Profilage des interactions en ligne entre les rédacteurs de la Wikipedia

Lydia-Mai Ho-Dac et Ludovic Tanguy

Université de Toulouse Jean Jaurès – CLLE

5 Dec. 2022, UE TAL

Plan

CMR – Communications Médiées par les Réseaux (*CMC*) et Wikipédia

- *CMC (Computer-Mediated Communications)* : variété de textes écrits "en réseau" via des "machines" (ordinateurs & smartphones) e.g. clavardages, SMS, forum de discussion, Twitter, Whatsapp, Facebook, commentaires Youtube, etc.
- Une quantité exponentielle de données produites
- De nouvelles formes de communication, nouveaux usages : variantes (ortho)graphiques, écrit spontané, informel, influence de l'ergonomie machine, conversations asynchrones, etc.
- CMC : interactions structurées en fils et messages (posts)
- Challenge : décrire ces nouvelles formes d'interaction
  - quels traitements automatiques pour ces données "non standards" ?
  - comment la connaissance est partagée ?
  - comment les échangent s'organisent ?
  - comment les discours se construisent ?

Un peu d'histoire : 2 conditions réunies dans les 90's

1– La technologie wiki

- 1990 : concept de *co-authoring* (C. M. Neuwirth)
- 1995 : ***WikiWikiWeb*** (W. Cunningham)

2– Grands projets d'encyclopédies ouvertes et libres [Sah15]

Encyclopédies collectives, gratuites, de référence, non soumises au copyright, accessibles sur internet, dont les auteurs seraient des rédacteurs volontaires

- 1993 : ***Interpedia*** (R. Gates) encyclopédie de référence non soumise au copyright
- 1997 : ***Distributed Encyclopedia*** (U. Fuchs, futur Wikipédien.de) contre le "poinds croissant de la sphère marchande sur le web"
- 1999 : ***GNUpedia***, puis *GNE's Not an Encyclopedia* (R. Stallman, logiciel libre) encyclopédie multilingue, pour "garder le savoir humain ouvert et librement disponible pour l'humanité"
- 2000 : ***Nupedia*** (J. Wales, *trader*) encyclopédie "libre" financée par la pub et écrite par des experts

Nupedia, parent de Wikipedia

Nupedia, Jimmy Wales alias @Jimbo et Larry Sanger (Dr philo.)

- encyclopédie participative écrite par des "experts" i.e. ayant un doctorat
- "écrit par des experts" donc nécessité d'un processus de "peer-reviewing" (comme dans le monde académique ← Larry Sanger)
  - Lourdeur du processus d'édition → faible productivité
  - *WikiWikiWeb* : une solution pour faciliter le processus
  - **15 janvier 2001, WP[EN] est ouvert** pour constituer le *brouillon* de *Nupedia* : "an article incubator for Nupedia." (Sanger 2006, *The Guardian*)

