Monologuer dans une discussion en ligne ? Profilage des interactions entre les rédacteurs de la Wikipédia

Ludovic Tanguy, Céline Poudat, Lydia-Mai Ho-Dac

Monologuer dans une discussion en ligne ? Profilage des interactions entre les rédacteurs de la Wikipédia

Ludovic Tanguy [1], Céline Poudat [2] et Lydia-Mai Ho-Dac [1]
[1] CLLE : CNRS & Université de Toulouse [2] BCL : CNRS & Université Côte d'Azur
ludovic.tanguy@univ-tlse2.fr, celine.poudat@univ-cotedazur.fr, lydia-mai.ho-dac@univ-tlse2.fr

Introduction

La présente étude s'inscrit dans un travail plus général de profilage des interactions en ligne entre les rédacteurs de la Wikipédia. Ces discussions prennent place dans des espaces spécifiques associés à chacun des articles de l'encyclopédie en ligne et constituent des échanges autour des différentes décisions que nécessite l'écriture collaborative. Elles permettent notamment aux contributeurs de se coordonner pour rédiger et mettre à jour l'article, de régler leurs éventuels différends.

Sur la base d'un corpus constitué de plus de 300 000 discussions associées aux articles de Wikipédia dans sa version française, nous proposons une investigation globale des interactions et des pratiques. Cette étude se concentre sur un phénomène relativement inattendu : les monologues. À travers un examen ciblé de ce type d'interaction, nous dégageons un ensemble de pratiques qui traduisent la spécificité des discussions Wikipédia par rapport, notamment, aux forums de discussion classiques.

Vue d'ensemble des discussions Wikipédia

Notre travail se base sur un corpus de discussions élaboré à partir du *dump* de la Wikipédia française téléchargé en septembre 2019. Nous avons sélectionné l'ensemble des pages de discussion associées aux articles, ainsi que leurs archives. Chaque page a été segmentée en fils de discussion selon les sections délimitées par les contributeurs et chaque fil en messages sur la base des signatures, indentations et autres marques de segmentation (le format *wiki* utilisé permet une grande souplesse). Seules les discussions comprenant au moins un message et 2 mots ont été conservées. À chaque message est associé l'identifiant du contributeur (ou l'adresse IP des contributeurs anonymes), sa date d'écriture et bien entendu le contenu textuel. Les données sont encodées au format XML selon la norme TEI dédiée aux communications médiées par les réseaux [Beißwenger et Lüngen, 2020]. Pour la présente étude nous avons également éliminé toutes les discussions faisant intervenir un robot (en nous basant sur l'identifiant de l'auteur).

Au final, nous disposons d'un corpus de 302 475 discussions exploitables pour un total de 769 880 messages. Ces discussions sont très hétérogènes en termes de taille, durée, nombre de participants, et bien entendu de contenu. Certaines particularités de ces discussions ont déjà été décrites dans de nombreuses études s'intéressant notamment aux conflits et aux actes de dialogues [Ferschke *et al.*, 2012, Yasseri *et al.*, 2012, Kittur et Kraut, 2010, Viégas *et al.*, 2004, Viégas *et al.*, 2007, Stvilia *et al.*, 2008, Wilkinson et Huberman, 2007]. Les principales caractéristiques sont résumées dans la table 1.

Caractéristique	Min.	Max.	Médiane	Moyenne
Nombre de messages par discussion	1	149	1	2,54
Nombre de participants par discussion	1	43	1	1,72
Durée (au moins 2 messages)	1 mn	16 ans	2,1 jours	184 jours

Table 1 – Caractéristiques des discussions

Schémas d'interaction

Afin de nous concentrer sur les interactions, nous avons identifié pour chaque discussion l'ordre dans lequel les différents participants intervenaient. Indépendamment de l'identité virtuelle de chaque participant, nous désignons par A l'auteur du premier message, par B le premier contributeur (distinct de A) qui intervient, etc. La figure 1 montre comment se déploient les discussions de notre corpus en se concentrant sur les 3 premiers messages des fils (le $ indique la fin de la discussion).

