# How to explore conflicts in French Wikipedia talk pages?

Céline Poudat, Laurent Vanni, Natalia Grabar

▶ **To cite this version:**

Céline Poudat, Laurent Vanni, Natalia Grabar. How to explore conflicts in French Wikipedia talk pages?. Statistics Analysis of Textual Data, Jun 2016, Nice, France. pp.645-656. hal-01359416

**HAL Id: hal-01359416**
**https://hal.science/hal-01359416**

Submitted on 2 Sep 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# How to explore conflicts in French Wikipedia talk pages?

Céline Poudat[1], Laurent Vanni[1], Natalia Grabar[2]

[1]UMR 7320 BCL, Université de Nice Sophia Antipolis – France

[2]UMR 8163 STL CNRS, Université de Lille 3 – France

## Abstract

With the exponential development of the Internet, new discourse genres and situations have expanded. These new web genres, which are still little described, are complex objects challenging our methodologies and our analysis tools: the encyclopedic project Wikipedia is one of these new objects which are part of Computer-mediated communication (CMC).
The present article concentrates on the exploration of conflicts in Wikipedia talk pages, using Hyperbase Web. Wikipedia data and CMC corpora have been little studied by French linguistics so far, and are still challenging text statistics, notably because of the complexity of such data (multiple annotations, consistent metadata, references between postings and user networks).
Based on the Wikiconflits corpus, which is already available and freely usable by researchers, we will propose some methodological avenues to explore Wikipedia data and CMC corpora.

## Résumé

Avec le développement exponentiel d'Internet, de nouveaux genres discursifs ont émergé, encore peu décrits et exploités du fait de leurs caractéristiques complexes (nombreuses métadonnées, références entre les posts, réseaux sociaux d'utilisateurs, multiples annotations).

Le présent article s'intéresse aux conflits dans les pages de discussion éditoriales collaboratives de Wikipédia. Après avoir dressé un état des lieux des difficultés que pose l'exploitation de Wikipédia comme corpus, il mobilise Hyperbase web pour proposer différentes pistes d'exploration.

**Key words :** Wikipedia, CMC corpora, Editorial conflicts, Text statistics

## 1. Discussions and conflicts in Wikipedia talk pages

With the exponential development of the Internet, new discourse genres and situations have expanded. These new web genres, which are still little described, are complex objects challenging our methodologies and our analysis tools: the encyclopedic project Wikipedia is one of these new objects which are part of Computer-mediated communication (CMC). 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 (almost 5M articles in English Wikipedia / more than 1,5M in the French version of the project by Oct. 2015) and the fact that it was the 5th most consulted website in France in 2015 (Médiamétrie).

While collaboration and cooperation in Wikipedia have been particularly described (e.g. Viegas et al. 2004, Brandes&Lerner 2007, Kittur et al. 2007&2009, Auray et al. 2009), the writing specificities of Wikipedia pages still remain an issue. Given its size, Wikipedia is indeed a very complex object, especially because of its tentacular organization: the project is far from being homogeneous and it contains various genres and types of pages: as a matter of

fact, Wikipedia, and wikis in general, are heterogeneous objects composed of a multiplicity of pages allowing users to write encyclopedic articles collaboratively but also:

(i) to observe text evolution and to compare various versions of the same text at different moments - in that respect, each modification generates a new page and the volume of all the pages associated to an article may be huge;

(ii) to have discussions on the articles: each article has a talk page where modifications and possible disagreements can be discussed. Wikipedia also offers other discussion places, such as user talk pages, mediation rooms, discussion places such as the French *bistro* page where users can exchange information or experience, or specific talk pages in case of conflicts;

(iii) to get information and help on Wikipedia policy: principles and rules have thus significantly developed as Wikipedia gained in size and scope.

This sprawling structure may explain why Wikipedia has been little studied by linguistics and corpus studies to date. Indeed, how to extract and organize Wikipedia data? How to build corpora and develop methods enabling researchers to capture linguistic and interactional phenomena as well as to observe collaborative writing processes? To what extent are text statistics methods relevant and useful to explore such corpora?

We hope that the present article will provide some clues to this vast issue and a further insight into the possible ways of exploring Wikipedia and CMC corpora using text statistics methods.

