Exploring Wikipedia Talk Pages for Conflict Detection

Lydia-Mai Ho-Dac, *University of Toulouse, CNRS*
Veronika Laippala, *University of Turku*
Céline Poudat, *University of Nice Côte d'Azur*
Ludovic Tanguy, *University of Toulouse, CNRS*

Abstract

The present study concentrates on Wikipedia talk pages, which are online discussions where the authors discuss the composition and content of Wikipedia articles. These pages provide new data for describing and analysing collaborative writing processes, which often involve conflicts. Previously, many studies have explored Wikipedia conflicts, highlighting opposite editing patterns in relation to cooperation, conflicts or quality. Most of these studies belong to the field of social sciences, and linguistic analyses are not very common in this context. Therefore, the linguistic characteristics of Wikipedia conflicts in talk pages are still little described in the literature. In this context, our objective is to analyse relevant linguistic cues which may help identify and characterize conflicts on Wikipedia talk pages. To this end, we apply two automatic methods. The first one consists of the supervised automatic classification of conflicting vs. harmonic discussion threads. In the second we apply multidimensional analysis to the data to help profile the Wikipedia talk genre, enabling us to highlight key features and oppositions at a global level. The analyses are carried out on the WikiTalk corpus, a resource based on the French Wikipedia talk pages (160M words, 3M posts, 1M threads). The corpus includes a wide range of metadata, providing extra-linguistic characterization of the Wikipedia discussions.

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

1 INTRODUCTION

The exponential development of the Internet has led to new communicative situations and genres. These new online genres, which are not yet fully characterized, are complex objects challenging the existing methodologies and analysis tools. In this context, the Wikipedia encyclopaedia project is one of the new textual objects that can be studied under the umbrella term Computer-Mediated Communication (CMC, see Herring et al. 2013). Wikipedia, which has now been available for more than 15 years, is an open and collaborative project, available in numerous languages. The success of this online encyclopaedia is indisputable, as evidenced by its huge size (5M articles in the English Wikipedia and 1.7M in the French Wikipedia, as of June 2016). In addition, Wikipedia is one of the 10 most consulted websites in the world.[1]

Over the last decade, Wikipedia has become a wealth of information which is increasingly used in the development of natural language processing (NLP) and text mining applications (Ferschke et al. 2013). It has also been the subject of many studies in social sciences. Indeed, since the quality of the encyclopaedia was first established by Giles (2005), a large number of studies have used Wikipedia to examine the coordination and collaboration processes that occur among people (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 editing, 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, most likely because of the sprawling structure of the site (its 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, article reverts (when users restore a previous version) have proven to be significant features in the detection of 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, and so on.

Such criteria serve as the basis for the automatic detection of quality articles (Wilkinson and Huberman 2007), pages that are the focus of conflicts (Kittur et

[1] https://www.alexa.com

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 editing, the linguistic characteristics of Wikipedia pages remain under-explored. In particular, talk pages are particularly interesting to observe as they are at the heart of the Wikipedia process. Each article is associated with a talk page, where most of the coordination work is done, and where 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 to an article, including sections to be rewritten or removed (Ferschke et al. 2012).

Wikipedia talks 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 this respect they might be compared to forum discussions, except that they rely on a specific Wiki technology 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 in particular to editing actions, conflict management and Wikipedia procedures (e.g. NPOV, i.e. Neutral Point of View, relevance, source, quality, and so on).

Conflicts are particularly interesting to observe on 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 common 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 signalling that a conflict is ongoing on the article or talk pages (e.g. NPOV or relevance disputes, "Calm talk" template). There are many examples of pages with such labels, such as *Abortion in Iran*, *Bengali cuisine*, and *Religion and sexuality*, to cite just a few. If a conflict grows in intensity and verbal abuse occurs, then the article and talk page may be blocked and some users may be banned; for instance if they write "toxic" comments by making personal attacks.[2] From Wikipedia's

[2] One of the policy of WP is to avoid any kind of personal attacks (see https://en.wikipedia.org/wiki/Wikipedia:No_personal_attacks).

point of view, conflicts must be regulated as they impact productivity, as noted in Wulczyn et al. (2016:2), "the Wikimedia foundation found that 54% those who had experienced online harassment expressed decreased participation in the project where they experienced the harassment".[3] Wulczyn et al. (2016) aimed to develop tools to identify toxic comments, and their first experiment on Wikipedia talk pages resulted in "Wikipedia DeTox",[4] an automatic detector of toxic comments. This automatic device is currently adapted to other CMC under the name "Perspective API," which provides the following definition of "toxic": "a rude, disrespectful or unreasonable comment that is likely to make you leave the discussion".[5] The relationship between toxicity, or verbal violence, and conflict is obvious, although verbal violence and toxicity are generally detected at the post level (Wulczyn et al. 2016), whereas conflicts are better observed and detected at the thread level, with threads corresponding to the sections of talk pages in this context.

The aim of the present study is thus twofold: (i) We would first like to explore the differences between the threads belonging to talk pages reported to be sources of conflict by Wikipedians, and the threads belonging to talk pages where no problems have been reported. Are the first set of threads clearly distinct from the second? With this in mind, we will perform an automatic classification on the WikiTalk corpus. (ii) At a descriptive level, we would like to contribute to the linguistic description of the discussions on Wikipedia talk pages, which have been little explored using linguistic criteria. Indeed, few linguistic studies have been conducted on French Wikipedia – see Denis et al. (2012) on the detection of conflicting threads and Poudat and Loiseau (2007) on the exploration of Wikipedia categories. In order to have a broader view of the linguistic characteristics of the French Wikipedia talk pages, we will propose a first profiling of the genre, using a mutidimensional analysis enabling us to highlight key features and oppositions at a global level. Threads that are the focus of conflicts will then be characterized within this global generic profile.

