Corpus DiscoWiki: Un corpus de Discussions Conflictuelles du Wikipédia francophone

Lydia-Mai Ho-Dac, Céline Poudat

Corpus DiscoWiki

Un corpus de **Dis**cussions **Co**nflictuelles du **Wiki**pédia 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 décision 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 accesibles : 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

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 Beißwenger (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 Ciara 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 : LHarmattan.

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.

VIEGAS, F., WATTENBERG, M., KRISS, J. AND VAn Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. In *40th Annual Hawaii International Conference on System Sciences*, HICSS 2007, pages 78-78.

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.

Les conflits dans la communauté Wikipédia

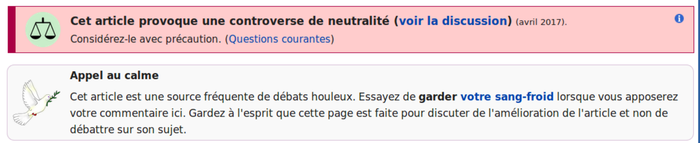
"Les premiers mois d'existence de l'encyclopédie sont marqués par l'apparition de conflits entre contributeurs. Selon Sanger [le fondateur du projet anglophone], des universitaires [...] écrivant dans leur domaine de spécialité pour Wikipédia, auraient abandonné l'encyclopédie [...] lassés de croiser 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

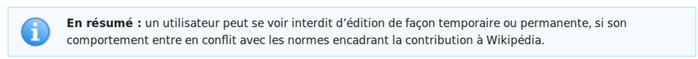
The Wikimedia foundation found that 54% those who had experienced online harassment expressed decreased participation in the project where they experienced the harassment

- 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

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 ».