Bearing this in mind, we chose to concentrate our efforts on Wikipedia article talk pages as they are the pages which raise the most difficult issues to text statistics, which were historically designed to process written corpora, i.e. literature, academic or political discourses. Several initiatives to process oral corpora (TXM team, Lyon) or different versions of a same text (Trameur team, Paris 3) have emerged within the last decade but to our knowledge, CMC corpora - and Wikipedia talk pages - are still challenging our field, notably because of the complexity of CMC data: dealing with complex annotations, consistent metadata, references between postings and user networks are issues needing thorough attention and development if we want our methods to remain at the cutting edge.

Wikipedia talk pages may indeed be considered as a *new discussion genre*: they share a common focus, i.e. article editing and improvement. They are not forums but open talks about article editing: discussions are quite specific and have become more and more regulated[1] as we have already pointed out. In spite of a high lexical variation, given the very wide range of topics and subjects covered, Wikipedia talk pages share common features referring for instance to editing actions or Wikipedia procedures (e.g. NPOV[2], relevance, source, quality etc.). Although they take the form of a dialogue, as observed in other CMC genres such as chats or forums, Wikipedia talks rely on a specific Wiki device which has direct consequences on the macrostructure (thread structure and posting formation, see 2.2.).

[1] https://fr.wikipedia.org/wiki/Aide:Discussion

[2] *neutral point of view*, which is a standard all editors are supposed to follow - see https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Neutralit%C3%A9_de_point_de_vue for more details.

In such talk pages, conflicts are objects which are particularly relevant to observe, as frontiers of the discussion process. Antagonistic edits and disagreements on article structure and content may indeed lead to conflicts and this corresponds to normal process when co-editing articles, before participants agree on more stable versions of articles. The conflicts usually occur when supporters of one point of view confront supporters of other points of view without finding common positions and ideas. Examples of such situations are quite numerous: cochlear implants, nocivity of chemical substances (phthalates, bisphenol A…), global warming and climate change to cite just a few.

To observe this, we carefully developed a corpus extracted from French Wikipedia within the national CoMéRé (***Co****mmunication* ***Mé****diée par les* ***Ré****seaux, Computer and Social network-mediated Communication*) project: Section 2 describes the *Wikiconflits* corpus, which is already available and freely usable by researchers, providing data reflecting different types of conflicts in the encyclopedic project. Note that we went further in the present paper and added another annotation at the thread level, signalling the existence or the absence of a conflict. Section 3 concentrates on the exploration of such a corpus, using methods from text statistics and Hyperbase Web[3].

## 2. Wikiconflits: a corpus to observe conflicts in French Wikipedia

### *2.1. Corpus design*

The Wikiconflits corpus[4] is part of the CoMéRé corpus (Chanier et al. 2014) which includes various corpora with mono- or multimodal interactions mediated through networks all structured in TEI-CMC and freely available[5]. Wikiconflits, whose primary objective is to enable researchers to study conflicts in French Wikipédia was developed to increase the representativity of CoMéRé.

Wikiconflits was designed in 3 stages:

1. talk page selection;
2. extraction of related pages;
3. structuring and annotation.

Talk pages were selected with the following principles in mind (for a more detailed description, see Poudat, Grabar, Paloque, 2016): the selected talk pages should (i) contain clearly expressed conflicts and be identified as conflictual by Wikipedia (NPOV or relevance labels, conflict discussed in mediation room etc.); (ii) be representative of different areas and classical conflicts within the field of Sciences and Techniques (Social sciences and Humanities, Natural sciences and Engineering…); (iii) correspond to a feasible treatment for the manual (and automatic) processing, and (iv) provide potentially interesting data and insights.

With these principles in mind, we preselected several sets of Wikipedia data (by Oct. 2013):

- Mediation room (salon de médiation): 73 articles
- Neutrality problem: 214 articles

[3] http://hyperbase.unice.fr/

[4] which is available here https://repository.ortolang.fr/api/content/comere/v2/cmr-wikiconflits.html

[5] https://repository.ortolang.fr/api/content/comere/v2/comere.html

- Relevance or content disputes: 546 articles
- Protected and semi-protected articles: 169 articles

All in all, 1,002 articles were analysed, discussed and filtered, and finally, seven talk pages meeting our criteria were selected, representative of four types of recurrent conflicts within the fields of Sciences and techniques: controversial scientific personalities (*Igor et Grichka Bogdanoff*), social controversies related to technological innovations (*Genetically modified organisms, Wind turbine*), pseudo-sciences and scientific legitimacy (*Intelligence Quotient, Psychoanalysis, Chiropratics*), relevance and sourcing problems (*History of logic*).