2 THE 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 these, as they may not be

[3] These findings are reported in a report called "Harassment Survey" made available by the Wikipedia Foundation at the url https://commons.wikimedia.org/w/index.php?title=File%3AHarassment_Survey_2015_-_Results_Report.pdf.

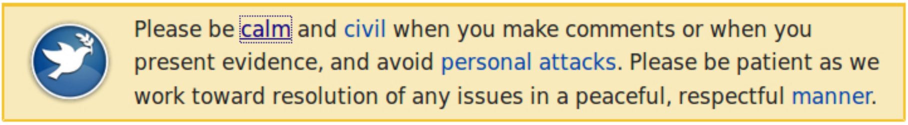
[4] https://tools.wmflabs.org/detox/

[5] http://www.perspectiveapi.com/

editorial discussions. Moreover, only 24% of the remaining talk pages contained more than two words.[6] The 365,612 remaining talk pages were associated with metadata, segmented into threads (i.e. headed sections) and posts (i.e. comments) and formatted according to the TEI-P5 guidelines.

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

1. “*discipline*” indicates the associated thematic portals,
2. “*avancement*” (progress) corresponds to the article’s quality scale based on Wikipedian assessments,[7]
3. “*interaction*” gives information about possible conflicts in the discussion. Such information may be manually inserted by Wikipedians via the template {{Calm talk}} which adds a dedicated banner to the top of the talk page (see Figure 1).[8]



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.

Figure 1: The {{Calm talk}} banner.

These metadata are encoded in the teiHeader in 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>
```

Automatic thread and post segmentation is based on the wikicode with the help of local grammars. Thread segmentation is achieved using the headings signalled in the wikicode by the pattern /==.*?==/. On the other hand, post segmentation is performed using both the signature manually inserted by the writer (such as:

[6] 1,013,791 (68%) talk pages were blank and 116,432 (8%) consisted in redirections to another talk page.

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

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

Viking59 10 mai 2009 à 17:16 (CEST), and the presence of a change in the interactional level indicated by the number of semi-colons (:) at the beginning of the post. Figure 2 illustrates the encoding of the wikicode into the TEI-P5 norm according to the following transformations: <div> for threads, <head> for topic titles, <post> and the three attributes: @who, @when and @interactionalLevel for posts.

Wikicode

```
== Jeux ==
Sinon, ce serait bien de retravailler la section […]
Fredscare 18 avril 2007 à 17:00 (CEST)
:J'ai retravailler la section […] Bouchette63 6 avril 2008 à
02:10 (CEST)
::J'ai vidé la section […] PV250X 15 avril 2009 à 20:39
(CEST)

