Procesamiento del Lenguaje Natural, Revista nº 46 marzo de 2011, pp 51-58 recibido 14-01-11 revisado 10-02-11 aceptado 24-02-11

Local Modifications and Paraphrases in Wikipedia's Revision History*

Modificaciones locales y paráfrasis de la historia de revisiones de Wikipedia

Camille Dutrey
INALCO
Paris, France
camille@dutrey.fr

Delphine Bernhard
LIMSI-CNRS
Orsay, France
delphine.bernhard@limsi.fr

Houda Bouamor and **Aurélien Max**
LIMSI-CNRS, Univ. Paris-Sud
Orsay, France
{houda.bouamor,aurelien.max}@limsi.fr

Resumen: En éste artículo, se analizan las modificaciones accesibles a través del historial de revisiones de Wikipedia en francés. Se define una tipologia de modificaciones basada en el estudio detallado de WiCoPaCo, un recurso gratuito construido a través de un estudio del historial de revisiones de Wikipedia. Conforme a ésta tipologia, detallamos el estudio de la anotación manual de un subconjunto del corpus, con la intención de evaluar la dificultad de la tarea de identificación automática de paráfrasis en el mismo corpus. Finalmente, evaluamos una herramienta de identificación de paráfrasis a base de reglas.
Palabras clave: Wikipedia, revisiones, identificación de paráfrasis

Abstract: In this article, we analyse the modifications available in the French Wikipedia revision history. We define a typology of modifications based on a detailed study of WiCoPaCo, a freely-available resource built by automatically mining Wikipedia's revision history. Based on this typology, 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.
Keywords: Wikipedia, text revisions, paraphrase identification

1 Introduction

Wikipedia keeps growing to be the world's largest and busiest free encyclopedia, in which articles are collaboratively written and maintained by volunteers online. The huge amounts of quality data in this encyclopedia have motivated many works on automatic acquisition of resources, e.g. acquisition of lexical-semantic knowledge (Zesch, Müller, and Gurevych, 2008).

Obviously, a majority of these studies use only the most recent version of the articles. In fact, besides the latest version, Wikipedia provides the entire revision history of each of its articles which are iteratively amended and refined by multiple Web users. This resource has been exploited previously for different tasks and applications.

Nelken and Yamangil (2008) exploit Wikipedia's revision history to acquire voluminous training data for three separate text processing tasks at different levels of linguistic granularity: collection of textual errors and their correction (single word level), training data for sentence compression algorithms (sentence level) and bootstrapping data for text summarization systems (document level) by comparing adjacent versions of the same article. Yatskar et al. (2010) utilise edit histories in Simple English Wikipedia to extract lexical simplifica-

* The authors thank Julien Boulet, Martine Hurault-Plantet and Guillaume Wisniewski for their participation in the creation of the WiCoPaCo corpus, as well as the contributors to the Multitrad corpus. This work was supported by a grant from LIMSI.

ISSN 1135-5948

tions. Max and Wisniewski (2010) describe WiCoPaCo (Wikipedia Correction and Paraphrase Corpus), a freely-available resource built by automatically mining Wikipedia's revision history and extracting local modifications made by human revisers which includes various types of corrections (such as spelling errors or typographical corrections) and rewritings. Finally, Zanzotto and Pennacchiotti (2010) use Wikipedia's revision history to extract a large set of textual entailment pairs and apply semi-supervised machine learning methods to make the extracted dataset homogeneous to the existing ones.

As Wikis and other collaborative information repository systems grow in popularity and use, issues concerning the trustworthiness of information become increasingly important. In particular, Hu et al. (2007) developed a revision history-based fragment trust model to compute and monitor the trustworthiness of Wikipedia articles and article fragments.

As we have shown, most previous works on Wikipedia's revision history focus on specific aspects of the resource and target well-defined applications such as text simplification, sentence compression or textual entailment. To our knowledge, there is no comprehensive overview of the local modification phenomena available in Wikipedia's revisions, although there is a large variety of types of local modifications that are of interest for many NLP applications.