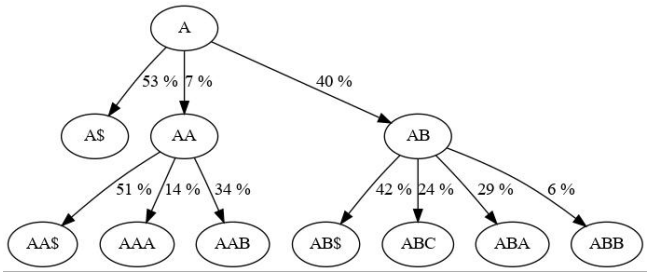

FIGURE 1 – Répartition des interactions en début de fil de discussion

Comme le montre la figure 1, la majorité des fils de discussion ne contient au final pas d'échange mais un seul message posté auquel personne n'a répondu (53% des fils discussions correspondent au schéma *A$*, phénomène classique dans les échanges en ligne, cf. [Beaudouin et Velkovska, 1999]). Le second schéma le plus récurrent renvoie à des situations plus classiques de dialogue entre A et B (40% des interactions) - l'échange majoritaire demeurant un simple *AB$*.

Nous avons été particulièrement interpellés par les situations qu'on pourrait appeler monologales, qui sont de notre point de vue les moins étudiées et les moins prévisibles : les interactions démarrant par AA représentent ainsi 7% des schémas identifiés, soit 19 947 fils de discussion. Comme indiqué dans la figure 1, ce second message sera le dernier dans 51% des cas ($AA\$$), sera suivi d'un troisième message du même auteur (14%, AAA) ou de l'intervention d'un interlocuteur (34%, AAB). Il n'est pas possible de conclure directement sur la fin d'une discussion, notamment parce que plus de la moitié des interventions n'entraînent pas de réponse. De plus, le format utilisé dans Wikipedia ne permet pas de déclarer qu'un fil est clos comme cela peut se faire sur d'autres plateformes d'échange, les discussions ayant la même durée de vie qu'une page de l'encyclopédie. Nous avons par exemple pu observer des cas où une réponse intervenait jusqu'à 15 années plus tard (soit presque l'empan temporel du corpus).

Si l'on se concentre sur les monologues purs, c'est-à-dire les fils dans lesquels un seul auteur s'est exprimé, ils représentent 57,5% des fils de discussions (174 009 sur 302 475). Parmi ceux-ci, seuls 7% (12 023) sont de véritables monologues avec plus d'un message, et 8,5% des fils avec plusieurs messages sont des monologues (les $AA\$$ et AAA de la figure 1). Le phénomène est donc loin d'être négligeable, sans même compter les séquences monologales au sein de discussions à plusieurs.

Nous avons décidé d'aborder l'étude des monologues par deux angles. Le premier est d'observer les situations les plus extrêmes, qui sont généralement bien plus simples à interpréter et qui permettent, au-delà de la simple anecdote, d'identifier des configurations que l'on peut par la suite retrouver avec des amplitudes plus réduites. Le second angle consiste à examiner les fils qui commencent par l'échange d'un utilisateur avec lui-même, afin d'étudier les conditions d'installation d'un monologue.

Monologues longs : étude des cas extrêmes

Nous avons donc dans un premier temps extrait les discussions monologales les plus longues : 136 fils contiennent ainsi plus de 5 messages (le plus long en contenant 28). Leur observation a permis d'identifier trois usages récurrents principaux. Le premier, que nous appellerons **tableau de bord**, correspond à l'utilisation d'un fil de discussion comme un outil de suivi des opérations réalisées ou à effectuer sur la page (vérifications, corrections, traductions etc.) [1]. Les messages de ces fils se distinguent par l'absence de marques d'interlocution. Deuxième usage, les **véritables monologues** qui se déclinent selon deux configurations puisque le monologue peut être subi ou choisi. Dans le premier cas, le monologue peut en fait être un monodialogue, dans lequel A tente d'entrer en contact avec un ou plusieurs autres Wikipédiens possiblement désignés explicitement qui ne lui répondent pas [2] Ces fils contiennent d'ailleurs une proportion significative de plaintes et de reproches. Dans le second

1. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Panth%C3%A9on_pyr%C3%A9n%C3%A9en#Faux_dieux_pyr%C3%A9n%C3%A9ens,_fausses_localisations,_etc.`

2. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Virginie_Grimaldi#%C2%AB_Chick_lit_%C2%BB_et_%C2%AB_feel_good_%C2%BB`

cas, *A* va écrire une série de messages dont l'enchaînement logique correspond à une réflexion ou une investigation [3]. Bien loin de la liste, ces messages sont de véritables textes exposant des faits, des interprétations, souvent argumentés et traduisant l'évolution du point de vue de *A*. On se rapproche alors de certaines formes de journaux extimes et on y note l'absence de marque d'une volonté interlocutrice : pas d'appel à un tiers ou à une communauté anonyme pour obtenir un avis, un complément ni même un assentiment. Le dernier usage observé correspond à des cas de **séries pures**, qui correspondent souvent à des listes d'items, rajoutés progressivement. Un cas extrême de série pure est celui d'un utilisateur qui, à de nombreuses reprises dans les pages de discussion associées à un acteur de cinéma non francophone, établit la liste des acteurs français qui ont assuré son doublage, avec un message pour chaque film - sans pour autant jamais éditer la page Article correspondante [4].