==Situation actuelle (2005 à aujourd'hui)==
Bonjour, […]
```

TEI-P5 encoding

```
[…]
<div id=“3” level=“1”>
<head>Jeux</head>
<post id=“5” who=“Fredscare” when=“18-04-2007-17:00”
interactionalLevel=“0”>
   <p id=“1”>Sinon, ce serait bien de retravailler la section
[…]</p>
</post>
<post id=“5” who=“Bouchette63” when=“06-04-2008-02:10”
interactionalLevel=“1”>
   <p id=“1”>J'ai retravailler la section […]</p>
</post>
<post id=“5” who=“PV250X” when=“15-04-2009-20:39”
interactionalLevel=“2”>
   <p id=“1”>J'ai vidé la section […]</p>
</post>
</div>
<div id=“4” level=“1”>
<head>Situation actuelle (2005 à aujourd’hui)</head>
<post who=“anonyme” bot=“no” when=“unknown”
interactionalLevel=“0”>
   <p id=“1”>Bonjour, […]</p>
[…]
```

Figure 2: From Wikicode to TEI-P5 encoding (extract from the “Sega” talk page).

Eight of the extracted talk pages, amounting to 413 posts and 47,284 tokens, were manually inspected to evaluate the extraction process. The 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. Table 1 gives a quantitative overview of the WikiTalk corpus.[9]

Table 1: Quantitative overview of the WikiTalk corpus.

#talk pages	#threads	#posts	#words
365,612	1,023,841	2,406,514	161,833,298

3 CLASSIFICATION OF CONFLICTING VS. NEUTRAL DISCUSSIONS

Are threads belonging to talk pages associated with conflicts significantly different from those belonging to harmonic or neutral pages? To answer this question, we carried out a data-driven comparison of the global linguistic characteristics of two classes of discussions, distinguished according to an experimental classification of "conflicting" vs. "neutral" talks. The selection criteria used for distinguishing between these two classes are based on alerts and reporting issued by Wikipedians.

3.1 Experimental DataSet for thread classification

An automatic classification of the WikiTalk corpus has already been tested for distinguishing Wikipedia talk pages from Wikipedia articles and other CMC, such as online forums (Ho-Dac and Laippala 2017). The results showed that these three text genres could be automatically detected on the basis of a simple bag of words. Unfortunately, we could not adopt the method proposed in Ho-Dac and Laippala (2017) for the following two reasons. First, in contrast with Ho-Dac and Laippala (2017), where talk pages, Wikipedia articles and online forum were clearly identified genres and large amounts of training data were easily available, there is no training data available for conflict detection, as no large-scale corpora with discussions annotated as conflicting or not exist. Secondly, as opposed to

[9] Soon available at http://redac.univ-_tlse2.fr/

Ho-Dac and Laippala (2017), where the analysis could be done over entire talk pages and Wikipedia articles, the thread level seems more suitable for detecting conflicts, as thus is used in this work.

As stated above, the development of a supervised machine learning system that would automatically classify threads requires a large amount of threads categorized as conflicting vs. neutral. In order to provide training data and because there is very little information at the thread level, we opted for an experimental classification of "conflicting" vs. "harmonic/neutral" talk pages, and then used this to assess the hypothesis that threads belonging to "conflicting" talk pages would be significantly different from those belonging to "harmonic/neutral" pages. The selection criteria used for distinguishing between these two classes are based on alerts and reporting issued by Wikipedians.

We considered that talk pages were conflicting when they were associated with metadata signalling the presence of a conflict, that is:

- <category type="interaction"> in teiHeader indicates that the "calm talk" template was inserted;
- a parallel talk page was created for discussing the article's neutrality;[10]
- the talk page is not a main page but a parallel talk page created for discussing the article's neutrality.

In contrast, talk pages associated with featured articles[11] were considered to be "neutral," based on the assumption that the acknowledged quality of these articles means that there is a consensus amongst the contributors. Criteria for *a priori* "neutral" talks are as follows:

- <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.

The resulting data set collected from the WikiTalk corpus based on these criteria is described in Table 2. Note that all the talk pages which contained less than 100 words were excluded.

[10] This possibility seems specific to the French Wikipedia.

[11] https://en.wikipedia.org/wiki/Wikipedia:Featured_articles

Table 2: Experimental dataset for the classifier: conflict vs. neutral discussions.

Selection criteria	#talk pages
More than 100 words in the talk page	152,931
Conflict discussions (11 M words)	**2,028**
Calm talk template in the header	39
Existence of a parallel NPOV talk page	1,782
Talk page is a “neutrality” talk page	207
Neutral discussions (8.8 M words)	**4,569**
A-class article mentioned in the header	1,099
Existence of a parallel talk page about A-ranking	3,470

3.2 Thread classification on the experimental DataSet

We trained a text classification model using the Vowpal Wabbit linear classifier (Agarwal et al. 2011), and tested it on a sub-part of the threads that were experimentally classified (henceforth “Experimental DataSet”), and also on the threads that were manually annotated (henceforth “Annotated DataSet”).