In this article we detail a typology of local modifications found in the WiCoPaCo corpus, with a particular focus on the phenomenon of *local paraphrases*, which are increasingly employed to improve the performance of several NLP applications such as Machine Translation or Information Retrieval systems. This article is organized as follows: we first describe the WiCoPaCo corpus that was used in this study in section 2, and we then present our proposed typology of local modifications in section 3. Section 4 details an annotation study based on this typology and section 5 reports on initial experiments for automatic rule-based paraphrase identification. Finally, section 6 contains concluding comments and references to our future work.

2 The WiCoPaCo Corpus

The acquisition of pairs of short text spans with equivalent meaning (*local paraphrases*) has attracted a lot of work on automatic mining of text corpora (see e.g. (Madnani and Dorr, 2010)). The text corpora that have been used can be roughly organized by the degree of correspondence between two units of text: pairs of sentential paraphrases, obtained e.g. by multiple translation (*monolingual parallel corpora*); pairs of sentences with similar content, obtained e.g. by filtering high-similarity pairs from groups of related documents (*monolingual comparable corpora*); pairs of phrases sharing common translations in other languages (*multilingual parallel corpora*). The first case is interesting as it involves pairs of text units which are supposed to convey the same content from which it is reasonable to assume that high-quality local paraphrases can be automatically acquired (Bouamor, Max, and Vilnat, 2010). Unfortunately, such corpora do not exist in large quantities and are costly to build. A major shortcoming of the other types of corpora used for paraphrase acquisition is that potential paraphrases are only observed indirectly via common translation or contextual similarity.

Another potential source of local paraphrases lies in the possibly numerous modifications that writers make when revising a text, some of them being intended not to alter the meaning of the text but to improve its quality, to make it more coherent, or to limit redundancy. Drafts of famous writers are for example used in *textual genetic criticism* (e.g. (Bourdaillet and Ganascia, 2007)) which studies the process of text creation. Unfortunately, such annotated documents are available in small quantities and are furthermore difficult to encode into electronic form. Moreover, such drafts often contain important textual reorganizations which are too difficult to exploit for paraphrase acquisition.

