



HAL
open science

Corpus DiscoWiki: Un corpus de Discussions Conflictuelles du Wikipédia francophone

Lydia-Mai Ho-Dac, Céline Poudat

► To cite this version:

Lydia-Mai Ho-Dac, Céline Poudat. Corpus DiscoWiki: Un corpus de Discussions Conflictuelles du Wikipédia francophone. Journée d'étude CORLI: Traitements et standardisation des corpus multimodaux et web 2.0., May 2018, Paris, France. hal-01873824

HAL Id: hal-01873824

<https://hal.science/hal-01873824>

Submitted on 13 Sep 2018

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.



Corpus DiscoWiki

Un corpus de Discussions Conflictuelles du Wikipédia francophone

Lydia-Mai HO-DAC (1), Céline POUDAT (2) (1) CLLE-ERSS – Université Toulouse UT2J hodac@univ-tlse2.fr; (2) BCL – Université Nice – Côte d'Azur celine.poudat@unice.fr



Un corpus de discussions médiées par les réseaux

- Projet TGIR Human-Num CORLI 2016
- Corpus issus de la rencontre entre 2 corpus aux objectifs complémentaires :

Disposer d'un large corpus de Communications Médiées par les Réseaux → **WikiTalk** (Ho-Dac et al. 2017a et 2017b)

Annoté au niveau de l'interaction et de l'apparition de conflits entre locuteurs → **Wikiconflits** (Poudat et al. 2017)

Les pages de discussion Wikipédia : un espace de discussion associé à chaque article

Un préalable à l'écriture collaborative (Viégas et al. 2007)

- Mis en place dès la naissance du projet Wikipédia, pour aider les co-auteurs à prendre des décisions et arriver à des consensus
- Un regard inédit sur les processus de construction de contenu, rédaction et de négociation

"From a scientific point of view, article Talk pages are a unique type of web discourse and a valuable resource for the humanities and writing sciences, since the discussions develop in parallel with the discussed articles and provide insights into the meta level of the collaborative writing process that normally remains hidden. With structured access to this resource, the linguists and researchers in the writing sciences have the unparalleled possibility to observe these hidden processes without having to conduct interviews or carrying out supervised field experiments." (Ferschke, 2014 : 111)

- Structure d'une page de discussion, voir figure ci-contre

- Chaque article est associé à une **page de discussion principale** (dont une partie peut être **archivée**) qui peut être complétée par des **pages parallèles** contenant des discussions sur des points particuliers



- Chaque page de discussion est composée de **fils de discussion** lancés par un contributeur qui poste un premier message et donne un **titre** au fil de discussion

- Chaque fil est composé de messages appelés « **posts** », **datés** qui sont **signés** ou laissés anonymes et peuvent présenter des **niveaux** d'inclusion plus ou moins complexes

- De nombreuses méta-données sont accessibles : le portail thématique et niveau d'avancement de l'article, le caractère controversé ou polémique de l'article, etc.

- Aperçu quantitatif du corpus WikiTalk

	#Pages	#Fils	#Posts	#Mots
WikiTalk	366 326	> 1 millions	> 3 millions	~ 160 millions

Les conflits dans la communauté Wikipédia

"Les premiers mois d'existence de l'encyclopédie sont marqués par l'apparition de conflits entre contributeurs. Selon Sanger [le fondateur du projet anglophone], des universitaires [...] écrivant dans leur domaine de spécialité pour Wikipédia, auraient abandonné l'encyclopédie [...] lassés de croiser le fer avec des "wiki-anarchistes" d'un niveau d'expertise bien moindre." (Sahut, 2015 : 241)

- Des conflits de niveaux d'expertise, de points de vue ou encore d'intérêt (cas d'utilisation de Wikipédia pour l'auto-promotion)
- Une nécessité indispensable de gérer les agressions et les conflits

The Wikimedia foundation found that 54% those who had experienced online harassment expressed decreased participation in the project where they experienced the harassment

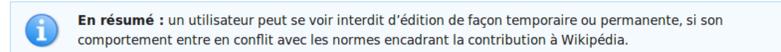
- Toxicité de certains posts/contributeurs : 11,7 % de posts « agressifs » (annotation de 115 000 posts, Wulczyn et al., 2017)

- Des outils manuels de prévention au sein de Wikipédia :

Bandeaux d'alerte :



Blocage en écriture



Structure d'une page de discussion Wikipédia

Discussion:Attentats du 11 septembre 2001 < titre de discussion

Autres discussions [liste]