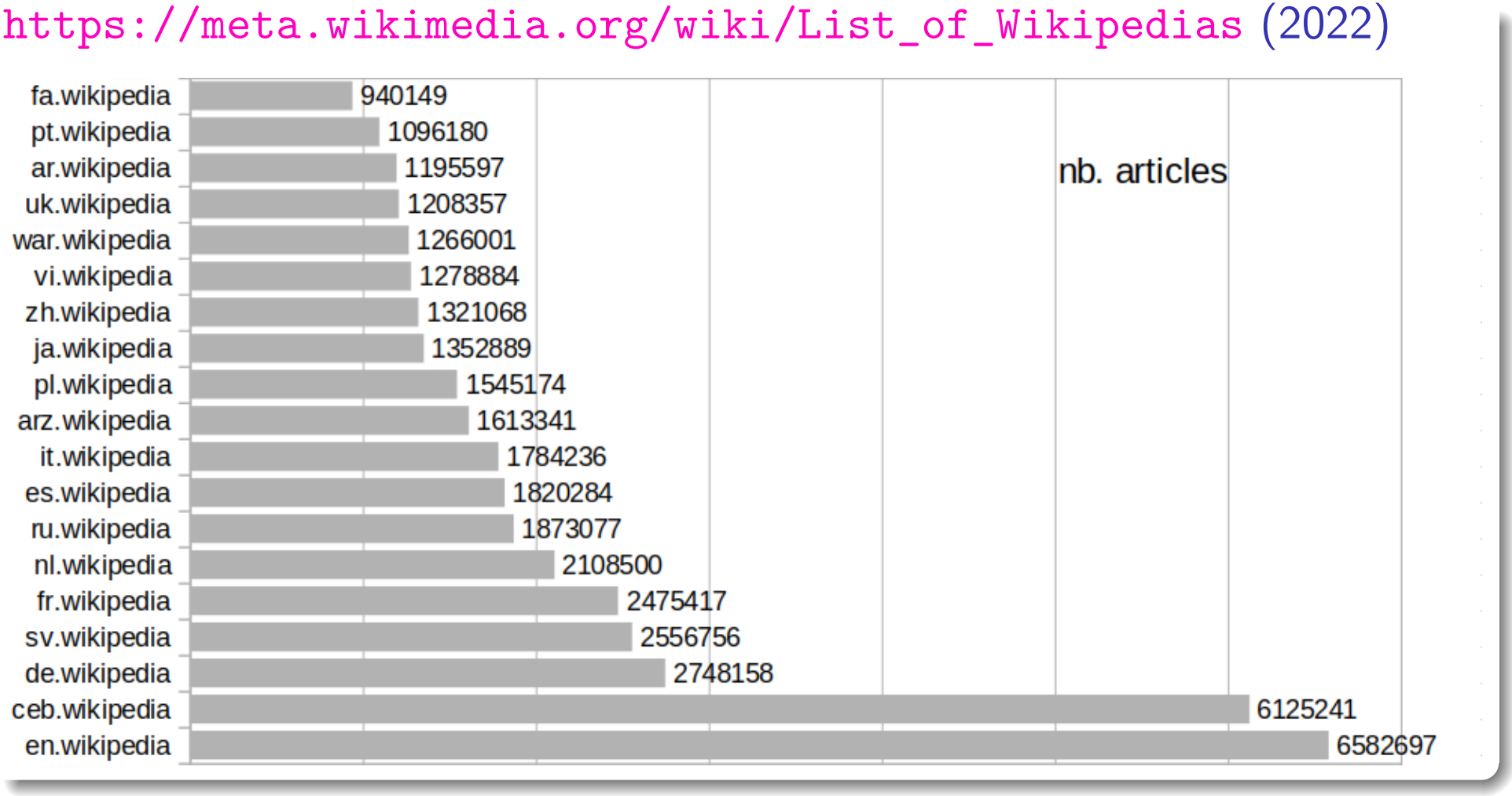
Naissance et premiers pas du projet WP

Un très gros boum

02.2001	600 articles sont créés
03.2001	1300
05.2001	3900
01.2002	20 000
09.2003	WP compte plus de 100 000 article *Nupedia* est abandonné (seuls 24 articles ont été publiés)
2022	WP est classée comme le 5e site le plus visité avec presque 7 G (milliards) de visite en oct. 2022 selon sem-rush.com derrière *Google 96 G, Youtube 77 G, Facebook 13 G, Twitter 8 G*

Aujourd'hui, un phénomène global (e.g. nombre d'articles)

https://meta.wikimedia.org/wiki/List_of_Wikipedias (2022)



Une ressource pour le TAL

Exploitation des articles

- extraction de connaissances [ZMG08, MMLW09]
- construction de ressources multilingues [KRPA10]

Exploitation des révisions

Un accès aux processus d'écriture [FDG13]

Exploitation des discussions

Un accès aux processus de négociations [FGC12, FDG13]

Exploitation des révisions et des discussions pour le TAL [FDG13]

Exploitation des révisions

DEF

! Règle WP des 3 reverts

- *diff* entre les versions pour extraire des paires de variantes orthographiques, les paraphrases, les simplifications/résumés
- persistance d'une phrase comme indice d'importance
- détection du vandalisme (environ 7% des révisions dans la WP anglaise selon [PSG08])

CF. `http://contropedia.net` : mesure de la "controversialité" d'un sujet par l'analyse du nombre d'édits dits "de désaccord" i.e. dont le résultat est la suppression de contenu, au moins 1 mot)

Exploitation des pages discussions

Un besoin originel de disposer d'espaces de discussion

Quelques jours après la création de WP, un des premiers contributeurs soulève le problème suivant : que faire des discussions concernant les articles ? Faut-il les déplacer dans une page à part ? Un autre contributeur pense que ce n'est pas nécessaire : sur un wiki tout se corrige de soi-même, les discussions trop étendues finiront par s'auto-réguler.

