Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone

Céline Poudat, Natalia Grabar, Camille Paloque-Bergès, Thierry Chanier, Jin Kun

▶ **To cite this version:**

Céline Poudat, Natalia Grabar, Camille Paloque-Bergès, Thierry Chanier, Jin Kun. Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. Ciara R. Wigham & Gudrun Ledegen. Corpus de communication médiée par les réseaux : construction, structuration, analyse, L'Harmattan, 2017, 978-2-343-11212-1. hal-01485427

HAL Id: hal-01485427
https://hal.science/hal-01485427v1

Submitted on 8 Mar 2017

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.

Distributed under a Creative Commons Attribution 4.0 International License

WIKICONFLITS : UN CORPUS DE DISCUSSIONS ÉDITORIALES CONFLICTUELLES DU WIKIPÉDIA FRANCOPHONE

Céline POUDAT, UMR 7320 BCL, Université de Nice Sophia Antipolis – France
Natalia GRABAR, UMR 8163 STL CNRS, Université de Lille 3 – France
Camille PALOQUE-BERGES, HT2S / DICEN-Idf, CNAM, Paris – France
Thierry CHANIER, LRL, Clermont Université – France
Kun JIN, LRL, Clermont Université – France

INTRODUCTION

Si Wikipédia (WP), qui fête aujourd'hui ses quinze ans, a donné lieu à de nombreuses études et projets de recherche qui ont permis de saisir différents aspects de son fonctionnement, de sa gouvernance ou encore des processus de réécriture à l'œuvre dans les articles, le projet encyclopédique a surtout été observé par les sciences sociales, et la question de l'écriture collaborative a été plutôt abordée du point de vue de la coopération (*e.g.* Viegas et al. 2004, Brandes & Lerner 2007, Kittur & Kraut 2008, Stvilia et al. 2008) que de celui de l'écriture, et des caractéristiques linguistiques et discursives particulières que le projet encyclopédique et son dispositif induisent. En effet, les études linguistiques se sont peu penchées sur Wikipédia, certainement du fait de la complexité de l'objet, qui multiplie les versions, les types de pages et les genres textuels. En revanche, lorsque les linguistes s'y intéressent, ils se focalisent sur les articles plutôt que les pages de discussions. En outre, malgré sa taille conséquente et sa relative stabilité, le projet Wikipédia de langue française reste sensiblement moins décrit que le projet de langue anglaise.

Le corpus Wikiconflits, qui est l'objet du présent article, a été développé pour pallier cette situation et encourager les études linguistiques sur le projet encyclopédique, du moins est-ce l'une de nos ambitions.

Wikiconflits s'articule ainsi autour des pages de discussion éditoriale associées aux articles encyclopédiques. Si le processus normal d'une édition d'article sur WP est collaboratif et constructif – c'est le cas de la grande majorité du WP anglophone, la coopération peut être plus ardue et entraîner des conflits éditoriaux. En tant que frontières de la discussion et la collaboration, les conflits nous semblent des objets particulièrement intéressants à aborder pour caractériser ce nouveau genre discursif de la page de discussion éditoriale et collaborative. Nous avons ainsi choisi de nous concentrer sur les articles ayant été le lieu de conflits, voire de guerres éditoriales.

L'objectif du présent article est de présenter le corpus Wikiconflits, de ses principes de constitution à sa construction, en explicitant également les perspectives de recherche dans lesquelles nous souhaitons le mobiliser.

Wikiconflits a été développé dans le cadre de l'action *Nouvelles collectes* du projet national CoMeRe issu du groupe de travail portant sur les *Nouvelles formes de communication* (Consortium corpus écrits - TGIR Huma-num, 2011-2015). L'objectif de CoMeRe visait à rendre accessible à la communauté un grand corpus CMC (Computer Mediated Communication) représentatif des échanges sur les réseaux en langue française. Dans cette perspective, les données, issues majoritairement des projets antérieurs des participants et donc hétérogènes, ont été harmonisées et annotées suivant un modèle d'annotation TEI-CMC standard (Beißwenger et al. 2012) avant d'être rendues librement accessibles à la communauté sur les serveurs Huma-Num Ortolang (CoMeRe, 2016). Construit sur un principe de variété, l'objectif de l'action *Nouvelles collectes* était d'accroître la représentativité du corpus global en l'augmentant de genres CMC dont l'absence aurait été

préjudiciable à la valeur d'échantillon de l'ensemble. C'est dans cette optique que d'autres corpus ont été développés : un corpus de tweets (Longhi et al. 2015, ce numéro) ainsi que le corpus Wikiconflits, qui constitue l'objet de notre article.