This may sound small but on the one hand, few pages met our criteria - for instance, most of the pages did not contain *conflictual talks*, as conflicts may be expressed through editing actions such as *reverts* (when users restore a previous version); series of antagonistic reverts are indeed well-known to signal editing wars and have often been used to explore conflicts (see Viegas et al. 2014 or Miller 2012 for instance). On the other hand, the number of pages was also determined given the following stages in the process: although talk pages are the very core of the corpus, all the pages somehow related to the chosen conflicts were indeed also extracted. As a matter of fact, Wikiconflits is not made of article talk pages only, but it also includes:

(i) other article talks which are hyperlinked to the talk page when they do exist:

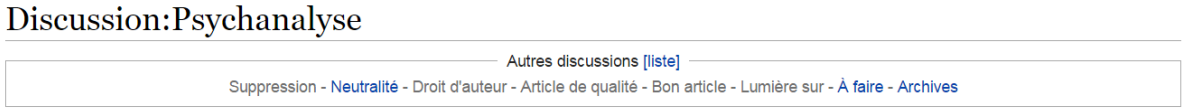


*Figure 1: other discussions in French Wikipédia*

(ii) the different versions of the concerned articles - so that discussions and conflicts may be related to the corresponding edits on the articles;

(iii) user pages, as talks may often spread from one talk page to another - with a restriction to the most active editors for each article. All in all, the seven talk pages with their related pages have a non-trivial size (330 Mo all corpus in the zip format).

In addition, we organized, structured and annotated set of pages according to the TEI-CMC recommendations, as described in the forthcoming section.

### *2.2. Structure and annotation*

Wikiconflits structure was defined and standardized within CoMéRé: talk pages (including article talks and other discussions) were all merged (see Poudat, Jin, Chanier 2014 for more details), and the datasets were converted into the TEI-CMC format (Chanier et al. 2015, Beißwenger et al. 2012, 2015), which is an extension of TEI-P5 guaranteeing standard and exchangeable data format. Besides, it is adapted to the specificities of the CMC corpora, such as interactions, time, level of answers, etc.

However, Wikipedia discussions and postings are slightly different from the standard situations as they are not related to any posting action, as Beißwenger et al. (2012) described:

[Postings are] stretches of text that an individual user produces in private and then passes on to the server through performing a "posting" action (usually by hitting the [enter] key on the keyboard or by clicking on a [send] or [submit] button on the screen).

As a matter of fact, in Wikipedia, users edit their postings and in spite of clear recommendations concerning the form of the postings (levels of answer, signing and date, etc.), talk pages often have a hybrid aspect, combining dialogues whose structure may not be obvious, and checklist elements. Besides, although the dialogue structure is threaded and may be divided into speech turns, there is no guarantee that the chronological and interaction order is respected. The situation is even worse: French Wikipedia contributors very often do not respect the writing conventions. This was the case in the discussions chosen for the study. Then, an important work consisted in manual correction of the threads in order to recover the order of interactions between the users. This is a very tedious task which required the reading of every post within a given thread and their reordering if necessary, on the basis of the post contents and their understanding by the researchers. When the information on time and user is available but not encoded with the Wikipedia conventions, the right coding was restored. A more difficult situation is when no such information is found explicitly. In that case, the reference to another user or reformulations in a given post were useful for performing this task, but it was often necessary to check revision histories to retrieve the complete information.

As we noticed, the checking and correction process has been done manually. A given talk page was processed by one annotator. The results of the manual correction were checked out again automatically and the necessary additional corrections (syntax errors, forgotten and wrong references…) were also corrected manually.

### *2.3. Assessing conflictuality: thread annotation*

For the purpose of the present study, an additional annotation has been performed using a binary variable (conflictual / non-conflictual) at the thread level. Discussions from two talk pages (*Psychanalyse* and *Frères Bogdanoff*) came through the process and were marked by one annotator each. Following the example of (Denis et al. 2012), the task was to determine whether a given thread (or discussion) was conflictual or not, given that our longer-term objective is to refine these results and to develop variables and scales with more granularity.