Monologues en début de discussion : pourquoi se répondre à soi-même ?

Afin de comprendre comment un monologue s'installe et se déploie, nous avons examiné un échantillon de 100 commencements de situations monologales, c'est-à-dire 100 fils dans lesquels l'auteur du premier message est le premier à intervenir à nouveau dans la discussion (schémas commençant par AA dans la figure 1), sélectionnés aléatoirement. Nous avons tenté de spécifier la fonction de chaque second message indépendamment de la suite du fil. Au final, plus de la moitié des débuts de monologues renvoient à deux cas de figure dominants : dans le premier cas, on observe un schéma **suggestion-action** (28%) dans lequel *A* suggère ou demande explicitement une action sur l'article qu'il indique ensuite avoir réalisée [5]. Le fil de discussion peut s'arrêter là ou se poursuivre, notamment par l'intervention d'un interlocuteur en désaccord avec l'action effectuée. Soulignons ici que la notification d'une *action* est propre au travail collaboratif visé par cet espace de discussion. Suggérer, demander, annoncer ou valider une modification de la page sont des motivations très courantes pour les discussions (*cf* la catégorie *explicit performative* de [Ferschke *et al.*, 2012] qui correspond à environ 60% des messages qu'ils ont annotés). Cette catégorie semble aller de pair avec celles des **tableaux de bord** ou deshttps ://www.overleaf.com/project/63d154d7eadd5aa11a7e4e48 **séries pures** (11% à elles deux), que nous avions identifiées dans les monologues longs : dans ce cas, les deux messages du même auteur forment une liste et sont généralement suivis d'autres items du même type. Ces listes peuvent concerner des remarques sur l'état de l'article, des problèmes identifiés, des actions à effectuer etc. On est ici dans une autre conception de la page de discussion Wikipédia qui s'apparenterait davantage à un tableau de suivi qu'à un outil de dialogue, ce qui nous semble en partie lié au format Wiki. En effet, les pages de discussion Wikipédia s'éditent comme l'article et prennent donc plus facilement la forme d'un texte écrit, qui peut avoir une certaine cohésion par rapports aux plateformes de discussion de type forum qui ne permettraient pas l'émergence de telles formes.

Une autre situation également propice à l'établissement d'un monologue a trait au second cas dominant relevé : la **complétion** (27%), dans lequel *A* complète sa remarque ou sa question initiale avec de nouvelles informations ou précisions [6]. Cet ajout peut s'accompagner d'une relance (ciblée ou non) afin d'obtenir une réponse ou une réaction tandis que dans d'autres situations les deux messages forment un raisonnement qui évoque une sorte de pensée à voix haute (avancée d'arguments contradictoires, évolutions de la position par rapport sur une constatation précédente, etc.). La discussion est alors rarement close et peut se poursuivre, prenant également potentiellement la forme d'un texte fragmentaire. Nous retrouvons ici les **véritables monologues** mentionnés supra.

Les autres situations relevées sont plutôt du côté de l'interaction et sont donc moins propices à l'installation d'un monologue. Outre la **relance** pure [7] (ciblée ou non) qui serait plus marginale (2%), nous avons observé deux cas qui semblent être des paires d'actes de langage associés, le deuxième message clôturant généralement le fil : (i) **message-rectification** (12%) : *A* rectifie son message

3. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:L%27H%C3%B4tel#%C2%AB_L%E2%80%99H%C3%B4tel_%C2%BB_situ%C3%A9_%C3%A0_l%E2%80%99emplacement_de_la_r%C3%A9sidence_de_Marguerite_de_France_?`

4. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Ben_Johnson_(acteur)#Voix_Fran%C3%A7aise`

5. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Romani#Nombre_de_locuteurs_par_pays`

6. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:S%C3%A9isme_et_tsunami_de_2004_dans_l%27oc%C3%A9an_Indien#Reprise_article_de_qualit%C3%A9?`

7. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Baruch_Goldstein#Comparatif_de_la_version_en_cours_et_la_version_propos%C3%A9e_par_Parmatus`

initial, généralement en admettant une erreur ou en ajoutant une information qui le rend caduc, ce qui met généralement fin à la discussion [8] ; et (ii) **question-réponse** (2%) : A répond lui-même à la question qu'il avait posée à la communauté, ce qui là aussi termine généralement l'échange [9].