Après avoir présenté les grands principes de constitution du corpus Wikiconflits, ce qui sera l'occasion de rendre compte de notre expérience de navigation dans un ensemble de données remarquablement complexe (section 1.), nous exposerons les grandes lignes des choix de structuration et des procédures d'annotation que nous avons appliquées (section 2.). Enfin, nous présenterons nos perspectives de recherche.

1 PRINCIPES DE CONSTITUTION DU CORPUS WIKICONFLITS

Constituer un corpus pose d'emblée la question de la représentativité de l'échantillon choisi. Nous avons fait le choix de sélectionner et de structurer en corpus un ensemble de pages ayant suscité des discussions conflictuelles dans le champ des sciences et des techniques. Il s'agit plus spécifiquement de conflits entre éditeurs (ou contributeurs) se déroulant dans les coulisses des pages d'articles, dans la section appelée "Discussion" ("Talk", en anglais).

Encadrés par des dispositifs de règles et de médiation spécifiques, les conflits ont une forme spécifique et régulée dans Wikipédia, que nous tenterons de restituer en 1.1. La section suivante 1.2. explicite le choix des pages, issues du champ des sciences et techniques.

1.1 LES CONFLITS DANS WIKIPÉDIA

Si la grande majorité des discussions éditoriales de Wikipédia sont non-conflictuelles et constructives, les consensus peuvent parfois être difficiles à dégager entre des contributeurs nombreux et hétérogènes, entraînant des conflits plus ou moins sévères. Du côté des sciences sociales, de

nombreuses études ont été menées dans la précédente décennie sur cette question, mettant en lumière des motifs conflictuels (Viegas et al. 2004) et des phénomènes remarquables : certains domaines génèreraient ainsi davantage de conflits, i.e. les pages rattachées aux catégories *Religion* et *Philosophie* seraient par exemple plus conflictuelles dans le Wikipédia anglophone (Kittur et al. 2007) ; par ailleurs, à l'heure où les Wikipédia les plus avancés (e.g. les versions anglophone ou francophone) se stabilisent, avec de moins en moins de nouveaux articles créés, les travaux de maintenance et les conflits autour des principes et des procédures mêmes de Wikipédia prennent de plus en plus d'importance, comme le soulignaient déjà Kittur et al. en 2007.

Les conflits sont généralement mesurés sur la base des *reversions* successives (ang. *reverts*, *i.e.* retour à une version antérieure de l'article) qui affectent parfois l'édition d'un article (Viegas et al. 2004, Brandes & Lerner 2008, Kittur et al. 2007, Suh et al. 2007, Kittur & Kraut 2010), générant ce qu'on appelle une *guerre d'édition*. La longueur des pages, le nombre de révisions des articles et des pages de discussion, ainsi que les insertions et suppressions de mots ou de caractères entre utilisateurs sont également des critères communément utilisés pour observer et détecter les conflits.

Les études linguistiques se sont de leur côté peu penchées sur l'étude des conflits dans Wikipédia, et sur le projet encyclopédique en général ; quelques travaux peuvent néanmoins être signalés, qui se concentrent sur les pages de discussion autour des articles et procèdent à l'annotation des fils de discussion conflictuels / non conflictuels (Denis et al. 2012) ou des messages suivant leur contenu (Schneider et al. 2010) ou les actes de langage que l'on y observe (Bender et al., 2011).

Les conflits prennent ainsi forme sur plusieurs lieux privilégiés (*namespaces*) loin d'être étanches puisqu'un conflit démarre généralement sur un article (éditions antagonistes), peut donner lieu à des discussions animées sur les différentes pages dédiées de l'encyclopédie, et éventuellement se solder par une médiation, voire un arbitrage, le comité d'arbitrage étant la

plus haute instance juridique de Wikipédia (Jacquemin et al. 2008).