The emergence and wide adoption of wikis has made collaborative writing a very common practice. The Wikipedia online encyclopedia, in particular, attracts many contributions on a broad range of subjects and in many languages. While some contributions consist in important changes (e.g. creation of an article, removal of a section, complete rewriting of a paragraph), a significant proportion of textual modifications are made on

```
<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: Example of a modification in the WiCoPaCo corpus.

short text spans by human contributors to correct, improve or augment the content of the encyclopedia. The revision history of such resources thus constitutes an important source of *naturally-occurring* rewriting phenomena, including local paraphrases in context.

The WiCoPaCo corpus[1] aims at collecting instances of such local modifications from Wikipedia's revision history, which is available as dumps with numerous metadata for most versions of Wikipedia, using an approach which is mostly language-independent. The construction of the corpus comprises four main steps: 1) selection of document pairs; 2) text normalization; 3) modification alignment; and 4) modification filtering. Any pair of consecutive revisions for a given article made by a human contributor can constitute a document pair, which implies a heavy computational load to process large dumps, but which can be straightforwardly parallelized. The text of these revisions is then normalized by removing some types of wiki elements such as tables, wiki text, and tokenizing the text if appropriate. An algorithm for finding longest common subsequences, as implemented in the standard `diff` tool, is then applied to extract all pairs of modifications involving 7 words at most. A modification can thus be described as a left and right context and a text span before and after the modification. The enclosing paragraph of the extracted modification is also extracted. As other modifications may have occurred at different places in the paragraph for the same revision, the full paragraphs before and after the modification are in fact recorded to describe the context of each modification. Various filters are then applied to filter out some types of small modifications involving e.g. case or punctuation modifications as well as some possibly significant rewritings by using a threshold on the ratio of common words in the enclosing sentence before and after the modification. An output XML file is finally produced with a new element for each individual modification, which is associated with various useful metadata, including complete reference to the original Wikipedia revision and the contributor of the modification, as illustrated in the example on Figure 1. In this work, the French version of WiCoPaCo, which contains 408,816 entries, was used.

[1]Freely downloadable from `http://wicopaco.limsi.fr/`

3 Typology of Local Modifications

We analysed the WiCoPaCo corpus in order to develop a detailed typology of local modifications in Wikipedia revisions (Dutrey et al., 2011).[2] The typology aims at representing all observable phenomena in the WiCoPaCo corpus and accounts for the degree of semantic variation between local modification segments. Therefore, it consists of two major categories distinguishing between two broad classes of semantic variation: the *weak semantic differences* class and the *strong semantic differences* class.

3.1 Weak Semantic Differences

Weak semantic differences encompass *surface corrections* (see Table 1) and *rephrasings* (see Table 2).

Surface corrections refer to surface changes which aim at improving the text so that it conforms to linguistic norms:

- *Typographical corrections* consist in changing the layout and format of the text, e.g., adding/removing blanks or punctuation marks, changing the case of a character, modifying the format of a date or an hour, writing a number in full or with numerals etc.

[2]Due to space constraints, we are not able to detail the full typology. We refer the interested reader to the full description available on the WiCoPaCo website: `http://wicopaco.limsi.fr`

Surface Corrections
Typographical corrections
⇒ e.g. a space replaced by a hyphen to correct a typographical error: *Le triceps brachial est un muscle extenseur de l' **[avant bras → avant-bras]** sur le bras.* eng: The triceps is an extensor muscle of the [**forearm**] on the arm.
Non-Word spelling corrections
⇒ e.g. an alphabetic character deleted to change a non-word into an attested word: *Ces trois parties se **[rejoingnent → rejoignent]** pour former une épaisse masse.* eng: These three parts [**come together**] to form a thick mass. ⇒ e.g. a diacritic replaced by another to change a non-word into an attested word: *L' **[ëglise → église]** gothique Sainte-Marie...* eng: The gothic **church** St. Mary...
Context-dependent word corrections
⇒ e.g. a diacritic replaced by another to correct a real-word error: *L'anathème pour le **[pêcheur → pécheur]** : ce dernier est privé de sépulture chrétienne.* eng: A curse for the [**fisherman → sinner**]: he is deprived of Christian burial. ⇒ e.g. a word replaced by another to correct a real-word error: *Il chante avec une **[voie → voix]** de troubadour.* eng: He sings with the [**way → voice**] of a troubadour.

