Catégorisation automatique d'actes de communication : les fils de discussion des pages de discussion Wikipédia

Anouk Birski

Université Toulouse II
Département Sciences du Langage

Master Linguistique, Informatique et Techniques de la Langue
2016 - 2017

Catégorisation automatique d'actes de communication
Les fils de discussion des pages de discussion Wikipédia

Anouk Birski
Mémoire de Master 1

Sous la direction de
Lydia-Mai Ho-dac & Ludovic Tanguy

Déclaration sur l'honneur
de non-plagiat

Je soussignée,
Birski Anouk

Régulièrement inscrite à l'Université de Toulouse II Jean Jaurès
N° étudiant 21606217
Année universitaire 2016-2017

Certifie que le document joint à la présente déclaration est un travail original, que je n'ai ni recopié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations intégrales sont signalées entre guillemets.

Fait à Toulouse

Le 14 juin 2017

Table des matières

Index des tableaux

Index des exemples

Introduction

L'élaboration collaborative de contenu a bénéficié ces dernières années d'un nouveau souffle grâce à l'apparition d'outils et de plate-formes numériques fondés sur ce processus particulier, qui diffère complètement de la rédaction individuelle.
Si de nombreuses études ont appréhendé ce phénomène en terme de qualité des contenus produits (Passig & Schwartz, 2007), d'influence de l'ergonomie du système sur le comportement des contributeurs (Chen, 1997) ou encore des différentes stratégies employées pour rédiger des documents de manière collaborative (Baecker et al., 1993), de plus en plus d'études s'intéressent de plus près à sa dimension purement interactive (Ferschke et al., 2012).

En effet, la collaboration entre plusieurs contributeurs passe forcément par une phase de communication, d'échange autour de la tâche à effectuer, qui est révélatrice de différents comportements en situation d'interaction à travers des outils numériques. L'analyse de ces interactions, afin de mieux comprendre leurs processus, leurs rouages ou encore leurs conséquences constitue un objet aussi riche que complexe à exploiter. En impliquant, d'une part, la prise en compte du caractère linguistique de ces données, et d'autre part, l'importance de l'automatisation de cette tâche en raison des quantités prolifiques générées chaque jour par les internautes, cette analyse fait directement appel aux problématiques auxquelles s'intéresse le Traitement Automatique des Langues (désormais TAL), et représente ainsi un enjeu particulier pour ce domaine.

Parmi les nombreux outils s'appuyant sur l'écriture collaborative, un modèle est particulièrement employé et répandu, c'est celui du *Wiki.* L'exemple le plus connu est l'encyclopédie participative Wikipédia[1], animée par des contributeurs qui se coordonnent afin de rédiger des articles, et qui fournit un espace dédié aux discussions liées au processus collaboratif : les pages de discussion. La Wikipédia fait d'ailleurs l'objet de nombreuses recherches, et notamment dans le domaine du TAL qui s'occupe particulièrement de ces pages de discussion (Ferschke et al., 2013).

Cette étude s'intéresse aux différents profils d'interactions qui peuvent être générés dans un contexte d'écriture collaborative au sein de la Wikipédia et qui se retrouvent dans les espaces de pages de discussion, afin de les identifier, mais aussi de les analyser automatiquement. Elle s'inscrit ainsi dans cette dynamique de recherche autour des productions collaboratives, et plus particulièrement celle qui se fonde sur la communauté active faisant vivre la Wikipédia, en mettant en place une approche d'analyse outillée des interactions des contributeurs de l'encyclopédie.

Le premier chapitre revient sur la dimension de communication dans le cadre de la plate-forme participative Wikipédia, et les enjeux auxquels tentent de répondre ce projet de recherche.

Le deuxième chapitre présente le corpus *WikiDisc* (Ho-Dac & Laippala, 2015) qui rassemble des pages de discussion issues de la Wikipédia francophone, données initiales utilisées dans cette recherche, ainsi que les premières manipulations effectuées dans le but d'affiner ces dernières et récupérer un corpus contenant des données plus appropriées pour l'étude des interactions.

1 https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Accueil_principal

Le troisième et quatrième chapitres reviennent sur l'identification de divers profils d'interactions entre les utilisateurs à partir des données sélectionnées dans le corpus. Cette caractérisation s'appuie sur une exploration approfondie du corpus par le biais de différents ensembles de traits dégagés grâce à une analyse outillée. Le troisième chapitre retrace la recherche de profils à travers des traits relatifs à des caractéristiques quantitatives des fils de discussion, et décrit l'identification de profils généraux, tandis que le chapitre suivant présente une collection de traits qui caractérisent des aspects linguistiques, temporels et interactionnels des fils de discussion, ainsi qu'une table d'annotation afin d'explorer de manière plus précise le contenu des fils et les types d'interactions. Cette analyse plus fine est amorcée par l'identification de deux profils précis reposant sur l'ensemble des traits dégagés.

Le cinquième et dernier chapitre discute les pistes à explorer afin d'améliorer ces analyses ainsi que les perspectives de cette recherche.

I. État de l'art : exploitation de la Wikipédia et étude des interactions entre utilisateurs

Cette section explique le choix de la plate-forme Wikipédia afin d'étudier les interactions inhérentes au processus de rédaction collective en présentant cette dernière, avant de s'intéresser à l'état des différentes recherches existant à ce sujet, pour enfin expliquer de manière détaillée les objectifs et la démarche de cette étude.

I.1 La Wikipédia : une communauté, des interactions

La Wikipédia (désormais WP) est une encyclopédie rédigée de manière collaborative par une communauté bénévole d'internautes. Cela signifie que les modifications sont libres, et tout internaute peut intervenir et modifier son contenu. Cette approche collaborative de la rédaction des articles est fondée sur un principe d'autorégulation au sein de la communauté active. Ainsi, comme le souligne Pierre-Carl Langlais, les prises de décision et les modifications s'appuient sur des choix consensuels, plutôt que sur l'appel à la majorité ou le choix d'une figure d'autorité (Langlais, 2014). Or, atteindre un consensus passe inévitablement par des processus de communication, d'échange, de négociation voire parfois de débat entre les différents contributeurs. Ainsi, la Wikipédia qui est principalement connue pour sa dimension encyclopédique et ses articles très fréquemment consultés par de nombreux utilisateurs qui ne participent pas à leur rédaction, est également le théâtre d'abondantes interactions entre les contributeurs impliqués dans son évolution.

Mais avant de nous attarder sur cette dimension, revenons brièvement sur la globalité de cette encyclopédie qui fait l'objet de nombreuses études, ainsi que son fonctionnement.

I.2 Une dynamique de recherche autour de la Wikipédia

Le développement d'internet a entraîné l'apparition de situations de communication nouvelles et de formes inédites de discours, couramment qualifiées, entre autre, de *Computer Mediated Communication* (CMC) en anglais, ou Communication Médiée par Ordinateur (CMO) en français (Herring et al., 2013). Avec pas moins de 30 millions d'articles au total, dont environ 1,9 million dans la WP française, 5,4 millions pour la version anglaise, ou bien plus de 200 000 en espéranto, la Wikipédia est l'un de ces objets d'étude, particulièrement attrayant de part la grande quantité de données qu'elle contient, son caractère multilingue, la conservation de toutes les modifications effectuées dans un historique ou encore le fait que l'accès au contenu de l'encyclopédie soit sous licence *Creative Commons BY-SA*[2], *i.e.* gratuit et libre d'exploitation sous réserve de respecter les conditions de la licence.

Ces particularités en ont fait un objet de recherche à part entière, autour duquel une réelle dynamique s'est installée. En témoignent les plus de 6 000 publications à son sujet recensées par *Wikipaper*[3], et dont les

2 https://creativecommons.org/licenses/by-sa/3.0/

3 http://wikipapers.referata.com/wiki/Main_Page

thèmes étudiés sont divers et variés. Notre travail s'inscrit dans cette dynamique de recherche, et plus particulièrement dans l'étude de l'interaction des contributeurs de la Wikipédia.

I.3 Étude de l'interaction des contributeurs de la Wikipédia

Comme évoqué dans la section précédente (*cf.* I.1 La Wikipédia : une communauté, des interactions), le processus collaboratif de l'encyclopédie passe par des phases d'échange et de communication en vue d'un consensus. Ainsi, pour chaque article de la Wikipédia, il est possible de créer une page de discussion (désormais PdD) dédiée aux échanges à propos des points à améliorer, des changements à effectuer, ou encore des précisions à apporter à l'article afin de le perfectionner, qu'il soit le plus complet et référencé possible.

Les PdD sont une source privilégiée pour l'étude des interactions entre utilisateurs, notamment sous leurs formes extrêmes. Jusqu'à maintenant les études menées à ce sujet se sont en effet penchées sur des événements particuliers, en partant d'hypothèses déterminées ou de phénomènes ciblés, telles que la détection de conflit dans les fils de discussion (Poudat et al., 2016), ou encore l'influence des différentes habitudes de communication en fonction du genre sur le comportement des utilisatrices (Sichler & Prommer, 2014). L'étude menée ici adopte une approche moins cadrée et plus guidée par les données du corpus. Ainsi, elle vise à faire émerger des profils de discussions, *i.e.* des situations récurrentes d'interactions entre des utilisateurs qui peuvent être identifiées et rassemblées dans des catégories types, à partir d'une analyse outillée des fils de conversation, et ce afin de pouvoir les identifier automatiquement par la suite.

Cette démarche se base sur l'analyse d'un ensemble de traits descriptifs, établis automatiquement pour chaque fil de conversation, qui offrent une vue synthétique de l'ensemble de ces fils afin d'identifier des profils d'interaction généraux. Ces caractérisations permettent, à terme, de cibler des cas plus précis et ce afin d'étudier de manière approfondie les comportements des utilisateurs. Le cheminement de la recherche effectuée peut se décomposer en trois stades distincts :

- Identification de cas où il n'y a pas d'interaction : fils de discussion vides ou n'impliquant qu'un seul utilisateur. Cette identification de ces profils permet de les écarter des données à explorer car ils ne correspondent pas à l'objet de notre recherche.

- Caractérisation des fils de discussion restants selon des traits tels que le nombre d'utilisateurs, le nombre de messages, le taux moyen de mots par message ou encore le nombre total de phrases du fil. Ces traits permettent de faire émerger des premiers profils de fils de discussion assez généraux, et considérés comme comportant un potentiel interactif plutôt restreint, ou bien du contenu problématique, tel que des séquences non produites par les auteurs dans le cadre de la discussion. Cette première vague d'identification permet poser un regard global sur les données ainsi que les différents types d'interaction, afin de mener une exploration approfondie des divers profils généraux établis. Dans le cadre de cette étude, l'exploration approfondie s'intéresse particulièrement aux fils de discussion n'étant pas encore caractérisés et considérés comme possédant un potentiel d'interaction élevé.

- Exploration approfondie du contenu linguistique des données à travers des traits tels que le taux de pronoms de la première personne ou le taux de phrases interrogatives afin de faire émerger des profils plus précis parmi les fils de discussions qui n'ont pas encore été caractérisés. Cette étape est

également l'occasion de projeter ces traits linguistiques sur les profils identifiés précédemment afin d'observer d'éventuelles corrélations avec des nouveaux traits, ce qui peut permettre de préciser ces profils généraux.

Ce projet ne prétend pas être en mesure de catégoriser précisément l'ensemble des fils de discussion des pages de discussion de la Wikipédia, mais développe une méthodologie qui s'inspire de l'observation des données (Tognini-Bonelli, 2001) (Biber, Egbert, & Davies, 2015) permettant d'identifier semi-automatiquement des profils généraux de discussions, fournissant ainsi une vue d'ensemble sur une grande quantité de données, et propose une méthodologie d'analyse plus approfondie afin de pouvoir, à terme, cibler des profils d'interactions plus précis.

II. Les données : pages de discussion et affinage du corpus WikiDisc en écartant les données non pertinentes

La première étape de ce projet repose sur l'élaboration, à partir d'une grande quantité de données, d'un corpus contenant des donnée plus pertinentes et centrées sur les interactions entre les utilisateurs.

Cette section décrit les deux stades de l'élaboration du corpus réduit, en présentant dans un premier temps l'objet principal de la recherche, les pages de discussion, ainsi que *WikiDisc*, le corpus de départ, avant de décrire la phase de filtrage qui permet d'écarter les données inappropriées afin d'établir un corpus dont l'analyse est explicitée dans les sections suivantes.

II.1 Les données initiales : WikiDisc, un corpus de pages de discussion

II.1.1 Les pages de discussion : une dimension méconnue de la Wikipédia

Afin d'assurer le bon déroulement du processus collaboratif de rédaction des articles, la communication entre les contributeurs est essentielle. Les pages de discussion sont donc des lieux indispensables au bon fonctionnement de l'encyclopédie collaborative. Il est ainsi possible de discuter directement avec un utilisateur sur les pages de discussion utilisateurs, d'interagir au sujet d'un article particulier sur les pages de discussion des articles, ou encore de discuter avec la communauté de wikipédiens sans pour autant viser un utilisateur ou un article particulier, au sein d'un des nombreux espaces de discussion[4]. Notre recherche vise plus particulièrement le second cas de figure, *i.e.* les discussions entre utilisateurs dans le cadre de la rédaction d'un article.

Chaque article possède donc une page dédiée à ces échanges : ce sont les pages de discussion, dont le lien est accessible par un onglet (Exemple 1) en haut de la page de l'article. Ces pages permettent aux internautes d'échanger sur les différents points à améliorer, les changements à effectuer, ou encore les précisions à apporter à l'article afin de le perfectionner, qu'il soit le plus complet et référencé possible. Chaque PdD peut être accompagnée de PdD parallèles qui ciblent un thème particulier de l'élaboration de l'article, comme par exemple la page *Suppression* qui peut exister si la pertinence de l'article est remise en cause, ou encore la page *Bon article* au sein de laquelle les contributeurs peuvent discuter des modifications à effectuer avant de demander le label *Bon article* au comité qui s'occupe de l'accorder. Par ailleurs, si un article fait partie d'un projet thématique, alors un bandeau en haut de la page de discussion précise cette information. Enfin, la plupart des pages de discussion possèdent un sommaire des fils de discussion qu'elle contient afin de faciliter la navigation dans la page.

4 https://fr.wikipedia.org/wiki/Wikip%C3%A9dia:Avenue_des_caf%C3%A9s_et_bistros

Exemple 1 : Discussion:Pierre Lambert de la Motte
PdD : 5993044
Exemple d'une page de discussion

Exemple 1 : Ici la page de discussion est reliée à trois pages de discussions parallèles existantes: *Article de qualité*, *Lumière sur* et *À faire*. Par ailleurs on peut constater que l'article fait partie des projets *Christianisme* et *Asie du Sud-Est*. Enfin le sommaire indique 7 fils de discussions.

Une page de discussion rassemble des fils de discussion créés par les contributeurs, correspondant chacun à une discussion sur un thème défini, qui est souvent précisé dans le titre du fil. Les fils de discussion (Exemple 2) sont eux alimentés par des messages postés par les internautes qui discutent, échangent, débattent ou négocient autour du sujet du fil de discussion. La structure du fil de discussion est censée illustrer le déroulement de la discussion, composée par des messages qui s'imbriquent à la suite les uns des autres, et où l'ordre est déterminé par l'enchaînement chronologique.

Exemple 2 : Discussion:Herpès
PdD : 3019839, FdD : 3
Fil « Remède de grand-mère » : enchaînement des messages

Exemple 2 : Ici quatre contributeurs abordent le sujet de la considération d'un « remède de grand-mère » : faut-il l'énoncer comme un traitement dont l'efficacité est avérée, ou bien prendre des distances ?

II.1.2 Le corpus WikiDisc

Le corpus *WikiDisc* (Ho-Dac & Laippala, 2015) constitue la base des données utilisées dans ce projet de recherche. Ce corpus rassemble des PdD issues de la Wikipédia francophone et a été constitué en plusieurs étapes :

- Extraction de toutes les pages de discussion (*cf.* II.1.1 Les pages de discussion : une dimension méconnue de la Wikipédia) de la sauvegarde[5] de la WP francophone à la date du 12 mai 2015 soit plus de 3,5 millions de pages.
- Conservation des pages liées à un article : 1 496 553 PdD (soit 43%)
- Conservation des pages contenant au minimum deux mots : 365 612 PdD (soit 24%)

Les pages conservées ont été encodées au format XML selon la norme TEI-P5[6].

Pages de Discussion	Fils de discussion	Messages	Mots
365 612	1 023 841	2 406 514	161 833 298

Tableau 1 : Le corpus WikiDisc en chiffres

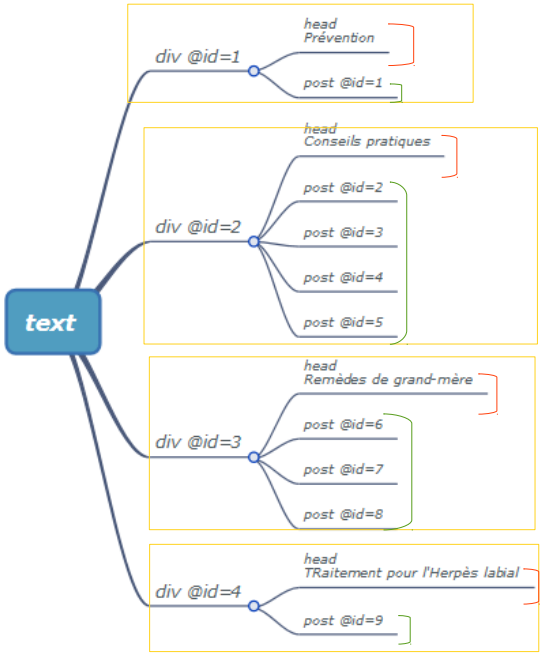
5 https://dumps.wikimedia.org/
6 http://www.tei-c.org/Guidelines/P5/

II.1.3 Structure des données de WikiDisc

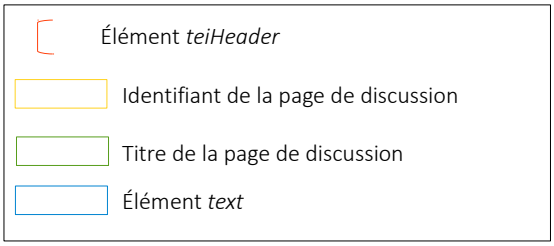
Le corpus WikiDisc est composé d'un ensemble de fichiers XML norme TEI-P5, et chaque fichier correspond à une PdD de la Wikipédia. Une PdD est composée d'un entête, l'élément **tei*Header*** (Exemple 3), qui renseigne des informations contextuelles du document conformément à la norme TEI-P5 et d'un élément ***Text*** (Exemple 4), dans lequel se trouve le contenu de la page de discussion.

