# French Wikipedia Talk Pages: Profiling and Conflict Detection

**Ho-Dac L.-M.(*), Laippala V.(**), Poudat C.(***) and Tanguy L.(*)**
(*) CLLE, University of Toulouse, CNRS, UT2J, 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 (Urieli, 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 huge size (5M articles in the English Wikipedia / 1.7M articles in the French Wikipedia as of June 2016). In addition, Wikipedia is one of the 10 most consulted websites in the world (Alexa, June 2016).

Over the last decade, Wikipedia has become a wealth of information which is more and more used by natural language processing (NLP) and text mining applications (Ferschke & al. (2013) propose an overview of the use of Wikipedia in NLP). It has also been the subject of many studies in social sciences. After the quality of the encyclopedia has been established by (Giles, 2005), a large number of studies use Wikipedia for describing human coordination and collaboration processes (Viegas et al., 2007; Brandes and Lerner, 2007; Kittur and Kraut, 2008; Stvilia et al., 2008) via the analysis of revisions and talk pages which provide evidence of collaborative edition, maintenance work, cooperation and conflict resolution (Kittur et al., 2007; Viégas et al., 2004).

Most of these studies do not focus on the linguistic and discursive aspects of Wikipedia pages, certainly because of the sprawling structure of Wikipedia (multiplicity of pages and versions), which makes corpus building quite difficult. As a consequence, these works mostly rely on network analysis or on statistical features extracted from article revision histories. For instance, an interesting result for our project is that article reverts (when users restore a previous version) are proven significant features to detect conflicts (Viégas et al., 2004; Brandes and Lerner, 2007; Kittur et al., 2007; Suh et al., 2007; Kittur and Kraut, 2010; Miller, 2012). Nevertheless, such features remain indirect markers of conflicts, as they may be interpreted differently, allowing no clear distinction between editorial conflicts and vandalism, for instance (Potthast et al., 2008; Yasseri et al., 2012; Adler et al., 2011). Other commonly used criteria include article and talk page length, number of revisions in article and talk pages, number of anonymous edits/users, character or word insertion or deletion between users, article labels, etc.

Such criteria serve as the basis for the automatic detection of quality articles (Wilkinson and Huberman, 2007), conflictual pages (Kittur et al., 2007; Vuong et al., 2008; Sumi et al., 2011) or topic categories which are more likely to generate conflicts, such as religion and philosophy according to (Kittur et al., 2009).

Although these studies have provided interesting insights on the evolution of Wikipedia's organization and collaborative edition, the linguistic characteristics of Wikipedia pages remain little explored. In particular, talk pages are specifically interesting to observe as they are at the heart of the Wikipedia device. Each article is associated with a talk page, where most of the coordination work is done, and where the potential conflicts are discussed and ultimately resolved in the best-case scenario (Viegas et al., 2007). Talk pages are the places where editors discuss the modifications to be made on the article, including sections to be rewritten or suppressed (Ferschke et al., 2012).

Wikipedia talk pages may be considered as a new discussion 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 editing 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 Wikipedians 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 mutidimensional 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[1].

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

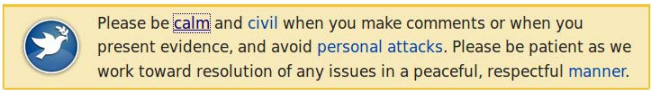
1. timestamp and an optional user signature, such as: *Viking59 10 mai 2009 à 17:16 (CEST)*; or
2. a change in the interactional level indicated by the number of semi-colons (:) in the beginning.

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.

```
<category type="discipline">
   <catDesc>Politique</catDesc>
   <catDesc>France</catDesc>
</category>
<category type="avancement">
   <catDesc>Featured</catDesc>
</category>
<category type="interaction">
   <catDesc>{{calm}}</catDesc>
</category>
```

Three kinds of metadata were automatically extracted to categorize and describe the discussions:

1. ”discipline” indicates associated thematic portals,
2. ”avancement” corresponds to article’s quality scale based on Wikipedian assessments[2],
3. ”conflictness” gives information about possible conflicts in the discussion. Such information may be manually inserted by Wikipedians via the template {{keep calm}} which adds the following banner at the top of the talk page[3]).