Four feature sets were tested: words, lemmas, character 5-grams and syntactic N-grams. While the first three sets are the one used in the traditional lexical approach, as in, for example, Scott and Tribble (2006), which proposes using keyword analysis to reflect thematic and stylistic features. Classification based on syntactic N-grams is less common (Kanerva et al. 2014, Goldberg et al. 2013). The syntactic N-grams we used are delexicalized *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 Figure 3.

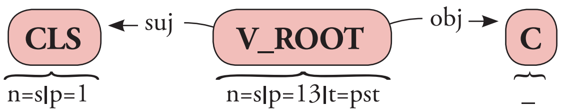


Figure 3: A delexicalized syntactic bi-arc describing a clitic+verb+conjunction as in the clause ‘I find that’.

Syntactic analysis and lemmatisation were provided by the Talismane toolkit (Urieli 2013). The classification method based on syntactic N-grams enables a more robust analysis based on text characteristics that does not depend on the text topic, but instead attempts to generalize the level of description beyond individual lexical topics to typical structures (Laippala et al. 2015).

The first classification experiment is performed using the stochastic gradient method with two-thirds of the Experimental DataSet used for training and the remaining for testing. Table 3 gives the precision (P) and recall (R) for detecting the "conflict" category by using the two feature sets on 46,690 threads.

Table 3: Comparison of different lexical vs. syntactic approaches for the automatic classification of conflicting threads and posts.

Features	P	R	F-measure
Words	0.83	0.65	0.74
Lemmas	0.84	0.60	0.72
Character 5-grams	0.82	0.72	0.77
Syntactic Bi-arcs	0.55	0.48	0.52
# threads	**46,690**		

The results show that character-based and lexical feature sets have good performance, while bi-arcs consisting of only syntax are not very useful. The best results are achieved by using lemmas. The 40 most distinctive lemmas for the conflicts, as estimated by the classifier, can be divided to two groups:

- words referring to the writing process, highlighting current sources of editorial conflicts, as well as (dis)agreement cues: *style, to hope, respect, version, way of writing, restructuring, reformulation, neutralisation, clumsy, uncoherent, respect, mistake, controversy, debate, ok;*
- words referring to the article topics: *rwanda, dictatorship, mandarin, quebec, islam, buddhism.*

These distinctive lemmas give a clear picture of the characteristics of the threads that the classifier identifies as conflicting. Importantly, we can assume that the first group of lemmas referring to the writing process may be common to all conflicts, regardless of the discussion topic. Considering our general aim of identifying conflicts in general, this is crucial. A closer look on the threads classified incorrectly or with a high probability is, however, necessary in future work in order to better understand the basis of the classification. The features which were selected are informative, but not necessarily explanatory of the ways in which conflicts arise or get resolved.

3.3 Thread classification on the annotated DataSet

The classifier model we obtained was then assessed on an Annotated DataSet, gathering the 215 threads of two talk pages. The two talk pages associated with the articles *Psychoanalysis* and *Bogdanoff brothers* were manually annotated using a binary variable, signalling the presence or absence of an ongoing conflicts in the thread (Poudat et al. 2016). As Table 4 shows, around one thread out of every two was deemed to be conflicting.

Table 4: Annotated DataSet : conflicting annotated threads in two talk pages.

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

Table 5 below gives the results of the classification of the annotated DataSet with the model trained on the experimental DataSet. The results indicate that the classifiers trained on the data deemed to be conflicting vs. neutral based on the metadata do not work for the manually annotated conflicts.

Table 5: Classifier results on the annotated DataSet.

Features	P	R	F-measure
Words	0.47	0.53	0.50
Lemmas	0.45	0.47	0.46
Character 5-grams	0.46	0.57	0.52
Syntactic Bi-arcs	0.53	0.45	0.49

As the classifier results on the experimental DataSet reported in Section 3 were decent, this difference indicates that the manually identified conflicts and the threads we assumed as conflicting based on the metadata differ.

In other words, conflict threads may need further linguistic analysis and manual evaluation to be properly detected, as Wikipedia metadata are obviously inadequate and insufficient for this purpose.

The next sections address these questions by proposing a range of new features for profiling threads in a bottom-up approach (Section 4), and presenting an ongoing project of manual conflict annotation in the WikiTalk corpus (Section 5).

4 A BOTTOM-UP APPROACH TO DISCUSSION PROFILING

The automatic classification method was supplemented by a second approach which uses exploratory data analysis techniques based on linguistic and structural features. Our objective is to highlight the structure and the profile of talk pages and threads in a bottom-up approach, without a specific focus on conflict. This method was applied to the whole dataset, i.e. 365,612 talk pages and 1,023,841 threads, using the *R FactoMineR* package dedicated to multivariate exploratory data analysis.[12] Four sets of features were calculated for each talk page and thread, named ***Global***, ***Thema***, ***Interact*** and ***DiscRel***.