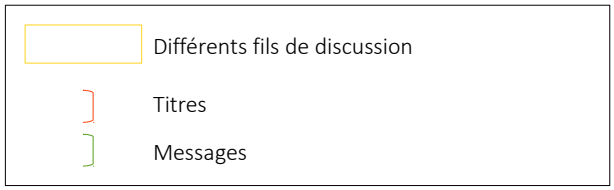
```
− <TEI>
  − <teiHeader>
    − <fileDesc>
      − <titleStmt>
          <title>3019839.xml</title>
        </titleStmt>
      − <publicationStmt>
          <publisher>CLLE-ERSS</publisher>
          <pubPlace>Toulouse, France</pubPlace>
          <date when="Tue May 3 11:06:42 2016"/>
        </publicationStmt>
      − <sourceDesc>
        − <biblStruct>
          − <analytic>
              <author>Wikipedia</author>
              <title>Discussion:Herpès</title>
              <idno type="Wikipedia">3019839</idno>
            </analytic>
          + <monogr></monogr>
          </biblStruct>
        </sourceDesc>
      </fileDesc>
    + <encodingDesc></encodingDesc>
    + <profileDesc></profileDesc>
    </teiHeader>
  + <text></text>
  </TEI>
```

Exemple 3 : Discussion:Herpès
PdD : 3019839
En-tête de la structure XML d'une page de discussion



Exemple 4 : Discussion:Herpès
PdD : 3019839
Schéma de la structure de l'élément text



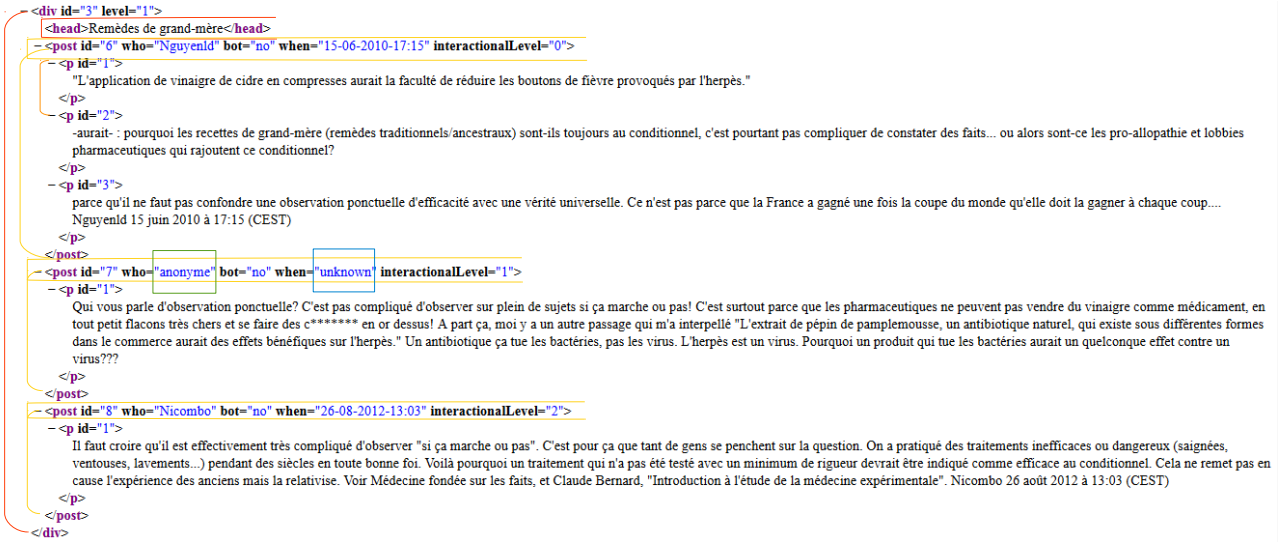


- Chaque PdD est composée d'un ou plusieurs **fils de discussion** (désormais FdD). Un fil correspond à une discussion à priori ciblée sur un sujet particulier et est créé par un internaute. Ce niveau est correspond à l'élément ***div*** de la structure XML, pour lequel est précisé l'identifiant dans l'attribut ***id***, ainsi que le niveau du fil de discussion, dans l'attribut ***level***.
  - Exemple : Pour la PdD associée à l'article Herpès (Exemple 4), les internautes ont lancé quatre fils de discussion intitulés *Prévention, Conseils pratiques, Remèdes de grand-mère* et *Traitement pour l'Herpès labial*.

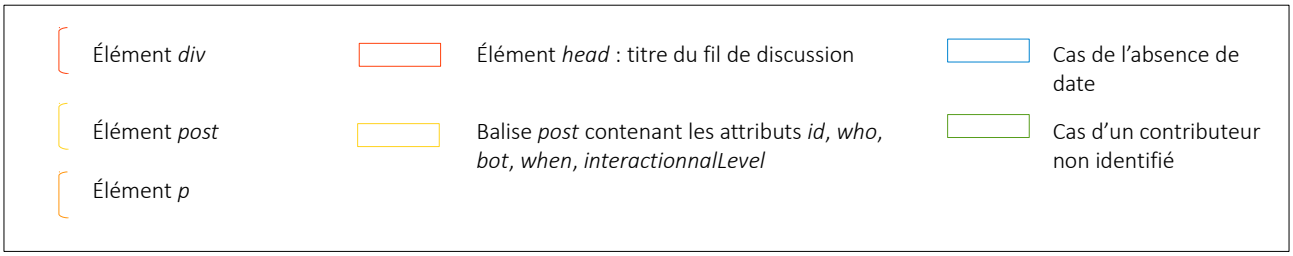
- Chaque FdD est composé d'un titre, parfois absent, ainsi que d'un ou plusieurs messages : les fils de discussions sont alimentés par les internautes qui participent à la discussion en postant des messages. Le niveau titre correspond à l'élément *head* de la structure XML, et les messages correspondent aux éléments *post.* Les messages peuvent être divisés en éléments *p* qui concordent avec les paragraphes.

- Chaque message est accompagné de méta-données précisant son identifiant, le nom de l'utilisateur qui l'a posté (ou *anonyme* le cas échéant) ainsi que la date et l'heure à laquelle il a été posté (ou *unknown* le cas échéant).
    - L'identifiant est indiqué par l'attribut *id*.
    - L'utilisateur est renseigné dans l'attribut *who* . Il peut arriver que l'utilisateur ne soit pas identifié, ou bien qu'un problème survienne lors de la récupération de la signature de l'auteur : dans le cas où l'auteur n'est pas identifié, il est considéré comme *anonyme (cf. post 19*, exemple 5).
    - La date , qui précise l'année, le mois, le jour, l'heure, et la minute auxquels le message a été posté, est indiquée dans l'attribut *when* de l'élément *post.* Cette information est parfois absente, et dans ce cas elle est indiquée comme *unknown,* comme l'illustre le *post 19* dans l'exemple 5.
    - L'attribut *bot* précise si le message a été posté par un robot (valeur *yes*) ou par un utilisateur (valeur *no*).
    - L'attribut *interactionalLevel* renseigne sur le niveau d'interaction du message. Ce niveau d'interaction est censé indiquer l'enchaînement discursif des messages, *i.e.* quels messages se suivent. L'exemple 5 montre bien que le *post 20* suit le *post 19* car son niveau d'interaction est supérieur. Cependant, certains problèmes survenus lors de la création même des messages apportent de fréquentes erreurs par rapport à cette information. Elle n'a donc pas été utilisée pour cette recherche.

L'objectif étant d'analyser les interactions entre les internautes, nous avons défini le niveau fil de discussion comme unité d'analyse, car il délimite le cadre d'un échange entre des contributeurs impliqués dans la même discussion. Bien que, parfois, les fils de discussion se répondent au sein de la page de discussion, ce phénomène n'est pas majoritaire, nous avons donc considéré que la page de discussion était un niveau trop large pour cibler ce type d'interactions.



Exemple 5 : Discussion:Herpès
PdD : 3019839, FdD : 3
Structure XML du fil de discussion Remèdes de grand-mère



II.1.3 Version étiquetée avec Talismane

L'analyse de certains traits (*cf.* Tableau 8 & Tableau 25) a nécessité la manipulation d'une version étiquetée (Exemple 6) des contenus des messages, récupérée grâce à l'analyseur syntaxique Talismane ((Urieli, 2013)Urieli, 2013). Cet outil permet de segmenter un texte brut en mots (phase *tokenizer*), en phrases (phase *sentence detector*), mais aussi d'étiqueter chaque mot en fonction de son rôle dans la phrase (*Part Of Speech tagger*). Enfin il permet d'identifier les liens de dépendance entre les différents éléments de la phrase (*syntax parser*).

Num	Mot (token)	Lemme	POS-tag	Catégorie gram.	Infos morpho syntaxiques							
1	La	la	DET	DET	n=s|g=f|	2	det	2	det	100.00	97.36	96.52
2	phrase	phrase	NC	NC	n=s|g=f|	7	suj	7	suj	100.00	97.93	99.16
3	que	que	PROREL	PROREL		5	obj	5	obj	100.00	95.44	93.22
4	vous	vous	CLS	CLS	n=p|p=2|	5	suj	5	suj	100.00	86.17	89.08
5	proposez	proposer	V	V	n=p|t=P|p=2|	2	mod_rel	2	mod_rel	100.00	75.89	97.01
6	n'	ne	ADV	ADV		7	mod	7	mod	100.00	96.88	91.75
7	est	être	V	V	n=s|t=P|p=3|	0	root	0	root	100.00	98.64	99.03
8	pas	pas	ADV	ADV		7	mod	7	mod	100.00	99.90	98.07
9	un	un	DET	DET	n=s|g=m|	10	det	10	det	100.00	98.67	99.58
10	résumé	résumé	NC	NC	n=s|g=m|	7	ats	7	ats	100.00	81.33	86.69
11	,	,	PONCT	PONCT		10	ponct	10	ponct	100.00	98.80	99.52
12	elle	elle	CLS	CLS	n=s|g=f|p=3|	13	suj	13	suj	100.00	99.53	96.94
13	est	être	V	V	n=s|t=P|p=3|	7	mod	7	mod	100.00	99.49	98.13
14	partielle	partiel	ADJ	ADJ	n=s|g=f|	13	ats	13	ats	100.00	98.25	93.55
15	.	.	PONCT	PONCT		14	ponct	14	ponct	100.00	100.00	99.09

Exemple 6 : Phrase étiquetée par Talismane

Ces informations supplémentaires sont utiles pour le calcul de certains traits statistiques tels que le nombre de mots contenus dans le fil ou le message, mais elles sont surtout essentielles pour effectuer une analyse linguistique approfondie du contenu des messages.

II.2 Première manipulation : pages de discussions parallèles et fils de discussion sans contenu interactif, un filtrage pour écarter les données non pertinentes

Dans un premier temps, le corpus a été étudié dans son intégralité. En prenant compte des objectifs d'observation de situations d'interactions entre utilisateurs dans les fils de discussion, plusieurs cas à écarter des données à observer ont été identifiés.

II.2.1Niveau pages de discussion : sélection des pages principales

Comme évoqué précédemment, les données étudiées sont divisées en plusieurs niveaux : tout d'abord, il y a le niveau "page de discussion", qui correspond à l'ensemble de la page de discussion dédiée à l'article Wikipédia associé. En réalité, chaque article peut avoir une page de discussion principale, mais il peut également exister des pages de discussion parallèles (Exemple 1), qui sont dédiées à un thème en particulier, d'après une liste de thèmes récurrents bien souvent liés à des questions de forme, de norme et d'organisation plus que de contenu. Le contenu de ces pages de discussion parallèles peut différer d'une page de discussion habituelle, car, en plus de contenir des fils de discussions développés, elles sont très fréquemment le lieu de vote sur le sujet auquel elles se rapportent. Ainsi, comme le montre l'exemple 7, sur la page parallèle *Article de qualité* de la discussion *Pierre Lambert de La Motte* (Exemple 1), l'admissibilité de l'article associé est débattue sous forme de vote.
Malgré la présence de situations d'interactions pertinentes dans les pages de discussion parallèles, nous avons fait le choix de ne pas les conserver dans les données à observer car elle sont aussi fréquemment le lieu d'interactions qui ne rentrent pas dans le type d'interaction que nous souhaitons analyser dans le cadre de cette étude.

Votes [modifier le code]

Format : Motivation, signature.

Article de qualité [modifier le code]

1. **Article de qualité** Proposant --Babouba Envie de me répondre ? 14 mai 2012 à 22:41 (CEST)
2. **Article de qualité** Très bon article (article de qualité quoi). **--Orikrin1998** 15 mai 2012 à 13:47 (CEST)
3. **Article de qualité** Très bon article en effet. Un bel esprit de synthèse, un style riche sans être lourd, une belle plume en somme.--Willuconquer (d) 16 mai 2012 à 10:53 (CEST)
4. **Article de qualité** Article de fond, très intéressant !--Eymery (d) 17 mai 2012 à 21:39 (CEST)
5. **Article de qualité** Très bon travail SC Lusoense 20 mai 2012 à 17:28 (CEST)
6. **Article de qualité** Excellent. Tant pis pour les jésuites .--Ps2613 (d) 24 mai 2012 à 03:09 (CEST)
7. **Article de qualité** Suivi depuis janvier. Excellente progression, article très intéressant sur une personnalité que je ne connaissais pas. Cordialement, Kertraon (d) 24 mai 2012 à 15:01 (CEST)
8. **Article de qualité** non seulement sans problème mais vraiment intéressant, tous les points importants détectés sont sourcés. Toutefois, alors qu'il est dans l'ensemble bien écrit (comme déjà relevé), il y reste quelques lourdeurs par répétition trop fréquente des noms des protagonistes. C'est du moins mon impression, et c'est une broutille par rapport à l'ensemble. Bonne continuation, c'est du super travail. --Acer11 Χαίρε 29 mai 2012 à 19:58 (CEST)
9. **Article de qualité** Article riche et bien écrit, intéressant même pour le néophyte, un excellent travail. Bibo le magicien (d) 14 juin 2012 à 08:48 (CEST)

Bon article [modifier le code]

Attendre [modifier le code]

Neutre / autres [modifier le code]

Exemple 7 : Discussion:Pierre Lambert de La Motte/Article de qualité
PdD : 6297650, FdD : 4
Vote des contributeurs

Exemple 7 : Les différents contributeurs votent dans le fil dédié pour indiquer s'ils sont d'accords pour attribuer le label *Article de qualité.*

Ce premier tri a ainsi permis de garder les pages de discussion principales, mais également les pages de discussion archivées, qui sont tout simplement les anciennes pages de discussion conservées sous forme d'archive afin de laisser de la place sur la page de discussion principale.

	Pages de discussion concernées	Fils de discussion concernés	Nombre de messages	Nombre de mots
Pages principales et archives	**159 841** **44 %**	**402 263** **39 %**	**1 133 421** **47 %**	**101 708 498** **63 %**
Pages parallèles	205 769 56 %	621 578 61 %	1 233 093 51 %	60 124 800 37 %

Tableau 2 : Distribution des pages de discussion et des fils de discussion selon les pages principales et les pages parallèles

II.2.2 Niveau fil de discussion : identification des fils n'ayant pas de potentiel d'interaction

Nous avons effectué un premier tri au niveau des pages de discussion afin d'écarter les fils de discussion qui sont issus de pages de discussion parallèles, car ces dernières contiennent régulièrement des fils de discussion au sein desquels se déroulent des votes (Exemple 7), situation d'interaction que nous ne souhaitons pas analyser lors de cette étude. Les données restantes sont ainsi composées de tous les fils de discussion qui proviennent de pages de discussion parallèles ou de pages de discussion archivées (*Tableau 2)*.

À partir de maintenant, les fils de discussion sont considérés comme unité d'analyse. Une première analyse des données restantes a fait émerger des profils de fils de discussion qui sont pauvres en interactions. Cette section décrit l'identification de ces situations à partir de traits établis grâce à l'analyse outillée des fils de discussion (Tableau 3) : le nombre d'utilisateurs impliqués dans le fil, le nombre de messages total dans le fil ainsi que l'utilisateur le plus actif du fil.

a. Présentation des traits

Cette première identification s'appuie sur trois traits qui caractérisent les fils de discussion, synthétisés dans le Tableau 3 :

- 01_*nbUser*
  - Technique
    Tous les noms d'utilisateurs du fil sont récupérés dans une table de fréquence qui, pour chaque utilisateur, recense le nombre de messages qu'il a posté. Ces informations sont issues des méta-données disponibles dans la structure XML (*cf.* II.1.3 Structure des données de WikiDisc). Le trait est calculé en comptabilisant le nombre d'utilisateurs différents de cette table de fréquence.

  - Description et Remarques
    Ce trait permet de connaître le nombre d'utilisateurs qui prennent part à la discussion au sein du FdD.
    Le cas des utilisateurs non identifiés, reconnus avec le pseudo *anonyme*, est traité en considérant qu'un utilisateur anonyme correspond à un seul utilisateur : ce cas de figure est explicité en section c.2 Cas des auteurs anonymes.

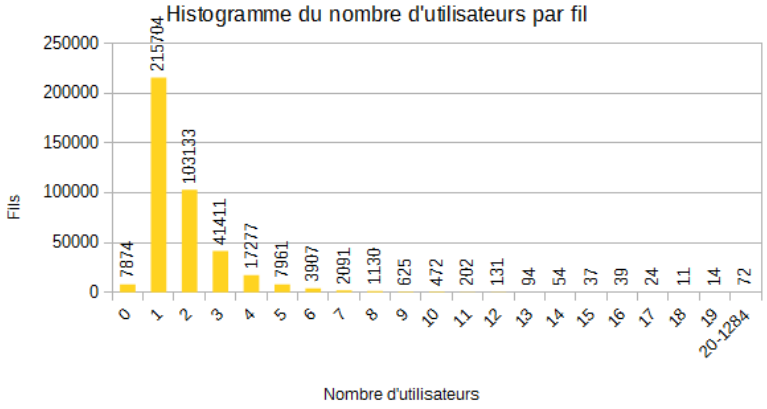


Figure 1: Histogramme du nombre d'utilisateurs par fil de discussion

La figure 1 permet d'observer que la plupart des fils du discussion n'impliquent qu'un seul utilisateur, et que la majorité des fils de discussion multi utilisateurs impliquent 2 ou 3 utilisateurs. Il apparaît également que certains fils n'impliquent aucun utilisateur : c'est le cas des fils vides, présentés en partie b.

- 02_*nbMsg*
  - Technique
    Ce trait est calculé en comptabilisant le nombre d'éléments *post* (*cf. II.1.3 Structure des données de WikiDisc*), qui correspondent aux messages, de l'élément *div*, qui correspond au fil de discussion.

  - Description et Remarques
    Ce trait permet de connaître le nombre total de messages qui constituent le fil de discussion.
    Ce paramètre est essentiel afin d'établir si un fil de discussion est actif ou s'il ne l'est pas. En effet, plus il contient de messages, plus son potentiel interactif est élevé.

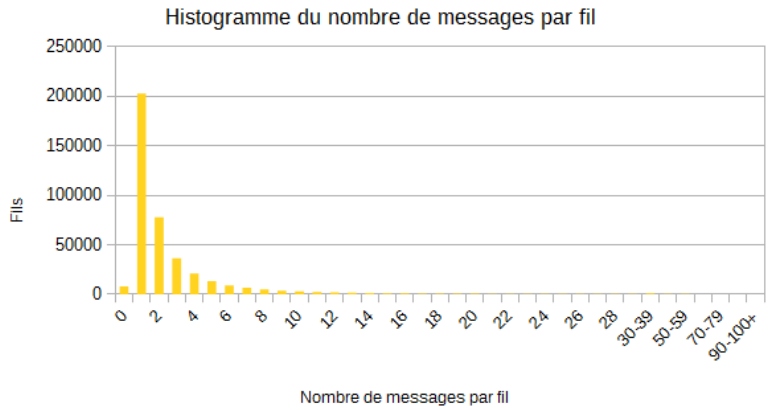


Figure 2 : Histogramme du nombre de messages par fil de discussion

Ici on peut voir qu'une grande majorité des fils de discussion ne comportent qu'un seul message. Concernant les fils contenant plusieurs messages, la plupart se limitent à deux, et on peut noter qu'il est très rare qu'un fil de discussion soit composé de plus de 5 messages. Par ailleurs on peut observer le fait que certains fils soient complètement vides, phénomène que l'on peut rallier à celui observé dans l'histogramme 1 qui fait ressortir les fils de discussion n'impliquant aucun utilisateur.

- 03_*actif*
  - Technique
    Ce trait est défini à partir de la table de fréquence de messages par utilisateurs. Le(s) utilisateur(s) ayant posté le nombre maximum de messages de la table est/sont considéré(s) comme le(s) plus actif(s).

  - Description et Remarques
    Ce trait permet d'identifier les utilisateurs qui postent le plus de messages au sein du fil. Il peut ainsi définir si le contributeur le plus actif est identifié comme anonyme (*cf.* c.2 Cas des auteurs anonymes*)* ou non.