The annotation task we performed provided a further insight in this context:

- Conflicts first need a certain length to be considered as conflictual. For instance, a thread containing only one or two postings cannot be annotated as conflictual, even if the postings are aggressive and thereby potentially conflictual. *Thread length* was indeed one of the most relevant criteria for conflict detection in Denis et al. (ibid), as a conflict is rarely resolved with less than five turns whereas it would need at least three turns to rise;
- Another difficulty raised by this annotation is clearly semantic: it was often hard to distinguish between *disagreement* and *conflict*. Disagreements are situations in which users express different points of view on a given subject or concerning an editing action. Conversations remain cordial and disagreements are resolved by the discussion process, as a solution is finally found. Such situations are the very purpose of Wikipedia talk pages. On the contrary, a conversation is conflictual when users cannot find common opinion or point of view. In such tense situation, users maintain their position and refuse shifting within a few turns at least. Mutual communication

and understanding are not possible. This situation is typically related to the conflicts. The corresponding threads are then annotated as conflictual;

- Certain users are clearly conflictual (e.g. *trolls*[6]) and users who have been banned from Wikipedia since then may be labelled with a specific variable. At least such a clue might be used when deciding on the conflictual nature of the discussions;
- And finally, another difficulty appears when some posts are removed by the users in talk pages, even if it was generally possible to rebuild the discussion - if the context was sufficient, such discussions were then annotated as conflictual.

Approximately half of the discussions were judged as conflictual, which is quite a high proportion confirming the page selection we performed in 2.1.

## 3. Adjusting text statistics techniques to explore conflicts within the Wikiconflit corpus

Made of different sets of related pages and multiplying annotations, relations and metadata, Wikiconflits is far from the traditional written corpora on the basis of which text statistics methods have developed. Exploring Wikiconflits, and thereby any CMC corpus, is indeed a real challenge for our field: how to explore such a corpus? Which contrasts shall we consider and which methods are particularly appropriate in our agenda?

Using Hyperbase web, our objective is to propose some methodological avenues for further analysis of this issue which is currently addressed to our community.

### *3.1. From Hyperbase to Hyperbase web*

Hyperbase, which was created and developed by Etienne Brunet a few decades ago (see Brunet, Poudat 2011), is one of the leading and widely used tools in text statistics. Among the existing software, Hyperbase is certainly the tool which is the most oriented towards computer-assisted reading and interpretation. This is certainly due to Brunet's academic background in French literary studies (see for instance Brunet, Mayaffre 2009).

The standalone version of Hyperbase is still freely downloadable[7] but no longer maintained. Hyperbase web, which is Hyperbase web version, is currently being developed in BCL, Nice, enabling users to analyse their corpora online. The goal of this platform is to provide a toolbox supplying a variety of text statistics methods for corpus exploration - the web version will ultimately provide most of the methods that were available in the standalone version. Interestingly, and this is the advantage of web tools, Hyperbase web does not depend on a host operating system to work properly (compatibility with every OS, such as Windows, Mac, Linux, or smartphone). The input data are by default simple text files with standard encodings, that users are free to partition using the metadata of their choice (e.g. author, date, genre etc.). Contrary to the standalone version of Hyperbase, the web version allows users to resort to as many metadata as they wish and the ancient text limit no longer applies: Hyperbase was indeed originally set up to explore corpora of literature, made of long texts including whole books or chapters, i.e. datasets which are quite far from CMC. Hyperbase web is thus a significant turning point, facing a double challenge: ensuring the continuity of

[6] https://meta.wikimedia.org/wiki/What_is_a_troll%3F

[7] http://logometrie.unice.fr/pages/logiciels/

Hyperbase as developed by Brunet, and developing new methods as well as adapting to new corpora, such as Wikiconflits and Wikipedia corpora.

### *3.2. How to explore Wikiconflits using Hyperbase web?*

Web 2.0. corpora have a complex and unusual structure for classical corpus methods . They are indeed hybrid objects, based on a participatory process which may take various forms (commentaries, evaluations or discussions etc.). Indeed, Web 2.0. objects are closely related to their representations. Within Wikipedia, encyclopedic articles and talk pages are for instance closely linked: articles may be studied separately and contrasted, but talk pages relate to article edits and content - and consequently, interpretation might be difficult without checking the associated articles. Moreover, Wikipedia talk pages involve interactions between users and user *postings* are grouped together into *threads*, corresponding to the topic divisions marked on each wiki page.

Given the complexity of the data, choosing relevant and comparable textual units to explore the corpus structure is often difficult. Shall we contrast threads or postings? How to handle interrelated textual units? For instance, *postings* are utterances varying significantly in size and content, whose characteristics notably depend (i) on their position within the thread (opening, closing of the thread, type of response, conflictual / harmonic thread etc.), (ii) on the talk page they are associated with (*e.g.* topic, tone of the page, conflictual / harmonic page etc.) and (iii) on stylistic parameters related to contributors (*e.g;* anonymous vs registered, small vs big contributors, mediators or administrators etc.).