Wikiconflits, dont la finalité est d'encourager les études sur Wikipédia en mettant à disposition de la communauté académique un corpus exploitable et dûment constitué, inclut ainsi l'ensemble des pages potentiellement concernées par chaque conflit retenu, i.e. les articles, leurs historiques et leurs différentes versions, ainsi que l'ensemble des discussions qui s'y rattachent, qui sont au cœur de nos intérêts et de notre corpus, et que nous détaillons précisément en 2.1.

1.2 DES CONFLITS DU CHAMP DES SCIENCES ET DES TECHNIQUES

Pour ce qui relève de la thématique des pages sur lesquelles les discussions portent, nous avons fait le choix d'opérer une sélection en lien explicite avec le champ des sciences et techniques, ce qui inclut les domaines des sciences naturelles, formelles, et de l'ingénieur aussi bien que ceux des sciences humaines et sociales.

Seules les pages rattachées à ce champ ont été conservées, excluant par exemple les articles consacrés aux hommes et femmes politiques, sportifs ou vedettes du cinéma, aux lieux ou à certaines entreprises et marques qui entraînent des formes différentes de conflits. En effet, les pages sélectionnées devaient contenir et déployer dans la discussion un nombre conséquent d'arguments à caractère scientifique. Nous ne nous sommes pas arrêtés sur la validité de ces arguments - les arguments fallacieux présentant la rhétorique de l'objectivité scientifique nous important tout autant.

En ceci, c'est la possibilité de croiser le cadre normatif des règles d'éditions de WP et celui de l'objectivité du discours scientifique qui nous a paru intéressante. Elle donne une forme de stabilité à la discussion, dans la mesure où les deux cadres normatifs se superposent et se répondent. Cette forme peut être

illusoire et n'empêche pas le conflit de se dérouler de manière désordonnée et sans fin. Mais cette rencontre normative est représentative des pratiques comme de l'ethos de Wikipédia.

Ensuite, un deuxième croisement nous a permis d'asseoir et de préciser ce choix, entre la culture du conflit de Wikipédia et les méthodologies d'analyse de la sociologie pragmatique, des sciences et de l'innovation. En effet, les conflits d'éditions sont célèbres dans le folklore d'Internet, et représentatifs de l'ethos démocratique des participants au projet encyclopédique numérique (Cardon, 2010). Or, les courants sociologiques cités proposent deux cadres d'analyse fertiles pour ce terrain. Tout d'abord, la méthode de la cartographie des *controverses scientifiques*, utilisée en particulier en *Science and Technology studies* (STS) permet de mettre en exergue les liens entre acteurs d'une dispute – considérée comme l'un des bricolages discursifs de la fabrication des faits scientifiques (Latour, 1995). On a donc cherché dans les conflits matière à controverses s'étant déroulées dans le champ des sciences et des techniques – ou présentant la potentialité à faire controverse. Ensuite, la théorie de l'*association hybride*, nous a semblé une hypothèse intéressante à tester sur ce type de corpus, dans la mesure où elle caractérise la variété des statuts et autorités sociales présentes dans les discussions et mobilisations collectives dans la production d'un savoir (Callon et al., 2011). Les discussions choisies sont ainsi des échantillons représentatifs de disputes d'autorités liées aux conflits d'édition – que cette autorité soit interne (s'appuyant sur une légitimité issue des règles de gouvernance de WP) ou externe (s'appuyant sur une légitimité issue des normes socio-professionnelles hors WP). Cette hypothèse est particulièrement intéressante à tester ici dans la mesure où elle impliquerait, non pas seulement une participation hybride, mais aussi la présence de protocoles formels (suivis ou disputés) qui encadrent la mobilisation.

Ces deux croisements ont en définitive une double dimension programmatique et méthodologique pour des

analyses futures de ces matériaux grâce aux corpus constitués, en dehors du domaine seul de l'analyse linguistique.

1.3 MODE DE CONSTITUTION DU CORPUS WIKICONFLITS

Le choix des pages conflictuelles elles-mêmes s'est appuyé sur une première pré-sélection des lieux les plus susceptibles d'abriter un conflit. Un ensemble de pages a ainsi été arrêté en date d'octobre 2013 pour évaluation et intégration dans le jeu de données final.