Please be calm and civil when you make comments or when you present evidence, and avoid personal attacks. Please be patient as we work toward resolution of any issues in a peaceful, respectful manner.

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[4].

| discussions | sections | posts | words |
|---:|---:|---:|---:|
| 365,612 | 1,023,841 | 2,406,514 | 161,833,298 |

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.

[2] https://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

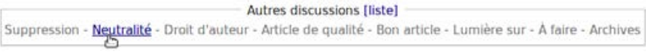
[3] https://en.wikipedia.org/wiki/Template:Calm

[4] Soon available at http://redac.univ-tlse2.fr/

## 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:

- `<category type="interaction">` in teiHeader indicates that the ”keep calm” template was inserted;
- a parallel talk page was created for discussing the article’s neutrality[5];



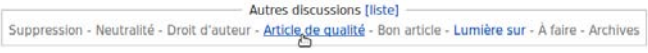
Autres discussions [liste]

Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives

- 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:

- `<category type="avancement">` in teiHeader indicates that the associated article was assessed to be ”Featured” or ”A-class”;
- a parallel talk page was created for deciding if the article deserves the ”featured” or ”A-class” status.



Autres discussions [liste]

Suppression - Neutralité - Droit d'auteur - Article de qualité - Bon article - Lumière sur - À faire - Archives

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 analysis for reflecting thematic and stylistic features. Classification based on syntactic features which considers texts as bags-of-syntactic N-grams more or less lexicalized is less common (Kanerva et al., 2014; Goldberg et al., 2013). This method enables a more robust analysis on text characteristics that does not depend on the text topic but attempts to generalize the level of description beyond individual lexical topics to typical structures (Laippala et al., 2015).

The classification is performed using the stochastic gradient 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.

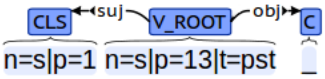


Figure 1: A delexicalized syntactic bi-arc describing a clitic+verb+conjunction as in the clause ‘I find that’.

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.

| | threads | | posts | |
|---|---|---|---|---|
| features | P | R | P | R |
| lemmas | 0.84 | 0.60 | 0.79 | 0.69 |
| bi-arcs | 0.55 | 0.48 | 0.63 | 0.59 |
| units | 46,690 | | 194,289 | |

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

[5] This possibility seems specific to the French Wikipedia

to multivariate exploratory data analysis[6]. Each talk page and thread was automatically described with four types of features:

- **THEMA**: portal sections of the associated article page knowing that an article may be categorized as belonging such as *Art, History, Sport*[7] 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 THEMA 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.

| Talk page's topic | # threads | # conflicts | % |
|---|---|---|---|
| Bogdanoff brothers | 75 | 37 | 49.3 |
| Psychoanalysis | 140 | 74 | 52.9 |
| Total | 215 | 111 | 51.6 |

Table 3: Conflicting annotated threads in two talk pages.

## 6. References

Adler, B. T., De Alfaro, L., Mola-Velasco, S. M., Rosso, P., and West, A. G. (2011). Wikipedia vandalism detection: Combining natural language, metadata, and reputation features. In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*, volume Part II of *CICLing'11*, pages 277–288, Berlin, Heidelberg. Springer-Verlag.

Agarwal, A., Chappelle, O., Dudik, M., and Langford, J. (2011). A reliable effective terascale linear learning system. *JMLR*, 15:1111–1133.

Brandes, U. and Lerner, J. (2007). Revision and corevision in wikipedia: Detecting clusters of interest. In *Proceedings of International Workshop Bridging the Gap Between Semantic Web and Web 2.0, 4$^{th}$ European Semantic Web Conference (ESWCÂ´07)*, Innsbruck, Austria.

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

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

[7] https://fr.wikipedia.org/wiki/Portail:Accueil

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

dans les communautés épistémiques en ligne. In *TALN-Actes de la Conférence sur le Traitement Automatique des Langues Naturelles-2012*.

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012). Behind the article: Recognizing dialog acts in wikipedia talk pages. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics.

