Les discussions Wikipedia : un corpus pour caractériser le genre “ discussion ”

Lydia-Mai Ho-Dac, Veronika Laippala

▶ **To cite this version:**

Lydia-Mai Ho-Dac, Veronika Laippala. Les discussions Wikipedia : un corpus pour caractériser le genre “ discussion ”. International Research Days: Social Media and CMC Corpora for the eHumanities, ATALA; Université de Rennes, Oct 2015, Rennes, France. hal-01271648

HAL Id: hal-01271648
https://hal.science/hal-01271648

Submitted on 16 Feb 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.

Les discussions Wikipedia : un corpus pour caractériser le genre « discussion »

Lydia-Mai Ho-Dac (Université de Toulouse, CLLE-ERSS) et Véronika Laippala (Université de Turku, TIAS)

Cette présentation propose une description des caractéristiques intra-linguistiques des discussions Wikipedia, forum de discussion associé à chaque article de l'encyclopédie Wikipedia. Après un exposé des propriétés qui font de ces textes un objet d'étude particulièrement intéressant pour les linguistiques de corpus, nous présenterons la procédure de constitution du corpus de discussion et une première description quantitative du corpus constitué. Nous finirons sur une présentation rapide d'un ensemble d'études linguistiques envisagées sur ce corpus.

Wikipédia est une encyclopédie libre et coopérative à laquelle tout internaute peut contribuer en modifiant ou créant un article ou encore en postant un message dans une page de discussion portant sur la structure, la pertinence, le contenu de l'article. Les contributeurs peuvent également participer à des forums portant sur la totalité du projet de Wikipedia, parmi lesquels les « cafés et bistrots »[1] (e.g. « Forum des Nouveaux » pour accueillir les nouveaux, « Le salon de médiation » pour « résoudre dans un cadre serein des conflits », etc.) ; ou encore des discussions autour des choix d'édition[2], des questions légales[3]...

Cette communauté fonctionne par le travail des internautes actifs qui sont ainsi amenés à acquérir un statut dans la communauté. Ils peuvent ainsi devenir « patrouilleurs » et avoir le droit et le devoir de « marquer une modification comme n'étant pas un vandalisme », ou encore « administrateur » dont le rôle est de « protéger et maintenir la qualité des éditions du projet »[4]. Tous ces rôles participent à la modération de la Wikipédia. En effet, tout ajout ou modification (que ce soit dans un article ou une discussion) est soumis à une phase de contrôle qui décide de sa publication.

Un corpus constitué de discussions Wikipedia représente un nombre important de caractéristiques avantageuses pour les linguistiques de corpus. Premièrement, il s'agit d'un forum de discussion libre de droits (licence Creative Commons by-sa) qui existe depuis 2001 et dans lequel les contributeurs interagissent autour d'une thématique explicite et détaillée soit dans l'article associé soit dans le type de

1 La liste des cafés et bistrots est donnée dans l'article http://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Avenue_des_caf%C3%A9s_et_bistros

2 http://fr.wikisource.org/wiki/Aide:Choix_%C3%A9ditoriaux, consulté le 12 mai 2015

3 http://fr.wikisource.org/wiki/Wikisource:Questions_l%C3%A9gales, consulté le 12 mai 2015

4 Une liste détaillée des statuts est donnée dans l'article http://fr.wikipedia.org/wiki/Aide:Statuts_des_utilisateurs

« café » (e.g. « Le salon de médiation »). Le contexte de production de ces discussions est ainsi beaucoup plus accessible que pour tout autre forum de discussion.

Autre point important, notamment pour l'application de techniques en TAL (traitement automatique des langues), de premiers travaux montrent que les discussions Wikipedia présentent relativement peu de déviance par rapport à la norme langagière. Les messages sont écrits de manière plutôt rigoureuse par rapport aux forums de discussion plus traditionnels : relativement peu de fautes d'orthographes et de grammaires, peu de recours à des modes de rédaction particuliers (lettres capitales PLUS, répétées ASSSSEEEZ, suite de ponctuation répétées !!!!). Nous présenterons quelques éléments de comparaison pour évaluer ce degré de déviance.

Troisièmement, les textes sont systématiquement associées à un nombre important de méta-données portant à la fois sur la thématique (portail thématique, article associé), le caractère subjectif de la discussion (caractère polémique, etc.) et le statut du locuteur (informations sur sa participation à la Wikipédia et sur son statut dans la communauté). Ces statuts ont fait l'objet d'un certain nombre d'études sur la corrélation entre le statut et le style langagier utilisé (Danescu-Niculescu-Mizli et al. 2012, 2013, Burke and Kraut 2008 inter alia).

