French Wikipedia Talk Pages: Profiling and Conflict Detection

Ho-Dac L.-M.(*), Laippala V.(**), Poudat C.(***), and Tanguy L.(*)

(*) CLLE, University of Toulouse, CNRS, UT2I, 5 allées A. Machado, 31058 Toulouse CEDEX 9, France
(**) TIAS, University of Turku, 0014 Turun yliopisto, Finland
(***) BCL, University of Nice Sophia Antipolis, 24, avenue des Diables bleus, 06357 Nice CEDEX 4, France

E-mail: hodac@univ-tlse2.fr, mavela@utu.fi, celine.poudat@unice.fr, tanguy@univ-tlse2.fr

Abstract

Wikipedia is a popular and extremely useful resource for studies in both linguistics and natural language processing (Yano and Kang, 2008; Ferschke et al., 2013). This paper introduces a new language resource based on the French Wikipedia online discussion pages, the WikiTalk corpus. The publicly available corpus includes 160M words and 3M posts structured into 1M thematic sections and has been syntactically parsed with the Talismane toolkit (Urieili, 2013). In this paper, we present the first results of experiments aiming at classifying and profiling the talk pages and threads in order to determine criteria for selecting discussions with conflicts.

Keywords: French Wikipedia talk pages, conflict detection, data-driven approaches

1. Introduction

With the exponential development of the Internet, new communicative situations and new genres have come about. The new web genres, which are not yet fully characterized, are complex objects challenging the existing methodologies and analysis tools: the Wikipedia encyclopedic project is one of these new textual objects that can be studied under the umbrella term Computer-Mediated Communication (CMC, (Herring et al., 2013)). Wikipedia, which celebrates its 15th birthday this year, is an open and collaborative project, available in numerous languages. The success of the web encyclopedia is indisputable, as evidenced by its number of visitors and number of articles. However, the structure of Wikipedia remains little explored. In particular, talk pages are complex objects challenging the existing methodologies and analysis tools: the Wikipedia encyclopedic project is one of these new textual objects that can be studied under the umbrella term Computer-Mediated Communication (CMC, (Herring et al., 2013)).

Wikipedia talk pages may be considered as a new discourse sub-genre. Wikipedia editorial talk pages are indeed quite specific: (i) they are directly related to the article they are associated with, and they share a common focus, i.e. article edition and improvement; (ii) they contain open asynchronous discussions that anyone may edit. In that respect, they might be compared to forum discussions except that they rely on a specific Wiki device which has direct consequences on the macrostructure: in spite of clear recommendations concerning the form of the postings (level of the answer, mandatory signature and date, etc.), talk pages are often hybrids, combining dialogues whose structure may not be obvious (as Wikipedia may for instance edit previous postings), and checklist elements; (iii) they share common features referring notably to editing actions, conflict
management and Wikipedia procedures (e.g. NPOV, i.e. Neutrality of Point of View, relevance, source, quality etc.). Conflicts are particularly interesting to observe in Wikipedia, since they can be considered as frontiers between collaboration and discussion. Antagonistic edits of the article structure and content may indeed lead to disagreements and this is quite usual when co-editing, before participants agree on a more stable version of the article. Disagreements may turn to conflicts when the editing process and/or the discussion process are deadlocked, which leads to an automated report. In such cases, pages are tagged with specific labels signaling that a conflict is ongoing on the article or talk pages (e.g. NPOV or relevance disputes, "Keep calm" banner). Examples of pages with such labels are quite numerous: Abortion in Iran, Bengali cuisine, List of Volvo trucks to cite just a few.

The aim of the present study is twofold: at a descriptive level, we would like to contribute to the linguistic description of Wikipedia talk pages, which have been little explored using linguistic criteria. In particular, few linguistic studies have been conducted on French Wikipedia (see Denis et al., 2012) on the detection of conflicting threads or (Poudat and Loiseau, 2007) on the exploration of Wikipedia categories. We will first perform an automatic classification on the entire set of French Wikipedia talk pages, which were gathered within the WikiTalk Corpus, making the most of the French "Appel au calme" (keep calm) label, signaling ongoing conflict(s) on the talk page. In order to have a broader view of the linguistic characteristics of the French Wikipedia talk pages, We will then propose a profiling of the genre, using a multidimensional analysis enabling us to highlight key features and oppositions at a global level. Conflicting threads and pages will be characterized within this global generic profile.