As a matter of fact, postings may hardly be contrasted without taking into account thread characteristics, which were besides annotated with a binary conflictual / harmonic variable (Section 2.3.), already giving us significant oppositions and insights.

The most significant specificities of the conflictual threads within the *Bogdanoff* and *Psychoanalysis* talk pages highlight different positions and tones (Figure 2). Indeed, topics and categories, which are besides associated with various levels of conflictuality (Kittur et al. 2009), involve different communities of users and consequently generate different types and styles of conflicts. The *Psychoanalysis* conflictual threads are therefore clearly more formal than the *Bogdanoff* ones: wikipedians involved in conflicts in the *Psychoanalysis* talk page significantly resort to the politeness formula *cordialement* ('Regards' in English) to close their postings, whereas the 2nd personal pronoun *tu* is on the contrary significantly used in *Bogdanoff* conflictual threads. Although conflictual threads naturally generate considerable discussion and debate, increasing the number of 2nd person pronouns, this opposition between *tu* and *cordialement* is the most significant stylistic difference between conflictual and non-conflictual threads within the two talk pages (Figure 2 and 3).

| écart | corpus | partition | mot |
|---|---|---|---|
| 15.98 | 146 | 131 | Alain |
| 13.12 | 137 | 113 | frère |
| 12.07 | 415 | 243 | leur |
| 10.27 | 101 | 80 | CNRS |
| 10.8 | 397 | 224 | scientifique |
| 9.75 | 105 | 80 | université |
| 9.12 | 97 | 73 | Igor |
| 7.56 | 46 | 39 | émission |
| 7.9 | 58 | 47 | doctorat |
| 7.28 | 118 | 75 | polémique |
| 6.28 | 61 | 43 | affaire |
| 6.48 | 34 | 29 | directeur |
| 6.99 | 184 | 103 | tu |
| 6.52 | 39 | 32 | notoriété |
| 6 | 38 | 30 | médiatique |
| 6.13 | 120 | 70 | thèse |
| 6.89 | 39 | 33 | Marianne |
| 5.11 | 133 | 70 | personne |
| 5.13 | 102 | 57 | site |
| 5.94 | 42 | 32 | journaliste |
| 5.97 | 40 | 31 | chercheur |
| 5.04 | 21 | 18 | jury |
| 5 | 76 | 45 | physique |

| écart | corpus | partition | mot |
|---|---|---|---|
| 22.42 | 2490 | 1774 | @card@ |
| 20.64 | 915 | 742 | psychanalyse |
| 12.34 | 216 | 192 | février |
| 10.54 | 2217 | 1381 | ( |
| 10.46 | 2248 | 1397 | ) |
| 9.78 | 89 | 86 | catégorie |
| 8.13 | 497 | 344 | -- |
| 8.78 | 148 | 126 | psychanalyste |
| 8.37 | 480 | 336 | critique |
| 7.98 | 109 | 95 | noir |
| 7.49 | 138 | 113 | — |
| 7.66 | 153 | 124 | cordialement |
| 6.81 | 53 | 50 | escroquerie |
| 6.62 | 74 | 65 | décembre |
| 6.52 | 2449 | 1425 | |
| 5.84 | 700 | 438 | source |
| 5.55 | 87 | 70 | psychanalytique |
| 5.46 | 185 | 132 | vue |
| 5.57 | 122 | 93 | mars |
| 5.04 | 201 | 139 | livre |
| 5.03 | 53 | 45 | psychanalyser |
| 5.24 | 68 | 56 | freudien |
| 5.21 | 103 | 79 | fiable |

*Figure 2: Most significant specificities of the Bogdanoff and Psychoanalysis conflictual threads*