Enfin, la base de données Wikipedia représente une masse de données imposante. Selon la page http://fr.wikipedia.org/wiki/Wikipédia:Statistiques (consulté le 12 mai 2015) : “Wikipédia en français compte 16192 contributeurs (Wikipédiens) ayant fait au moins une modification ces 30 derniers jours (hors utilisateurs sous IP). Parmi ceux-ci près de 5 000 contributeurs ont fait au moins 5 modifications et près de 800 ont fait au moins 100 modifications sur la même période.” La principale activité reste l'édition d'articles (1 622 066 articles au 11 mai 2015), mais la participation aux discussions est également très importante. Notre corpus est ainsi constitué de 366 326 discussions, 1 024 351 sections de discussion (topics internes à une discussion), 2 255 959 messages et 159 578 279 mots.

Afin de constituer notre corpus de discussions, plusieurs procédures automatiques ont été mises en place pour extraire et formater les discussions. L'extraction consiste à traiter la sauvegarde globale des pages courantes de la Wikipédia française (archive frwiki-20140331-pages-meta-current#.xml.bz2 diffusée librement sur la page http://dumps.wikimedia.org/frwiki/20140331/), d'y repérer les discussions et de les transformer en fichiers XML normés selon la TEI-P5.

75 % des discussions ont été exclues du corpus (1 130 227 sur les 1 496 553 contenues dans le Dump). Les critères d'exclusion sont les suivants :

- La discussion porte sur un utilisateur de la Wikipedia[5]
- Indication explicite d'une redirection vers une autre discussion :

  <text xml:space="preserve">#REDIRECT [[Discuter:Iolo Morganwg]]</text>

  Exemple d'indication en tête de la discussion Discussion:Yolo Morganwg déplacée vers la page Discussion:Iolo Morganwg suite à une erreur sur l'initiale du prénom.

  <text xml:space="preserve">Doublon de [[Son (physique)]].</text>

5 Exemple : http://fr.wikipedia.org/wiki/Discussion_utilisateur:Hashar

Exemple d'indication en tête de la discussion Discussion:Onde acoustique supprimée car en doublon de celle associée à la page Son (physique).

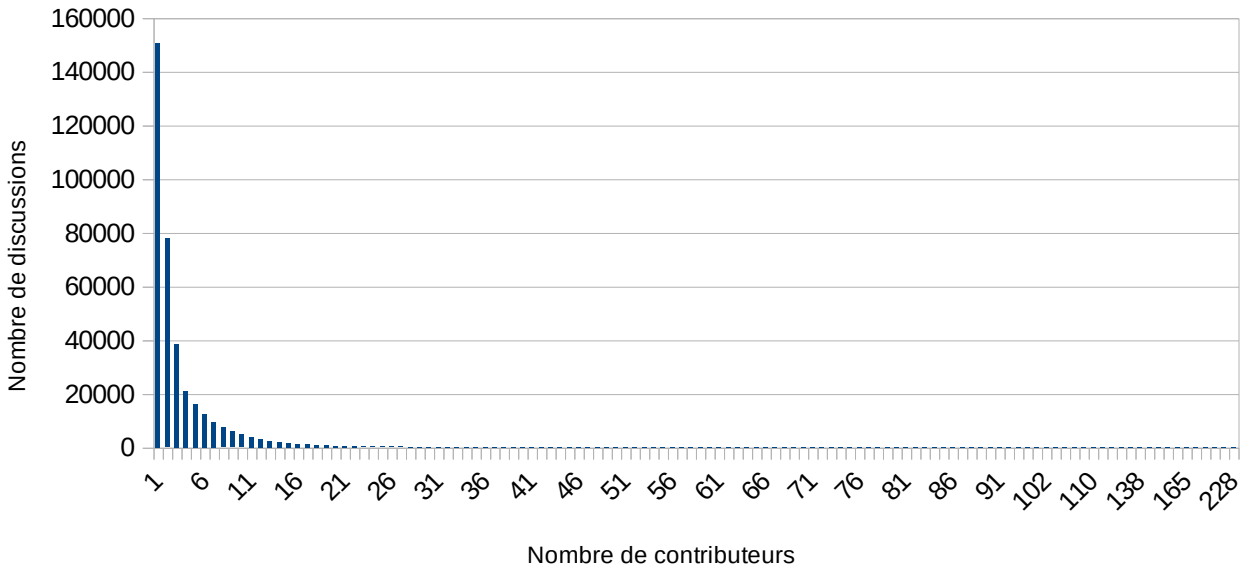
- Aucun message dans le corps de la discussion
- Moins de 2 mots dans le corpus de la discussion

Les discussions mono-contributeurs où un locuteur lance un sujet auquel personne ne répond ont été conservées. Elles sont au nombre de 150 603 (soit un peu plus de 40 % des discussions retenues).