4.1 Linguistic and structural features for profiling threads

The *Global* features correspond to general non-linguistic characteristics automatically extracted from the thread and talk page. Table 6 describes the eight *Global* features taken into account in this study.

Table 6: Global features for describing threads.

Label	Description
#words_log	Number of words in the thread (logarithm)
#threads	Number of threads in the page containing the thread
#posts	Number of posts in the thread
max_depth	Maximum depth, i.e., the highest interactional/hierarchical level of a post in the thread
#users_thread	Number of different participants in the thread by considering all anonymous (i.e., unregistered) users as a single participant
%anonymous	Percentage of anonymous posts in the thread, either unsigned or signed by an unregistered user
A-class	Binary feature indicating if the talk page (and by extension the thread) is linked to an A-class article
Keep_calm	Binary feature indicating if the talk page (and by extension the thread) has been tagged with a “calm talk” template

The ***Thema*** features give details of the main topics of the talk pages, based on the portal sections of the associated article. The French Wikipedia comprises 11 portals:[13] Art, Geography, History, Leisure, Medicine, Politics, Religion, Science, Society, Sport and Technology. Geography is the most important portal in the context of this study (119,359 talk pages). Figure 4 gives an overview of the amount of talk pages per portal, although it should be noted that an article (and its associated talk page) may belong to several portals.

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

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

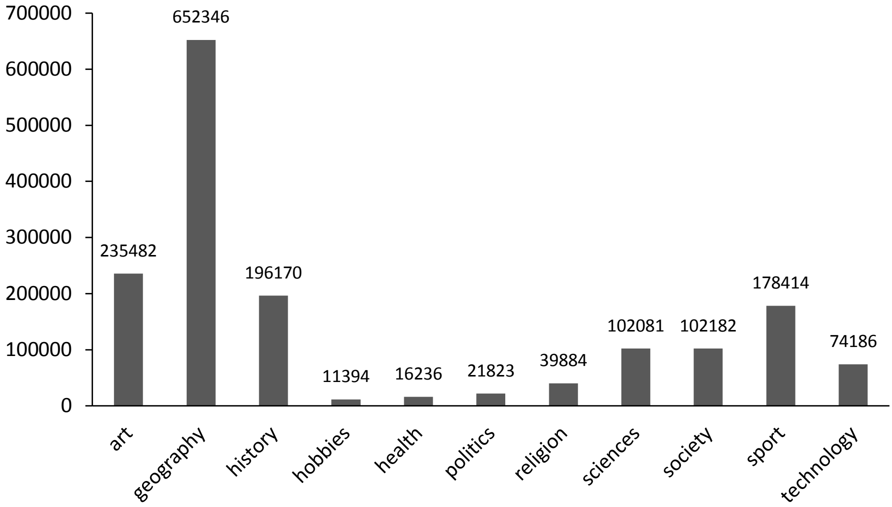


Figure 4: Amount of talk pages per portal.

More than 56% of the articles are categorized in at least two portals (44% in exactly two, with a maximum of six portals for a single article). We thus defined 11 binary features, one for each portal.

The *Interact* features correspond to the relative frequency of a range of basic interaction cues, related to agreement, disagreement and politeness. The counting was performed at the thread level, and 11 different types of cues were automatically identified with simple regular expressions (see Table 7).

Table 7: Interact features for describing threads.

Label	Description
politeness	*thanks, hello, goodbye, hi, sincerely, cheers, please, would you,* etc.
agreement	*OK, agree, yes, no, actually,* etc.
question	question mark (*?*)
je	1st singular person pronouns + the adverb *personally*
tu	2nd sing. pers. pronouns, informal "*you*"
vous	2nd plur. pers. and formal "*you*" pronouns
nous	1st plur. pers. Pronouns
on	Informal "we" (indefinite 3rd sing. pers. pronoun)
WP	Explicit reference to the Wikipedia project ("*Wikipedia*" or "*WP*")
pour	Sentence-initial *For* or *I'm for*
contre	Sentence-initial *Against* or *I'm against*

Table 8 gives the number of cues and the proportion of threads in which these *Interact* features were automatically detected. Agreement cues, questions and first singular person mentions occur in more than 25% of the total threads. The rarest features are the formal "we," "pro" and "against." These two latter features are actually very specific to threads dedicated to voting "for" or "against" editorial acts (e.g., article removal or article A-class ranking).

Table 8: Number and proportion of threads with Interact features.

Interact features	#cues	#threads with	%threads with
politeness	317,532	159,924	15.9
agreement	659,291	270,233	26.9
question	751,878	271,237	27.0
je	946,736	386,833	38.5
tu	400,052	106,427	10.6
vous	886,460	217,715	21.7
nous	120,560	79,328	7.9
on	630,616	201,656	20.1
WP	241,510	153,260	15.2
pour	142,785	85,871	8.5
contre	6,987	4,513	0.4
Total		1,005,592	100.0

The last type of feature, called ***DiscRel***, gives an idea of the rhetorical structures occurring in a thread. Using LexConn (Roze et al. 2012), "a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey," we projected these 328 connectives on each thread and measured the cumulative frequency for each discourse relation as defined in LexConn. Twenty-two discourse relations are defined in the LexConn database. When a connective is polysemous, all possible relations were considered. As for *Interact* features, the frequency was normalized on the number of words in the thread.

Table 9 gives the number and proportion of threads and connectives associated with each discourse relation (relation names are those used in the LexConn resource). The two columns labelled "Connectives" provide the number of connectives detected for each relation and proportion it covers among all the discourse relations. The two columns labelled "Threads with" indicate the number and proportion of the threads in which at least one connective expressing the relation occurs.

Table 9: Number and proportion of threads and connectives associated with each discourse relation.

Discourse Relations	Connectives		Threads with	
	#	%	#	%
alternation	583,585	4.9	**317,971**	**31.6**
background	512,690	4.3	189,967	18.9
commentary	25,581	0.2	21,740	2.2
concession	647,056	5.5	248,271	24.7
condition	1,483,308	12.5	496,852	49.4
consequence	162,213	1.4	123,036	12.2
continuation	1,462,713	12.4	469,608	46.7
contrast	528,004	4.5	240,919	24.0