Table 1: Weak Semantic Differences: surface corrections

- *Non-word spelling corrections* affect non-word spelling errors and involve e.g. the modification of diacritics (accented characters) or the replacement of one or several characters.
- *Context-dependent word corrections (real words)* resolve real-word spelling errors which can only be detected and corrected by taking the context into account.

Rephrasings correspond to more significant changes which modify the lexical and syntactic choices made by the previous contributor without strongly altering the text's meaning:

- *Lexical rephrasings* consist in e.g., replacing an acronym by its full name, translating a foreign or loan word, replacing a regional variant by its standard variant, changing the part of speech of a word, etc.
- *Syntactical rephrasings*, which e.g., modify the order of clauses, tranform to active or passive voice or change the type of a clause.
- *Semantic rephrasings* consist in e.g. using hypernyms or hyponyms, performing encyclopaedic normalisation, using synonyms or adding extraneous information.

3.2 Strong Semantic Differences

Strong semantic differences encompass factual corrections and vandalism (see Table 3). In this case, the meaning of the text is strongly affected and can be totally changed.

Factual corrections correspond to any modification which induces a strong change in meaning. They consist in e.g. replacing a word by an antonym or changing the tense of a verb so that the meaning of the sentence is modified. Factual corrections aim at ameliorating Wikipedia's contents.

Vandalism refers to modifications which deliberately alter or destroy contents in order to damage Wikipedia's quality. Obvious vandalism is characterised by the insertion of non-words or insults while subtle vandalism is harder to detect since contextually inconsistent real words are inserted.

4 Manual Annotation

We designed an annotation schema based on the typology previously described. The goal of the annotation study was to assess the difficulty of manually identifying paraphrases within local modifications. The annotation was therefore driven by our target application which is automatic paraphrase identification.

In our typology, paraphrases roughly correspond to rephrasings, within the broader class of weak semantic differences. They have to be distinguished from surface correc-

Rephrasings
Lexical rephrasings
⇒ e.g. an acronym replaced by its full form: *L'Autriche est membre de l'**[UE]** → L'Autriche est membre de l'**[Union Européenne]** ...* eng: Austria is a member of the [**EU**] → Austria is a member of the [**European Union**] ...
Syntactical rephrasings
⇒ e.g. switching between segments on the syntagmatic axis: ***[l'Invention de l'Europe d'Emmanuel Todd → Emmanuel Todd, L'Invention de l'Europe]**.* eng: [**The Invention of Europe by Emmanuel Todd → Emmanuel Todd, The Invention of Europe**]. ⇒ e.g. a circumstantial clause changed into a relative clause: *Un infomercial pseudo-scientifique **[en exposant → qui expose]** grossièrement...* eng: A pseudoscientific infomercial [**in** roughly **outlining** → **which** roughly **outlines**]...
Semantic rephrasings
⇒ e.g. a word replaced by another from the same lexical field (infra hyponymy): *Il fonde le **[journal → quotidien]** francophone "Le Tunisien" en 1907.* eng: He founded the French-speaking [**newspaper → daily paper**] "Le Tunisien" in 1907. ⇒ e.g. paraphrasing serving different purposes (infra precision of meaning): *Ce vers de Nuit rhénane d'Apollinaire **[qui paraît presque sans structure rythmique → dont la césure est comme masquée]**...* eng: This verse from Apollinaire's Nuit Rhénane [**which seems almost without rhythmic structure → whose cesura is as if hidden**]...

Table 2: Weak Semantic Differences: rephrasings

Factual Corrections
⇒ e.g. a word replaced by an antonym: *Un catalyseur solide (phase **[liquide → solide]**) avec de l'hydrogène (phase gazeuse).* eng: A solid catalyst ([**liquid → solid**] phase) with hydrogen (gas phase). ⇒ e.g. a segment replaced by another segment without semantic link: *représente pour eux **[l'Occident chrétien → la supériorité de la race celto-germanique]**.* eng: represents for them [**the Chistian West → the superiority of the Celtic-Germanic race**].
Vandalism
⇒ e.g. an inserted string producing a non-word (obvious vandalism): *L'Autriche a été occupée **[par → psh ! ! ar]** les Romains.* eng: Austria was occupied [**by → bsh ! ! y**] the Romans. ⇒ e.g. a word replaced by another which doesn't make sense in context (subtle vandalism): *Devant la Cour de **[Cassation → Castration]**...* eng: In front of the Court of [**Cassation → Castration**]...

Table 3: Strong Semantic Differences

tions, which are not relevant for the task of paraphrase identification, and strong semantic differences which induce a major change in meaning. To this aim, we developed an annotation schema consisting of four main classes:

- *Surface corrections*, which encompass all modifications which aim at making the text compliant with language standards.
- *Rephrasings*, which correspond to different kinds of paraphrases, including reformulations, precisions and simplifications.
- *Strong semantic variations*, including vandalism and revisions.