Num	Nom	Intitulé	Description	Minimum	Maximum
01	01_*nbUser*	Nombre d'utilisateurs du fil	Ce trait correspond au nombre d'utilisateurs qui participent au fil de discussion.	0	1 284
02	02_*nbMsg*	Nombre de messages du fil	Ce trait correspond au nombre total de messages postés dans le fil de discussion.	0	228
03	03_*actif*	Utilisateur(s) le(s) plus actif(s) du fil de discussion	Ce trait indique le(s) contributeur(s) ayant posté le plus de message dans le fil de discussion.	_	_

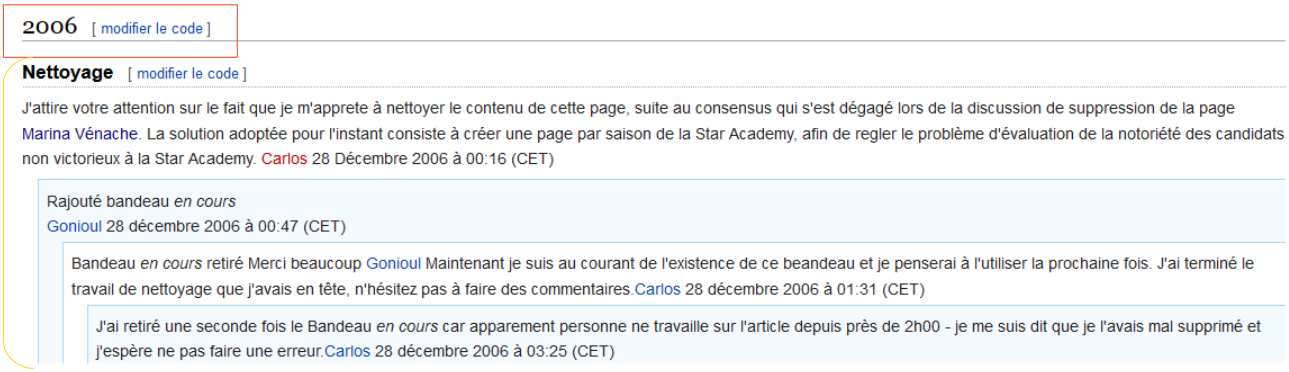
Tableau 3 : Tableau 3: Synthèse des traits descriptifs utilisés pour, dans un premier temps, effectuer l'identification des fils de discussions dépourvus d'interaction, puis par la suite caractériser les profils de d'interaction au sein des données restantes

b. Fils vides et mono message

Le deuxième niveau de structure des données est celui du fil de discussion, comme expliqué en section II.1.3 Structure des données de WikiDisc. Un premier aperçu de l'ensemble des fils de discussion, grâce à leurs statistiques quantitatives, met rapidement en évidence des cas qui risquent de ne pas être exploitables.

Tout d'abord, certains FdD sont vides, c'est à dire qu'ils ne contiennent aucun message. Une observation manuelle de 25 fils vides a fait ressortir que plusieurs situations peuvent expliquer cette absence de message. En voici les principales :

- Problème lors de la création du fil / l'élaboration d'un message : l'utilisateur n'a pas bien respecté la structure de la WP, et a posté un message en tant que fil, par exemple.
- Le titre du fil de discussion est utilisé comme titre de niveau supérieur pour des fils de discussion de hiérarchie inférieure (Exemple 8).
- Le titre du fil est utilisé comme section de vote où les utilisateurs souhaitant voter pour cette section sont censés le signaler sous forme de message dans ce fil. Parfois la section ne génère aucun vote, et le fil reste donc vide.



Exemple 8 : Discussion:Star Academy (France)
PdD : 501634, FdD : 1, 2
Exemple de fil vide - Fil "2006" utilisé comme titre de niveau supérieur

Exemple 8 : Ici le fil « 2006 » est utilisé comme titre de niveau hiérarchique supérieur au fil « Nettoyage » qui suit. Il ne contient aucuns messages.

Les fils ne contenant pas de messages sont donc dépourvus de contenu où des utilisateurs interagissent. Dans un premier temps il a été envisagé de conserver ceux qui font office de titre de niveau supérieur, car leur contenu sémantique aurait pu être pertinent à étudier, notamment pour les questions de recouvrement lexical. Cependant, la distinction des différents cas de figure étant relativement complexe, ces fils vides ont été définitivement écartés des données à étudier.

Par ailleurs, le cas de fils de discussions constitués d'un seul message a été observé. Ces messages isolés étant également des situations dans lesquelles il n'y a pas d'échanges entre les utilisateurs, les cas de FdD *mono message* ont eux aussi été mis de côté.

	Nombre de PdD concernées	Nombre de fils	Nombre de messages	Nombre de mots
Fils ayant au moins deux messages	**81 538** **51 %**	**191 640** **48 %**	**930 689** **82 %**	**84 288 277** **83 %**
Fils vides	5 138 3 %	7 874 2 %	0 0 %	0 0 %
Fils *mono message*	115 365 72 %	202 732 50 %	202 732 18 %	17 420 137 17 %

Tableau 4 : Proportion de fils vides, fils mono message et fils composés d'au moins deux messages.

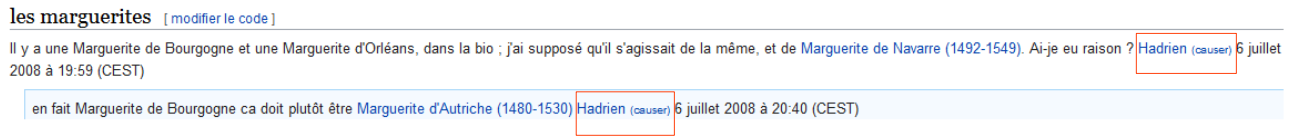
C . Fils monologue

Certains fils de discussion sont caractérisés par le trait 01_*nbUsr* = 1, ce qui signifie qu'*a priori* un utilisateur fait un monologue. En croisant l'observation avec le trait 03_*actif*, deux cas de figure ont émergé : soit le l'utilisateur est identifié, soit l'utilisateur est considéré comme anonyme, ce qui signifie que plusieurs contributeurs peuvent potentiellement se cacher sous le pseudonyme « anonyme ».

c.1 Cas des auteurs identifiés

L'observation de 5 fils de discussion n'impliquant qu'un seul contributeur identifié a fait ressortir plusieurs objets de monologue :

- Le contributeur pose une question, fait une remarque, et y répond lui même (Exemple 9).
- Le contributeur fait une *check list* des modifications qu'il effectue.
- Le contributeur exprime plusieurs remarques, qui ne sont pas forcément articulées entre elles, comme si le fil était un journal de bord.



Exemple 9 : Discussion:Henri-Corneille Agrippa de Nettesheim
PdD 995566, FdD 2
Fil de discussion « les marguerites », exemple de monologue

Un seul et même auteur

Exemple 9 : Ici un contributeur, « Hadrien », poste deux messages : le premier pour faire une remarque et poser une question, le second pour indiquer qu'il répond à la question qu'il a posé dans le message précédent.

	Pages de discussion concernées	Fils	Messages	Mots
Monologue auteurs identifiés	94 076 59 %	141 936 35 %	152 835 13 %	11 888 161 12 %

Tableau 5 : Proportion de fils monologue dont l'auteur est identifié

c.2 Cas des auteurs anonymes

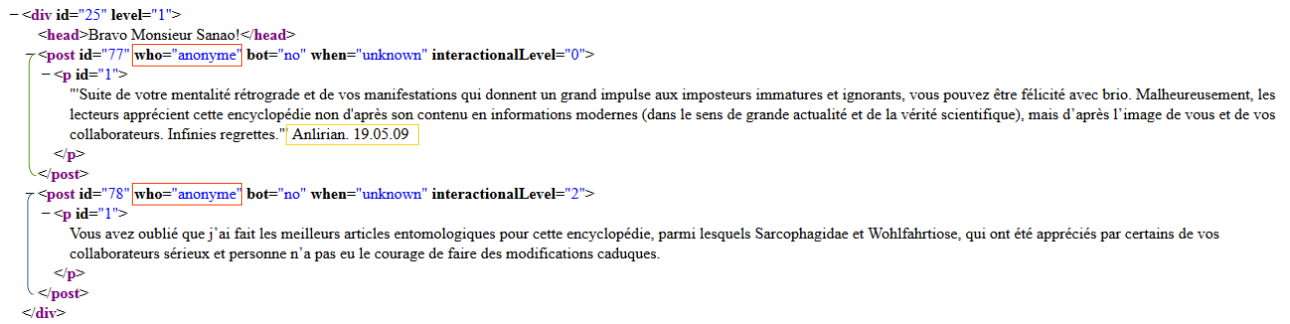
Dans le cas où l'utilisateur seul est considéré comme « anonyme », il n'est pas évident de conclure que le fil de discussion n'implique effectivement qu'un seul utilisateur, car tous les utilisateurs anonymes se retrouvent considérés comme un seul et même utilisateur. Il est ainsi possible qu'un fil de discussion ayant plusieurs utilisateurs anonymes soit caractérisé par 01_nbUser = 1. Le Tableau 6 synthétise l'observation de 10 fils ayant comme trait *01_nbUser* = 1, *03_actif* = « anonyme » et *02_nbMsg* > 1.

Référence				02_ *nb Msg*	Nombre d'utilisateurs identifiés par l'annotation manuelle
Num PdD	Titre PdD	Num FdD	Titre FdD		
5010271	Discussion:Mireille Faugère	1	Untitled	5	**2**
2270498	Discussion:Bernard Dubourg/archive2	120	réponses à "par ordre" de MLL	6	1
471074	Discussion:Dessein intelligent	1	La formulation...	4	**2**
5491006	Discussion:Affaire Dominique Strauss-Kahn	7	Couverture médiatique du procès	2	1
1865840	Discussion:Bengalia	18	Valeur de l'article ?	3	**2**
1865840	Discussion:Bengalia	25	Bravo Monsieur Sanao!	2	**2**
1905	Discussion:Liste des pays du monde	25	Longueur de forme	2	1
5486756	Discussion:Dominique Strauss-Kahn/Archive 2	19	Affaire Tapie	2	1
501407	Discussion:Liste des personnages de Bleach	5	Couleur de cheveux. ^^	5	**2**
131810	Discussion:Surf music	4	S. Koenig est passé par là..	3	**2 ou 3**

Tableau 6 : Observation du nombre réel d'utilisateurs lorsque le nombre indiqué est 1 et que l'utilisateur est identifié comme anonyme

Globalement il apparaît que plusieurs, souvent deux, utilisateurs sont identifiés lorsque le fil contient plus de deux messages. Cette exception sera donc prise en compte dans le tri des fils de discussion à un seul utilisateur.

Les fils concernés par le cas *anonyme* conservés *i.e.* constitués de plus de deux message, seront considérés comme des fils de discussion impliquant deux utilisateurs.



Exemple 10 : Discussion:Bengalia
PdD : 1865840, FdD : 25
Fil « Bravo Monsieur Sanao ! » : un seul utilisateur anonyme mais plusieurs utilisateurs identifiés par annotation manuelle

Par ailleurs, cette situation est bien entendu possible dans tous les FdD comprenant un utilisateur identifié *anonyme* correspondant à plusieurs utilisateurs, parmi d'autres utilisateurs identifiés. Ce cas de figure étant particulièrement complexe à déceler, la situation expliquée précédemment est la seule pour laquelle le cas *anonyme* a subi un traitement particulier. Désormais, tous les cas *anonyme* seront considérés comme un seul utilisateur lorsque la caractéristique *01_nbUser* sera prise en compte.

Dans la mesure où une situation d'échange entre utilisateurs ne peut exister que si plusieurs, donc au moins deux, utilisateurs sont en interaction, les fils de discussion identifiés comme monologue, ainsi que les fils de discussion identifiés comme monologue_anonyme ne sont pas conservés dans les données à étudier car ils sont dépourvus de potentiel d'interaction. Les données restantes, le corpus_v1, est composé de tous les fils ne discussion qui ne correspondent pas au profil vide, mono message, monologue ou monologue_anonyme (Tableau 7).

Profil	Caractéristiques	Proportion			
		PdD	FdD	Messages	Mots
01_*vide*	02_*nbMsg* = 0	5 138 3 %	7 874 2 %	0 0 %	0 0 %
02_*mono message*	02_*nbMsg* = 1	115 365 72 %	202 732 50 %	202 732 18 %	17 420 137 17 %
03_*monologue*	01_*nbUser* = 1 03_*actif* = un auteur identifié	94 076 59 %	141 936 35 %	152 835 13 %	11 888 161 12 %
04_ *monologue_anonyme*	01_*nbUser* = 1 03_*actif* = anonyme e 02_*nbMsg* ≤ 2	2 924 2 %	3 346 0,8 %	6 692 0,6 %	649 878 0,6 %

Tableau 7 : Synthèse des profils pauvres en interaction

III. Une première caractérisation : des fils de discussion facilement identifiables

Le premier affinage a permis d'identifier les fils de discussion n'ayant pas de potentiel d'interaction. Cette section revient sur l'analyse du corpus_v1 basée sur les traits présentés précédemment (Tableau 3) ainsi que de nouveaux traits relatifs à des caractéristiques quantitatives des fils (Tableau 8). Cette étape permet de faire émerger des premiers profils d'interactions entre les utilisateurs.

III.1 Méthodologie de l'observation

Une analyse outillée a permis de calculer des traits quantitatifs pour chaque fil de discussion, tels que leur taille en mots, en phrases, ou encore la moyenne de messages par utilisateur. Cet ensemble de trait fournit ainsi une base pour l'observation du corpus_v1, qui est réalisée selon la méthodologie suivante :

1. Observer les traits établis par l'analyse outillée : hypothèse et repérage d'un trait quantitatif pouvant être significatif pour l'hypothèse.
    - Exemple : Le trait "nombre d'utilisateurs" : les interactions entre deux utilisateurs ne sont pas les mêmes que lorsque le nombre d'utilisateurs est plus élevé.

2. Établir un seuil pour isoler les fils de discussion concernés par l'hypothèse.
    - Exemple : Récupérer tous les fils comprenant uniquement 2 utilisateurs pour observer les interactions "duo" : premier profil dégagé .

3. Une fois un premier seuil choisi, observer d'autres traits parmi les données concernées par le seuil pour affiner l'hypothèse.
    - Exemple : Dans les conversation duo, y-a-t-il une différence entre les conversations ayant un format Message1/Message2, et celles plus prolifiques ?

4. Annotation d'un échantillon des données dégagées afin de déterminer les corrélations et les disparités entre les différents fils sélectionnés. Cette observation peut amener à préciser à nouveau d'autres seuils ou bien à établir un profil.

III.2 Une observation assistée par une analyse outillée : présentation des traits calculés

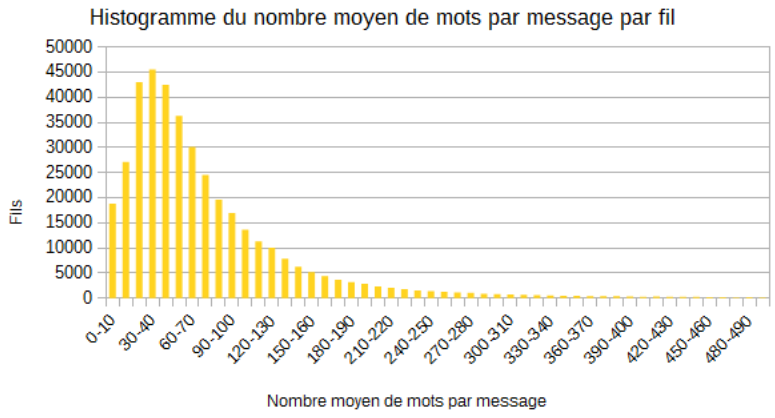
La manipulation du corpus_v1 est assistée par un ensemble de traits (Tableau 8) qui font ressortir des caractéristiques quantitatives des fils de discussion, permettant ainsi de rassembler ces derniers selon certains seuils des traits afin de distinguer des profils types d'interactions :

Les histogrammes 3, 4, 5 et 6 illustrent la répartition des fils de discussion selon les différents traits dans l'ensemble des fils de discussion issus de pages principales et archivées, soit 402 263 fils.

- 04_nbMots
  - Technique
    Dans la version étiquetée par Talismane (*cf.* II.1.3 Version étiquetée avec Talismane), le contenu des messages est segmenté en mots. Le nombre de mots total du fil est calculé en comptabilisant toutes les unités mots identifiés dans la totalité du fil.
  - Description et Remarques
    Ce trait permet de connaître le nombre total de mots dans le fil de discussion. Il n'est pas employé directement pour l'analyse, mais utilisé pour calculer d'autres traits.

- 05_nbPhrases
  - Technique
    Dans la version étiquetée par Talismane (*cf.* II.1.3 Version étiquetée avec Talismane), le contenu des message est également segmenté en phrases, et chaque phrase est séparée d'une autre par une ligne vide. Le nombre de phrases est calculé en comptabilisant le nombre de lignes vides du fil.
  - Description et Remarques
    Ce trait permet de connaître le nombre total de phrases dans le fil de discussion. Il n'est pas employé directement pour l'analyse, mais utilisé pour calculer d'autres traits.

- 06_moyMotsMsg
  - Technique
    Diviser 04_*nbMots* par 02_*nbMsg*.
  - Description et Remarques
    Ce trait permet de récupérer le nombre moyen de mots par message du fil. Il permet d'établir si les messages échangés sont plutôt développés ou plus succincts. Ce constat peut être représentatif de types différents d'interaction.



La majorité des messages comporte entre 10 et 100 mots, même si on peut observer une quantité non négligeable de messages très courts (moins de 10 mots), et certains messages très longs (plus de 100 mots) mais minoritaires.

Figure 3 : Histogramme du nombre moyen de mots par message – Jusqu'à 500 mots

- 07_moyPhrasesMsg

- Technique
  Diviser 05_*nbPhrases* par 02_*nbMsg*.
- Description et Remarques
  Ce trait permet de récupérer le nombre moyen de phrases par message du fil. Il permet d'établir si les messages échangés sont plutôt développés ou plus succincts. Ce constat peut être représentatif de types différents d'interaction.
-

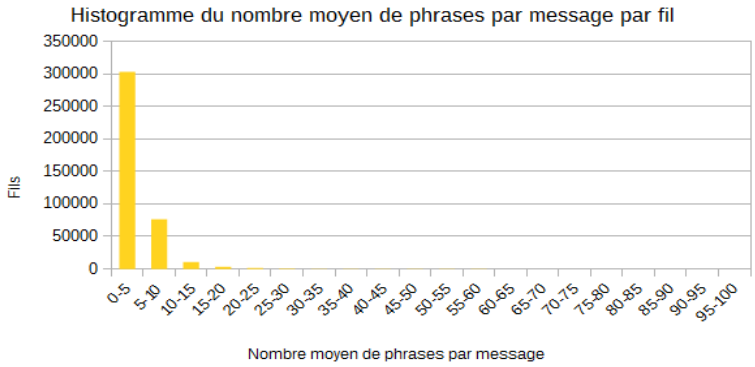


Figure 4 : Histogramme du nombre moyen de phrases par message - Jusqu'à 100 phrases

On peut observer que presque tous les messages sont composé d'au maximum 10 phrases. On peut avancer que les messages situés au-delà de ce seuil sont considérablement longs.

- 08_moyMotsPhrase
  - Technique
    Diviser 06_*moyMotsMsg* par 07_*moyPhrasesMsg*.
  - Description et Remarques
    Ce trait permet de connaître le nombre moyen de mots par phrase du fil. Il permet d'observer si les phrases du fil sont globalement longues, moyennes ou courtes. Dans un cas comme dans l'autre, les extrêmes peuvent être révélateurs de contenu ou d'interactions particuliers. Par exemple, des phrases très courtes en mots peuvent signaler un vote, ou au contraire, des phrases contenant énormément de mots peuvent être symptomatique d'un contenu non rédigé par l'auteur, comme une liste.

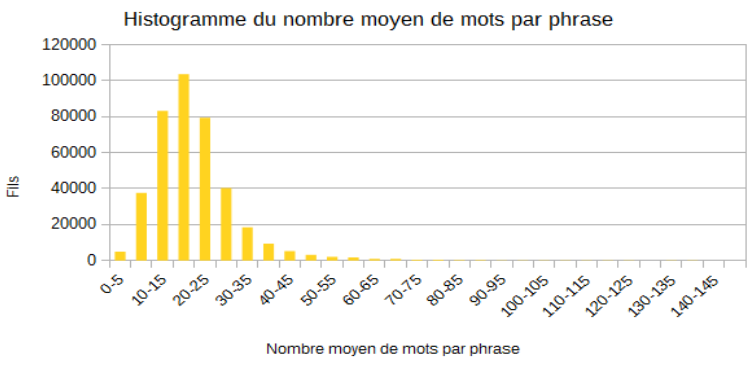


Figure 5 : Histogramme du nombre moyen de mots par phrase - Jusqu'à 150 mots

Globalement, les phrases font entre 5 et 30 mots. Ainsi, une phrase sera considérée comme très courte en deçà de ce seuil, et les phrases dépassant les 30 mots seront considérées comme très longues.