detachment	32,297	0.3	27,487	2.7
elaboration	151,878	1.3	99,880	9.9
evidence	55,707	0.5	43,146	4.3
explanation	1,358,509	11.5	483,269	48.1
flashback	159,759	1.4	102,979	10.2
goal	749,597	6.3	**381,776**	**38.0**
narration	288,718	2.4	151,711	15.1
opposition	1,100,550	9.3	**330,437**	**32.9**
parallel	489,105	4.1	215,176	21.4
rephrasing	158,407	1.3	102,922	10.2
result	657,081	5.6	255,064	25.4
summary	17,858	0.2	15,636	1.6
time	905,059	7.6	**447,176**	**44.5**
unknown	301,741	2.6	157,851	15.7
Total	**11,831,416**	**100.0**	**1,005,592**	**100.0**

Table 9 shows strong variations and extremely frequent relations. Two groups of relation may be distinguished:

- The Condition, Continuation and Explanation relations, which each represent about 12% of all discourse relations, and appear in almost 50% of the total threads (49.4%, 46.7%, 48.1% respectively);
- The Alternation, Goal, Opposition and Time relations, which each represent a smaller percentage of all discourse relations (from 4.9% to 9.3%), but are also detected in a large proportion of the total threads (from 31.6% to 44.5%).

The occurrence of the first group of relations should be linked to the number of words in the thread (the more words, the more of these relations).

4.2 Exploring the threads with PCA

In order to observe how these different features interact with each other, and to help us identify the different thread profiles, we performed a standard multidimensional statistical analysis, and thus a Principal Components Analysis (PCA) was applied on the 1,023,841 threads. As we focus on the linguistic aspects of the discussions, we used the *Interact* and *Discrel* sets of cues as active variables to highlight the structure of the corpus and its main dimensions. The other features were projected afterward as illustrative variables in the reduce-dimension vector space resulting from the PCA.

This first two dimensions explain more than 20% of the total variance, the third one analysed here adding another 5%. Figures 5 and 6 show the first two factor maps, illustrating the main correlations among the features.

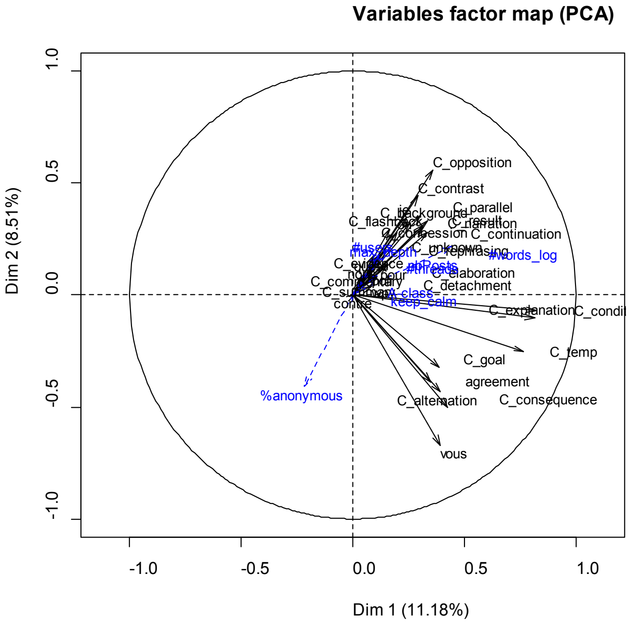


Figure 5: First factor map (dimensions 1 and 2) resulting from the PCA performed by taking into account the linguistic features. Additional features are shown in blue.

The first dimension, explaining around 12% of the total variance, is related to the size of the text units: the more words the threads contains, the more users participate, and the more features there are. As a consequence, the most frequent features (e.g. *Je*, *Vous*, Continuation, Condition and Explanation relations) are also the most significant.

We should also mention that the proportion of anonymous posts is higher for short threads. Let us also note that portals are not associated with significant linguistic cues.

The second and third dimensions are more clearly associated with linguistic features. The second dimension explains more than 8% of the total variance and opposes:

- threads with agreement cues (ok, agree, of course, yes, no, etc.), formal you and a significant presence of consequence, alternation and goal discourse relations (at the bottom of Figure 5); and
- threads containing a substantial amount of I ("je"), formal we/indefinite pronoun ("on") and connectives related to opposition and contrast (at the top of Figure 5).

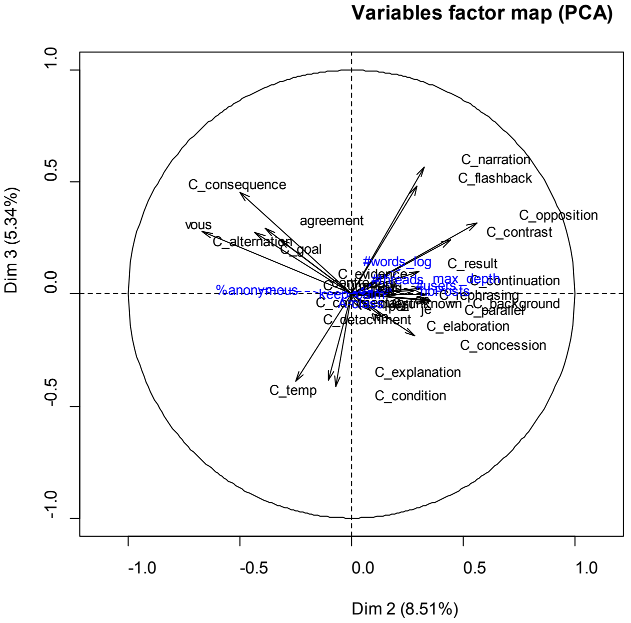


Figure 6: Second factor map (dimensions 2 and 3) resulting from the PCA performed by taking into account the linguistic features. Additional features are shown in blue.

The third dimension, which explains more than 5% of the total variance, opposes threads characterized by a significant presence of narrative relations (at the top of Figure 6), and threads including connectives expressing condition and explanation relations.

A closer look at the threads which are situated at the borders of dimensions 1, 2 and 3 provides a better understanding of the structure of the data, and the profiles of the threads they may relate to. The most extreme threads that dimension 1 opposes are very short ones that are usually made of anonymous posts. Actually, these threads may be described as very poor in terms of interaction, such as in example (1), a thread extracted from the talk page for "Protoplaste".

(1) ***techniques de l'obtention des protoplastes (technical criteria to obtain protoplasts)***

en cours (in progress)

On the other hand, we also found threads containing much more connectives and linguistic cues. Among these, dimension 2 may oppose threads characterized by a significant use of agreement markers as in example (2), to threads resorting to *I* ("je")*,* informal *we ("on")* and connectives expressing opposition, such as in example (3).[14]

(2) ***D'accord*** *pour rapporter les "controverses" scientifiques, mais sans négliger le style cf Wikipédia:Style encyclopédique. (**I agree** to report scientific "controversies" but without neglecting the encyclopedic style, see Wikipédia :Style) Les anglais me semblent plus pragmatiques de n'avoir traité que de l'"affaire". Pour résumer restons : neutre, impersonnel, clair, précis, compréhensible, non académique et moins "people". Bien à **vous** (kind regards).*