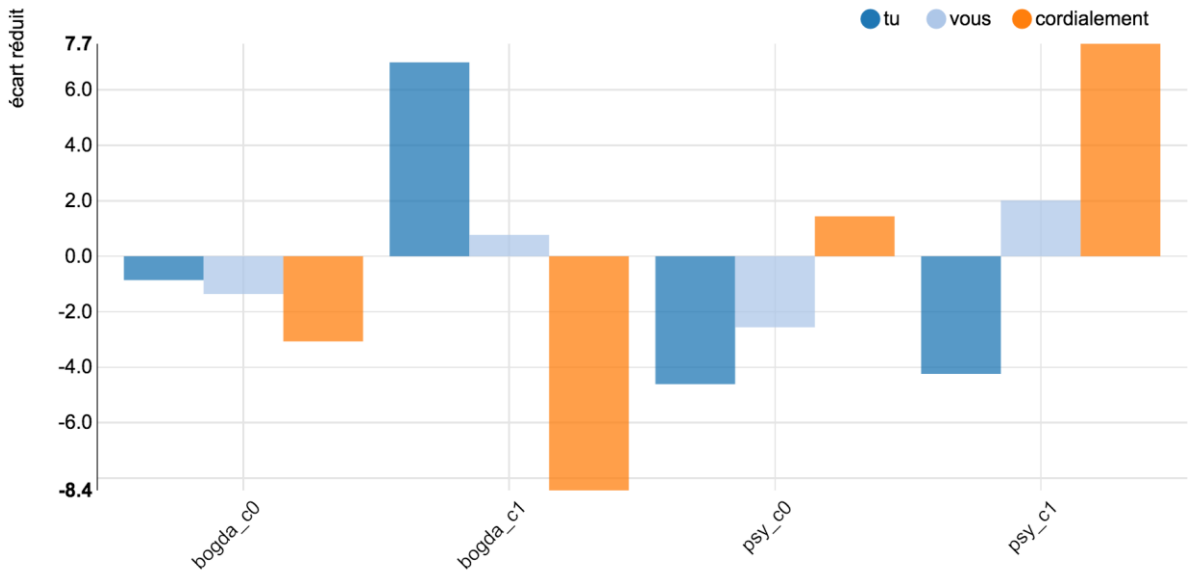


*Figure 3: Distribution of **tu**, **vous** and **cordialement** in conflictual and non-conflictual threads within Bogdanoff and Psychoanalysis discussions*

Although this binary opposition is obviously valuable to explore conflicts as a preliminary step and to contrast postings, we also resorted to further variables relating postings and

threads. On the one hand, we added basic variables enabling to contrast postings regarding their positions within threads (see Figure 4): *length*, which refers to the position of the message in the thread chaining - for instance enabling us to contrast first postings to the others; and *depth*, referring to the position of the message regarding the number of reactions raised in a thread at a given point.

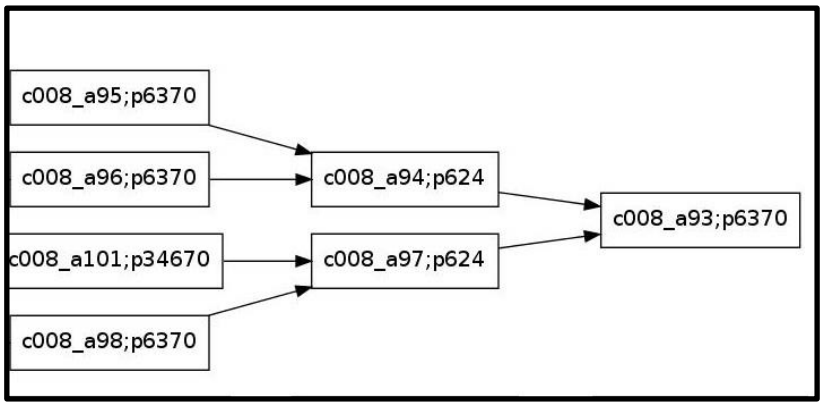


*Figure 4: Example of thread - lenth 3; depth: 4. Arrows = responses.*

Nevertheless, the resulting contrasts were often hard to interpret as postings of length 2 or 3 were quite heterogeneous. This is why we added other contrastive variables, related to *thread* global characteristics and shapes. For instance,

- *thread length*, enabling us to contrast postings within threads of comparable size,
- *thread depth*, enabling us to contrast postings within threads of comparable depth.

These two measures were particularly interesting regarding conflict explorations, as conflictual threads need a minimal length to raise (Denis et al. 2012) and, sometimes, to be resolved. On the other hand, threads with high depth contain postings generating lots of reactions (i.e. numerous subsequent postings) and this may enable us to obtain an approximate measure of the *reactivity rate* of the thread. As shown in Figure 5, the question of *sources* is for instance significantly over-used in threads with high depth.