2. WikiTalk Corpus

The WikiTalk corpus is composed of talk pages extracted from the French Wikipedia dump dated May 12th 2015 which contains 3.5M talk pages. Only 365,612 pages were kept in the released WikiTalk Corpus. Indeed, 57% of the talk pages were user pages and we chose to remove them, even if these talk pages are basically online discussions. Only 24% of the remaining talk pages contained more than two words. The 365,612 remaining talk pages were segmented into threads and posts based on the wikicode. Threads correspond to divisions delimited by (sub)headings signaled by the wiki markup: /==.*?==/. Posts are delimited according to

1. timestamp and an optional user signature, such as: Viking59 10 mai 2009 à 17:16 (CEST); or
2. a change in the interactional level indicated by the number of semi-colons (:) in the beginning.

Once threads and posts were delimited, all discussions were formatted according to the TEI-P5 guidelines. Metadata are encoded in the teiHeader as illustrated below with the <classDecl> element.

Discussion structure is encoded according to the following TEI elements:

- <div> for threads
- <head> for topic titles and
- <post who="user" when="timestamp" interactionalLevel="#"> for posts.

Table 1 gives a quantitative overview of the WikiTalk corpus.

<table>
<thead>
<tr>
<th>discussions</th>
<th>sections</th>
<th>posts</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>365,612</td>
<td>1,023,841</td>
<td>2,406,514</td>
<td>161,833,298</td>
</tr>
</tbody>
</table>

Table 1: Quantitative overview of the WikiTalk corpus.

Eight of the extracted talk pages, amounting to 413 posts and 47,284 tokens, were manually inspected to evaluate the extraction process. Results show that 23 posts were not extracted at all and 33 posts were wrongly delimited, among which 25 merged several posts in one. As a result, the extraction process has an estimated precision of 0.92 and a recall of 0.95. Post attribute values (@who, @when and @interactionalLevel) were only checked for one talk page but indicated 100% accuracy.

1. 1,013,791 (68%) talk pages were blank and 116 432 (8%) consisted in redirections to another talk page.


3. Classification of Conflicting vs. Peaceful Talk Pages

The first tested method consisted in a data-driven comparison of the global linguistic characteristics of two classes of talk pages, distinguished according to an experimental classification of "conflicting" vs. "peaceful" talks. The selection criteria used for distinguishing between these two classes are based on the Wikipedians’ assessment of the article’s quality and the Wikipedians’ alert regarding conflict or impoliteness in a talk page. Moreover, only talk pages containing more than 100 words were taken into account. Among those, 2,028 a priori “conflicting” talks (11M words) were selected according to the following criteria:

- a parallel talk page was created for discussing the article’s neutrality;
- the page itself is a parallel talk page created for discussing the article’s neutrality.

Criterion for selecting 4,569 a priori "peaceful" talks (8.8M words) are the following:

- a parallel talk page was created for discussing the article’s "featured" or "A-class";
- a parallel talk page was created for discussing the article’s neutrality.

For the purpose of evaluating our distinction between these two classes while also determining features that may be used for selecting talk pages where conflicts may occur, we trained a text classification model using the Vowpal Wabbit linear classifier (Agarwal et al., 2011). In addition to being fast and easily adjustable to large corpora, it has the advantage of generating a list of the most significant features and their relative weights.

Two feature sets were tested for the classification task: lexical features and syntactic features. Classification based on lexical features which considers texts as bags-of-words or bags-of-lemmas is the traditional approach, as for example (Scott et al., 2006) which propose a keyword definition method with two-thirds of the corpus used for training and the remaining for testing. As lexical features we use lemmas; as syntactic features we use unlexicalized bi-arcs composed of two syntax dependencies between tokens with the actual lexical information deleted but with all other information on the syntactic dependency, Part-of-Speech and other morphological features, as illustrated in Fig. 1.

![Figure 1: A delexicalized syntactic bi-arc describing a clitic+verb+conjunction as in the clause ‘I find that’](image)

Syntactic analysis and lemmatisation were provided by the Talismane toolkit (Urieli, 2013). Two levels of text segments were considered: threads and posts. Entire pages were not taken into account because a conflict usually happens inside a thread. In addition, our previous experiments on the page-level have already shown higher scores for the bag of words method (Ho-Dac and Laippala, 2015). In the analysis, we consider, however, that all the posts and threads in a page labeled as conflicting / peaceful are in the same category. Table 2 gives the precision (P) and recall (R) for detecting the "conflict" category by using the two feature sets on threads and posts.

<table>
<thead>
<tr>
<th>features</th>
<th>threads</th>
<th></th>
<th>posts</th>
</tr>
</thead>
<tbody>
<tr>
<td>lemmas</td>
<td>0.84</td>
<td>0.60</td>
<td>0.79</td>
</tr>
<tr>
<td>bi-arcs</td>
<td>0.55</td>
<td>0.48</td>
<td>0.63</td>
</tr>
<tr>
<td>units</td>
<td>46,690</td>
<td>194,289</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Comparison of lexical vs. syntactic approaches for the automatic classification of conflicting threads and posts.