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

pages parallèles >

Avancement Importance pour le projet

Maximum	États-Unis (discussion • critères • liste • stats • hist. • comité)
	New York (discussion • critères • liste • stats • hist. • comité)
	Histoire (discussion • critères • liste • stats • hist. • comité)
	Politique (discussion • critères • liste • stats • hist. • comité)
	Criminologie (discussion • critères • liste • stats • hist. • comité)
Elevée	Wikipédia 1.0/Les plus consultés (discussion • critères • liste • stats • hist. • comité)
	Sélection transversale (discussion • critères • liste • stats • hist. • comité)
	Aéronautique (discussion • critères • liste • stats • hist. • comité)
Faible	Islam (discussion • critères • liste • stats • hist. • comité)

meta-données >

- portail
- avancement

liste des fils >

titre d'un fil >

post niveau 1 >

post niveau 2 >

post niveau 3 >

post niveau 4 >

post niveau 5 >

post niveau 6 >

Sommaire [masquer] archivage des anciens fils > Archives

1 Réactions de la défense aérienne

2 le titre

3 Théorie du complot

4 Enquête sur les faux passeports des terroristes en Afghanistan. Précisions passeports retrouvé dans les décombrés

Wikipédia se ridiculise [modifier le code]

La dernière phrase de l'introduction est tirée d'une vidéo Dailymotion sur un talk show people de Fox News « Certaines personnes ou groupes de personnes contestent la version officielle... » Wikipédia se ridiculise en s'abaissant à ce niveau sur un sujet comme le 11 septembre.

> +1. Pas besoin de se faire l'écho en introduction de théories complotistes fantaisistes et minoritaires. Surtout en renvoyant vers un article aussi PoV que **Théories du complot à propos des attentats du 11 septembre 2001** et en sourçant avec une vidéo FoxNews hébergée sur Dailymotion (j'espère par ailleurs qu'avoir simplement donné cet avis ne me vaudra pas un blocage, hein) SM ++ 23:24 ** 21 novembre 2010 à 12:18 (CET)

signature : pseudo + date

Qu'on y ajoute du crédit ou pas (et personnellement, je trouve que ces théories sont toutes plus ridicules les unes que les autres), il n'en reste pas moins qu'il existe un gros battage médiatique autour de ces "théories" et qu'un nombre non négligeable de personnes semblent les considérer plausibles. Au point que c'est devenu un indéniable phénomène de société. Ce qu'il faut qu'il est difficile que Wikipedia les passe sous silence. Au demeurant, une section de l'article y est consacrée avec renvoi vers "l'article" ad hoc. C'est sur cette considération que j'avais **modifié** la phrase mise en introduction tout en restant réservé par rapport à la référence que je n'avais pas pu consulter à ce moment. L'ayant vue entre-temps, il me semble qu'il faut conserver la phrase en l'état (elle ne fait de rendre compte d'une réalité - les théories du complot - sans prendre position) tout en supprimant la référence (il n'est du reste pas requis de sourcer ce qui figure dans une introduction), qui est loin d'être ce qui se fait de mieux en la matière. --Lebob (d) 21 novembre 2010 à 13:03 (CET)

> j'entends bien. Tu as parfaitement raison, Wikipédia ne peut pas ignorer l'existence des théories du complot, et la section qui y est consacrée a toute sa place. Simplement, je ne suis pas du tout convaincu par la pertinence d'en parler dès l'introduction. Encore moins avec la source actuelle qu'il faut, dans tous les cas, retirer SM ++ 23:24 ** 21 novembre 2010 à 13:54 (CET)

+1 wikignome (d)

> Bien d'accord sur la question de la source: d'une part il n'est pas exigé de mettre de sources en introduction et, d'autre part, celle-là est particulièrement mauvaise. Pour le reste, je suggère d'attendre d'autres avis sur le maintien (ou pas) de cette phrase ou d'une possible reformulation d'icelle. --Lebob (d) 21 novembre 2010 à 14:03 (CET)

> ✓ Source retirée, sans toucher au reste pour l'instant, faute de consensus sur ce qu'il convient de faire de cette phrase: SM ++ 23:24 ** 21 novembre 2010 à 14:12 (CET)

Composition du corpus DiscoWiki

Article associé	# Pages*	# Fils	# Posts	# Mots
Chiropratique	2 (+1)	38	285	47 414
Quotient Intellectuel	1 (+0)	45	169	22 815
Psychanalyse	7 (+1)	136	729	80 912
OGM	8 (+1)	373	3 173	345 482
Eolienne	1 (+1)	43	145	15 792
Igor et Grichka Bogdanoff	3 (+1)	129	727	137 809
Histoire de la Logique	1 (+0)	9	45	3 850
Vladimir Poutine	1 (+1)	68	359	50 664
Attentats du 11/09	3 (+1)	134	1 149	135 103
DiscoWiki	27 (+7)	975	6 781	839 841