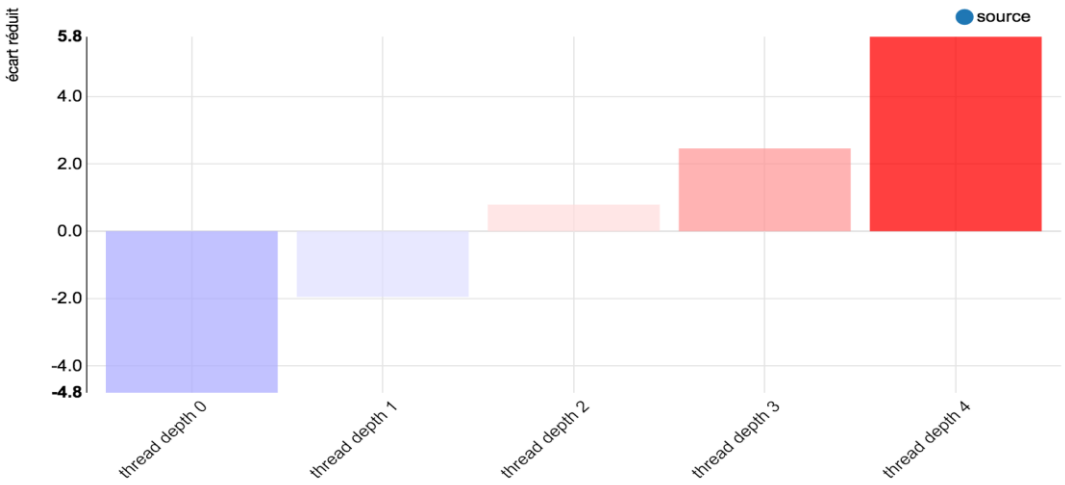


*Figure 5: Distribution of the lemma source - Thread depth from 0 to 4*

Sources are indeed at the heart of Wikipedia substantial and procedural principles and *source* is significantly over-used in conflictual threads (see Figure 2). Figure 5 clearly shows the high impact of the word *source* in interactions.

This observation is also coherent with the data presented in Figure 2. For instance, posts from the Bogdanoff discussion pages (on the left), contain several descriptors related to the source issues: *CNRS, scientifique, université, doctorat, directeur, notoriété, thèse, chercheur, jury*. Indeed, in the Bogdanoff-related discussions, the validity of the PhD work of the Bogdanoff brothers, the scientific value of this work, and the reputation of their universities and advisors are the main issues. Hence, an important part of the conflictual threads is devoted to these questions and interestingly, anonymous users were the less concerned about sources, while registered users were the ones proposing to lay down more objective and explicit principles to assess the validity of this work: for instance, do not consider the mediatic reputation of the brothers, but rather give attention to the notoriety of universities and advisors, to the publications performed by the Bogdanoff brothers, and the citations of their work.

## Conclusion and perspectives

Wikipedia data have been little studied by French linguistics so far and we have tried exposing the main difficulties raised by Wikipedia and CMC corpora processing and exploration. Our perspectives are threefold:

1. contribute to Wikipedia linguistic description, and more particularly, to the characterization of collaborative editorial talk pages, as a new genre. In this perspective, we will go further our study of editorial conflicts, as frontiers within talk pages. Among our objectives, we would like to determine types and levels of editorial conflicts, refine the annotation we performed and distinguish more clearly between conflicts and disagreements;
2. contribute to Wikipedia and CMC corpora exploration. Although the question remains wide open, CMC data, which are more and more developing, need our attention and further methodological work. In particular, we are currently assessing the interest of combining graph representations and contrastive methods;
3. perform cross-linguistic study of conflicts in Wikipedia. Some of the issues to be studied are for instance: whether the same topics are conflictual across languages, what are the distinctive conflictual features in these languages (politeness, call for the Wikipedia foundational principles, number of contributors...);
4. propose linguistic models of Wikipedia-specific conflicts and then study the possibility of an early detection of conflicts, which may permit to provide a more constructive and cordial context for collaborative writing of the Wikipedia articles;
5. promote linguistic research on Wikipedia. It is in this perspective that we have developed and released the Wikiconflits corpus, and we hope that our efforts will be successful. The already built corpus is freely available for the research purposes[8].

## Acknowledgments

This work was supported by the TGIR Huma-Num - *Corpus écrits* consortium. We are also thankful to the anonymous reviewers for their comments on a previous version of this paper.