Les pages de discussions WP

Une discussion en ligne derrière les articles...

- https://fr.wikipedia.org/wiki/Glyphosate
- https://fr.wikipedia.org/wiki/Discussion:Glyphosate

Les pages de discussions WP : une nécessité venant des WP non anglaises

EN système de modération par un processus de consensus pur et de "dictateur bienveillant" (Wales, sorte d'éditeur encyclopédique)

- Wales : "les discussions sur la nature de Wikipédia ne peuvent figurer qu'en dehors de WP (de préférence sur la liste de diff.)"

WP français puis allemand et italien, un autre fonctionnement

Mise en place d'espaces de discussion et de processus de négociation (le dictateur bienveillant ne parle qu'anglais !) [Lan14]

FR octobre 2002, création d'un page "prise de décision" par Florence Dévouard (@Anthère) où "le choix final dépend d'un vote consensuel et non d'un simple consensus"

IT *Sondaggio*, "mécanisme simple et rapide pour résoudre facilement les problèmes actuels."

DE *Meinungsbilder*, "clarifier les questions sur lesquelles aucun consensus n'a pu être atteint."

Les pages de discussion WP

Un enregistrement d'interactions inédit

- Forum de discussion systématiquement associé à chaque article où les Wikipédiens peuvent discuter du processus d'écriture de l'article
- Un observatoire des coulisses de la collaboration entre humains

> *From a scientific point of view, article Talk pages are a unique type of web discourse and a valuable resource for the humanities and writing sciences, since the discussions develop in parallel with the discussed articles and provide insights into the meta level of the collaborative writing process that normally remains hidden. With structured access to this resource, the linguists and researchers in the writing sciences have the unparalleled possibility to observe these hidden processes without having to conduct interviews or carrying out supervised field experiments. [FGC12].*

Example : https://fr.wikipedia.org/wiki/Psychanalyse

Les pages de discussion WP

- des discussions pour co-construire un contenu (révisions antagonistes)
- des discussions pour respecter l'idéologie Wikipedia (contenu objectif – NPOV – et sourcé)
- des discussions pour trouver un consensus en cas de désaccord
- des discussions (conflictuelles) entre spécialistes et *wiki-anarchistes*

  > *Des universitaires, venus du projet Nupedia, écrivant dans leur domaine de spécialité pour WP, auraient abandonné l'encyclopédie [...] lassés de croiser le fer avec des "wiki-anarchistes" d'un niveau d'expertise bien moindre. [Sah15, 241]*

- des accusations contre l'auto-promotion (31% de professionnels de la communication ont contribué à l'écriture de la page de leur produit [DiS12]) → *Wikiscanner*, (V. Griffith) pour identifier les origines des adresses IP (*Pepsi*, *CIA*, *Exxon*, *partis politiques*, etc.)
  mais parfois inefficace e.g. *Bogdanoff*, *Trump*, *Zemour*, etc.)

Le corpus EFG_talkPages

1. Extraction des discussions depuis les dump

   sauvegarde globale des pages courantes (archive diffusée librement sur la page http ://dumps.wikimedia.org/frwiki/)
2. Sélection des discussions "à garder"
   - au moins 1 post et 2 mots
   - les discussions courantes, archivées et portant sur la notion de neutralité
3. Détection des segments constitutifs de chaque discussion : sections (fils), messages
4. Conversion selon la TEI–CMC

	EFG Dumps, 09.2019		EFG corpus	
lang.	#articles	#discussions	#disc. retenues	% retenu
WP.en	14856106	7903148	2025888	30
WP.fr	3729677	1852689	266699	14
WP.de	3920295	769091	713485	93
	22506078	10524928	3340527	32