(3) ***Par contre*****,** ***je*** *doute qu'**on** puisse "ignorer" l'existence de ce rapport et qu'au minimum, le contenu qui a été diffusé par d'autres media soit admissible mais **j'**attends l'avis d'autres wikipédiens à ce sujet. (**However, I** doubt that **anyone** may "ignore" the existence of this report and I think that the material disseminated through the media is admissible but **I** await the opinions of other Wikipedians on this question.)*

This closer look at threads positioned on the extremities of the factors provides another view of the data, but does not permit us to identify precise and interpretable profiles of conflict threads. The next step is the projection of the annotated conflict threads through the three-dimensional vector space resulting from the PCA.

[14] Example 2 and 3 are extracted from the talk page about the Bogdanoff brothers.

4.3. Annotated conflict threads through the factor map

Figure 7 gives the location of the 215 annotated threads of the Annotated DataSet (Section 3.3) through the factor map resulting from the PCA. It seems that the best dimension for describing conflict threads is dimension 2. Conflict threads (red crosses) appear to be mainly situated on the positive side of this dimension. According to the PCA, these conflicting threads may be defined as those with more *I* ("je"), informal *we ("on")* and connectives expressing opposition and contrast discourse relations, and fewer agreement cues and formal "you."

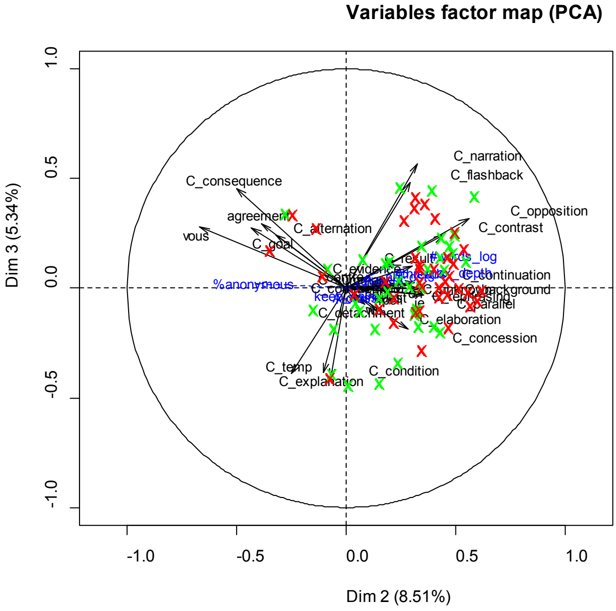


Figure 7: Second factor map (dimensions 2 and 3) with annotated threads located in the PCA and shown by red crosses for conflicting and green crosses for non-conflicting.

Example (4) illustrates one such profile, with the heading and the beginning of three posts of a thread annotated as conflicting in the talk page about Psychoanalysis ("Psychanalyse") (all the significant features are in bold)

(4) *<head> Citation et citations (Lacan et ses exégètes) </head>*

*<post>**je** propose des sources hors du champ de la critique psychanalytique pour exclure les débats LLNDLP ou Onfray etc (**I** propose sources outside the field of the criticism of psychoanalysis to exclude debates on LLNDLP or Onfray etc.) [...]</post>*

> *<post>Apparemment **on** oubli les politesse(s) avec Vous G de gonja…, **j'**invite chacun à jeter un oeil à ceci : (Politeness is not a virtue with you G. de gonja…, **I** encourage everyone to have a look at this) [...]</post>*
>
> *<post> 'None' * **Je** ne vois pas bien ce que le commentaire de G de Gonjasufi apporte : personne n'a jamais nié que Lacan ait employé le terme. (I don't really see what G de Gonjasufi's comment provides) **En revanche,** ce que nous disons c'est qu'il ne s'agit pas d'une qualification de la psychanalyse dans son (**In contrast**, what we are saying is that it is not a disqualification of psychoanalysis as a whole) [...]</post>*

5 CONCLUSION

We have proposed different ways to explore Wikipedia talk pages in this paper, motivated by the notion that CMC genres are indeed complex objects that challenge our traditional methods, and thus 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 conflict talk pages are still fairly inaccurate. In addition, our definition of a conflict discussion should be more specific. Data mining methods and first results in thread profiling give us some leads that must be followed up in this regard, and we are currently exploring relevant features to describe the thread level. We will notably use other categories to characterize talk pages and threads, combining, for instance, the article labels signalling conflicts, the talk page labels and the talk page type. On the linguistic level, the list of connectives and the discourse relation they express must be refined in order to distinguish discourse markers from conjunctions, and to get a better manage handle on polysemy (as for example, 17 connectives are associated with contrast in LexConn, including the very polysemous uses of "but" and "while").

In addition, other interaction features must be taken into account, including, for example, thread headings, timeline and context features. We are also concentrating on the first and the last posts of the threads, which generally play a key role in conflicts arising and being resolved. As such, we are currently annotating speech acts and politeness cues in these posts. Another avenue of investigation concerns the relation between disagreement and conflict: disagreement is quite common on Wikipedia, and although many conflicts arise from a disagreement, all disagreements do not naturally lead to conflict. What are the specificities of such disagreements / such conflicts? One of the main differences between disagreements and conflicts is certainly the presence of verbal violence, and we are currently exploring this question. In any case, it seems obvious that the most pressing need for identifying thread types is to provide a dataset of annotated threads according to interaction, politeness and conflict.

References

Adler, Thomas B., Luca de Alfaro, Santiago M. Mola-Velasco, Paolo Rosso, Andrew G. West, 2011: Wikipedia vandalism detection: Combining natural language, metadata , reputation features. *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing*. Berlin, Heidelberg. 277–288.

Agarwal, Alekh, Olivier Chappelle, Miroslav Dudik, John Langford, 2011: A reliable effective terascale linear learning system. *JMLR* 15. 1111-1133.

Brandes, Ulrik and Jürgen Lerner, 2007: Revision and co-revision in Wikipedia: Detecting clusters of interest. *Proceedings of International Workshop Bridging the Gap Between Semantic Web and Web 2.0*. Innsbruck, Austria.