* indique le nombre pages principales (courantes et archivées) et, entre parenthèses, le nombre de pages parallèles "Neutralité" associées

Annotation manuelle des fils de discussion

- Objectifs : (1) proposer une description de l'expression linguistique des conflits dans les discussions Wikipédia (et par ext., les CMC) ; (2) développer des systèmes d'observables opératoires pour explorer et caractériser les discussions Wikipédia (et par ext., les CMC)

- Grille d'annotation des fils composés de plus de 2 posts

- présence d'une médiation au cours du fil

Bonsoir, J'ai bloqué 3 jours Dujo pour passage en force. Il n'est pas acceptable que le travail collaboratif de discussion qui a lieu ici soit perturbé par des renouveaux de guerre d'édition. Je vous souhaite de continuer à débattre dans le respect mutuel, comme suggéré par ConradMayhew à 7:43. Bien cordialement,

- présence de marqueurs d'agression/attaque verbale au cours du fil

Mais je vous emmerde espèce de grosse vache stalinienne. Je suis biologiste, allez donc terminer vos primaires arriéré prétentieux que vous êtes...

- degré de politesse du premier post

- topique du premier post : sources de l'article, structuration de l'article, neutralité de point de vue, contenu de l'article (définition, contenu erroné ou imprécis etc.), autre

- capacité du dernier post de clore la discussion / résoudre le conflit

Merci pour vos commentaires, très intéressants.

- Résultats préliminaires sur 384 fils (Pages associées aux articles « Quotient Intellectuel », « Éolienne » et « Histoire de la logique »)

Annotation	Valeur	%
Médiation	= 1 (vs. 0)	2
Agression verbale	> 0 (= 1 ou 2 selon le degré)	5
Politesse du 1er post	= 1 (vs. 0)	18
Topique**	Source de l'article	20
	Structuration de l'article	11
	Neutralité de point de vue	13
	Contenu de l'article	62
	Autre	7
Dernier post	= clôt/résout la discussion/le conflit	31

** L'annotation permet d'associer plusieurs topiques à un premier post. Pour 14 % des fils annotés, le premier post a été associé à plus d'un topique. Dans la grande majorité des cas, il s'agit d'une combinaison impliquant le topique « contenu ».

FERSCHKE, O. (2014). The Quality of Content in Open Online Collaboration Platforms : Approaches to NLP-supported Information Quality Management in Wikipedia. PhD thesis, Technische Universität, Darmstadt, 2014.

HO-DAC, L.-M., LAIPPALA, V., POUDAT, C. AND TANGUY, L. (2017a) Exploring Wikipedia talk pages for conflict detection. In Darja Fišer and Michael Bellwenger (dir.) Investigating Computer-Mediated Communication : Corpus-Based Approaches to Language in the Digital World, Ljubljana University Press, Faculty of Arts, p.146-168.

HO-DAC, L.-M. ET LAIPPALA, V. (2017b). Le corpus WikiDisc : ressource pour la caractérisation des discussions en ligne. In Clara R. Wigham et Gudrun Ledegen (dir.) Corpus de communication médiée par les réseaux : construction, structuration, analyse., L'Harmattan, coll. "Humanités numériques", p.107-124.

POUDAT, C., GRABAR, N., PALOQUE-BERGES, C., CHANIER, T. et KUN, J. (2017). Wikiconflits : un corpus de discussions éditoriales conflictuelles du Wikipédia francophone. In Wigham, C.R & Ledegen, G., Corpus de communication médiée par les réseaux : construction, structuration, analyse. Collection Humanités numériques. Paris : L'Harmattan.

SAHUT, G. (2015). Wikipédia, une encyclopédie collaborative en quête de crédibilité : le référencement en questions. Thèse de doctorat en Sciences de l'information et de la communication, Université Toulouse Jean Jaurès, Université de Toulouse.

VEIGAS, F., WATTENBERG, M., KRISSE, J. AND VAN HAM, F. (2007). Talk Before You Type: Coordination in Wikipedia. In 40th Annual Hawaii International Conference on System Sciences, HICSS 2007, pages 78-79.

WULCZYN, E., THAIN, N. AND DIXON, L. (2017). Ex machina: Personal attacks seen at scale. In Proceedings of the 26th International Conference on World Wide Web, ACM, p. 1391-1399.7.