La délimitation des différents messages et contributeurs s'appuie sur un ensemble de règles, notamment la présence nécessaire de la date de publication du message. L'évaluation manuelle de 7 discussions comptabilisant 413 messages et 47 284 mots montre une précision de 0,92 (3 messages vides ; 5 messages scindés en 2 ; 25 messages fusionnant 2 ou 3 messages) et un rappel de 0,95 (23 messages absents).

Le modèle utilisé pour représenter les différents sujets de discussion et messages s'inspire de la norme dédiée à la fois aux dialogues oraux et aux pièces de théâtre. Ainsi, chaque message est balisé correspondant aux messages avec pour attribut "who" le nom de l'utilisateur, "when" la date de publication du message et "interactionalLevel" son niveau d'interaction (réponse au précédent message, réponse à une réponse, etc.) Chaque message est ensuite découpé en paragraphe (<p>) ou des éléments de liste (<item>).

Les discussions et les messages présentent de grandes variations de taille : entre 1 et 1 103 messages par discussion et entre 1 et 3 428 mots par message. En moyenne, une discussion implique 5 contributeurs identifiés différents, avec là aussi, de fortes variations, allant de 1 à 228 contributeurs différents pour une même discussion. Le graphique ci-dessous indique le nombre de contributeurs impliqués dans les discussions. Parmi ce décompte des contributeurs, tous les anonymes non inscrits



Nombre de contributeurs par discussion

sont regroupés, ce qui représente autour de 4,5 % des contributeurs, chiffre relativement stable.

Concernant les caractéristiques intra-linguistiques, nous proposerons un premier inventaire assez large du contenu de ce corpus : n-grams (morpho-)syntaxiques typiques (cf. Laippala et al. 2015), expression de la subjectivité (projection du lexique FEEL, Abdaoui et al. 2014), formule d'ouverture et de fermeture des messages, expression de l'accord et du désaccord.

Enfin un cas d'étude linguistique plus ciblée sera présenté autour du phénomène des noms sous-spécifiés (Schmid 2000), marqueurs de l'organisation discursive dont le rôle est susceptible de fortement varier selon les types de texte. Leur analyse contrastive permettra de tester leur pertinence en tant qu'indice distinctif entre discussions et articles.

Bibliographie :

Abdaoui, A., Azé, J., Bringay S. et Poncelet. P. (2014) FEEL: French Extended Emotional Lexicon. ISLRN: 041-639-484-224-2

Burke, M., and Kraut, R. (2008) Mopping up: Modeling Wikipedia promotion decisions. ACM CSCW 2008: Conference on Computer-Supported Cooperative Work. 27-36. PDF

Danescu-Niculescu-Mizil C., Sudhof M., Jurafsky D., Leskovec J., Potts C. (2013) A computational approach to politeness with application to social factors, in *Proceedings of ACL*2013.

Danescu-Niculescu-Mizil C., Lee L., Pang B. and Kleinberg J.. (2012) Echoes of power: Language effects and power differences in social interaction, in *Proceedings of WWW*2012.

Denis, A. Quignard, M. Fréard, D, Détienne, F., Baker, M. & Barcellini, F. (2012) Détection de conflits dans les communautés épistémiques en ligne. In G. Antoniadis et H. Blanchon (Eds.) *Actes du 19ème congrès de Traitement Automatique des Langues Naturelles*, 4-8 Juin, Grenoble, France.

Laippala, V., Kanerva, J., Ginter, F. (2015) Syntactic Ngrams as Keystructures Reflecting Typical Syntactic Patterns of Corpora in Finnish, Procedia - Social and Behavioral Sciences, Volume 198, 24 July 2015, Pages 233-241, ISSN 1877-0428, http://dx.doi.org/10.1016/j.sbspro.2015.07.441.

Schmid H-J. (2000) *English abstract nouns as conceptual shells : from corpus to cognition*, Topics in English Linguistics, 34, Berlin & New York: Mouton de Gruyter