Quatre ensembles de pages ont été retenus et examinés :

1. les 73 pages faisant l'objet d'une procédure de **médiation** à cette date sur la page « salon de médiation »[2] qui répertorie les conflits en cours de médiation ;
2. les 214 pages faisant l'objet d'un **désaccord de neutralité**, signalé par un bandeau apposé sur les articles, la neutralité de point de vue (NPV) étant l'un des principes fondateurs de Wikipédia. Les articles se doivent effectivement d'être neutres : l'ensemble des points de vue pertinents émis sur un objet doit être restitué, avec leurs auteurs et leurs contextes, et sans privilégier un point de vue sur un autre, afin notamment d'éviter les conflits d'intérêt. Lorsque les éditeurs estiment que ce principe n'est pas respecté, on parle de *désaccord de neutralité* ;
3. les 546 pages faisant l'objet d'un **désaccord de pertinence** : un bandeau indique la présence d'un désaccord relatif au contenu de l'article, par exemple lorsque certains contributeurs s'interrogent sur la pertinence d'un article, d'une section ou d'un point de vue jugé minoritaire ou controversé ;

[2] https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Le_salon_de_m%C3%A9diation

4. les 169 articles faisant l'objet d'une **protection**[3] du fait d'une guerre d'édition en cours, ou d'une **semi-protection longue**[4], *i.e.* les articles régulièrement et durablement vandalisés du fait de la sensibilité de leur thème (politique, religion, sexe…) ont également été examinés.

Au final, 1002 pages ont été examinées et deux principes ont régi le choix des pages retenues : (i) l'inscription des articles dans le champ des sciences et des techniques, selon les critères que nous avons détaillés dans la section 1.2. ; et (ii) la présence de conflits dans les pages de discussion. Dans cette perspective, le niveau et la forme des conflits des fils de discussion dans les pages de discussion rattachées aux articles ont naturellement été considérés ; comme nous l'avons déjà évoqué, un conflit peut parfaitement être acté, à travers des éditions ou des reversions successives de l'article, sans être nécessairement discuté. Ont donc été écartées les pages qui contenaient peu de fils et d'interactions conflictuelles, qui représentaient d'ailleurs la majorité des pages que nous avons examinées.

Enfin, il nous faut mentionner que des contraintes techniques liées au volume total du corpus et aux modalités de traitement des pages, de la conversion du corpus en XML-TEI à l'annotation et la correction semi-manuelle des fils de discussion (voir 2.3.) nous ont amenées à réduire drastiquement la taille du corpus et le nombre de pages sélectionnées.

Des 1002 pages candidates collectées dans les lieux de conflits de WP n'ont finalement été retenus que sept articles et leurs révisions et discussions associées. Ce passage de 1002 à 7 peut paraitre étonnant et très limité, mais il faut souligner que nous souhaitions disposer d'un corpus correctement annoté en termes d'interactions ; or, et c'est lié aux spécificités de la syntaxe Wiki, une part importante de correction manuelle a été requise sur les fils de discussion (2.2). Nous avons donc choisi

[3] https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Article_prot%C3%A9g%C3%A9
[4] https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Page_en_semi-protection_longue

de livrer un corpus plus limité, mais propre, qui nous permettra, ainsi qu'aux collègues travaillant sur les interactions, de travailler sur les fils et les formes des interactions. L'ensemble retenu reste conséquent (4,456 posts rédigés par 3,971 contributeurs, soit 489,000 tokens – 330 Mo au format ZIP).

Notre sélection tient compte des thématiques récurrentes représentatives des conflits dans le champ des sciences et techniques :

- trois pages s'inscrivent dans les débats autour des **pseudo-sciences** : *Quotient Intellectuel, Psychanalyse, Chiropratique* ;
- deux pages portent sur des problématiques de **technosciences** : *Eolienne* et *Organismes Génétiquement Modifiés (OGM)* ;
- une page a trait à la **légitimité académique et scientifique d'individus** étant par ailleurs des personnalités médiatiques : *Igor et Grichka Bogdanoff* ;
- une page s'inscrit dans les conflits relatifs à **l'histoire et l'épistémologie d'une discipline** : *Histoire de la logique*.