Denis, Alexandre, Matthieu Quignard, Dominique Fréard, Françoise Détienne, Michael Baker and Flore Barcellini, 2012: Détection de conflits dans les communautés épistémiques en ligne. *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles*. 351–358.

Ferschke, Oliver, Iryna Gurevych and Yevgen Chebotar, 2012: Behind the article: Recognizing dialog acts in wikipedia talk pages. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 777–786.

Ferschke, Oliver, Johannes Daxenberger and Iryna Gurevych, 2013: A survey of NLP methods and resources for analyzing the collaborative writing process in Wikipedia. Gurevych, Iryna and Jungi Kim (eds.): *The People's Web Meets NLP: Collaboratively Constructed Language Resources*. Springer.

Giles, Jim 2005: Internet encyclopaedias go head to head. *Nature* 438/7070. 900–901.

Goldberg, Yoav and Orwant, Jon, 2013: A dataset of syntactic-n grams over time from a very large corpus of English books. *Proceedings of the Second Joint Conference on Lexical and Computational Semantics* (*SEM).

Herring, Susan, Dieter Stein and Tuija Virtanen 2013: *Pragmatics of computer-mediated communication* 9. Berlin: De Gruyter.

Ho-Dac, Lydia-Mai and Veronika Laippala, 2017: Le corpus WikiDisc, une ressource pour la caractérisation des discussions en ligne. Wigham, Ciara and Gudrun Ledegen (eds.): *Corpus de communication médiée par les réseaux : construction, structuration, analyse*. Collection Humanités Numériques. Paris : L'Harmattan. 107–124.

Kanerva, Jenna, Juhani Luotolahti, Veronika Laippala and Filip Ginter, 2014: Syntactic n-gram collection from a large-scale corpus of internet Finnish. *Proceedings of the Sixth International Conference Baltic HLT*.

Kittur, Aniket and Robert E. Kraut, 2008: Harnessing the wisdom of crowds in Wikipedia: quality through coordination. *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. 37–46.

Kittur, Aniket and Robert E. Kraut, 2010: Beyond Wikipedia: coordination and conflict in online production groups. *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. 215–224.

Kittur, Aniket, Bongwon Suh, Bryan A. Pendleton and Ed H. Chi, 2007: He says, she says: conflict and coordination in Wikipedia. *Proceedings of the SIGCHI conference on Human factors in computing systems*. 453–462.

Kittur, Aniket, Ed H. Chi and Bongwon Suh, 2009: What's in Wikipedia?: Mapping topics and conflict using socially annotated category structure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1509–1512.

Laippala, Veronika, Jenna Kanerva and Filip Ginter, 2015: Syntactic n-grams as key structures reflecting typical syntactic patterns of corpora in Finnish. *Procedia - Social and Behavioral Sciences*. 233–241.

Miller, Nathaniel, 2012: Characterizing conflict in Wikipedia. Mathematics, *Statistics , Computer Science Honors Projects* 25.

Potthast, Martin, Benno Stein and Robert Gerling, 2008: Automatic vandalism detection in Wikipedia. *Advances in Information Retrieval*. Springer. 663–668.

Poudat, Céline and Sylvain Loiseau, 2007: Représentation et caractérisation lexicale des sciences dans Wikipédia. *Revue française de linguistique appliquée* 12/2. 29–44.

Poudat, Céline, Laurent Vanni and Natalia Grabar, 2016: How to explore conflicts in French Wikipedia talk pages? *JADT*. 645–656.

Roze, Charlotte, Laurence Danlos and Philippe Muller, 2012: Lexconn: A French lexicon of discourse connectives. *Discours* 10. 1–15.

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

Stvilia, Besiki, Michael B. Twidale, Linda C. Smith and Les Gasser, 2008: Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology* 59/6. 983–1001.

Suh, Bongwon, Ed H. Chi, Bryan A. Pendleton and Aniket Kittur, 2007: Us vs. them: Understanding social dynamics in Wikipedia with revert graph visualizations. *Visual Analytics Science and Technology* 2007. IEEE. 163–170.

Sumi, Róbert, Taha Yasseri, András Rung, András Kornai and János Kertész, 2011: Characterization and prediction of Wikipedia edit wars. *Proceedings of the ACM WebSci'11*. Koblenz, Germany. 1–3.

Urieli, Assaf, 2013: Analyse syntaxique robuste du français: concilier méthodes syntaxiques et connaissances linguistiques dans l'outil Talismane. Ph.D. thesis, Université de Toulouse – Jean Jaurès.

Viégas, Fernanda B., Martin Wattenberg and Kushal Dave, 2004: Studying cooperation and conflict between authors with history flow visualizations. *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 575–582.

Viegas, Fernanda B., Wattenberg, Martin, Jesse Kriss and Frank van Ham, 2007: Talk Before You Type: Coordination in Wikipedia. *40th Annual Hawaii International Conference on System Sciences*. 78–78.

Vuong, Ba-Quy, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, Hady Wirawan Lauw and Kuiyu Chang, 2008: On ranking controversies in Wikipedia: Models and evaluation. *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM. 171–182.

Wilkinson, Dennis M. and Bernardo A. Huberman, 2007: Cooperation and Quality in Wikipedia. *Proceedings of the 2007 International Symposium on Wikis*. ACM. 157–164.

Wulczyn, Ellery, Nithum Thain and Lucas Dixon, 2017: Ex machina: Personal attacks seen at scale. *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee: 1391–1399.

Yasseri, Taha, Robert Sumi, András Rung, András Kornai, János Kertész, 2012: Dynamics of conflicts in Wikipedia. *PloS one* 7/6.