes et documents du LIMSI 2011-01, LIMSI-CNRS.

F. CHEVALIER, S. HUOT ET J-D. FEKETE (2010). Visualisation de mesures agrégés pour l'estimation de la qualité des articles Wikipédia. In *Extraction et gestion des connaissances (EGC'2010), Actes, Hammamet, Tunisie, 26 au 29 janvier 2010*.

GERMANN U. (2008). Yawat : Yet Another Word Alignment Tool. In *Proceedings of the ACL-08 : HLT Demo Session*.

GLICKMAN O., DAGAN I. & KOPPEL M. (2005). A probabilistic classification approach for lexical textual entailment. In *Proceedings of the 20th national conference on Artificial intelligence (AAAI'05)* : AAAI Press.

HU M., LIM E.-P., SUN A. & LAUW, HADY WIRAWANAND VUONG B.-Q. (2007). Measuring article quality in Wikipedia : models and evaluation. In *CIKM '07 : Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Lisbon, Portugal : ACM.

MADNANI N. & DORR B. J. (2010). Generating Phrasal & Sentential Paraphrases : A Survey of Data-Driven Methods. *Computational Linguistics*, **36**(3).

MAX A. (2008). Génération de reformulations locales par pivot pour l'aide à la révision. In *Actes de TALN*, Avignon, France.

MAX A. (2010). Example-based paraphrasing for improved phrase-based statistical machine translation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*.

MAX A. & WISNIEWSKI G. (2010). Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In *Proceedings of LREC 2010*, Valletta, Malta.

NELKEN R. & YAMANGIL E. (2008). Mining Wikipedia's Article Revision History for Training Computational Linguistic Algorithms. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence :An Evolving Synergy*.

SUH B., CHI E., KITTUR A. & PENDLETON B. (2008). Lifting the veil : improving accountability and social transparency in Wikipedia with Wikidashboard. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems* : ACM.

VIÉGAS F., WATTENBERG M. & DAVE K. (2004). Studying Cooperation and Conflict Between Authors With History FlowVisualization. *Proceedings of the SIGCHI conferenceonHuman Factors in Computing Systems (CHI'04)*.

WISNIEWSKI G., MAX A. & YVON F. (2010). Recueil et analyse d'un corpus écologique de corrections orthographiques extrait des révisions de Wikipédia. In *Actes de TALN 2010*, Montréal, Canada.

YATSKAR M., PANG B., DANESCU-NICULESCU-MIZIL C. & LEE L. (2010). For the sake of simplicity : Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of the NAACL*.

ZANZOTTO F. M. & PENNACCHIOTTI M. (2010). Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the 2nd Workshop on Collaboratively Constructed Semantic Resources*.

ZESCH T., MÜLLER C. & GUREVYCH I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Language Resources and Evaluation(LREC'08)*, Marrakech, Morocco.