Deux précisions doivent être énoncées pour comprendre la complexité de l'argumentation conflictuelle sur ces pages. Tout d'abord, on trouve certains arguments transversaux à ces thèmes. Parmi les questions transversales :

- celle de l'autorité et du statut socioprofessionnel d'un individu cité comme référence dans un article, par exemple pour les théoriciens de la logique à citer dans l'histoire de la discipline, ou pour les personnalités médiatiques que sont les frères Bogdanoff ; ou d'un individu intervenant dans la discussion : les Bogdanoff, ou leur représentant, prenant part au conflit et à la guerre d'édition sur leur propre article, mais

aussi divers experts ou scientifiques ou se proclamant comme tels ;

- celle des rapports entre science et société : l'alliance des sciences avec des intérêts politiques ou industriels, en particulier dans les conflits autour de ce qu'on peut appeler les « technosciences » ; les problèmes de reconnaissance d'une discipline dans l'espace public restreint (académique) ou large (médiatique), comme dans le cas des débats autour des pseudo-sciences.

Ensuite, dans les termes mêmes du débat, des glissements de registre argumentatif sont fréquents : de l'épistémologique au sociologique, à l'idéologique… Notre sélection n'a pas seulement pour critère le sujet des articles, mais aussi le type de discussion engagée dans les coulisses de l'article. Dans le cadre d'un sujet controversé, la limite entre débat d'ordre scientifique et technique et débat d'ordre politique et idéologique est évidemment très poreuse (la controverse telle que définie par les STS se fonde d'ailleurs notamment sur cette incertitude). C'est pourquoi nous avons, parmi la première sélection des articles à conflit, ciblé ceux dont les discussions illustrent un recours explicite des participants à des types de raisonnement relatifs aux langages scientifiques et techniques et au discours de la preuve (débat sur une théorie, un concept, une méthode, une démonstration, une discipline…) – quelles que soient la qualité du raisonnement ou des arguments avancés, ou les prises de position politiques ou dérives idéologiques flagrantes les nourrissant.

2 STRUCTURATION ET ANNOTATIONS

Une fois les sept pages retenues suivant les principes de représentativité explicités dans la section précédente, s'est posée la question de la structuration et de l'annotation du corpus.

2.1 LES CLUSTERS DE PAGES

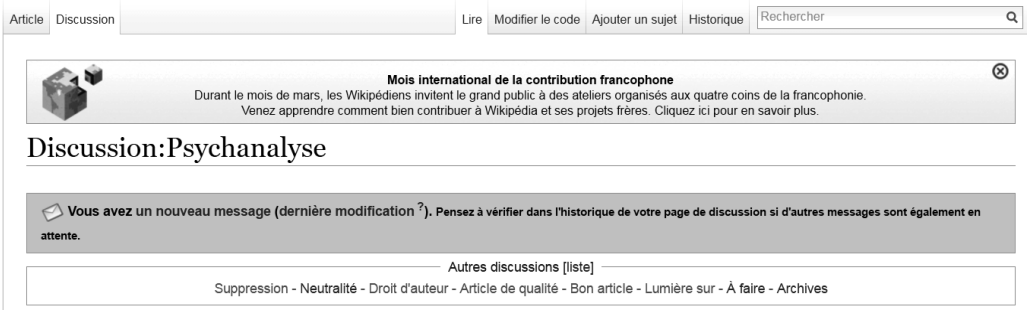
Dans la mesure où un conflit en cours peut traverser plusieurs pages, comme nous avons pu l'observer dans quelques centaines de pages, nous avons veillé à extraire l'ensemble des pages susceptibles d'abriter un conflit autour d'une page, que nous appellerons désormais *clusters de pages*.

Deux ensembles de pages ont ainsi été pris en compte et sont regroupés au sein de chaque cluster :

1. **Les lieux de révision** : les révisions se manifestent sur la page de l'article et la page historique liée, pointant sur les différentes versions de l'article et livrant une mémoire fiable de son évolution. Si nécessaire, ces pages supplémentaires peuvent donc être consultées et analysées pour étayer l'analyse des pages de discussion. Comme nous l'avons déjà évoqué, d'autres traces des conflits en cours sont présentes sur la page de l'article, sous la forme de bandeaux signalant des conflits potentiels ou avérés dans le texte (désaccords de pertinence ou de neutralité) ;
2. **Les lieux de discussion** : chaque article possède donc une page de discussion également éditable, où le conflit prend une forme verbale et s'argumente. La discussion prend toutefois régulièrement une forme non linéaire, d'une part parce que les discussions sont régulièrement archivées par les utilisateurs[5] et d'autre part parce que les bandeaux apposés sur l'article génèrent également des pages de discussions spécifiques. Il en va par exemple ainsi de *Psychanalyse*, qui a été taxée de *non neutralité* ; aussi peut-on voir un lien vers une page