- 09_moyMsgUser
  - Technique
    Diviser 02_*nbMsg* par 01_*nbUser.*

  - Description et Remarques
    Ce trait permet de récupérer le nombre moyen de messages postés par les utilisateurs. Il peut révéler si un fils de discussion est entretenu par un groupe d'auteurs impliqués et actifs, ou plutôt par une participation ponctuelle de plusieurs utilisateurs.

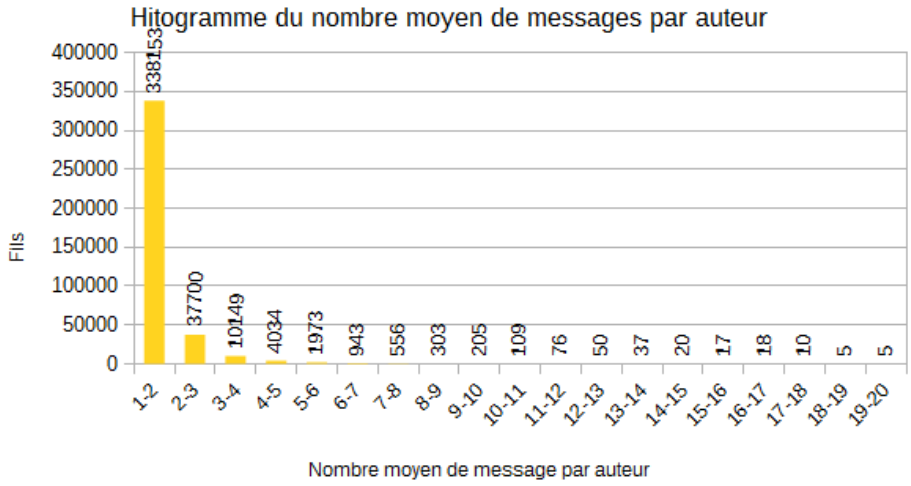


Figure 6 : Histogramme du nombre moyen de messages par auteur

On peut observer qu'il est peu fréquent qu'un auteur rédige plus de 2 messages. Ce seuil pourrait permettre de définir qu'un auteur est particulièrement impliqué s'il rédige au moins 3 messages, par exemple.

Num	Nom	Intitulé	Description	Minimum	Maximum
04	04_*nbMots*	Nombre total de mots du fil	Ce trait correspond au nombre total de mots qui composent le fil de discussion.	0	39 969
05	05_*nbPhrases*	Nombre total de phrases du fil	Ce trait correspond au nombre total de phrases qui composent le fil de discussion.	0	1 860
06	06_*moyMotsMsg*	Nombre moyen de mots par message	Ce trait correspond à la taille moyenne des messages du fil en mots.	0	39 969
07	07_*moyPhrasesMsg*	Nombre moyen de phrases par message	Ce trait correspond à la taille moyenne des messages du fil en phrases.	0	1 212
08	08_*moyMotsPhrase*	Nombre moyen de mots par phrase	Ce trait correspond à la taille moyenne des phrases du fil en mots.	0	998
09	09_*moyMsgUser*	Nombre moyen de messages postés pour chaque utilisateur	Ce trait correspond au nombre moyen de messages postés par les contributeurs qui participent au fil.	1	56.25

Tableau 8 : Synthèse des traits descriptifs employés pour accompagner l'observation du corpus_v1

III.3 Des profils de fils de discussion généraux

En terme d'interaction, le nombre de participants semble être une première caractéristique permettant de distinguer plusieurs types de rapport entre les utilisateurs. En partant de ce postulat, nous avons décidé de diviser le corpus principal en deux sous-ensembles afin d'étudier de plus près les différentes situations de communication au sein de ces deux profils plus généraux : les interactions *Duo* entre deux utilisateurs, et les interactions *Pluri* entre plus de deux utilisateurs (Tableau 9).

	Pages de discussion concernées	Fils	Messages	Mots
Deux utilisateurs	59 681 37 %	103 690 26 %	287 888 25 %	23 741 398 23 %
Plus de deux utilisateurs	34 651 22 %	75 544 19 %	616 141 54 %	58 433 047 57 %

Tableau 9 : Proportion de fils Duo et de fils Pluri

III.3.1 Profil Duo : exactement deux utilisateurs

Ce premier profil encore assez général permet d'observer des interactions n'impliquant que deux utilisateurs. Pour observer ce cas particulier, tous les fils ayant exactement 2 utilisateurs comptabilisés, ainsi que les fils ayant un seul utilisateur du cas *anonyme* (*cf. c.2 Cas des auteurs anonymes*) ont été relevés (Tableau 9), mais répartis dans deux sous corpus différents (Tableau 10): *cDuo2usr* et *cDuo2usrAnonyme*. Toutes les observations primaires ont été effectuées sur *cDuo2usr*, puis les profils identifiés ont été appliqués

à *cDuo2usrAnonyme* afin de constater si le cas anonyme est vraiment différent, ou s'il peut être relié à *cDuo2usr*.

	Pages de discussion	Fils	Messages	Mots
2 utilisateurs exactement	59 182 37 %	103 133 26 %	285 456 25 %	23 504 385 23 %
Cas anonyme	499 0,3 %	557 0,1 %	2 432 0,2 %	237 013 0,2 %

Tableau 10 : Répartition des sous-ensembles cDuo2usr et cDuo2usrAnonyme

a. cDuo2usr

- **Nombre de messages**

Ce paramètre permet d'identifier si l'interaction est très courte et donc peu développée, ou bien plus prolifique, et donc potentiellement plus susceptible de contenir des interactions complexes.

La première observation vise à déterminer si les rapports sont effectivement différents selon la quantité de messages produits, et si oui, à partir de quel seuil. Pour cela, nous commençons par observer les fils très sommaires qui ne contiennent que 2 messages afin d'établir des profils d'interactions récurrentes, puis observer les fils avec un seuil plus élevé et comparer les interactions.

	Pages de discussion concernées	Fils	Messages	Mots
cDuo2usr 2 messages	44 833 28 %	66 562 16,5 %	133 124 12 %	9 867 617 10 %
CDuo2usr plus de 2 messages	24 453 15 %	36 571 9 %	152 332 13 %	13 636 768 13 %

Tableau 11 : Proportion de fils correspondant à cDuo2usr de 2 messages ou plus de 2 messages

- **Deux messages exactement**

L'observation de 20 fils de discussion contenant uniquement 2 messages a permis d'identifier plusieurs types d'interaction :

- **A Échange succinct**
  Cet échange est caractérisé par un structure binaire dans laquelle le premier message pose une question peu développée ou une remarque courte, et le second message amène une réponse succincte, une confirmation ou un désaccord avec le précédent message. Il n'y a pas ou peu de négociation dans ce type d'interaction.

- **B Échange riche**
  L'échange riche semble assez similaire à l'échange succinct en terme de structure, mais son contenu

est beaucoup plus développé par les auteurs.

- C Contenu non discursif
  Cette situation est rencontrée lorsque un ou plusieurs messages du fil contiennent des contenus qui ne sont pas rédigés par l'auteur lui même sur le moment de l'interaction, et qui sont issus d'une source extérieure et produits de manière différée par rapport à la discussion.

Ces types étant souvent corrélés avec le nombre moyen de mots par messages (06_*moyMotsMsg*), trois seuils ont ainsi été déterminés dans le but d'isoler les cas concernés afin d'en observer un échantillon :

- Moins de 50 mots : le seuil de 50 a été déterminé à partir de deux points :
  - L'analyse qualitative a fait ressortir que les fils type A avaient une moyenne inférieure à 50,
  - Le calcul de la médiane des moyennes du nombre de mots/message de l'ensemble des fils (60,5) a confirmé que le seuil de 50 était un palier bas, et que les fils situés en dessous faisaient bien parti des fils ayant des messages plus courts.

- Plus de 1000 mots : ce seuil a été déterminé arbitrairement, mais il est complètement au dessus de la moyenne des moyennes du nombre de mots/messages qui est de 124.

- Entre 50 et 1000 mots : le reste des messages sont supposés de type B. Une observation plus approfondie permettra de le confirmer ou le contredire.

- Cas A Échange succinct : *06_moyMotsMsg* < 50

L'annotation manuelle de 10 messages à ce seuil a confirmé que les fils concernés contiennent peu de contenu interactif. La structure discursive Question/Remarque > Réponse rapide/Confirmation de modification/Contradiction de la remarque est très récurrente. Le contenu n'est pas très interactif et très succinct : comme le montre l'exemple 11, le premier contributeur poste un message pour poser deux questions concises, et le second contributeur répond à ses questions de manière très succincte, en moins de 10 mots . Le profil *Échange succinct* est donc identifié selon les seuils

- *01_nbUser* = 2
- *02_nbMsg* = 2
- *06_moyMotsMsg* < 50

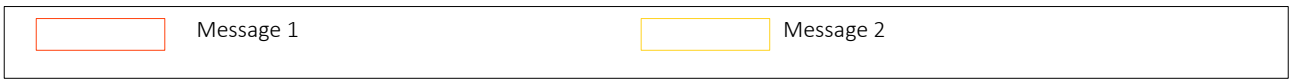
Autonomes [modifier le code]

Pourquoi est-ce que l'extrême-gauche ne regrouperait qu'une partie des autonomes ? Il y a des autonomes qui ne sont pas d'extrême-gauche ?

Alphonse Wagner (d) 21 février 2009 à 23:04 (CET)

oui par ex : 'les anars autonomes' --Leslib (d) 22 février 2009 à 22:14 (CET)

Exemple 11 : Discussion:Extrême gauche
PdD : 45229, FdD : 37
Fil de discussion « Autonomes » : échange de deux messages courts entre deux utilisateurs



Exemple 11 : Ici les messages sont très courts : le premier contient deux questions succinctes, qui n'est pas très développée par l'auteur. Le second message répond à cette question de manière rapide : la phrase n'est même pas vraiment construite.

• **Cas B** Échange riche 50 < 06_*moyMotsMsg* < 1 000

Ce sous ensemble présente beaucoup plus de matière à étudier, et l'observation fait ressortir des interactions concernant le contenu de l'article, mais aussi les comportements des utilisateurs. La structure de l'interaction est presque similaire à celle de duo A *i.e.* Question/Remarque → Réponse/Confirmation de modification/Contradiction de la remarque, mais le contenu est bien plus développé. Cependant le fait qu'il n'y ait pas de retour possible après le second message limite l'interaction entre les deux utilisateurs.
Le profil *question/réponse développé* est donc identifié selon les seuils :

- *01_nbUser* = 2
- *02_nbMsg* = 2
- *50 < 06_moyMotsMsg* 50 < 1 000

• **Cas C** contenu « déviant » 06_*moyMotsMsg* > 1 000

Ce seuil a fait émerger un sous ensemble de messages ne suscitant pas l'échange entre les utilisateurs, et contenant bien souvent des séquences non discursives, tels que des historiques de suppression ou encore des listes extrêmement longues (Exemple 12).
Le profil *non discursif* est donc identifié selon les seuils :

- *01_nbUser* = 2
- *02_nbMsg* = 2
- *06_moyMotsMsg* > 1 000


Bonjour,
On [5] m'a que coucou pour me faire naturigoaler belge je devais organiser chez moi un concert de soutien à Noël Godin en invitant Arolde, Plastic Bertrand, Jan Bucquoy, Jean-Luc Fonck de Sttellla, Stefan Liberski, Marka et les Snuls. Bon, Noël Godin, je sais c'est qui parce qui l'a chanté *Le mur* [6] sur l'album *Il faut tourner l'Apache*, l'année où on a été champion du monde dans ma salle de concert. Mais les autres, je sais pas c'est qui. Est-ce que quelqu'un qui ~~peut~~ sait m'aider ? [7]
Merci d'avance pour me soutenir. --Johnny Hallyday 8 avril 2007 à 01:43 (CEST)

Si vous me permettez d'intervenir [8], j'ai dans mes tiroirs une vieille recette qui pourrait être intéressante pour cette soirée [9]. Je l'ai adaptée à ma façon, afin qu'il vous soit facile de la mettre en œuvre. D'autre part, je suis prêt à vous aider pour que cette discussion soit la première promue discussion de qualité dans Wikipédia. J'ai, dans un autre de mes tiroirs, une étoile dont je ne me sers plus [10] et que je vous offrirai bien volontiers en cas de succès. --Avec tout mon respect gourmant, Pierre Wynants (discuisson) 10 avril 2007 à 02:44 (CEST)

Desserts, sauces, pâtes et biscuits

TARTE À LA FOU' FOUNE

INGRÉDIENTS

Pour 5 personnes

Pâte
- 1 œuf fermier entier pondu le matin
- 200 g de farines pures moulues la veille
- 1 pincée de sel du moulin de Guérande
- 100 g de beurre fermier fondu à feu non sucré mais doux
- 75 g de sucre glace à température ambiante

Garniture
- 200 g de cassonade rousse ou blonde tirlemontoise
- 2 œufs fermiers entiers pondus pendant l'heure du midi
- 25 cl de Fou' Foune

Ustensiles
- Un fouet
- Une toque
- Une paire de mains

PRÉPARATIFS (13 min. 18 sec.)

Contenu non rédigé dans le cadre de la discussion : Recette

Exemple 12 : Discussion:Noël Godin (extrait)
PdD : 395870, FdD : 7
Fil « Recette de la tarte aux gémonies » : contenu non discursif

Exemple 12 : Ici un des contributeurs a inséré une recette, vraisemblablement issue d'une source extérieure, et donc qui n'a pas été rédigée dans le cadre de la discussion.

L'analyse qualitative des fils de discussion impliquant deux utilisateurs, avec un message par utilisateur a fait ressortir trois profils distincts : le *question/réponse succinct*, le *question/réponse développé* et le *non discursif*.

- Plus de deux messages

Une première observation de 20 fils de ce sous ensemble a fait ressortir que le contenu des fils était beaucoup plus interactif, notamment en terme de débat et/ou de collaboration par rapport à la rédaction.
Nous avons tout de même observé certains seuils extrêmes afin d'exclure d'éventuels fils qui pourraient ne pas être pertinents. Ainsi, la moyenne du nombre de mots par messages (06_*moyMotsMsg*) a été observée selon les cas extrêmes :

- Extrême haut : le seuil a été établi après observation de la moyenne totale des moyennes de mots par message du fil. Celle-ci étant de 107 mots par message, et la médiane de 80 mots par messages, un échantillon de fils ayant une moyenne nettement supérieure à 107 ont été observés (entre 300 et le maximum, 1269).
- Extrême bas : ce seuil a été défini en observant la moyenne de mots par message minimale, de 6 messages. Peu de fils ayant une moyenne aussi basse, un échantillon de fils ayant une moyenne situées entre le minimum et 50 ont été observés.

• Extrême haut : 06_*moyMotsMsg* > 300

L'observation manuelle de 5 fils parmi les extrêmes a fait émerger plusieurs cas :

- Cas 1 : L'ensemble ou une partie du contenu n'est pas rédigé par l'utilisateur, et agrandit sensiblement la taille du message. Ce type de contenu est issu d'une source extérieure : comme l'illustre l'exemple 13, un des messages postés contient un historique, ce qui n'est pas une séquence produite par un des auteurs.
  - Exemples :Citation, Exemples, Liste (historique)
- Cas 2 : Les messages qui constituent le fil sont riches et produits par les contributeurs.

Afin de différencier ces différents cas, deux autres traits ont été observés : le nombre total de messages du fil, ainsi que la moyenne de phrases par message du fil.
À noter que la moyenne de phrase a été rapportée à la moyenne de mots afin qu'elle soit plus facilement interprétable.
Le Tableau 12 reprend ces indices, et pour chaque fil précise s'il concerne le cas de contenu externe (cas 1), ou bien le cas de messages effectivement rédigés par les utilisateurs (cas 2).

Référence				5_moyMotsMsg	2_nbMsg	6_moyPhraseMsg	7_moyMotsPhrase	Cas 1	Cas 2
Num PdD	Titre PdD	Num FdD	Titre FdD						
1374039	Discussion:Ayahuasca	4	Discussion sur les sources	1282.33	3	52,67	24,3		X
112483	Discussion:Boycott	4	Appel au boycott interdit en France ?	1081	3	12,67	85,3	X	
6776017	Discussion:Matérialisme dialectique/archive1	2	Remarques et Bibliographie toute récente	380	3	14,67	25,9		X
5014854	Discussion:Heinrich Harrer/archives	21	TI	351	3	9	39	X	
105760	Discussion:Action française	45	Politique extérieure : soutien au fascisme italien et antigermanisme	301	7	19	15,84		X
432283	Discussion:Bug (informatique)	13	Exemples	330	3	3,67	89	X	
1552170	Discussion:Chinua Achebe	1	Vieux motard que j'aimais, euh je veux dire Mieux vaut tard que jamais : revert vandalisme	344	7	2,14	160	X	
471074	Discussion:Dessein intelligent	11	Pourquoi ne pas etre clair, à partir d'un certain moment ?	339	4	17	19,9		X
330409	Discussion:Bande de Gaza	51	Point de vue de GastelEtzwane	394	3	16,33	24,12		X
1337072	Discussion:Corée du Nord/Archive 01	32	Utlisation de la page "Corée du Nord" sur le wikipédia anglophone	401	4	10,75	37,3		X
250067	Discussion:Démocratie athénienne	10	Recyclage de la page nécessaire?	539	6	25,33	21,2		X

Tableau 12 : Observation d'un échantillon de fils ayant une moyenne de mots / message élevé.

Vieux motard que j'aimais, euh je veux dire Mieux vaut tard que jamais : revert vandalisme [modifier le code]

Bonjour à tous. Et en particulier aux utilisateurs "sérieux" ayant collaboré à cet article. Je m'aperçois seulement maintenant que l'article Chinua Achebe a été vandalisé de fond en comble. Je suis retourné à la version de Utilisateur:Loveless du 30/12/2008. J'ai peur qu'il y aura désormais du travail pour tout remettre en ordre.

Erasmus.new (d) 29 mai 2010 à 11:31 (CEST)

(actu | diff) 29 mai 2010 à 11:18 Erasmus.new (discuter | contributions) (9 211 octets) (MIEUX CVAUT TARD QUE JAMAIS : article vandalisé le 18 janvier 2009 à 12:01 Fatima59 (en rouge) -- > retour à la version du 30 décembre 2008 à 09:15 de Loveless) (défaire)
(actu | diff) 29 mai 2010 à 10:58 Erasmus.new (discuter | contributions) (10 644 octets) (catégorisation) (défaire)
(actu | diff) 29 mai 2010 à 10:57 Erasmus.new (discuter | contributions) (10 609 octets) (→Thématiques : *A man of the people* est une frAsque politique ---> *A man of the people* est une frEsque politique !!!!!!!!!!!!!!!!!!!) (défaire)
(actu | diff) 29 mai 2010 à 10:55 Erasmus.new (discuter | contributions) (10 605 octets) (typographie ; wikification ; style ; orthographe) (défaire)
(actu | diff) 7 mai 2010 à 14:28 TXiKiBoT (discuter | contributions) m (10 579 octets) (robot Ajoute: fa:چینوا آچیبه) (défaire)
(actu | diff) 22 avril 2010 à 21:45 82.234.26.108 (discuter) (10 552 octets) (→Vie et oeuvres) (défaire)
(actu | diff) 5 avril 2010 à 17:05 RibotBOT (discuter | contributions) m (10 552 octets) (robot Ajoute: ca:Chinua Achebe) (défaire)

Exemple 13 : Discussion:Chinua Achebe (extrait)
PdD : 1552170, FdD : 1
Extrait du fil de discussion « Vieux motard que j'aimais, euh je veux dire Mieux vaut tard que jamais : revert vandalisme » correspondant au cas 1 : contenu non rédigé par l'utilisateur

Contenu non produit par l'utilisateur

Exemple 13 : Ici un contributeur a inséré l'historique de révision de l'article.