Results show that the best method for detecting conflict seems to be a classification of threads by using a lexical approach. A closer look on the threads classified with high probability and on typical bi-arcs used by the classifier is necessary for better understanding.

Even if the precision of more than 80% seems encouraging, we must admit that these results lead us to question both the features used for classification and our a priori definition of a conflicting talk. Next sections begin to address these questions by proposing a range of new features for profiling Talk pages in a bottom-up approach and presenting a current project of conflict manual annotation in the WikiTalk corpus.

4. A Bottom-Up Approach to Talk Page Profiling

The automatic classification was supplemented by a second approach which uses statistical techniques based on linguistic features and portals information for discovering talk pages and thread profiles in a bottom-up approach, without a focus on conflict. This method considered all the 366,612 talk pages and used the R package FactoMineR dedicated
to multivariate exploratory data analysis6. Each talk page and thread was automatically described with four types of features:

- **TEMA**: portal sections of the associated article page knowing that an article may be categorized as belonging such as Art, History, Sport7 up to 7 of the 11 possible Wikipedia sections (these 11 variables were binarised);
- **GLOBAL**: general quantitative characteristics (number of words and posts) and, for entire talk pages, amount of threads and different contributors, proportion of anonymous posts;
- **INTERACT**: the frequency of a wide range of interaction and politeness cues per talk pages and threads (social deixis, marks of agreement and disagreement);
- **DISCREL**: the frequency of connectives for each discourse relations as defined in the LEXCONN, "a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey" (Roze et al., 2012).

A Principal Components Analysis on talk pages and threads extracted 5 dimensions that explain around 30% of the total variance (29.2% for entire talk pages, 32.4% for threads). The first dimension is simply related to the size of the text units. The second dimension is more interesting and the correlated features differ between talk pages and threads. As for talk pages, it opposes

- talk pages with politeness cues (thanks, hello, cheers, please, etc.), formal you (vous) and we (nous) and discourse relations expressing concession, condition and temporal relations; to
- talk pages with more discourse relations expressing contrast, background/narration and causality.

As for threads, dimension 2 opposes

- threads with agreement cues (ok, agree, of course, yes, no, etc.), formal you and discourse relations expressing alternation, consequence, goal and temporal relations; to
- threads with more I, informal we (on) and discourse relations expressing contrast.

A third dimension that may be relevant gathers together talk pages (as threads) in which more connectives expressing narrative relations (then, later, once, before, etc.) and consequence relations (in this case, in this respect, etc.) occur. We may also notice that no TEMA features are significant for any dimensions.

More precise details defining these profiles will be presented during the presentation, with a focus on extreme talk pages and threads on each dimension. Our next goal is to locate conflicting threads in this 5 dimensional space.

### 5. Perspective: Exploring Conflicts at the Thread Level

In this paper, we have proposed different ways to explore Wikipedia talk pages; CMC genres are indeed complex objects that challenge our traditional methods and we assume that such objects require different levels of investigation. The profiling step still needs further analysis but is already quite promising.

The results of the automatic classification show that the features taken into account and the parameters used for detecting conflicting talk pages are still fairly inaccurate. In addition our definition of a conflict discussion must be revised. Several paths are currently being followed, including (i) using other criteria, starting with the dimensions with identified in the profiling step; (ii) using more detailed categories, combining the article labels signaling conflicts, and the talk page labels; and (iii) using a dataset of manually annotated talk pages. We are currently annotating the threads of 30 talk pages extracted from the WikiTalk corpus in terms of conflicts (degree, intensity, type) thanks to a CORLI grant 8.

We just led a first annotation experience, following the example of (Denis et al., 2012), which enabled us to bring interesting contrasts to light (Poudat et al., 2016).

For the moment, two talk pages have been annotated, totalling 255 threads for which coders have just to indicate if the thread is conflict or not with a very basic definition. As Table 3 shows, around one thread on 2 was annotated as conflicting.

<table>
<thead>
<tr>
<th>Talk page’s topic</th>
<th># threads</th>
<th># conflicts</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bogdanoff brothers</td>
<td>75</td>
<td>37</td>
<td>49.3</td>
</tr>
<tr>
<td>Psychoanalysis</td>
<td>140</td>
<td>74</td>
<td>52.9</td>
</tr>
<tr>
<td>Total</td>
<td>215</td>
<td>111</td>
<td>51.6</td>
</tr>
</tbody>
</table>

Table 3: Conflicting annotated threads in two talk pages.

### 6. References


Denis, A., Quignard, M., Fréard, D., Détienne, F., Baker, M., and Barcellini, F. (2012). Détection de conflits

6[http://factominer.free.fr/index.html](http://factominer.free.fr/index.html)


8TGIR Huma-Num CORLI (Corpus, Languages and Interactions, French National Consortium)