- *Misalignments* which correspond to cases where the local modifications identified present a default in their alignment. However, even with a misaligment a segment might contain a local modification.

An annotation covers the entire segment identified as a local modification (denoted by `m` XML elements in the corpus, as illustrated on Figure 1): the goal is to determine the modification's type from a pair of segments but not to re-align words under those segments. Moreover, it was possible to assign several labels to the same modification segment.

For the annotation, we used the Yawat (Germann, 2008) tool originally designed for the alignment of parallel texts at the word and phrase level. We adapted the tool's annotation scheme for our multi-level annotation. The annotation was performed by four trained annotators[3] on 200 pairs of modification segments taken from a filtered version the WiCoPaCo corpus. As punctuation modifications are frequent, only modification segments with a Levenshtein edit distance of at least 4 were considered for annotation.

4.1 Annotation Results

Table 4 describes the inter-annotator agreement for our annotation, as indicated by the κ statistics.[4] The inter-annotator agreement ranges from moderate to substantial, depending on the class. Overall, the κ values are close to the values reported by Dolan and Brockett (2005) for paraphrase identification (κ of 0.62) and by Glickman, Dagan, and Koppel (2005) (κ of 0.6) for textual entailement.

We also report the number of identical annotations assigned by 2 to 4 annotators, as well as annotations assigned by a unique annotator (see Table 5). This makes it possible to roughly quantify the phenomena available in the corpus. Interestingly, rephrasings have the largest number of occurrences, followed by strong semantic variations. This tends to prove that Wikipedia revisions constitute a well-suited corpus for the automatic acquisition of paraphrases. Moreover, misalignements are quite few, which shows that the alignment method used for building the WiCoPaCo corpus is precise enough to provide useful modifications.

The annotation study highlighted some potential problems for the automatic identification of the classes described in our typology. First, several phenomena may occur simultaneously, e.g. a diathesis (grammatical voice) transformation may include a correction on a non-word error. In this case, an automatic classifier should be able to assign several classes to a modification segment. Second, the sentential context provided by the WiCoPaCo corpus is sometimes not sufficient to make a decision about a specific modification type. A larger context could therefore be useful for automatic classifiers. Third, correctly typing a modification may necessitate some external knowledge about the contributor's intentions. This kind of information is sometimes available in the comments associated with a revision, however it may be hard to interpret automatically.

5 *Rule-based Paraphrase Identification*

After this annotation study, we developed an identification method designed to distinguish paraphrases from other modifications. To this goal, we used an adaption of the `Fastr` tool (Christian Jacquemin, 1994), which was originally developed for the recognition of term variants.

`Fastr` identifies term variants in a text corpus using pre-defined transformation rules (defined via metarules applied on term rules) relying on POS tags assigned by the TreeTagger.[5] More precisely, we created a new set of metarules for paraphrase recognition,[6] but we reused the standard resources (morphological and semantic families) provided with the software. The goal of this experiment was to assess whether a rule-based tool is suited for the identification of paraphrases in Wikipedia revisions. We used a different corpus for the development of metarules, in order to verify whether the rules are general enough to be applied to different corpora. The corpus used for rule development was taken from the MultiTrad dataset which was built by collecting several translations for the same input text during a web-based experiment (Bouamor, 2010).

Figure 2 displays an example metarule. This rule applies to a source segment with

[3] Co-authors of the present article.

[4] We used the online κ calculator for multiple annotators and multiple classes available at: `http://cosmion.net/jeroen/software/kappa/`

[5] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger`

[6] The original set of metarules was deemed inappropriate for our study as it was developed with the objective of high recall for term variant recognition.

Type	Avg. κ	Interpretation	Maximum κ	Minimum κ
Strong semantic variation	0.65	Substantial agreement	0.71	0.61
Rephrasing	0.60	Moderate agreement	0.71	0.51
Surface correction	0.54	Moderate agreement	0.81	0.40
Misalignement	0.48	Moderate agreement	0.62	0.28

Table 4: Inter-annotator agreement for the annotation of Wikipedia revisions.

	4 ann.	3 ann.	2 ann.	unique	Total
Surface correction	9	2	7	23	41
Rephrasing	60	33	24	15	132
Strong semantic variation	47	15	13	32	107
Misalignment	2	4	8	6	20

Table 5: Number of identical annotations assigned by 2, 3 or 4 annotators and annotations made by one annotator only (unique).

MultiTrad
décrit dans la proposition ↔ proposé
objectif ultime ↔ but ultime
docteurs ↔ médecins
WiCoPaCo
décéda ↔ mourut
abritant ↔ qui abrite
standardisation ↔ normalisation