[5] L'archivage n'étant pas automatique, les critères d'archivage, de même que la forme ou l'emplacement-même des archives (comparer l'archivage des pages *psychanalyse* et *chiropratique* - les archives étant accessibles via un hyperlien en haut de la page de discussion pour cette dernière page) varient beaucoup d'une page à l'autre.

de discussion *Neutralité* dans la rubrique *Autres discussions* de la page (figure 1).



1 La page de discussion Psychanalyse

En outre, la discussion, et donc les conflits, peuvent potentiellement s'étendre au-delà de l'espace éditorial propre au dispositif, déjà complexe, de l'onglet discussions lié directement à l'article, notamment dans les pages utilisateurs des participants ou dans l'historique des révisions - chaque révision peut en effet être commentée, ce qui génère des discussions potentielles.

Pour chaque article sélectionné, nous avons donc extrait quatre types de pages :

1. la page *article* et son historique wiki avec les diffs ;
2. la page *discussion* de l'article et son historique wiki avec les diffs ;
3. les éventuelles autres archives et les pages de discussion *Neutralité,* le plus souvent accessibles depuis le lien dans l'encadré *Autres discussions* de chaque page de discussion. Les pages de *discussion* (article et neutralité) ainsi que les archives existantes ont été regroupées au sein d'une même page *discussion* ;
4. enfin, ont été également extraites les pages *discussion* des contributeurs principaux de l'article (au moins 10 éditions de l'article).

2.2 DE L'INTÉRÊT ET DES LIMITES DU FORMAT WIKI POUR DISCUTER

Les pages de discussion Wikipédia ont cette spécificité d'être, à l'instar des articles, en format wiki. Ceci génère des pratiques d'interaction et de structuration des fils et des messages qui diffèrent des formes plus standard de type forum. De fait, les messages postés par les contributeurs ne sont pas associés à une action spécifique, comme le décrivent Beißwenger et al. (2012) :

> [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).

Malgré des recommandations claires concernant le format des messages[6] (niveau de réponse, signature et date etc.), les pages de discussion Wikipédia ont un aspect hybride, souvent à mi-chemin entre conversation et liste écrite et annotée de propositions ou de points à traiter. En outre, la structuration des messages et des fils de discussion n'est pas toujours limpide puisque les wikipédiens restent libres d'éditer les fils et les messages à leur convenance. Un contributeur peut par exemple choisir de diviser son message en différentes sections, que l'on interprétera comme autant de fils tandis que l'ordre des tours de parole n'est pas toujours linéaire. Les éditeurs étant susceptibles de ne pas dater ni signer leurs messages (ce qui était le cas de 10% des messages de notre corpus), ou encore d'intervenir au milieu d'un message précédent, il n'y a en effet aucune garantie que l'ordre chronologique, ou l'ordre des interactions soit respecté.

A fortiori, nous avons pu constater que les wikipédiens francophones s'affranchissaient régulièrement des règles d'édition des discussions ; du moins était-ce le cas d'une large

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

majorité des pages que nous avons examinées dans le cadre de ce projet.

Dans la mesure où les discussions éditoriales pouvaient difficilement être observées sans une délimitation correcte des interactions, nous avons adopté une procédure semi-automatique : les messages et les fils ont été annotés automatiquement sur la base de l'annotation Wiki (Poudat et al. 2014) avant d'être corrigés manuellement.

La correction manuelle s'est avérée particulièrement fastidieuse, nécessitant la lecture attentive de chaque message et de chaque fil, leur réagencement le cas échéant et la vérification régulière de l'historique des discussions pour récupérer les métadonnées (noms et dates) des messages non signés. Dans quelques cas, ces dernières méta-informations étaient étonnamment absentes des historiques et ont dû être induites du contenu des messages et des indications temporelles qui se trouvaient dans les messages précédents ou suivants - la date du message restant dans ce dernier cas relative (*before_date-d'un-autre-message* ou *after_date-d'un-autre-message*).