Structuration et encodage du corpus

Des métadonnées et des liens inter-langues

```
      </monogr>
      <relatedItem type="langLink">
        <ref target="https://de.wikipedia.org/wiki/Fl%C3%BCchtlingskrise%20in%20Europa%20ab%202015" targetLang="de">Flüchtlingskrise in Europa ab 2015</ref>
      </relatedItem>
      <relatedItem type="langLink">
        <ref target="https://en.wikipedia.org/wiki/European%20migrant%20crisis" targetLang="en">European migrant crisis</ref>
      </relatedItem>
      <relatedItem type="articleLink">
        <ref n="9366469" target="https://fr.wikipedia.org/wiki/Crise%20migratoire%20en%20Europe" targetLang="fr">Crise migratoire en Europe</ref>
      </relatedItem>
    </biblStruct>
  </sourceDesc>
</fileDesc>
<profileDesc>
  <creation>
    <date type="last-change">2018-07-30T17:58:12Z</date>
  </creation>
  <textClass>
    <classCode scheme="https://fr.wikipedia.org/wiki/Portail:Accueil"><ref target="https://fr.wikipedia.org/wiki/Portail:G%C3%A9ographie" targetLang="fr">
    Portail:Géographie</ref><ref target="https://fr.wikipedia.org/wiki/Portail:Politique" targetLang="fr">Portail:Politique</ref><ref target=
    "https://fr.wikipedia.org/wiki/Portail:Europe" targetLang="fr">Portail:Europe</ref><ref target=
    "https://fr.wikipedia.org/wiki/Portail:Relations%20internationales" targetLang="fr">Portail:Relations internationales</ref><ref target=
    "https://fr.wikipedia.org/wiki/Portail:Ann%C3%A9es%202010" targetLang="fr">Portail:Années 2010</ref><ref target=
    "https://fr.wikipedia.org/wiki/Portail:Humanitaire%20et%20d%C3%A9veloppement" targetLang="fr">Portail:Humanitaire et développement</ref><ref target=
    "https://fr.wikipedia.org/wiki/Portail:Europe" targetLang="fr">Portail:Europe</ref><ref targetLang="fr">Portail:Organismes</ref><ref targetLang="fr">
    Portail:Thèmes</ref></classCode>
    <classCode scheme="https://fr.wikipedia.org/wiki/Catégorie:Accueil"><ref target="https://fr.wikipedia.org/wiki/Catégorie:Guerre%20civile%20syrienne"
    targetLang="fr">Catégorie:Guerre civile syrienne</ref><ref target="https://fr.wikipedia.org/wiki/Catégorie:Crise%20migratoire%20en%20Europe" targetLang
    ="fr">Catégorie:Crise migratoire en Europe</ref></classCode>
    <keywords>
      <term>Europe</term>
      <term>Union européenne</term>
      <term>Moyen-Orient</term>
      <term>Afrique</term>
      <term>Relations internationales</term>
      <term>Humanitaire et développement</term>
    </keywords>
    <classCode scheme="https://fr.wikipedia.org/wiki/Projet:Évaluation"><ref target="https://fr.wikipedia.org/wiki/Wikipédia:Article_d%27avancement_BD"
    targetLang="fr">bon début</ref></classCode>
  <classCode scheme="https://en.wikipedia.org/wiki/Category:Contents"><ref target="https://en.wikipedia.org/wiki/Category:Syrian%20Civil%20War" targetLang=
  "en">Category:Syrian Civil War</ref><ref target="https://en.wikipedia.org/wiki/Category:European%20migrant%20crisis" targetLang="en">Category:European
  migrant crisis</ref></classCode><classCode scheme="https://en.wikipedia.org/wiki/Wikipedia:Contents/Portals"><ref target=
  "https://en.wikipedia.org/wiki/Portal:Geography" targetLang="en">Portal:Geography</ref><ref target="https://en.wikipedia.org/wiki/Portal:Politics"
```

Structuration et encodage du corpus

Corps d'une CMC encodée selon la TEI-CMC

idéalement, une interaction se définit comme

- un fil de discussion ( ) autour d'un sujet (<head>) qui donne lieu à plusieurs messages qui se répondent les uns aux autres et se succèdent dans le temps
- chaque messages (<post>) devrait être associé à
  - une signature indiquant <signed>
    - son auteur (anonyme ou inscrit) @who
    - sa date de publication @when−iso
  - sa position dans le fil de la discussion @xml:id
  - son "rôle" dans l'échange @indentLevel