L'observation de ces résultats permet de remarquer que l'indice significatif écartant les fils de discussion extrêmes à cause de contenu non rédigé par les utilisateurs est le taux mots par phrase. Ainsi, au-delà du seuil de 39 mots par phrase, le fil de discussion a beaucoup plus de chance de contenir des éléments non rédigés, souvent copiés collés, ou bien un historique.
Le profil *contenu externe* est donc identifié par les seuils :

- 01_*nbUser* = 2
- 02_*nbMsg* > 2
- 06_*moyMotsMsg* > 300
- 08_*moyMotsPhrase* > 39

	Pages de discussion	Fils	Messages	Mots
Profil contenu externe	616 0,4 %	680 0,2 %	2 855 0,25 %	1 248 328 1,2 %

Tableau 13 : Proportion du profil contenu externe

•Extrême bas : 06_*moyMotsMsg* < 50

L'observation d'un échantillon de 9 fils a fait ressortir que globalement, en dessous de 20 mots par messages l'interaction était vraiment pauvre. L'exemple 14 montre l'échange de 3 messages concis, qui ne font pas plus d'une phrase chacun. Les messages 1 et 2 ont été postés par le même auteur, qui propose une suppression de bandeau et demande l'avis d'autres contributeurs impliqués dans la rédaction de l'article, puis en l'absence de réponse réitère son intention de supprimer le bandeau. Le message 3 est posté par un second auteur qui précise qu'il n'est pas opposé à sa proposition.
Le profil *interaction succincte* est donc identifié par les seuils :

- 01_*nbUser* = 2
- 02_*nbMsg* > 2
- 06_*moyMotsMsg* < 50

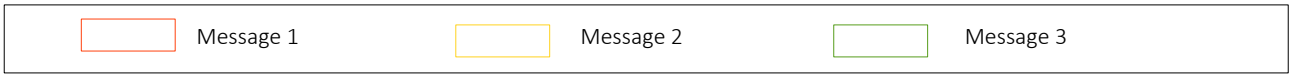
sources [modifier le code]

J'ai pas l'impression que l'article soit pauvre en sources, cela gène quelqu'un si on vire le bandeau ? Papillus (d) 21 septembre 2008 à 19:18 (CEST)

Sauf avis contraire je supprime le bandeau dans une semaine. Papillus (d) 24 septembre 2008 à 22:27 (CEST)

Pas d'"objections -- SerSpock à l'inter...ŧ&ŧ& 25 septembre 2008 à 11:10 (CEST)

Exemple 14 : Discussion:SUD Étudiant/Archive3 (extrait)
PdD : 5132579, FdD : 3
Extrait du fil de discussion « sources » : type messages courts entre deux utilisateurs avec plus de deux messages



L'analyse du corpus Duo, impliquant exactement 2 utilisateurs a fait émerger les profils suivants :

Profil	Caractéristiques	Proportion			
		PdD	Fils	Messages	Mots
05_ *échange succinct*	*01_nbUser* = 2 *02_nbMsg* = 2 *06_moyMotsMsg* < 50	26 807 16,8 %	32 350 8 %	64 700 5,7 %	2 002 498 2 %
06_*échange riche*	*01_nbUser* = 2 *02_nbMsg* = 2 *50 < 06_moyMotsMsg* < 1 000	23 902 14,95 %	33 751 8,4 %	67 502 5,95 %	7 505 222 7,4 %
07_ *non discursif*	*01_nbUser* = 2 *02_nbMsg* = 2 *06_moyMotsMsg* > 1 000	83 0,05 %	83 0,02 %	166 0,01 %	322 061 0,3 %
08_ *contenu externe*	01_*nbUser* = 2 02_*nbMsg* > 2 06_moyMotsMsg > 300 08_moyMotsPhrase > 39	616 0,4 %	680 0,2 %	2 855 0 ,25 %	1 248 328 1,2 %
09_ *interaction succincte*	01_*nbUser* = 2 02_*nbMsg* > 2 06_moyMotsMsg < 50	10 548 6,6 %	12 237 3 %	46 110 4 %	1 641 918 1,6 %

Tableau 14 : Synthèse des profils généraux identifiés dans cDuo

Ces premiers profils ont été identifiés au fil des différentes observations d'échantillons, et les seuils définis selon les situations observées. Il serait intéressant d'appliquer les seuils observés dans chaque sous-ensemble (ici cDuo 2 messages et cDuo + de 2 messages) afin de voir si des profils similaires émergent.

Tous les autres fils de discussion de cDuo qui ne rentrent pas dans l'une des catégories identifiées sont retenus dans la seconde version du corpus qui a pour objectif d'être analysée de manière plus approfondie.

b. cDuo2usrAnonyme

Une fois plusieurs profils établis, nous avons projeté les seuils caractéristiques sur le sous-ensemble cDuo2usrAnonymes afin de déterminer si les cas étaient similaires.

- b.1 01_ *question/réponse succincte*

Ce profil n'est pas compatible avec le cas anonyme, car ce dernier requiert comme seuil 02_*nbMsg* = 2, alors que les fils du profil anonyme sont composés d'au minimum 3 messages (c.2 Cas des auteurs anonymes).

- *b.2 02_question/réponse développé*

La situation est idem : le seuil 02_*nbMsg* = 2 n'est pas compatible.

- *b.3 03_ non discursif*

Ce profil a également comme seuil 02_nbMsg = 2 non compatible avec les caractéristiques du sous-ensemble cDuo2usrAnonyme, mais nous avons tout de même essayé d'appliquer le second seuil de ce profil : *06_moyMotsMsg* > 1 000.

Ce seuil a fait ressortir 4 fils de discussion, dont l'annotation manuelle (Tableau 15) a permis d'identifier deux cas :

- Échange de messages très longs et argumentés, sous forme de présentation approfondie des points de vue, mais ne menant pas vraiment à de la collaboration.
- Messages contenant de longues séquences issues de sources externes (article lié, site web, mail).

Références				02_*nbMsg*	Nombre de contributeurs identifiés	Profil
Num PdD	Titre PdD	Num FdD	Titre FdD			
2816512	Discussion:Faïence de Rouen	1	Untitled	3	2	Échange de messages très longs
15366	Discussion:Jacques Cheminade	1	Neutralisations	3	2/3	**Source externe**
130648	Discussion:Peul	24	paragraphe a retraiter	3	1	**Source externe**
1082466	Discussion:Droits de l'Homme en Iran/Archives	60	Réponse de Pentocelo (Mailée à Alithia, en raison d'un blocage de mon IP)	3	1	**Source externe**

Tableau 15 : Échantillon de fils de discussion corpus cDuo2UsrAnonyme avec le seuil 03_non discursif

La majorité des profils identifiés sont plutôt en adéquation avec le profil non discursif dans la mesure où ce sont des messages en grande partie non rédigés par les utilisateurs et n'impliquant pas de réels échanges entre les utilisateurs, mais ils correspondent en réalité plus au profil 04_*contenu externe*. Cette observation interroge sur l'éventualité de fusionner les profils 03_ *non discursif* et 04_ *contenu externe.*

- *b.4 04_ contenu externe*

Les seuils de ce profil ont fait émerger 26 fils de discussion. L'échantillon d'observation constitué de 5 fils de discussion (Tableau 16) a mis en évidence deux cas :

- Cas de contenu externe correspondant au profil
- Cas de d'échanges de messages rédigés par les utilisateurs, et ne contenant pas principalement des séquences issues de sources externes.

Références				02_*nbMsg*	Nombre de contributeurs identifiés	Profil
Num PdD	Titre PdD	Num FdD	Titre FdD			
3247743	Discussion:Thierry Groensteen	2	Rappelons les règles de base	6	2	Discussion
2140397	Discussion:Kenadsa	7	Références de Frenchinmorocco	3	1	**Contenu externe**
160512	Discussion:Quasiturbine	4	Avis critique sur les fondements scientifiques du concept	3	3	Discussion
595636	Discussion:Biométrie	6	Suppression en cours, de ma participation à l'article Biométrie	4	1	**Contenu externe**
992658	Discussion:Politique linguistique de la France	9	A propos de la langue bretonne	4	?	**Contenu externe**

Tableau 16 : Échantillon de fils de discussion corpus cDuo2UsrAnonyme avec le seuil 04_contenu externe

- *b.5 05_ interaction succincte*

Les seuils de ce profil ont fait émerger 195 fils de discussion. L'analyse qualitative d'un échantillon de 13 fils de discussion (Tableau 17) a fait ressortir quatre situations :

  - A : Correspondante au profil interaction succincte
  - B : Un seul utilisateur
  - C : Interaction développée
  - D : Liste

Références				02_*nbMsg*	Nombre de contributeurs identifiés	Profil
Num PdD	Titre PdD	Num FdD	Titre FdD			
3874669	Discussion:Étienne Raffort	1	ABS	3	1	Un seul utilisateur
2072036	Discussion:Saïd Taghmaoui	1	Nationalité	3	3	**Interaction succincte**
1534406	Discussion:Troupes coloniales	3	[[Troupes coloniales (France)|Troupes Coloniale]]	6	2	**Interaction succincte**
1223967	Discussion:Abalone (jeu)/Archives01	4	c dommage	4	3	**Interaction succincte**
1027523	Discussion:Xavier Niel	9	la Pravda ?	3	2 / 3	**Interaction succincte**
1014145	Discussion:Ortie	1	Purin d'ortie, illégal ?	3	2 / 3	**Interaction succincte**
2036945	Discussion:Séparation des variables	1	Erreur manifeste dans l'exemple n°2 (http://fr.wikipedia.org/wiki/Séparation_des_variables#Exemple_n.C2.B01)	6	2	**Interaction succincte**
3028590	Discussion:Rapport Grin	3	le nom de grin	3	2	**Interaction succincte**
55456	Discussion:Théorèmes d'incomplétude de Gödel	16	Sur la notion de vérité, plus en détail	18	2	Interaction poussée
8663041	Discussion:Exo-noyau	26	Les LibOSes	7	2	**Interaction succincte**
58337	Discussion:Ministres du premier gouvernement de Jean-Pierre Raffarin	3	Ministres délégués	12	1	Liste
2371046	Discussion:Typhlopidae	2	Les différentes espèces	12	1	Liste
1647737	Discussion:International Bank Account Number	1	Modulo	15	2 / 3	Interaction poussée

Tableau 17 : Échantillon de fils de discussion corpus cDuo2UsrAnonyme avec le seuil 05_interaction succincte

Le point intéressant qui ressort de cette observation est la prise en compte du trait 02_*nbMsg*, qui permet clairement d'identifier les cas d'interaction développée et de liste. Ce seuil serait à vérifier pour préciser le profil dans le corpus cDuo2Usr.

Dans tous les cas, les différentes observations ont fait ressortir qu'il n'y pas exactement deux utilisateurs impliqués, le cas anonyme n'est donc pas assez régulier pour être intégré au corpus c*Duo2usr* en considérant qu'il y a systématiquement deux utilisateurs. Une observation ultérieure pourrait permettre de mieux repérer le nombre réel d'utilisateurs impliqués dans ce type de fils de discussion.

III.2.2 Profil Pluri : Plus de deux utilisateurs

Ce sous-ensemble a été établi en récupérant les fils de conversation pour lesquels le nombre d'utilisateurs était supérieur à deux.
Les échanges ciblés entre deux utilisateurs diffèrent forcément d'échanges au sein d'une interaction impliquant plus de deux utilisateurs, c'est pourquoi l'observation du sous ensemble *pluriUser* a été réalisé séparément du sous ensemble *Duo* (Tableau 9).

Par ailleurs, la moyenne d'utilisateurs dans ce sous ensemble est située entre 4 et 5 utilisateurs. Deux sous ensembles ont donc été distingués pour un première observation :

- Pluri moyen: 5 utilisateurs maximum
- Pluri surpeuplé : au dessus de la moyenne (de 6 à 42)

a. Pluri moyen : 5 utilisateurs maximum

	Pages de discussion	Fils	Messages	Mots
Entre 2 et 5 utilisateurs	32 797 20,5 %	66 649 16,5 %	447 555 39,5 %	42 072 876 41,4 %

Tableau 18 : Proportion de fils Pluri moyen impliquant entre 2 et 5 utilisateurs

L'observation de 15 fils de discussion issus de ce sous-ensemble a fait émerger plusieurs situations de communication qui peuvent être regroupées en deux grandes catégories :

- La première comprend des fils de discussion au sein desquels les utilisateurs participent, débattent et collaborent : l'interaction entre les utilisateurs y est donc particulièrement présente.
- La seconde, catégorie *composite* (Exemple *15*), correspond plutôt aux fils de discussion où chacun des utilisateurs apportent leur remarque, sans forcément interagir avec celles des autres utilisateurs. Cette catégorie est souvent marquée par le fait que chaque utilisateur n'a posté qu'un message, ce qui explique d'ailleurs le manque d'interaction, car le dialogue ne se crée pas vraiment lorsqu'il n'y a pas de réponse.

PG [modifier le code]

Je rappelle que le PCF a décidé de soutenir les listes PS. Il n'y a donc aucune raison de parler de liste FDG ou FG. Thémistocle (discuter) 29 novembre 2013 à 22:14 (CET) Et j'ajoute que les précisions du genre "enseignante au chômage" ou "militante contre la précarité" n'ont aucun intérêt encyclopédique. Thémistocle (discuter) 29 novembre 2013 à 22:19 (CET)

Les listes se revendiquaient comme telles, en avait le logo, y compris sur le bulletin de vote. N'y figurait d'ailleurs aucune autre mention de parti. Un certain nombre de tête de listes, étaient, PCF (Paris 17e), ensemble (19e, et d'autres) ou non encarté, la dénomination PG, est donc tout aussi contestable.

> Certaines en avaient le logo à la suite du conflit qui a opposé Mélenchon à Laurent. Mais objectivement, il n'y a pas eu de listes FDG à Paris en 2014. Il y a eu les listes PG menées par Simonnet et c'est tout. Le PCF, mené par Brossat, ayant rejoint Hidalgo. Celette (discuter) 19 avril 2014 à 22:58 (CEST)

Exemple 15 : Discussion:Élections municipales de 2014 à Paris
PdD : 7187405, FdD : 3
Extrait du fil de discussion « pg » : type composite

Exemple 15 : Ici trois contributeurs (« Thémistocle », « Celette » et un anonyme) font des remarques et manifestent leur point de vue, mais n'interagissent pas vraiment avec les messages précédents : il n'y a pas de connections interactionnelles entre eux.

Ainsi, le cas *composite* est identifié par les seuils suivants :

- $2 < 01\_nbUser \leq 5$
- $09\_moyMsgUser = 1$

	Pages de discussion	Fils	Messages	Mots
Cas composite	15 514 9,7 %	20 279 5 %	66 759 5,9 %	4 961 320 4,9 %

Tableau 19 : Proportion de cas composite : entre 2 et 5 utilisateurs, 1 seul message par utilisateur en moyenne

b. Pluri surpeuplé : plus de 5 utilisateurs

	Pages de discussion	Fils	Messages	Mots
Profil pluri surpeuplé	5 173 3,2 %	8 895 2,2 %	168 586 14,9 %	16 360 171 16 %

Tableau 20 : Proportion de fils Pluri surpeuplé impliquant plus de 5 utilisateurs

Les 15 fils observés dans ce sous-ensemble ont révélé que certains fils pouvaient être le théâtre de vote (Exemple 16), ces derniers se réalisant normalement le plus souvent au sein des pages parallèles *(cf. II.2.1Niveau pages de discussion : sélection des pages principales*). Par ailleurs, le cas similaire des fils *composite* a également été observé. Dans ces deux situations, chacun apporte son avis de manière indépendante, sans vraiment prendre part à une discussion ni interagir avec les autres.

Roms [modifier le code]

1. ~~Il est vrai que je trouve très étrange ce nom de Rroms, très peu usité il me semble. Roms me semblerait plus juste, à la fois commun et encyclopédique.~~ ~~El Comandante~~Hasta ∞ ~~28 mai 2006 à 18:09 (CEST) Changement de vote suite à la lecture du commentaire de Joub ci dessus, et d'une vérification Google.~~ ~~El Comandante Hasta ∞ 8 septembre 2006 à 18:57 (CEST)~~ Nouveau changement d'avis suite aux nouvelles recherches que j'ai effectuées sur ce sujet (cf. Commentaire 1 ci-dessus). El Comandante Hasta ∞ 30 janvier 2008 à 19:15 (CET)
2. Semble être le terme le plus employé tout en étant exact. Felipeh | *hable aquí* 16 juin 2006 à 09:40 (CEST)
3. --Alaiche 30 juin 2006 à 00:58 (CEST)
4. On peut préciser (ou *Rroms*) dans le chapeau. L'usage doit primer. Popo le Chien ouah 24 juillet 2006 à 14:33 (CEST)
5. Dans la mesure ou c'est une encyclopédie Francophone et non Kalophone, je serait pour l'orthographe qui corespond à la prononciation selon l'alphabet français. Donc, ce serait sympa si quelqu'un pouvait metre fichier multimédia donnant la pronouciation "corecte" (on va dire parmi le R(r)oms francophone). En attendant, vote blanc --Madlozoz 25 juillet 2006 à 14:29 (CEST)
6. Il s'agit du terme qui est le plus général. Ce peuple se désigne lui-même par ce vocable. O2 27 juillet 2006 à 16:33 (CEST)
7. Roms me semble le plus commun. Des #redirect et un paragraphe sur les différentes dénominations éliminera les confusions. • Sherbrooke (✎) 8 août 2006 à 15:27 (CEST)
8. **Rom** est le terme le plus employé tout en étant exact. --Diligent 10 août 2006 à 18:39 (CEST)

Exemple 16 : Discussion:Rom
PdD : 508619, FdD : 5
Extrait du fil de discussion « Roms » faisant l'objet d'un vote

Exemple 16 : Ici les internautes votent pour désigner un terme : presque tous les messages sont postés par un contributeur différent. Chacun exprime son avis sur le sujet.

La table qui suit synthétise ces observations ainsi que certains traits descriptifs qui peuvent être révélateurs selon les trois cas observés :

- **Cas A** : Vote
- **Cas B** : Composite
- **Cas C** : Interaction

Références				A	B	C	09_moyMsgUser	06_moyMotsMsg	07_moyPhrasesMsg	01_nbUsers	02_nbMsg
Num PdD	Titre PdD	Num FdD	Titre FdD								
8104739	Discussion:Guerre de	41	Conférence du caire et réactions			X	3,38	54	3	6	23

	Gaza de 2014		du Hamas								
7298738	Discussion:Opposition au mariage homosexuel en France/1	69	Pape François et les lois			X	8,8	99.61	3.04	12	97
817049	**Discussion:Serbie-et-Monténégro**	**3**	**Fusion abandonnée entre [[Serbie-et-Monténégro]] et [[République fédérale de Yougoslavie]]**	**X**			**1**	**51,57**	**2,71**	**7**	**7**
2794380	**Discussion:Organisme génétiquement modifié/Archive 7**	**7**	**Risques environnementaux et sanitaires**		**X**		**1**	**77.5**	**3.5**	**6**	**6**
51982	**Discussion:Or**	**1**	**Discussion anciennes**		**X**		**1**	**45.7**	**3.2**	**10**	**10**
1722630	Discussion:Socialisme/Archive avril 2005-mai 2007	20	[[Club de l'horloge]]			X	2	87	4,4	11	22
8651951	**Discussion:Attentat contre Charlie Hebdo**	**73**	**Fusion avec l'article "Attentats de janvier 2015 en France"**	**X**			**1,,27**	**45**	**3,3**	**15**	**19**
508619	**Discussion:Roms**	**5**	**Roms**	**X**			**1,5**	**53,3**	**2,75**	**16**	**24**
7733069	**Discussion:Dieudonné/Archive9**	**18**	**Renommage**	**X**			**1,36**	**55,4**	**3,4**	**25**	**34**
13088	**Discussion:Tokyo**	**18**	**Renommer [[Tōkyō]] en [[Tokyo]]**	**X**		**X**	**3,76**	**75,8**	**4,2**	**42**	**158**
136740	Discussion:Augusto Pinochet	45	Une analyse			X	4,7	106,7	4,8	10	47
8651951	Discussion:Attentat contre Charlie Hebdo	47	Islamophobie			X	3,44	95,7	4,16	18	62
6724169	Discussion:Opération Pilier de défense	49	Évolutions après le cessez-le-feu			X	4,33	44	3,6	6	26
448900	**Discussion:Pie XII/Archives**	**67**	**Vote**	**X**			**1,7**	**30,39**	**1,72**	**10**	**17**
1722630	Discussion:Socialisme/Archive avril 2005-mai 2007	65	ilestinadmissibleque			X	1,7	164	6,7	10	17

Tableau 21 : Synthèse de l'observation d'un échantillon de pluri surpeuplé pour distinguer trois cas récurrents.