2.3 ANNOTATION TEI-CMC

Dans un dernier temps, afin de fournir un corpus normé et facilement utilisable par d'autres chercheurs, le corpus a été converti au format TEI-CMC (Chanier et al. 2015, Beißwenger et al. 2012, 2015), qui est une extension de la TEI-P5, format d'encodage largement reconnu et utilisé par les humanités numériques. L'encodage de Wikiconflits, et plus largement, du corpus CoMeRe en TEI-CMC facilite donc à la fois l'échange des données entre chercheurs et la construction de corpus de référence.

Si les articles ont également été convertis en TEI-P5 (Poudat et al. 2014), nous ne détaillerons ici que les spécificités des pages de discussion (figure 2).

Les pages de discussion Wikipédia étant segmentées en différents sujets de discussion que l'on comprend comme autant de fils de discussions, ceux-ci sont délimités par des éléments ** **.

En accord avec les travaux du groupe européen TEI-CMC[7], les travaux déjà entamés par les collègues allemands sur Wikipédia (Margaretha & Lüngen 2014) et dans la continuité des autres corpus CoMeRe, chaque message est balisé par l'élément **<post>**, qui est l'élément principal utilisé dans tous les projets CoMeRe.

@xml:id	identifiant du message. Sa valeur est unique dans le fichier XML entier.
@when-iso	date de post, sa valeur est sous format ISO 8601, *e.g.* « 2006-04-29T19:57 » pour un message rédigé le 29 avril 2006 à 19:57
@when-custom	date personnalisée, qu'on utilise au cas où un post n'a pas de date, la valeur par défaut étant « unknown »
@who	auteur du message - à noter que dans Wikiconflits, les noms des contributeurs sont représentées par un identifiant - notamment pour pallier le problème des utilisateurs qui changeraient d'alias. Les contributeurs (alias et id) sont livrés dans un fichier séparé.
@n	présente le niveau d'indentation du message.
@ref	identifiant du message auquel répond le message courant. Lorsque le niveau d'indentation est « 0 », cet attribut n'apparaît pas.
<p>	Un message contient un ou plusieurs <p>
<signed>	Signature de message. Un message contient un ou zéro élément <signed>

2 Attributs et sous-éléments XML associés aux messages <post>

[7] http://www.tei-c.org/Activities/SIG/CMC/

CONCLUSION ET PERSPECTIVES DE RECHERCHE

Nous avons présenté notre travail de constitution d'un corpus CMC, réalisé dans le cadre de l'action *Nouvelles collectes* du projet national CoMeRe issu du groupe de travail portant sur les *Nouvelles formes de communication*. Nous avons décrit le cadre et la démarche suivie pour construire le corpus Wikiconflits, le normaliser et le rendre accessible et utilisable par d'autres chercheurs - Wikiconflits est comme l'ensemble des corpus CoMeRe librement disponible sur la plateforme de la TGIR Ortolang Huma-num (Poudat et al., 2015).

Nous exploitons à l'heure actuelle ce corpus en poursuivant différentes perspectives de recherche : Wikiconflits est d'une part naturellement exploité pour étudier les conflits dans Wikipédia et différentes pistes sont d'ores et déjà privilégiées, comme par exemple les traits distinctifs de la langue des conflits, leur détection automatique et surtout leur détection préventive, ce qui permettrait de les déceler et de les résoudre plus rapidement. Dans cette optique, une annotation de la conflictualité des fils de discussion est nécessaire : une page de discussion conflictuelle ne contient pas que des fils et des messages conflictuels, et l'annotation des fils nous permettra de mettre au jour les spécificités des discussions conflictuelles en les contrastant aux fils plus pacifiques.

D'autre part se pose également la question de l'exploration de ce type de corpus, que nous avons déjà commencé à interroger, en mobilisant les méthodes classiques des statistiques textuelles (Poudat et al. 2016). Les genres Web 2.0. mettent en effet au défi les méthodes classiques d'exploration de corpus et d'analyse de données textuelles : comment décrire ces objets hybrides et tentaculaires, articulant différents genres et de nombreuses pages en lien les unes avec les autres ? Comment prendre en compte la dimension temporelle de l'écriture et des discussions ?

Comme nous l'avons déjà évoqué, nous espérons enfin que notre corpus participera, aux côtés d'autres initiatives (Ho-Dac & Laippala, ce volume) au développement des études linguistiques sur Wikipédia en France.