Table 6: Example paraphrase pairs identified by `Fastr`.

a noun followed by an adjective and identifies variants with a verb, followed either by an article, a pronoun or a preposition, followed by a noun and an adjective. The rule also integrates some morphological and semantic constraints which specify (i) that the noun in the source segment and the verb in the target segment share an identical morphological root and (ii) that the adjectives in the source and target segments are synonymous. Note that this approach is substantially comparable to that implemented in the work of Deléger and Zweigenbaum (2009) on the extraction of lay paraphrases of specialized expressions. Overall, we developed 83 metarules for paraphrase identification. We first assessed the coverage of the manually built rules on a sub-part of the Multitrad corpus (206 sentence pairs) which was not used for rule development. `Fastr` was able to identify 185 paraphrase candidates, some of which are illustrated in Table 6.

In order to further evaluate the rules on Wikipedia revisions, we manually built a corpus of positive and negative paraphase examples (200 of each type) from the WiCoPaCo corpus. `Fastr` identified 31 pairs of candidate paraphrases in the positive corpus. Among these, 22 (70%) are correct (i.e. the whole modification is identified as a paraphrase), 7 (22.5%) correspond to a subpart of the modification, and 2 (6%) do not exist in the reference (i.e. they cover another part of the context sentence). In the negative corpus, only 4 candidates were identified, among which only 1 exists in the reference modification corpus. These preliminary results show that morpho-syntactic rewriting patterns can achieve a good precision to identify local paraphrases in Wikipedia revisions. However, their coverage is very limited by the range of phenomena, which is too wide to be captured by rules developed on a different corpus. Moreover, examination of several examples revealed that the morphological and semantic resources used by `Fastr` could be enriched to provide better coverage for our task.

6 Conclusions and Future Work

In this article, we defined a detailed typology of the local modification phenomena which are present in WiCoPaCo, a corpus of natural rewritings extracted from Wikipedia's revision history. This typology should help fostering future research on this dataset. We also performed a manual annotation of a subset of the corpus. This study showed that a substantial amount of modifications correspond to rephrasings with weak semantic differences, i.e. paraphrases. Finally, we evaluated a rule-based approach to paraphrase identification in Wikipedia's revision history. While this approach yields very precise results, it is not able to account for the diver-

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 eng: constant protection → protect in a permanent fashion

Figure 2: An example `Fastr` metarule

sity of phenomena available in the corpus.

In the future, we plan to combine the rule-based approach with machine-learning methods for the automatic identification of paraphrases in order to constitute a large-scale resource of paraphrases extracted from local modifications in Wikipedia.

References

Bouamor, Houda. 2010. Construction d'un corpus de paraphrases d'énoncés par traduction multiple multilingue. In *Actes de RÉCITAL 2010*.

Bouamor, Houda, Aurélien Max, and Anne Vilnat. 2010. Comparison of Paraphrase Acquisition Techniques on Sentential Paraphrases. In *Proceedings of the 7th International Conference on NLP (IceTAL 2010)*.

Bourdaillet, Julien and Jean-Gabriel Ganascia. 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*.

Deléger, Louise and Pierre Zweigenbaum. 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*.

Dolan, William B. and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Dutrey, Camille, Houda Bouamor, Delphine Bernhard, and Aurélien Max. 2011. Typologie des modifications dans les révisions de wikipédia. Notes et documents du LIMSI 2011-01, LIMSI-CNRS.

Germann, Ulrich. 2008. Yawat: Yet Another Word Alignment Tool. In *Proceedings of the ACL-HLT 2008 Demo Session*.

Glickman, Oren, Ido Dagan, and Moshe Koppel. 2005. A probabilistic classification approach for lexical textual entailment. In *Proceedings of AAAI 2005*.

Hu, Meiqun, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. 2007. Measuring article quality in wikipedia: models and evaluation. In *Proceedings of CIKM '07*.

Madnani, Nitin and Bonnie J. Dorr. 2010. Generating Phrasal & Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*.

Max, Aurélien and Guillaume Wisniewski. 2010. Mining Naturally-occurring Corrections and Paraphrases from Wikipedia's Revision History. In *Proceedings of LREC 2010*.

Nelken, Rani and Elif Yamangil. 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*.

Yatskar, Mark, Bo Pang, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2010. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proceedings of NAACL 2010*.

Zanzotto, Fabio Massimo and Marco Pennacchiotti. 2010. Expanding textual entailment corpora from Wikipedia using co-training. In *Proceedings of the 2nd Workshop on Collaboratively Constructed Semantic Resources*.

Zesch, Torsten, Christof Müller, and Iryna Gurevych. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of LREC 2008*.