Le cas B *composite* est clairement délimité par le trait 09_*moyMsgUser* lorsqu'il est à 1, tandis que le cas A *vote* peut être repéré par le croisement du même trait avec le trait 06_*moyMotsMsg*. Ainsi, lorsqu'il y a peu de messages par utilisateurs et que ces messages sont globalement courts, alors le fil a plus de chance d'être un fil de discussion dédié à un vote.

Le cas *composite* dans le sous-ensemble Pluri impliquant plus de 5 utilisateurs est donc identifié par les seuils suivants :

- 01_*nbUser* > 5
- 09_*moyMsgUser* = 1

Pour le cas *vote*, il est caractérisé par les seuils suivants :

- 01_*nbUser* > 5
- 1 < 09_*moyMsgUser* < 2
- 06_*moyMotsMsg* < *60*

	Pages de discussion concernées	Fils	Messages	Mots
Cas Vote	1 241 0,8 %	1 365 0,3 %	14 792 1,3 %	655 858 0,6 %
Cas Composite	690 0,4 %	726 0,2 %	5 049 0,4 %	253 199 0,2 %
Cas Discussion	3 242	6 804	148 745	15 451 114

Tableau 22 : Répartition des cas Discussion et des cas Vote et Composite issus de l'observation du sous-ensemble Pluri Surpeuplé

L'observation du sous-ensemble Pluri a donc fait émerger d'autres profils de fils de discussion (Tableau 23). Ceux qui ne sont pas concernés par ces profils viennent compléter le corpus_v2 pour une analyse approfondie.

Profil	Caractéristiques	Proportion			
		PdD	Fils	Messages	Mots
10_*composite*	01_*nbUser* > 2 09_moyMsgUser = 1	16 204 10,1 %	21 005 5,2 %	71 808 6,3 %	5 214 519 5,1 %
11_*vote*	01_*nbUser* > 5 1 < 09_moyMsgUser < 2 06_moyMotsMsg < *60*	1 241 0,8 %	1 365 0,3 %	14 792 1,3 %	655 858 0,6 %

Tableau 23 : Synthèse des profils généraux identifiés dans Pluri

A l'image du cas 06_composite qui s'est retrouvé dans les deux sous-ensembles analysés (Pluri moyen et Pluri surpeuplé), il serait intéressant de vérifier si le profil 07_*vote* émerge dans le premier sous-ensemble dans lequel il n'avait pas été repéré en lui appliquant ses seuils.

III.3 Corpus_v2 : Profils retenus pour une observation approfondie

Cette première caractérisation a permis d'identifier des profils généraux de fils de discussion. L'objectif de ce projet étant de s'intéresser particulièrement aux situations impliquant une réelle collaboration entre les utilisateurs, cette première phase a été l'occasion de caractériser des profils que l'on ne souhaite pas, dans un premier temps, analyser de manière plus approfondie, car ils ont un potentiel d'interaction relativement pauvre, ou bien contiennent des séquences qui n'ont pas été produites par les utilisateurs, et donc risquent de fausser la valeurs de certains traits.
En partant du postulat que les interactions diffèrent selon le nombre d'interlocuteurs impliqués, cette étape a également permis de répartir les données restantes dans deux sous-ensembles qui différencient les interactions entre deux locuteurs exactement, et entre plus de deux locuteurs. Ces deux sous-corpus qui composent le corpus_v2 sont ainsi constitués des fils de discussion qui n'ont pas encore été caractérisés par

l'un des profils identifiés précédemment, et sont destinés à être observés de manière plus approfondie afin de faire émerger des nouveaux profils d'interactions, plus subtils et moins généraux.

- Sous-corpus *cDuo*
  - 2 utilisateurs exactement
  - Plus de 2 messages dans la totalité du fil
  - Si les messages sont longs (plus de 300 mots/message en moyenne) : seuls les fils ayant une moyenne inférieure ou égale à 39 mots/phrases sont conservés.
  - Si les messages sont courts (moins de 50 mots/message en moyenne) : seuls les fils ayant une moyenne supérieure à 20 mots/message sont conservés

- sous-corpus *cPluri* :
  - Plus de 2 utilisateurs
  - Les fils ayant en moyenne plus de 1 message par utilisateur sont conservés
  - Les fils ayant une moyenne de messages par utilisateur comprise entre 1 et 2 ne sont conservés que si leur moyenne de mots/message est inférieure à 60.

	Pages de discussion concernées	Fils	Messages	Mots
cDuo	24 027 15 %	24 368 6 %	149 060 13,15 %	13 583 323 13,35 %
cPluri	24 164 15,1 %	53 168 13,2 %	443 148 39 %	41 944 445 41,2 %

Tableau 24 : Proportion des données de cDuo et cPluri

IV. Faire émerger des profils : exploration approfondie du corpus_v2

Après avoir affiné le corpus WikiDisc en excluant les fils de discussion jugés non pertinents (*cf.* II.2 Première manipulation : pages de discussions parallèles et fils de discussion sans contenu interactif, un filtrage pour écarter les données non pertinentes) les données restantes de la première version du corpus ont été analysées selon des traits descriptifs (Tableau 8) relevant de certaines caractéristiques quantitatives des FdD, afin de faire émerger des profils d'interaction (Tableau 23). Les données n'ayant pas encore été identifiées, *i.e.* le corpus_v2, ont été réparties dans deux sous-corpus, *cDuo* et *cPluri*, afin d'en faciliter l'analyse.

La prochaine étape consiste ainsi à explorer de manière plus poussée les données n'ayant pas encore été catégorisées, selon les traits déjà employés au cours des analyses précédentes (Tableau 3 & Tableau 8), mais aussi selon de nouveaux traits qui caractérisent le contenu linguistique ainsi que la dimension interactive et temporelle des fils (Tableau 25), comme par exemple le taux de certains pronoms personnels, la proportion de phrases exclamatives, ou encore la symétrie de la répartition des tours de paroles. Ces traits permettent d'isoler certains échantillons selon des seuils établis. L'analyse qualitative de ces échantillons est guidée par une grille d'annotation inspirée des différentes situations qui ont été annotées manuellement lors des étapes précédentes. .

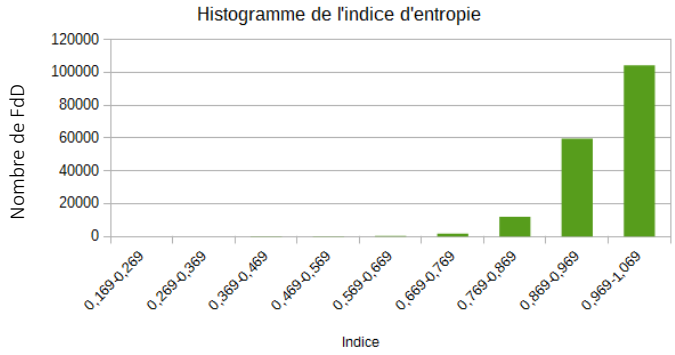
IV.1 Explorer le contenu des fils de discussion : des traits plus précis et une grille d'annotation

IV.1.1 Présentation des traits employés pour explorer le contenu des fils de discussion

Pour l'étape d'analyse approfondie des données, de nouveaux traits ont été introduits afin de visualiser certaines dimensions linguistiques et interactives des fils de discussion. À l'instar des précédents, ces traits sont tous issus d'une analyse outillées des données, et ont été calculés sur le corpus_v2, qui correspond à

- 10_*entropie*
  - Technique
    Elle est calculée à partir du nombre total de messages du fils par rapport au nombre de messages postés par chaque contributeur, puis normalisée afin d'obtenir un indice plus facilement comparable.
  - Description et Remarques
    Cet indice, compris entre 0 et 1, permet d'observer la répartition des messages par auteur. Plus

l'indice se rapproche de 1, plus la répartition est symétrique, c'est à dire que les contributeurs impliqués dans la discussion ont posté un nombre égal de message. Plus l'indice se rapproche de 0, plus la répartition est inégale et il y a de chances qu'un contributeur soit beaucoup plus actif que les autres.

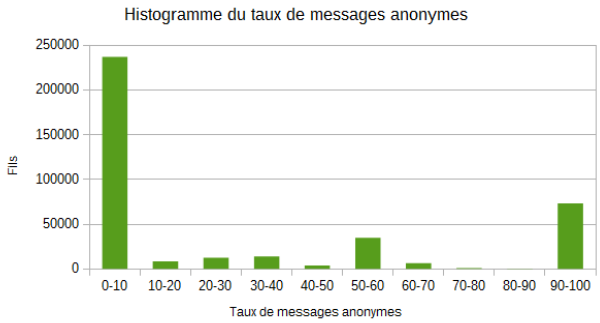


L'indice d'entropie concerne les fils comprenant au moins 2 utilisateurs, soit 178 676 fdd.

Figure 7 : Histogramme de l'indice d'entropie

L'histogramme montre que les tours de parole sont majoritairement bien répartis (pour les fils ayant un indice de plus de 0.97). On peut considérer qu'en dessous de ce seuil les tours de parole commencent commencent à être inégaux.

- 11_*tauxAnonyme*
  - Technique
    Ce trait est calculé d'après le nombre de messages postés par un utilisateur « anonyme » disponible dans la table de fréquence des messages postés par les utilisateurs du fil, par rapport au nombre total de messages (*cf.* Tableau 8).
  - Description et Remarques
    Comme il n'est pas possible de récupérer le nombre de contributeurs anonymes, cet indice peut être significatif dans la prise en compte du critère d'anonymat dans le fil de discussion. Cette indice permet de renseigner sur le taux de messages anonymes postés dans l'ensemble du fil de discussion. Les messages anonymes pouvant fausser certains calculs d'indices (notamment ceux impliquant le nombre d'utilisateurs), ce paramètre peut être intéressant pour établir un seuil pour écarter les fils ayant trop de messages anonymes. Il peut également être utilisé pour considérer le fort taux de messages anonymes comme un trait distinctif pour certains types de discussions.



Le taux d'anonymat a été calculé sur l'ensemble des fdd des pdd principales et archivées, soit 402 263 fdd.

Figure 8 : Histogramme de taux de messages anonymes

On peut observer le caractère extrême du taux de messages anonymes : il y a un très grand nombre de fils contenant très peu de messages anonymes (de 0 % à 10 %), mais de nombreux fils contiennent un taux élevé de messages anonymes (plus de 90%).

- 12_*phExcl*, 13_*phExclExpr*, 14_*phInterro*, 15_*phInterroExpr*
  - Technique
    Ces taux sont calculés en comptabilisant les différentes ponctuations de fin de phrases, repérées à partir de motifs, dans la version étiquetée par Talismane du corpus (*cf. II.1.3 Version étiquetée avec Talismane*) :
    - Les phrases exclamatives sont identifiées par le motif *!*
    - Les phrases exclamatives expressives sont identifiées par le motif *!/2 ou +*
    - Les phrases interrogatives sont identifiées par le motif *?*
    - Les phrases interrogatives expressives sont identifiées par le motif *?/2 ou +*
  - Description et Remarques
    La présence de plusieurs ponctuations à la suite marque le désir du locuteur d'appuyer ce qu'il exprime dans son message, nous avons donc décidé de faire une distinction entre une ponctuation simple et une séquence de plusieurs ponctuations en considérant une version « expressive » pour chacune des deux types de phrases (interrogatives et exclamatives).
    A noter que, même si la fonction de base d'une phrase exclamative est de mettre en relief le contenu de la phrase, l'intensité est beaucoup plus forte lorsque plusieurs points d'exclamations se suivent, et donc nous avons conservé une version « normale » et une version « expressive » de ce type de phrase.

- 16_*jacPrem*, 17_*jacPrec*, 18_*jacFil*
  - Technique
    Un indice Jaccard est calculé entre chaque paire d'objet (message/message ou message/totalité du fil) du fil de discussion, et tous les indices du fil sont conservés dans une table. Les traits correspondent à la moyenne de l'ensemble des indices Jaccard pour chaque paire d'objet.

L'indice Jaccard permet de calculer la similarité et la diversité entre deux objets et se calcule selon la formule suivante :

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

- Description et Remarques
  Ces indices indiquent le recouvrement lexical moyen entre deux messages ou un message et l'intégralité du fil. Ils permettent ainsi de visualiser l'évolution du lexique du fil de discussion et d'observer si le vocabulaire varie beaucoup ou non.
  L'interprétation des indices est néanmoins délicate, car ces derniers sont très sensibles aux variations de taille du message. Ainsi, logiquement, plus le message est long, plus son indice a de chances d'être élevé. Ils ne sont donc pas infaillibles, mais permettent toutefois d'avoir un premier aperçu de la cohésion et l'évolution lexicale d'un fil de discussion.

- 19*_pronoms1PS*, 20*_pronoms2PS*, 21*_pronoms1PP*, 22*_pronoms2PP*
  - Technique
    Ces traits sont récupérés en comptabilisant le nombre de pronoms ciblés par rapport au nombre total de pronoms du fil. Les pronoms sont identifiés à partir de la version étiquetée du corpus.
  - Description et Remarques
    L'observation des personnes employées dans la conjugaison des verbes peut être révélatrice du ton de l'interaction. Ainsi une prédominance de deuxième personne du singulier indique que le tutoiement est établi dans l'interaction, ce qui peut faire supposer un ton plus familier, ou bien un ton trop familier. Un prédominance de première personne du singulier peut indiquer le caractère individualiste de la discussion, alors qu'une forte présence de première personne du pluriel peut signifier une vigoureuse cohésion entre les interlocuteurs.

- 23*_marqueInterro*
  - Technique
    Ce trait est établi en comptabilisant les étiquettes indiquant les formes interrogatives (adverbe interrogatif, déterminant interrogatif, pronom interrogatif) sur la version étiquetée du corpus.
  - Description et Remarques
    L'interrogation, lorsqu'elle n'est pas rhétorique, est un marqueur fort de l'interaction entre des contributeurs qui se sondent pour obtenir une réponse, un avis, un conseil, etc. Or cet acte de langage, qui est souvent marqué par la tournure et la ponctuation interrogative, peut également être exprimé à travers des mots interrogatifs, tels que *comment, quelle, quand, pourquoi,* etc. Ainsi, leur prise en compte est pertinente afin de repérer de manière plus complète les actes d'interrogation.

- 24*_vbImperatif*
  - Technique
    Le nombre de verbes conjugués à l'impératif est récupéré en comptabilisant le nombre d'étiquettes indiquant les verbes à l'impératif dans la version étiquetée du corpus.

- Description et Remarques
  L'emploi de l'impératif exprime une injonction qui peut réaliser une requête, un conseil, une invitation, ou encore un ordre. Ainsi, ce mode est la marque claire que le contributeur qui l'emploie s'adresse de manière directe à d'autres contributeurs, ce qui induit un potentiel d'interaction à observer et peut révéler des rapports particuliers entre les contributeurs.

- 25_*tempMoy*, 26_*tempMed*, 27_*tempPremDer*
  - Technique
    Ces traits sont calculés à partir de l'information temporelle précisée dans l'attribut @*when* de l'élément *post* (*cf.* II.1.3 Structure des données de WikiDisc), qui permet d'établir une étendue temporelle entre chaque paire contiguë de message du fil. La moyenne et la médiane sont calculées à partir de l'ensemble des étendues temporelles récupérées et préalablement converties en heures.

  - Description et Remarques
    Le temps de réponse global entre chaque message peut être représentatif de l'implication des contributeurs, ou bien de l'instantanéité des messages postés, qui peuvent avoir demandé un temps de rédaction plus ou moins long. L'étendue temporelle de l'ensemble de la discussion permet d'observer sur quelle durée elle a mobilisé les contributeurs. Une étendue temporelle totale très courte peut être le signe d'une discussion avortée, ou bien d'une discussion concise et efficace. Le cas d'une étendue très longue peut marquer le fait que la tâche discutée est laborieuse, ou encore, ce qui est le fréquemment le cas lorsque l'étendue est extrêmement longue, que la discussion est relancée par un auteur plusieurs semaines, mois voire années après la création du fil.
    Comme évoquée précédemment (*cf. II.1.3 Structure des données de WikiDisc*), la structure imbriquée n'est pas forcément respectée à la lettre par les contributeurs, ce qui peut amener à des cas où les messages qui se suivent dans la structure ne sont pas les messages qui se suivent réellement dans le temps. Ce type de situation entraîne des erreurs lors du calcul des étendues temporelles, car certaines paires de messages peuvent être composées d'un message plus récent en tant que premier message, ce qui aboutit à des étendues temporelles négatives. Ces étendues temporelles déviantes ne sont pas prises en compte.

Num	Nom	Intitulé	Description	Minimum	Maximum	Type
10	10_*entropie*	Taux d'entropie normalisée du fil de discussion	Permet d'observer si la répartition des messages postés par les utilisateurs du fils est équilibrées ou non. Plus il est proche de 1, plus la répartition est symétrique..	0,169	1	indice
11	11_*tauxAnonyme*	Taux de messages anonymes du fil	Permet d'observer la proportion de messages postés par des utilisateurs non identifiés dans le fil de discussion.	0	100	%
12	12_*phExcl*	Taux de phrases exclamatives	Représente la proportion de phrase se terminant par un point d'exclamation.	0	100	%
13	13_*phExclExpr*	Taux de phrases exclamatives expressives	Représente la proportion de phrase se terminant par plusieurs point d'exclamation.	0	100	%
14	14_*phInterro*	Taux de phrases interrogatives du fil	Représente la proportion de phrase se terminant par un point d'interrogation.	0	100	%

15	15_*phInterroExpr*	Taux de phrases interrogatives expressives	Représente la proportion de phrase se terminant par plusieurs points d'interrogation.	0	75	%
16	16_*jacPrem*	Moyenne des indices Jaccard par rapport au premier message	Moyenne des indices Jaccard calculés entre chaque message et le premier message du fil. Permet d'observer le recouvrement lexical moyen entre le premier message et chaque message du fil.	0	0,62	indice
17	17_*jacPrec*	Moyenne des indices Jaccard par rapport au message précédent	Moyenne des indices Jaccard calculés entre chaque message et celui qui le précède . Permet d'observer le recouvrement lexical moyen entre chaque paire contiguë de messages.	0	0,62	indice
18	18_*jacFil*	Moyenne des indices Jaccard par rapport à l'ensemble du FdD	Moyenne des indices Jaccard calculés entre chaque message et l'ensemble du fil de discussion. Permet d'observer le recouvrement lexical moyen entre la totalité du fil de discussion et chaque message du fil.	0,01	0,81	indice
19	19_*pronoms1PS*	Taux de pronoms de la première personne du singulier	Permet d'observer la proportion de pronoms de la première personne du singulier par rapport à tous les pronoms du fil de discussion.	0	100	%
20	20_*pronoms2PS*	Taux de pronoms de la deuxième personne du singulier	Permet d'observer la proportion de pronoms de la deuxième personne du singulier par rapport à tous les pronoms du fil de discussion.	0	78	%
21	21_*pronoms1PP*	Taux de pronoms de la première personne du pluriel	Permet d'observer la proportion de pronoms de la première personne du pluriel par rapport à tous les pronoms du fil de discussion.	0	75	%
22	22_*pronoms2PP*	Taux de pronoms de la deuxième personne du pluriel	Permet d'observer la proportion de pronoms de la deuxième personne du pluriel par rapport à tous les pronoms du fil de discussion.	0	100	%
23	23_*marqueInterro*	Marqueurs de l'interrogatif	Permet d'observer le nombre de marques interrogatives qui ne relèvent pas de la ponctuation.	0	74	fréquence
24	24_*vbImperatif*	Verbes à l'impératif	Permet d'observer le nombre de verbes à l'impératif employés dans le fil de discussion.	0	23	fréquence
25	25_*tempMoy*	Étendue temporelle moyenne entre chaque message	Permet d'observer le temps de réponse moyen en heures du fil de discussion.	-91576	95085	Durée (h)
26	26_*tempMed*	Étendue temporelle médiane entre chaque message	Permet d'observer le temps de réponse médian en heures du fil de discussion.	-91575,68	95085	Durée (h)
27	27_*tempPremDer*	Étendue temporelle entre le premier et le dernier message	Permet de connaître la durée totale en heures de la discussion.	-91576	97584	Durée (h)

Tableau 25 : Synthèse des traits utilisés pour analyser le contenu linguistique et la dimension interactive des fils de discussion

IV.1.2 Une grille d'annotation pour accompagner l'observation

La première phase de tri présentée dans la section précédente a également été l'occasion d'observer manuellement un échantillon conséquent de fils de discussion, permettant ainsi, en plus de déterminer la pertinence de certains types de FdD, d'identifier des caractéristiques récurrentes. Ces caractéristiques ont ainsi servi à élaborer une grille d'annotation pour guider l'observation manuelle de la seconde phase du projet de recherche : celle de l'identification de profils types d'interaction. Cette grille d'annotation a pour but d'accompagner l'analyse des échantillons de fils de discussion sélectionnés à partir de certains seuils des traits caractéristiques présentés précédemment, afin d'être examinés manuellement, mais elle n'est pas considérée comme un schéma strict d'annotation à respecter de manière rigoureuse.