Ferschke, O., Daxenberger, J., and Gurevych, I. (2013). A survey of nlp methods and resources for analyzing the collaborative writing process in Wikipedia. In *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Goldberg, Y., , and Orwant, J. (2013). A dataset of syntactic-n grams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), 1. Association for Computational Linguistics*.

Herring, S., Stein, D., and Virtanen, T. (2013). *Pragmatics of computer-mediated communication*, volume 9. Walter de Gruyter.

Ho-Dac, L.-M. and Laippala, V. (2015). Les discussions wikipedia : un corpus pour caractériser le genre "discussion". In *International Research Days Social Media and CMC Corpora for the eHumanities*, Rennes, France, october.

Kanerva, J., Luotolahti, J., Laippala, V., , and Ginter, F. (2014). Syntactic n-gram collection from a large-scale corpus of internet finnish. In *Proceedings of the Sixth International Conference Baltic HLT*.

Kittur, A. and Kraut, R. E. (2008). Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, pages 37–46. ACM.

Kittur, A. and Kraut, R. E. (2010). Beyond wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224. ACM.

Kittur, A., Suh, B., Pendleton, B. A., and Chi, E. H. (2007). He says, she says: conflict and coordination in wikipedia. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 453–462. ACM.

Kittur, A., Chi, E. H., and Suh, B. (2009). What's in wikipedia?: Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 1509–1512, New York, NY, USA. ACM.

Laippala, V., Kanerva, J., and Ginter, F. (2015). Syntactic ngrams as keystructures reflecting typical syntactic patterns of corpora in finnish. *Procedia - Social and Behavioral Sciences*, 198:233 – 241.

Miller, N. (2012). Characterizing conflict in wikipedia. *Mathematics, Statistics, and Computer Science Honors Projects*.

Potthast, M., Stein, B., and Gerling, R. (2008). Automatic vandalism detection in wikipedia. In *Advances in Information Retrieval*, pages 663–668. Springer.

Poudat, C. and Loiseau, S. (2007). Représentation et caractérisation lexicale des sciences dans wikipédia. *Revue française de linguistique appliquée*, 12(2):29–44.

Poudat, C., Vanni, L., and Grabar, N. (2016). How to explore conflicts in french wikipedia talk pages? In *JADT*, pages 645–656.

Roze, C., Danlos, L., and Muller, P. (2012). Lexconn: A french lexicon of discourse connectives. *Discours*, 10.

Scott, M., , and Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Philadelphia, PA, USA: John Benjamins Publishing Company.

Stvilia, B., Twidale, M. B., Smith, L. C., and Gasser, L. (2008). Information quality work organization in wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001, April.

Suh, B., Chi, E. H., Pendleton, B. A., and Kittur, A. (2007). Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 163–170. IEEE.

Sumi, R., Yasseri, T., Rung, A., Kornai, A., and Kertész, J. (2011). Characterization and prediction of wikipedia edit wars. In *Proceedings of the ACM WebSci'11*, pages 1–3, Koblenz, Germany, June 14-17 2011.

Urieli, A. (2013). *Analyse syntaxique robuste du français : concilier methods syntaxiques et connaissances linguistiques dans l'outil Talismane*. Ph.D. thesis, Université de Toulouse - Jean Jaurès.

Viégas, F. B., Wattenberg, M., and Dave, K. (2004). Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582. ACM.

Viegas, F., Wattenberg, M., Kriss, J., and van Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. In *40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, pages 78–78, January.

Vuong, B.-Q., Lim, E.-P., Sun, A., Le, M.-T., Lauw, H. W., and Chang, K. (2008). On ranking controversies in wikipedia: Models and evaluation. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 171–182, New York, NY, USA. ACM.

Wilkinson, D. M. and Huberman, B. A. (2007). Cooperation and Quality in Wikipedia. In *Proceedings of the 2007 International Symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA. ACM.

Yano, T. and Kang, M. (2008). Taking advantage of wikipedia in natural language processing term project report. *Language and Statistics*, II:11–762.

Yasseri, T., Sumi, R., Rung, A., Kornai, A., and Kertész, J. (2012). Dynamics of conflicts in wikipedia. *PloS one*, 7(6):e38869.