Enfin, une particularité des discussions Wikipédia par rapport à d'autres types d'interactions en ligne tient à leur adossement aux articles en cours de rédaction qui correspondent en quelque sorte à un monde extralinguistique de référence ; ainsi les actions sur l'article peuvent être annoncées et commentées par leur auteur comme nous l'avons vu. Mais un autre phénomène nous a intéressé, à savoir la **réaction** de A à un événement (16%), par exemple une édition de l'article par un tiers. Cette réaction peut prendre différentes formes : celle du remerciement ou de l'encouragement [10] (rare), de la critique [11] (plus fréquent) ou encore de la demande d'information ou de complément [12]. À noter que ces situations sont souvent complexes à suivre, car le contenu du deuxième message n'est pas toujours explicite et peut faire référence à d'autres espaces de discussion (un autre article ou la page personnelle d'un contributeur). Ces configurations sont le produit de la complexité d'un écosystème communicationnel multicanal comme celui de la Wikipédia, mais font aussi écho à des modalités d'interaction dans des contextes d'interaction autour de manipulations partagées entre deux interlocuteurs (par exemple [Mondada, 2006, Koester, 2006]).

Ce dernier phénomène reste à explorer de façon plus systématique. Plus globalement nombre des situations spécifiques aux discussions autour des pages Wikipédia que nous avons mises au jour dans l'étude des monologues peuvent se retrouver au sein de situations plus variées entre différents locuteurs.

8. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Marie-Th%C3%A9r%C3%A8se_de_France_(1778-1851)#Ses_m%C3%A9moires`

9. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Comt%C3%A9_de_Hudson#Union_City,_une_ville_du_comt%C3%A9,_est_la_plus_peupl%C3%A9e_des_%C3%89tats-Unis.`

10. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Art_tib%C3%A9tain`

11. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Musique_classique/archive1#Ah,_a_propos_de_musique_%22savante%22`

12. Voir par exemple `https://fr.wikipedia.org/wiki/Discussion:Harry_Potter_et_le_Prince_de_sang-m%C3%AAl%C3%A9_(film)#Incoh%C3%A9rences_majeures_par_rapport_au_livre`

Bibliographie

[Beaudouin et Velkovska, 1999] Beaudouin, V. et Velkovska, J. (1999). Constitution d'un espace de communication sur internet (forums, pages personnelles, courrier électronique...). *Réseaux. Communication-Technologie-Société*, 17(97):121–177.

[Beißwenger et Lüngen, 2020] Beisswenger, M. et Lüngen, H. (2020). CMC-core : a schema for the representation of CMC corpora in TEI. *Corpus*, 20.

[Ferschke *et al.*, 2012] Ferschke, O., Gurevych, I. et Chebotar, Y. (2012). Behind the article : Recognizing dialog acts in Wikipedia talk pages. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics.

[Kittur et Kraut, 2010] Kittur, A. et Kraut, R. E. (2010). Beyond Wikipedia : coordination and conflict in online production groups. *In Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 215–224. ACM.

[Koester, 2006] Koester, A. (2006). *Investigating workplace discourse*. Routledge.

[Mondada, 2006] Mondada, L. (2006). Interactions en situations professionnelles et institutionnelles : de l'analyse détaillée aux retombées pratiques. *Revue française de linguistique appliquée*, 11(2):5–16.

[Stvilia *et al.*, 2008] Stvilia, B., Twidale, M. B., Smith, L. C. et Gasser, L. (2008). Information quality work organization in Wikipedia. *Journal of the American Society for Information Science and Technology*, 59(6):983–1001.

[Viégas *et al.*, 2004] Viégas, F. B., Wattenberg, M. et 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*, pages 575–582. ACM.

[Viégas *et al.*, 2007] Viégas, F., Wattenberg, M., Kriss, J. et van Ham, F. (2007). Talk Before You Type : Coordination in Wikipedia. *In 40th Annual Hawaii International Conference on System Sciences, 2007. HICSS 2007*, pages 78–78.

[Wilkinson et Huberman, 2007] Wilkinson, D. M. et Huberman, B. A. (2007). Cooperation and Quality in Wikipedia. *In Proceedings of the 2007 International Symposium on Wikis*, WikiSym '07, pages 157–164, New York, NY, USA. ACM.

[Yasseri *et al.*, 2012] Yasseri, T., Sumi, R., Rung, A., Kornai, A. et Kertész, J. (2012). Dynamics of conflicts in Wikipedia. *PloS one*, 7(6):e38869.