Caractéristiques	Description	Paramètres
Comportement des utilisateurs		
Interaction	Les messages échangés sont adressés aux autres utilisateurs.	**Ciblée** : Les messages visent des utilisateurs ciblés. **Universelle** : Les messages s'adressent à l'ensemble des utilisateurs qui participent au fil de discussion, ou plus largement à toute la communauté WP susceptible de répondre.
Coopération	Les utilisateurs communiquent, échangent et prennent en compte les remarques et idées des autres utilisateurs.	**Écoute** : Le point de vue des autres est pris en compte. **Conseil** : Une suggestion ou un recommandation est faite à un utilisateur. **Question** : Interroger les contributeurs pour demander leur point de vue, leur aide, une précision, etc.
Résistance	Les contributeurs rejettent en bloc les interventions des autres utilisateurs, aucune interaction n'est possible.	
Confrontation	Les participants ont des points de vue différents : proposition d'idées, réaction aux idées des autres.	**Discussion** : Chacun expose son point de vue, et discute sans opposition. **Débat** : Chacun défend son point de vue de manière plus appuyée. **Conflit** : Chacun défend de manière affirmée son point de vue et attaque le point de vue des autres.
Modification	Les utilisateurs évoquent une modification concrète.	**Proposition** : Propose d'effectuer une modification. **Annonce** : Évoque une modification qui a été effectuée.
Confirmation	Un message confirme que la ou les modifications et changements discutés ont été effectués.	
Consensus	Les participants ont trouvé un accord suite à leurs échanges.	
Sujet de l'échange		
Contenu de l'article	La discussion tourne autour de points concernant l'article lié à la page de discussion.	Fond : Le contenu rédactionnel de l'article es discuté . Forme : La structure de l'article est discutée.
Comportement	La discussion tourne autour du comportement d'un ou de plusieurs utilisateurs (modifications effectuées sans l'accord des autres, ton employé, etc).	

Tableau 26 : Grille d'annotation pour accompagner la phase d'exploration manuelle

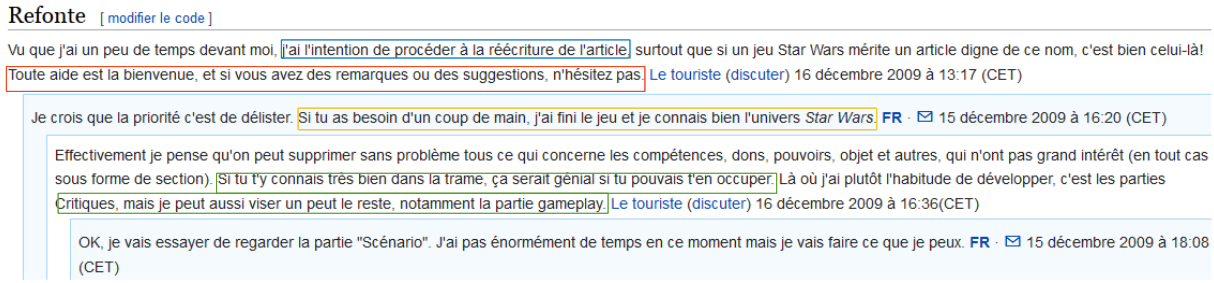
IV.2 Amorce de l'identification de profils plus précis

L'objectif de cette étape est d'identifier des profils d'interaction plus précis et subtils que les profils généraux dégagés auparavant à partir d'un ensemble plus complet de trait et d'une grille d'annotation pour guider l'analyse qualitative d'échantillons. Cette section revient ainsi sur l'application de cette méthodologie sur des données sélectionnées et réparties dans les sous-ensembles *cDuo* et *cPluri*, car n'ayant pas encore été catégorisées.

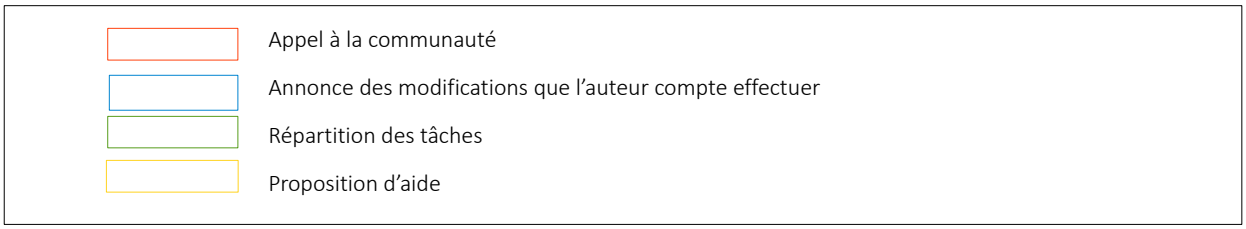
Un échantillon de 40 fils de discussion a été constitué de manière aléatoire à partir du sous-ensemble cDuo pour la phase d'annotation manuelle. Tous les fils n'ont pas été catégorisés, mais cette analyse a mis en relief deux types d'interactions entre les utilisateurs.

- Collaboration harmonieuse

Ce profil concerne les fils de discussion dans lesquels les utilisateurs sont réellement impliqués et coopérèrent, avec pour objectif de modifier et améliorer l'article dans son ensemble ou certains points précis. Les échanges sont centrés sur la tâche à effectuer, et le bon déroulement du processus de rédaction collaborative passe par des interrogations, des conseils, et la prise en compte des remarques entre les contributeurs. Dans ce type d'interaction, les utilisateurs se répartissent fréquemment les tâches, ils proposent des modifications concrètes à effectuer, et atteignent un consensus. Dans l'exemple 17, le premier contributeur s'adresse à la communauté et invite de potentiels contributeurs à se joindre à lui, puis les deux contributeurs impliqués se répartissent bien les tâches et font des des propositions concrètes.



Exemple 17: Discussion:Star Wars: Knights of the Old Republic
PdD : 870164, FdD : 3
Extrait du fil « Refonte » : type collaboration harmonieuse



L'analyse des traits des 5 fils de discussions correspondant à ces observations a révélé certains seuils récurrents. En ce qui concerne les valeurs temporelles, les valeurs négatives ont été ignorées dans l'observation car elles sont déviantes (*cf. IV.1.1 Présentation des traits employés pour explorer le contenu des fils de discussion*) et ne peuvent pas être interprétées correctement.

Le premier paramètre qui apparaît, c'est que tous les contributeurs sont identifiés *i.e.* le taux d'anonymat est nul. Par ailleurs on peut observer que le taux d'emploi de la première personne du singulier est plutôt élevée, avec une moyenne de 41,8 % et une médiane de 46,4 %, et que l'emploi du vouvoiement est assez récurrent (dans 4 cas sur 5). L'entropie est globalement élevée, même si la répartition des messages par utilisateurs n'est pas totalement symétrique pour tous les fils, ce qui signifie que le nombre de messages postés par les utilisateurs est plutôt équivalent, et qu'il n'y a pas un auteur beaucoup plus actif qu'un autre. Par ailleurs, l'emploi du mode interrogatif est lui aussi plutôt récurrent, dans 4 cas sur 5, mais n'est pas très élevé (avec un maximum de 13,04 % de phrases interrogatives). En revanche il y a peu de phrases exclamatives et pas du tout de phrases dites expressives, ce type d'interaction ne semble donc pas prêter à des effusions d'humeurs. Enfin, on remarque que le temps de réponse médian ne dépasse jamais 24h, ce qui peut signifier que les contributeurs impliqués dans ce type de discussion sont plutôt réactifs.

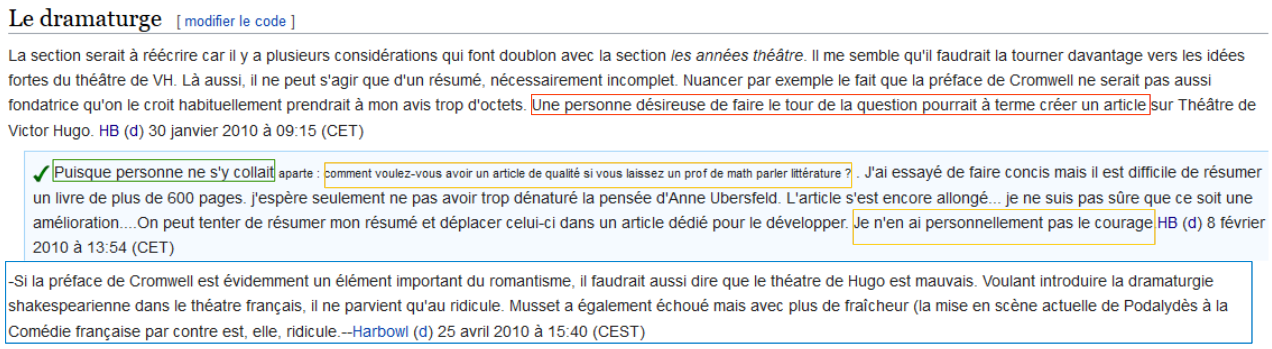
Références				02_ nb Msg	11_ taux Ano nym e	10_ entr opie	19_ pro no m1 PS	20_ pro no m2 PS	21_ pro no ms1 PP	22_ pro no ms2 PP	13_ phI nter ro	12_ phE xcl	15_ phI nter roEx pr	14_ phE xclE px	25_ tem pM ed	23_ mar quel nter ro
Num PdD	Titre PdD	Num FdD	Titre FdD													
4873758	Discussion:Instinctothérapie/Archives/04	2	Réactions sociétales	4	0	1	46,4	0	0	35,71	3,57	0	0	0	18,25	0
250067	Discussion:Démocratie athénienne	10	Recyclage de la page nécessaire?	6	0	1	30,16	0	11,9	7,94	5,77	1,92	0	0	5,57	8
5913081	Discussion:Témoins de Jéhovah/archive7	15	Effectifs	5	0	0,971	56,52	4,35	0	0	8	4	0	0	4,62	0
870164	Discussion:Star Wars: Knights of the Old Republic	3	Refonte	4	0	1	57,14	19,05	0	4,76	0	7,14	0	0	(- 20,95)	0
7344047	Discussion:Pierre Sellier (Salamandre)/Archives	19	Plaintes	3	0	0,918	18,75	0	6,25	6,25	13,04	0	0	0	1,35	0
Moyenne				4,4	0	0,98	41,8	4,68	3,63	10,9	6,1	2,612	0	0	7,448	1.6
Médiane				4	0	1	46,4	0	0	6,25	5,77	1,92	0	0	5,095	0
Écart type				1,14	0	0,036	16,9	8,25	5,36	14,16	4,9	3,022	0	0	7,425	3.57

Tableau 27 : Analyse des 5 FdD type collaboration harmonieuse

- Entreprise solitaire forcée

Ce profil a été observé dans 4 FdD, dans lesquels un auteur lance un projet de modification de l'article et un appel implicite (Exemple 18) ou explicite (Exemple 19) à la contribution d'autres auteurs, mais personne ne prend vraiment part au projet. Ainsi l'intervention des autres contributeurs peut juste être de l'ordre de l'approbation de la proposition de modification (Exemple 19), ou bien une remarque sur un point, mais ils ne répondent pas aux questions posées ou ne proposent pas de s'impliquer réellement. Parfois

l'auteur « solitaire » peut pointer le fait que personne ne s'est impliqué à ses côtés et qu'il va devoir travailler seul (Exemple 18), et peut également montrer des signes de déception et de découragement (Exemple 18).



Le dramaturge [modifier le code]

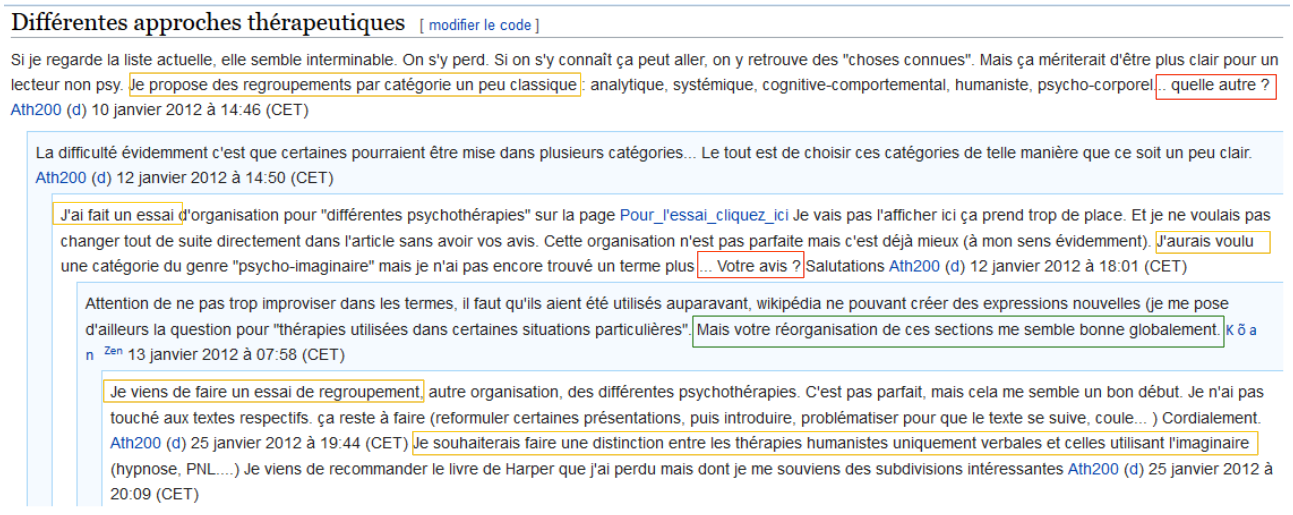
La section serait à réécrire car il y a plusieurs considérations qui font doublon avec la section *les années théâtre*. Il me semble qu'il faudrait la tourner davantage vers les idées fortes du théâtre de VH. Là aussi, il ne peut s'agir que d'un résumé, nécessairement incomplet. Nuancer par exemple le fait que la préface de Cromwell ne serait pas aussi fondatrice qu'on le croit habituellement prendrait à mon avis trop d'octets. Une personne désireuse de faire le tour de la question pourrait à terme créer un article sur Théâtre de Victor Hugo. HB (d) 30 janvier 2010 à 09:15 (CET)

✓ Puisque personne ne s'y collait aparte : comment voulez-vous avoir un article de qualité si vous laissez un prof de math parler littérature ? . J'ai essayé de faire concis mais il est difficile de résumer un livre de plus de 600 pages. j'espère seulement ne pas avoir trop dénaturé la pensée d'Anne Ubersfeld. L'article s'est encore allongé... je ne suis pas sûre que ce soit une amélioration....On peut tenter de résumer mon résumé et déplacer celui-ci dans un article dédié pour le développer. Je n'en ai personnellement pas le courage HB (d) 8 février 2010 à 13:54 (CET)

-Si la préface de Cromwell est évidemment un élément important du romantisme, il faudrait aussi dire que le théatre de Hugo est mauvais. Voulant introduire la dramaturgie shakespearienne dans le théatre français, il ne parvient qu'au ridicule. Musset a également échoué mais avec plus de fraîcheur (la mise en scène actuelle de Podalydès à la Comédie française par contre est, elle, ridicule.--Harbowl (d) 25 avril 2010 à 15:40 (CEST)

Exemple 18 : Discussion:Victor Hugo
PdD : 1871518, FdD : 26
Extrait du fil de discussion « Le dramaturge » : type entreprise solitaire forcée

	Remarque sur le fait que personne ne s'implique
	Signes de découragement
	Demande implicite de participation
	Intervention d'un autre auteur sous forme de remarque, mais pas de proposition de participation



Différentes approches thérapeutiques [modifier le code]

Si je regarde la liste actuelle, elle semble interminable. On s'y perd. Si on s'y connaît ça peut aller, on y retrouve des "choses connues". Mais ça mériterait d'être plus clair pour un lecteur non psy. Je propose des regroupements par catégorie un peu classique : analytique, systémique, cognitive-comportemental, humaniste, psycho-corporel... quelle autre ? Ath200 (d) 10 janvier 2012 à 14:46 (CET)

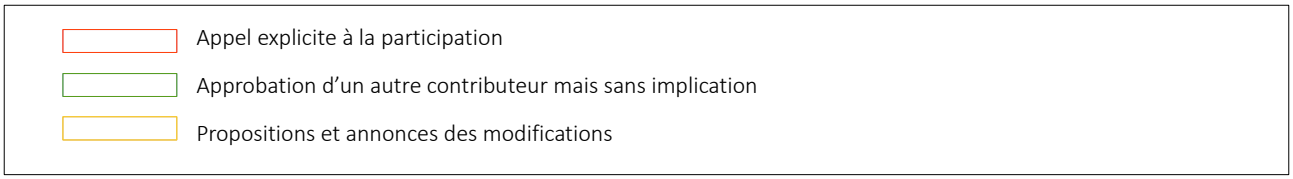
La difficulté évidemment c'est que certaines pourraient être mise dans plusieurs catégories... Le tout est de choisir ces catégories de telle manière que ce soit un peu clair. Ath200 (d) 12 janvier 2012 à 14:50 (CET)

J'ai fait un essai d'organisation pour "différentes psychothérapies" sur la page Pour_l'essai_cliquez_ici Je vais pas l'afficher ici ça prend trop de place. Et je ne voulais pas changer tout de suite directement dans l'article sans avoir vos avis. Cette organisation n'est pas parfaite mais c'est déjà mieux (à mon sens évidemment). J'aurais voulu une catégorie du genre "psycho-imaginaire" mais je n'ai pas encore trouvé un terme plus... Votre avis ? Salutations Ath200 (d) 12 janvier 2012 à 18:01 (CET)

Attention de ne pas trop improviser dans les termes, il faut qu'ils aient été utilisés auparavant, wikipédia ne pouvant créer des expressions nouvelles (je me pose d'ailleurs la question pour "thérapies utilisées dans certaines situations particulières". Mais votre réorganisation de ces sections me semble bonne globalement. K õ a n Zen 13 janvier 2012 à 07:58 (CET)

Je viens de faire un essai de regroupement autre organisation, des différentes psychothérapies. C'est pas parfait, mais cela me semble un bon début. Je n'ai pas touché aux textes respectifs. ça reste à faire (reformuler certaines présentations, puis introduire, problématiser pour que le texte se suive, coule...) Cordialement. Ath200 (d) 25 janvier 2012 à 19:44 (CET) Je souhaiterais faire une distinction entre les thérapies humanistes uniquement verbales et celles utilisant l'imaginaire (hypnose, PNL....) Je viens de recommander le livre de Harper que j'ai perdu mais dont je me souviens des subdivisions intéressantes Ath200 (d) 25 janvier 2012 à 20:09 (CET)

Exemple 19 : Discussion:Psychothérapie
PdD : 283633, FdD : 16
Extrait du fil de discussion « Différentes approches thérapeutiques » : type entreprise solitaire forcée



L'analyse des traits de ces 4 fils de discussion met en évidence que le taux de message anonymes est, dans 3 cas sur 4, nul, et que dans le 4ème cas, l'auteur anonyme n'est pas celui qui est le plus actif. Ensuite l'indice d'entropie est plutôt bas, situé en 0,722 et 0,918, ce qui montre bien que la répartition des messages n'est pas symétrique est qu'un auteur est plus impliqué dans la discussion. Par ailleurs la durée totale (étendue temporelle entre le premier et le dernier message) du fil est globalement élevée, largement supérieure à 24h, et varie entre 95h et 2046h. En ce qui concerne les pronoms employés, la première personne du singulier prédomine, avec un taux stable, 40 % en moyenne et 40 % en médiane, et l'emploi de la deuxième personne du pluriel n'est pas systématique, ce qui peut illustrer le fait que certains auteurs font un appel implicite à la contribution, mais également le manque d'échanges inter-contributeurs dû au profil solitaire de la discussion. Par ailleurs, le taux de phrases interrogatives est tangible quoique plutôt bas dans 3 cas sur 4, variant de 6,25 à 17,65 %, mais le 4ème cas pour lequel ce taux est nul présente tout de même un mode interrogatif à travers la présence de deux mots interrogatifs.