Structuration et encodage du corpus

```
<div type="talk">
  <head>Crise migratoire en Europe</head>
  <note creation="template" type="header">
  <div type="thread" xml:id="i.9394134_1">
    <head>Titre de l'article : &amp;quot;réfugiés&amp;quot; ou &amp;quot;migrants&amp;quot;? les grands
    Francais et wikipedia anglais disent migrants</head>
    <post indentLevel="0" mode="written" when-iso="2015-09-05T05:52+02" who="WU00004600" xml:id="i.9394
      <p> Le titre "Crise des réfugiés en Europe" ne me parait pas approprie. Les grands journaux franç
      "migrants" (cf: le monde, Libération, Le point, le figaro)</p>
      <p> Un réfugié est une personne fuyant la guerre/persécution. beaucoup de personne arrivant en eu
      pas la guerre (certains oui), migrant est une définition plus large regroupant toute personne en
      <p> L'article anglais de wikipeidai utilise aussi le terme de migrant. <signed type="signed"><ref
      "https://fr.wikipedia.org/wiki/Utilisateur:Killy-the-frog"><name full="yes">Killy-the-frog</name>
      septembre 2015 à 05:52 (CEST)</date></signed></p>
    </post>
    <post indentLevel="1" mode="written" when-iso="2015-09-05T10:27+02" who="WU00005017" xml:id="i.9394
      <p><name creation="template" type="notif">Killy-the-frog</name>. Je me faisais la même réflexion
      https://www.google.fr/search?q=%22crise+des+migrants%22ie=utf-8oe=utf-8gws_rd=crei=UKbqVYTkA6KR7A
      +des+migrants%22tbm=nws "crise des migrants" contre
      https://www.google.fr/search?q=%22crise+des+r%C3%A9fugi%C3%A9s%22ie=utf-8oe=utf-8gws_rd=crei=WKbq
      q=%22crise+des+r%C3%A9fugi%C3%A9s%22tbm=nws "crise des réfugiés". "Crise des migrants" est plus u
      semble plus exact. Il faudrait renommer en Crise des migrants en Europe ou Crise migratoire en Eu
      (actuellement une redirection). D'autres avis en faveur, ou contre un renommage ? --<signed type=
      target="https://fr.wikipedia.org/wiki/Utilisateur:Pitth%C3%A9e"><name full="yes">Pitthée</name></
      septembre 2015 à 10:27 (CEST)</date></signed></p>
    </post>
    <post indentLevel="2" mode="written" when-iso="2015-09-06T15:59+02" who="WU00045669" xml:id="i.9394
    <post indentLevel="3" mode="written" when-iso="2015-09-06T19:46+02" who="WU00005017" xml:id="i.9394
    <post indentLevel="4" mode="written" when-iso="2015-09-07T10:12+02" who="WU00003750" xml:id="i.9394
    <post indentLevel="5" mode="written" when-iso="2015-09-07T10:56+02" who="WU00004475" xml:id="i.9394
    <post indentLevel="6" mode="written" when-iso="2015-09-07T13:31+02" who="WU00003750" xml:id="i.9394
    <post indentLevel="7" mode="written" when-iso="2015-09-07T13:39+02" who="WU00003750" xml:id="i.9394
    <post indentLevel="8" mode="written" when-iso="2015-09-08T14:18+02" who="WU00016557" xml:id="i.9394
    <post indentLevel="9" mode="written" when-iso="2015-09-08T15:43+02" who="WU00006582" xml:id="i.9394
    <post indentLevel="1" mode="written" when-iso="2015-09-07T13:58+02" who="WU00000130" xml:id="i.9394
    <post indentLevel="2" mode="written" when-iso="2015-09-07T14:13+02" who="WU00005017" xml:id="i.9394
```

Structuration et encodage du corpus

Idéalement, un fil de discussion autour d'un sujet donne lieu à plusieurs échanges signés qui se répondent les uns aux autres et se succèdent dans le temps, mais...

who certains messages écrits par des robots, ne sont pas signés, ou l'indication de l'auteur a été effacée

when certains messages non datés (date omise ou effacée), datés manuellement et incorrectement (ex : 2009-11-47T12:14+00 ou 2007-02-02T27:39+00) ou avec une date postéditée (insertion d'un autre message avec une autre date, troncation de la partie contenant la signature, remplacement de la signature par une autre)

id la position initiale du message peut avoir été modifiée (insertion/suppression de messages)

indentLevel le déroulé de l'échange initial se retrouve chamboulé et parfois difficile à reconstruire...