[8] https://repository.ortolang.fr/api/content/comere/v2/cmr-wikiconflits.html

## References

Auray, N., Hurault-Plantet, M., Poudat, C., & Jacquemin, B. (2009). La négociation des points de vue : une cartographie sociale des conflits et des querelles dans le Wikipédia francophone. In *Réseaux* 2/2009, n° 154: 15-50.

Beißwenger M., Chanier T., Ehrhardt E., Herold A., Lüngen H., Poudat C. and Storrer A. (2015). TEI across corpora, languages and genres: Towards a standard for the representation of social media and computer-mediated communication. In *Text Encoding Initiative: connect, animate, innovate. 2015 Annual Conference and Members' Meeting of the TEI Consortium*, Oct 2015, Lyon, France. 2015, <halshs-01222982>

Beißwenger M., Ermakova M., Geyken A., Lemnitzer L. and Storrer A. (2012). A TEI Schema for the Representation of Computer-mediated Communication. *Journal of the Text Encoding Initiative (jTEI)*. 3 (Nov. 2012). <http://jtei.revues.org/476>

Brandes, U., & Lerner, J. (2007). *Revision and Co-revision in Wikipedia Detecting Clusters of Interest*. Konstanz: Bibliothek der Universität Konstanz.

Brunet É. (auteur), Mayaffre D. (éd.) (2009). *Comptes d'auteurs*. *tome I - Études statistiques de Rabelais à Gracq*. Préface de Henri Béhar. Collection *Lettres numériques*, Honoré Champion, Paris.

Brunet É. (auteur), Poudat C. (éd.) (2011). *Ce qui compte. Écrits choisis, tome II – Méthodes statistiques*. Préface de Ludovic Lebart. Collection *Lettres numériques*, Honoré Champion, Paris.

Chanier T., Poudat C., Sagot B., Antoniadis G., Wigham C., et al. (2014). The CoMeRe corpus for French: structuring and annotating heterogeneous CMC genres. *JLCL - Journal for Language Technology and Computational Linguistics*, 2014, 29 (2), pp.1-30.<http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf>

Chanier T., Poudat C. and Wigham C. (2015). The CoMeRe French CMC corpora and their modeling in TEI. ird-cmc-rennes: Social Media and CMC Corpora for the eHumanities., Oct 2015, Rennes, France. 2015, <http://ird-cmc-rennes.sciencesconf.org/>. <halshs-01222979>

Denis, A., Quignard, M., Fréard, D., Détienne, F., Baker, M., & Barcellini, F. (2012). Détection de conflits dans les communautés épistémiques en ligne. Presented at the TALN - Actes de la Conférence sur le Traitement Automatique des Langues Naturelles - 2012. Retrieved from https://hal.inria.fr/hal-00766397/document

Kittur, A., Chi, E. H., & 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 (pp. 1509–1512). New York, NY, USA: ACM. http://doi.org/10.1145/1518701.1518930

Kittur, A., Suh, B., Pendleton, B. A., & 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 (pp. 453–462). New York, NY, USA: ACM. http://doi.org/10.1145/1240624.1240698

Miller, N. (2012). Characterizing Conflict in Wikipedia. Mathematics, Statistics, and Computer Science Honors Projects. Retrieved from http://digitalcommons.macalester.edu/mathcs_honors/25

Poudat, C., Grabar, N. Kun, J. & Paloque-Berges, C. (2015). Corpus wikiconflits, conflits dans le Wikipédia francophone. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. http://hdl.handle.net/11403/comere/cmr-wikiconflits. 7 conflictual topics ; 3971 contributors ; 4456

posts / contributions in discussions ; 489 000 tokens in discussions (articles not counted) ; 330 Mo (7 sub-corpora zip)

Poudat, C., Grabar, N., & Paloque-Berges, C. (2016). *Wikiconflits: un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In* Ledegen G. and Wigham C., Médias sociaux et corpus de communication médiée par les réseaux, collection Humanités numériques, L'Harmattan (à paraître).

Poudat, C., Jin, K., & Chanier, T. (2014). Manuel du corpus wikiconflits (cmr-wikiconflits-tei-v1-manuel.pdf). In Corpus Wikiconflits extraits de Wikipedia. Dans banque de corpus CoMeRe.org. Ortolang.fr: Nancy. http://www.ortolang.fr/11403/comere/, cmr-wikiconflits-tei-v1

Viégas, F. B., Wattenberg, M., & 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 (pp. 575–582). New York, NY, USA: ACM. http://doi.org/10.1145/985692.985765