Références				02_nbMsg	11_tauxAnonyme	10_entropie	19_pronom1PS	20_pronom2PS	21_pronoms1PP	22_pronoms2PP	13_phInterro	12_phExcl	15_phInterroExpr	14_phExclEpx	25_tempPremDer	23_marqueInterro
Num PdD	Titre PdD	Num FdD	Titre FdD													
2270498	Discussion:Bernard Dubourg/archive2	130	Vaneigem 3	4	25	0,811	40	0	0	10	17,65	0	0	0	95	1
283633	Discussion:Psychothérapie	16	Différentes approches thérapeutiques	5	0	0,722	52,38	0	0	0	8	0	0	0	365	1
1871518	Discussion:Victor Hugo	26	Le dramaturge	3	0	0,918	30,77	0	0	7,69	6,25	0	0	0	2046	1
5016025	Discussion:Mouvement social contre la réforme des retraites en France de 2010	20	Une phrase dont on peut faire l'économie dans la partie "contexte"	3	0	0,918	40	0	0	0	0	0	0	0	192	2
Moyenne				3,75	6,25	0,84	40,79	0	0	4,423	7,975	0	0	0	674,5	1
Médiane				3,5	0	0,86	40	0	0	3,845	7,125	0	0	0	278,5	1
Écart type				0,96	12,5	0,095	8,89	0	0	5,19	7,30	0	0	0	921,12	0,8

Tableau 28 : Analyse des 4 FdD type entreprise solitaire forcée

Les indices de recouvrement lexical sont en revanche plus difficilement interprétables, car hormis le fait que le recouvrement lexical par rapport au fil (courbe jaune, Figures 9,12,10,11) est constamment plus élevé, et ce dans tous les cas observés, peu de constantes apparaissent. Cela peut être notamment dû au fait que les fils de discussion observés sont plutôt courts en termes de taille de messages.

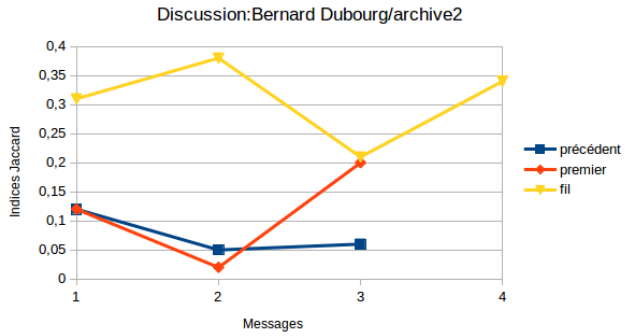


Figure 9 : Recouvrement lexical du fil « Vaneigem 3 »

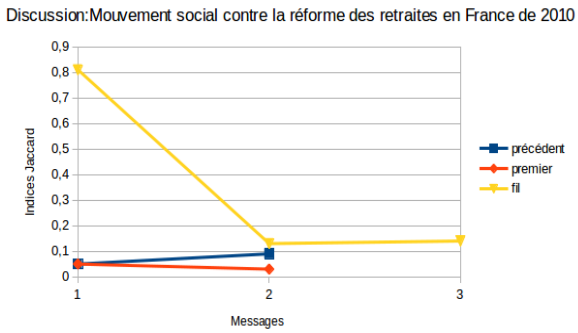


Figure 10 : Recouvrement lexical du fil « Une phrase dont on peut faire l'économie dans la partie "contexte" »

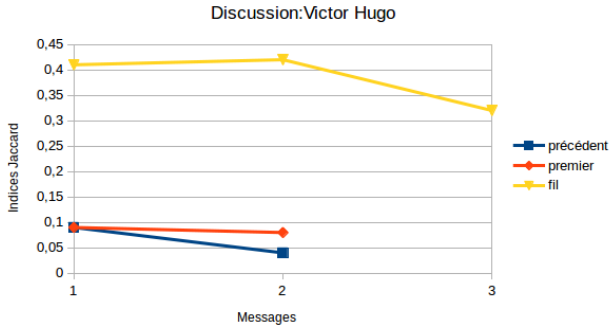


Figure 11 : Recouvrement lexical du fil « Le dramaturge »

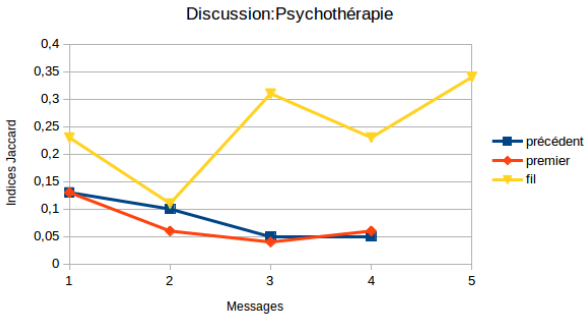


Figure 12 : Recouvrement lexical du fil « Différentes approches thérapeutiques »

V.Synthèse des profils identifiés

V.1 Profils généraux

Ces profils sont issus d'une première observation basée sur des traits statistiques (Tableau 3 & Tableau 8), et ont été considérés comme assez pauvres en interaction ou contenant des séquences qui n'ont pas été produites par les contributeurs dans le cadre de l'échange . Nous avons cependant appliqué certains traits plus fins qui caractérisent le contenu linguistique et interactif des fils de discussion (Tableau 25) afin d'observer si ils pouvaient déterminer ces profils déjà identifiés.

Profil	Caractéristiques	Proportion			
		PdD	FdD	Messages	Mots
01_*vide*	02_*nbMsg* = 0	5 138 3 %	7 874 2 %	0 0 %	0 0 %
02_*mono message*	02_*nbMsg* = 1	115 365 72 %	202 732 50 %	202 732 18 %	17 420 137 17 %
03_*monologue*	01_*nbUser* = 1 03_*actif* = un auteur identifié	94 076 59 %	141 936 35 %	152 835 13 %	11 888 161 12 %
04_ *monologue_anonyme*	01_*nbUser* = 1 03_*actif* = anonyme e 02_*nbMsg* ≤ 2	2 924 2 %	3 346 0,8 %	6 692 0,6 %	649 878 0,6 %
05_ *échange succinct*	*01_nbUser* = 2 *02_nbMsg* = 2 *06_moyMotsMsg* < 50	26 807 16,8 %	32 350 8 %	64 700 5,7 %	2 002 498 2 %
06_*échange riche*	*01_nbUser* = 2 *02_nbMsg* = 2 *50 < 06_moyMotsMsg* < 1 000	23 902 14,95 %	33 751 8,4 %	67 502 5,95 %	7 505 222 7,4 %
07_ *non discursif*	*01_nbUser* = 2 *02_nbMsg* = 2 *06_moyMotsMsg* > 1 000	83 0,05 %	83 0,02 %	166 0,01 %	322 061 0,3 %
08_ *contenu externe*	01_*nbUser* = 2 02_*nbMsg* > 2 06_moyMotsMsg > 300 08_moyMotsPhrase > 39	616 0,4 %	680 0,2 %	2 855 0 ,25 %	1 248 328 1,2 %
09_ *interaction succincte*	01_*nbUser* = 2 02_*nbMsg* > 2 06_moyMotsMsg < 50	10 548 6,6 %	12 237 3 %	46 110 4 %	1 641 918 1,6 %
10_*composite*	01_*nbUser* > 2 09_moyMsgUser = 1	16 204 10,1 %	21 005 5,2 %	71 808 6,3 %	5 214 519 5,1 %
11_*vote*	01_*nbUser* > 5	1 241	1 365	14 792	655 858

	1 < 09_moyMsgUser < 2 06_moyMotsMsg < *60*	0,8 %	0,3 %	1,3 %	0,6 %

Tableau 29 : Synthèse des profils généraux

Ainsi, appliquer le troisième ensemble de traits (Tableau 25) sur ces profils généraux permis de distinguer ou rassembler certains profils entre eux en faisant ressortir plusieurs de leurs caractéristiques communes.

Pour commencer, un taux assez élevé de pronoms de la première personne est récurrent dans les profils, mais certaines variations permettent de regrouper certains profils (Tableau 30).

- La plupart des profils (groupe 1) ont un taux moyen de première personne situé en entre 32,51 et 39,82, un taux médian entre 28,55 et 33,, et un écart type entre 24,27 et 37,65.
- Le profil 07_*non discursif* (groupe 2) se démarque en affichant un taux assez bas, avec une moyenne de 15,55, une médiane de 8 et une dispersion, relativement basse par rapport au groupe majoritaire, de 22.
- Au contraire, le profil 09_*interaction succincte* (groupe 3) se caractérise par un taux assez élevé avec une moyenne de 43,43 et une médiane de 42,86 et une dispersion de 29,23, qui se situe dans la fourchette moyenne du groupe majoritaire.
- Enfin les profils 10_*composite* et 08_*contenu externe* (groupe 4) s'apparentent avec une dispersion nettement plus basse que les autres, avec, respectivement 17,28 et 15,10. Par ailleurs, dans ces deux cas, les indices de moyenne et de médiane sont très proches : 36,02 de moyenne et 35,71 de médiane pour le premier profil, et 29 de moyenne et 29,63 de médiane pour le second profil.

Profil	Seuil			Groupe
	Moyenne	Médiane	Écart-type	
02_*mono message*	36,30	30,77	37,60	1
03_*monologue*	39,61	33,33	37,65	1
04_ *monologue_anonyme*	32,51	28,57	25,07	1
05_ *échange succinct*	39,82	33,33	36.86	1
06_*échange riche*	34,37	33,33	25,74	1
07_ *non discursif*	15,55	8	22,52	2
08_ *contenu externe*	29	29,63	15,10	4
09_ *interaction succincte*	43,43	42,86	29,23	3
10_*composite*	36,02	35,71	17,28	4
11_*vote*	36,57	33,33	24,27	1

Tableau 30 : Trait 1PS appliqué aux profils généraux

En ce qui concerne le taux de messages anonymes, hormis les profils qui ont déjà été caractérisés par un tel paramètre (03_*monologue* et 04_*monologue anonyme*), certains groupes se distinguent également (Tableau 31). Les profils 10_*vote* et 11_*composite* ont en effet un taux relativement bas, avec moins de 10 % en moyenne et en médiane, et une dispersion de 10,37 maximum.

Profil	Seuil		
	Moyenne	Médiane	Écart-type
02_*mono message*	34,44	0	45,16
05_ *échange succinct*	22,72	0	24,89
06_*échange riche*	21,16	0	24,70
07_ *non discursif*	38,24	50	21,10
08_ *contenu externe*	27,96	16,67	30,13
09_ *interaction succincte*	20,52	0	28,68
10_*composite*	6,99	0	7,73
11_*vote*	8,53	6,67	10,37

Tableau 31 : Trait Taux de messages anonymes appliqué aux profils généraux

Le nombre de marque interrogatives est généralement très bas, voire nul, mais deux profils affichent une moyenne et une médiane plus haut que les autres profils : 07_*non discursif* et 08_*contenu externe* (Tableau 32). Cette particularité peut amenée par le contenu extérieur qui se retrouve dans ces fils de discussion, ainsi que leur taille relativement élevée. Ce cas a déjà été observé lors d'une des phases d'exploration précédentes : le cas d'un poème en portugais issu d'une source extérieure contenait de nombreux « que », considérés par Talismane comme des pronoms interrogatifs (Exemples 20 et 21). La présence de ces marques interrogatives doit donc être interprétée avec prudence, car elles ne sont pas forcément le signe d'un mode interrogatif récurrent dans ce type de profils.

Profil	Seuil		
	Moyenne	Médiane	Écart-type
07_non discursif	4,86	2	12,83
08_contenu externe	3,32	2	3,69

Tableau 32 : Trait nombre de marques interrogatives appliqué aux profils généraux

```
<#p idno='119387>
1   Que que PROWH   PROWH       2   obj 2
2   vibra   vibrer  V   V   n=s|t=J|p=3|
3   os  os  NC  NC  g=m|    2   suj 2   suj
4   feros   _   NPP NPP _   3   mod 3   mod
5   raios   _   NPP NPP _   3   mod 3   mod
6   de  de  P   P       3   dep 3   dep 100.
7   Vulcano _   NPP NPP _   6   prep    6
8   ,   ,   PONCT   PONCT       7   ponct
```

Exemple 20: Discussion:Les Lusiades
PdD : 119387, FdD : 2
Extrait du contenu étiqueté

Étiquette de pronom interrogatif

22
Estava o Padre ali, sublime e dino,
Que vibra os feros raios de Vulcano,
Num assento de estrelas cristalino,
Com gesto alto, severo e soberano;
Do rosto respirava um ar divino,
Que divino tornara um corpo humano;
Com ũa coroa e ceptro rutilante,
De outra pedra mais clara que diamante.

Exemple 21 : Discussion:Les Lusiades
PdD : 119387, FdD : 2
Extrait du fil « Effectivement c'est pas très instructif...faudrait developper... »

V.2 Profils affinés

Nous avons également amorcé une identification affinée de profils d'interactions plus précis à partir des traits linguistiques et temporels (*cf.* IV.2 Amorce de l'identification de profils plus précis), en dégageant deux profils : 12_*collaboration harmonieuse* et 13_*entreprise solitaire* forcée. Voici une synthèse de ces deux profils caractérisés par des statistiques descriptives de leurs traits.

Statistiques Descriptives	02_nbMsg	11_tauxAnonyme	10_entropie	19_pronom1PS	20_pronom2PS	21_pronoms1PP	22_pronoms2PP	13_phInterro	12_phExcl	15_phInteroExpr	14_phExclEpx	25_tempMed	26_tempPremDer	23_marqueInterro
12_*collaboration harmonieuse*														
Minimum	3	0	0,918	18,75	0	0	0	0	0	0	0	1,35	3	0
Maximum	6	0	1	57,14	19,05	11,9	35,71	13,04	7,4	0	0	18,25	347	8
Moyenne	4,4	0	0,98	41,8	4,68	3,63	10,9	6,1	2,612	0	0	7,448	118.5	1.6
Médiane	4	0	1	46,4	0	0	6,25	5,77	1,92	0	0	5,095	62	0
Écart type	1,14	0	0,036	16,9	8,25	5,36	14,16	4,9	3,022	0	0	7,425	154.9	3.57
13_*entreprise solitaire forcée*														
Minimum	3	0	0.722	30.77	0	0	0	0	0	0	0	0,08	95	1
Maximum	5	25	0.918	52.38	0	0	10	17.65	0	0	0	1023,21	2 046	2
Moyenne	3,75	6,25	0,84	40,79	0	0	4,423	7,975	0	0	0	287,6	674,5	1
Médiane	3,5	0	0,86	40	0	0	3,845	7,125	0	0	0	63,53	278,5	1
Écart type	0,96	12,5	0,095	8,89	0	0	5,19	7,30	0	0	0	492	921,12	0,8

Tableau 33 : Synthèse de deux profils précis dégagés et statistiques descriptives de leurs traits

VI. Synthèse et perspectives

Cette recherche dont l'objectif est d'étudier les différentes interactions entre des contributeurs impliqués dans un travail de rédaction collaborative, a mis en place une méthodologie afin de faire émerger des profils larges de situations d'interaction. Elle est basée sur l'analyse manuelle d'échantillons de fils de discussion établis à partir de seuils de traits caractérisant la taille des fils (Tableau 8) ou bien leur contenu linguistique, interactif et leur étendue temporelle (Tableau 25). Cette analyse a fait émerger des profils d'interaction généraux, repérés comme contenant des interactions assez réduites ou bien des contenus non rédigés par les auteurs dans le cadre de la discussion (Tableau 29), mais propose également une première exploration approfondie des données afin d'identifier des interactions plus précises (Tableau 33).

Cette étude a permis une première exploration des données brutes afin d'avoir une vision globale de l'interaction au sein des fils de discussion. Cette étape ouvre la voie à une analyse plus fine des fils de discussion afin de faire émerger des profils d'interactions plus précis et ciblés. Dans cette perspective il pourrait être pertinent de compléter et d'ajuster le travail effectué sur certains points.

Pour commencer, certains traits, notamment linguistiques se sont révélés être assez difficiles à interpréter. Par exemple les marques de l'impératif sont extrêmement rares, et il en est de même pour les versions « expressives » des phrases interrogatives et exclamatives. Il semble que les modalités d'expression dans le cadre des pages de discussion de la Wikipédia ne se prêtent pas à l'emploi de ces formes. Ces traits pourront peut être servir de seuil complémentaire pour une identification très précise. Il serait donc bénéfique d'inclure de nouveaux traits afin d'explorer le contenu des données, notamment en mettant à profit les nombreuses autres informations renseignées par la version étiquetée du corpus (*cf. II.1.3 Version étiquetée avec Talismane*).

Par ailleurs, l'analyse qualitative d'autres échantillons pourrait permettre de vérifier et d'ajuster les seuils établis pour le moment, mais serait également l'occasion d'appliquer certains seuils établis sur un sous-corpus à l'ensemble des données. Ces observations peuvent, en outre, permettre de compléter la grille d'annotation (*cf.* IV.1.2 Une grille d'annotation pour accompagner l'observation), qui pourrait ainsi être suivie de manière plus rigoureuse pour les phases d'annotation manuelle et faciliter l'émergence et la caractérisation des profils d'interaction.

Bibliographie

- Baecker, R. M., Nastos, D., Posner, I. R., & Mawby, K. L. (1993, avril). The User-Centred Iterative Design Of Collaborative Writing Software. *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems*, New York, USA.

- Biber, D., Egbert, J., & Davies, M. (2015). Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora*, 10(1), 11-45. doi : 10.3366/cor.2015.0065

- Chen, C. (1997). Writing with Collaborative Hypertext: Analysis and Modeling. *Journal of the American Society for Information Science*, 48(11), 1049.

- Ferschke, O., Daxenberger, J., & Gurevych, I. (2013). A Survey of NLP Methods and Resources for Analyzing the Collaborative Writing Process in Wikipedia. Dans I. Gurevych & J. Kim (dir.), *The People's Web Meets NLP: Collaboratively Constructed Language Resources* (p. 121-160). Berlin, Heidelberg: Springer. doi : 10.1007/978-3-642-35085-6_5

- Ferschke, O., Gurevych, I., & Chebotar, Y. (2012). Behind the Article: Recognizing Dialog Acts in Wikipedia Talk Pages. Dans *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (p. 777-786). Stroudsburg, PA, USA : ACL.

- Herring, S., Stein, D., & Virtanen, T. (2013). Introduction to the pragmatics of computer-mediated communication. Dans *Handbook of pragmatics of computer-mediated communication* (p. 3-31). Berlin: Mouton de Gruyter.

- Ho-Dac, L.-M., & Laippala, V. (2015, octobre). *Les discussions Wikipedia : un corpus pour caracériser le genre « discussion »*. Présenté à International Research Days: Social Media and CMC Corpora for the eHumanities, Rennes, France.

- Langlais, P.-C. (2014). Negociation vs. democracy: the Wikipedia case. *Négociations*, Démocratie et négociations(21), 21-34. doi : 10.3917/neg.021.0021

- Passig, D., & Schwartz, G. (2007). Collaborative Writing: Online Versus Frontal. *International Journal on E-Learning*, 6(3), 395-412.

- Poudat, C., Vanni, L., & Grabar, N. (2016, juin). How to explore conflicts in French Wikipedia talk pages? Dans Mayaffre, D., Poudat, C., Vanni, L., Magri, V. and Follette, P. (dir.), *Proceedings of 13 th International Conference on Statistical Analysis of Textual Data* (Vol. 2, p. 645-656). Nice, France : Presses de FacImprimeur.

- Sichler, A., & Prommer, E. (2014). Gender Differences within the German-Language Wikipedia. *ESSACHESS Journal for Communication Studies, 7*(2), 77-93.

- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing.

- Urieli, A. (2013). *Robust French syntax analysis: reconciling statistical methods and linguistic knowledge in the Talismane toolkit* (Thèse de doctorat). Toulouse II le Mirail, Toulouse.