Retrouver le déroulement initial (URL)

Un article biaisé écrit au masculin [modifier le code]

"Les femmes et les religions" est un prèche qui n'a rien à voir avec le féminisme, ou de surcroit ce qui dérange la perspective catholique est habilement supprimé : "Les exactions commises par l'inquisition sur les femmes pour lesquelles **jamais aucune excuses n'ont été faites**".

Il n'en sera rien. Cet article sera passé au rouleau compresseur par les différentes tendances antiféministes existentes parmi les contributeurs. Pour ma part, je cesse de contribuer sur cet article, j'ai d'ailleurs cessé de contribuer en général.--fl0 3 août 2005 à 12:26 (CEST)fl0 [répondre]

Les wikipédiennes représentent moins de 20% des contributeurs, il eut été bien naturel qu'*au moins* sur ce sujet elles soient les principales contributrices.

> C'est parfaitement inepte. Considerait on comme naturel que l'article sur le socialisme soit principalement redige par des socialistes, l'article sur le catholiscisme par des catholiques, l'article sur le sionisme par des juifs (et celui sur l'anti-sionisme par des anti-sioniste) ? Par ailleurs tout le monde peut participer a wikipedia. CdC 28 mai 2006 à 12:52 (CEST) [répondre]

> La phrase que tu cites plus haut n'a pas été supprimée ; je ne la trouve pas très heureuse, mais elle ne me dérange pas.
> Sinon, je suis bien d'accord qu'un article concernant un mouvement minoritaire doit acquérir un minimum d'immunité vis-à-vis des déprédations causées par ses détracteurs. C'est dans ce sens qu'il faut comprendre mon intervention sur cet article. Gemme 3 août 2005 à 12:57 (CEST) [répondre]

> Ce qui fait la qualité d'un article, c'est la compétence des contributeurs et des contributrices, pas leur sexe. Apokrif 3 août 2005 à 23:44 (CEST) [répondre]

> > oui mais vous etes les plus nombreux et on ne peut pas dire que proportionnellement vos compétencent vous fassent honneur. Aussi pourriez vous avoir un minimum d'humilité que vous n'avez pas.--fl0 3 août 2005 à 23:59 (CEST)fl0 [répondre]

Qui interpellez-vous avec ce « vous » ? Apokrif 7 août 2005 à 23:50 (CEST) [répondre]

Nous ne demandons pas votre douteuse "protection" pour parler à notre place, mais la liberté d'expression.--fl0 3 août 2005 à 14:32

Structurer, avec les moyens du bords le déroulement initial

```
<div type="thread" xml:id="i.1360139_11">
  <head>Un article biaisé écrit au masculin</head>
  <post indentLevel="0" mode="written" when-iso="2005-08-03T12:26+02" who="WU00002748" xml:id="i.1360139_11_1">
    <p> "Les femmes et les religions" est un prêche qui n'a rien à voir avec le féminisme, ou de surcroit ce qui dérange la perspective catholique
    est habilement supprimé : "Les exactions commises par l'inquisition sur les femmes pour lesquelles "'jamais aucune excuses n'ont été faites"'".</p>
    <p> Il n'en sera rien. Cet article sera passé au rouleau compresseur par les différentes tendances antiféministes existentes parmi les
    contributeurs. Pour ma part, je cesse de contribuer sur cet article, j'ai d'ailleurs cessé de contribuer en général.--<signed type="signed"><ref
    target="https://fr.wikipedia.org/wiki/Utilisateur:Floreal"><name full="yes">Floreal</name></ref> fl0 <date>3 août 2005 à 12:26 (CEST)
    </date></signed>fl0</p>
  </post>
  <post indentLevel="0" mode="written" who="WU00000000" xml:id="i.1360139_11_2">
    <p> Les wikipédiennes représentent moins de 20% des contributeurs, il eut été bien naturel qu"'au moins" sur ce sujet elles soient les
    principales contributrices.</p>
  </post>
  <post indentLevel="1" mode="written" when-iso="2006-05-28T12:52+02" who="WU00002668" xml:id="i.1360139_11_3">
    <p>C'est parfaitement inepte. Considerait on comme naturel que l'article sur le socialisme soit principalement redige par des socialistes,
    l'article sur le catholiscisme par des catholiques, l'article sur le sionisme par des juifs (et celui sur l'anti-sionisme par des anti-sioniste)
    ? Par ailleurs tout le monde peut participer a wikipedia. <signed type="signed"><ref target="https://fr.wikipedia.org/wiki/Utilisateur:CdC"><name
    full="yes">CdC</name></ref><date>28 mai 2006 à 12:52 (CEST)</date></signed></p>
  </post>
  <post indentLevel="1" mode="written" when-iso="2005-08-03T12:57+02" who="WU00001822" xml:id="i.1360139_11_4">
    <p>La phrase que tu cites plus haut n'a pas été supprimée ; je ne la trouve pas très heureuse, mais elle ne me dérange pas.</p>
    <p> Sinon, je suis bien d'accord qu'un article concernant un mouvement minoritaire doit acquérir un minimum d'immunité vis-à-vis des déprédations
    causées par ses détracteurs. C'est dans ce sens qu'il faut comprendre mon intervention sur cet article. <signed type="signed"><ref target=
    "https://fr.wikipedia.org/wiki/Utilisateur:Gemme"><name full="yes">Gemme</name></ref><date>3 août 2005 à 12:57 (CEST)</date></signed></p>
  </post>
  <post indentLevel="1" mode="written" when-iso="2005-08-03T23:44+02" who="WU00000384" xml:id="i.1360139_11_5">
    <p>Ce qui fait la qualité d'un article, c'est la compétence des contributeurs et des contributrices, pas leur sexe. <signed type="signed"><ref
    target="https://fr.wikipedia.org/wiki/Utilisateur:Apokrif"><name full="yes">Apokrif</name></ref><date>3 août 2005 à 23:44 (CEST)
    </date></signed></p>
  </post>
  <post indentLevel="2" mode="written" when-iso="2005-08-03T23:59+02" who="WU00002748" xml:id="i.1360139_11_6">
    <p>oui mais vous etes les plus nombreux et on ne peut pas dire que proportionnellement vos compétencent vous fassent honneur. Aussi pourriez vous
    avoir un minimum d'humilité que vous n'avez pas.--<signed type="signed"><ref target="https://fr.wikipedia.org/wiki/Utilisateur:Floreal"><name full
    ="yes">Floreal</name></ref> fl0 <date>3 août 2005 à 23:59 (CEST)</date></signed>fl0</p>
  </post>
  <post indentLevel="0" mode="written" when-iso="2005-08-07T23:50+02" who="WU00000384" xml:id="i.1360139_11_7">
    <p> Qui interpellez-vous avec ce « vous » ? <signed type="signed"><ref target="https://fr.wikipedia.org/wiki/Utilisateur:Apokrif"><name full="yes"
    >Apokrif</name></ref><date>7 août 2005 à 23:50 (CEST)</date></signed></p>
  </post>
  <post indentLevel="0" mode="written" when-iso="2005-08-03T14:32+02" who="WU00002748" xml:id="i.1360139_11_8">
    <p> Nous ne demandons pas votre douteuse "protection" pour parler à notre place, mais la liberté d'expression.--<signed type="signed"><ref target=
    "https://fr.wikipedia.org/wiki/Utilisateur:Floreal"><name full="yes">Floreal</name></ref> fl0 <date>3 août 2005 à 14:32 (CEST)</date></signed>fl0
    </p>
  </post>
  <post indentLevel="2" mode="written" when-iso="2005-08-03T14:41+02" who="WU00001822" xml:id="i.1360139_11_9">
```

Plan

Objectif : Décrire les interactions dans les discussions WP

- Description linguistique des discussions Wikipedia pour avancer dans la description du genre "discussion en ligne"
- Distinguer les modes d'interactions dans ces CMR : intentions des auteurs, type de contenu discuté (e.g. encyclopédique, idéologique, éditorial), aboutissement de la discussion, type de texte (narratif, argumentatif, instructionnel, descriptif)
- Détection de comportements ciblés (e.g. comportements toxiques)

Méthodes

- Approches top-down : partir de catégories prédéfinies et les appliquer au corpus (annotation et description des caractéristiques linguistiques des catégories)
- Approche bottom-up : partir des données pour identifier des classes capables de "trier" les données dans des catégories interprétables

Question 1 : Qui intervient ?

- Des inscrits (sous pseudo, avec plus ou moins de droits e.g. (super)-administrateur, patrouilleur)
- Des anonymes : *Si 90% des participants sont anonymes, ils ne sont à l'origine que de 20% des contributions. Probablement parce que les vandales et les robots spammeurs se créent rarement un compte, les IPs font l'objet d'une surveillance particulière. Ils sont considérés avec suspicion par les participants actifs, qui vérifient leurs écrits et s'échangent des scripts qui ne font apparaître que les révisions anonymes. De manière remarquable, nous avions également observé que la moitié de ces révisions (évaluées en nombre de caractères insérés) était supprimée – alors que 80% des insertions réalisées par les 100 plus grands contributeurs de Wikipédia persistaient. Les IPs sont ainsi associés aux conflits dans Wikipédia.* [**?**]
- Des robots (InternetArchiveBot, Revert-Statistik, SignaturBot, TheCowBot, etc.)
- Des inconnus

Aperçu des interactions dans le corpus EFG_talkPages

	E	F	G	EFG
Discussions	2025888	266699	713485	3006072
archives	95105	1740	23052	119897
neutralité	5	3096	0	3101
Fils	6636783	608857	2121852	9367493
Msg	20025945	1832416	6938168	28796529
Écrits par un bot	1306737	55868	213004	1575609
(% msg)	6,5	3,0	3,1	5,5
Non signé	3232039	451613	824151	4507803
(% msg)	16,1	24,6	11,9	15,7
Msg humain et signé	15487169	1324935	5901013	22713117
(% msg)	77,3	72,3	85,1	78,9

Table – Intervenants dans les discussions du corpus EFG_talkPages

Question 2 : Pourquoi un contributeur intervient-il ?

Ce que caractériser les discussions pourrait signifier

administration

wording and point of view

content

No article protection??? [edit]

As an IP editor, I cannot f***ing believe this article is open to IP editors! The abuses are ongoing, and flagrant.184.145.42.19 (talk) 03:01, 12 Febr (UTC)

> What abuses?--Nowa (talk) 15:24, 12 February 2017 (UTC)
>
> @184.145.42.19: It's really not as bad as some other articles on my watchlist. In any case, anyone (**you**) can open a request for article protecti --**BurritoBazooka** Talk Contribs 15:42, 12 February 2017 (UTC)
>
> The last ip edit to be reverted is from December 26 and it wasn't rampant vandalism in that case either, just a lack of oversight of WP:NPOV (if y count the recent good faith edit reverted for lack of consensus). Saturnalia0 (talk) 17:09, 12 February 2017 (UTC)

Wording of page [edit]

I am concerned about the wording of the title of this article. A migrant is someone who willingly moves to another country for a better life e.g. bette economic prospects. A refugee is someone who has no choice but to flee their country because their life is at risk and they are being persecuted. A not a migrant and a migrant is not a refugee. The two words have got conflated in recent times, but in principle remain entirely different concepts. seen that the page "European refugee crisis" redirects to this one. Should there not be a separate page for "European refugee crisis", because this suitable. — Preceding unsigned comment added by Kats987124 (talk • contribs) 11:55, 10 March 2017 (UTC)

This is a wiki encyclopedia. Not a SJW politically correct soap box. 151.225.204.78 (talk) —Preceding undated comment added 19:20, 15 July 2017 (UTC)

The crowds include both migrants and refugees from various countries. Their lack of documentation makes establishing their point of origin and the motivation difficult. Dimadick (talk) 23:10, 19 March 2017 (UTC)

> @Kats987124: There have been a few discussions already about the name of the article regarding "refugee" and "migrant". See: Current title (1 september 2015), Migrants and refugees, Requested move 19 March 2016, Wording of page. See there for some of the reasons why the article is called European *migrant* crisis. In regards about splitting the page, see this page about splitting articles. **Seagull123** Φ 17:17, 2 April 201

The whole article is full of right wing propaganda. i deleted/changed some things like "most are economic migrants" (which is bullshit)... — Preceding comment added by 92.217.63.215 (talk) 18:50, 25 June 2017 (UTC)

Going forward with editing, has a decision been made regarding using "migrant" vs "refugee"?gmousalimas (talk) 22:34, 15 February 2018 (UTC)

Islamic state agents among refugees [edit]

Question 2 : Pourquoi un contributeur intervient-il ?

Ce que caractériser les discussions pourrait signifier

content and contrary points of view

Islamic state agents among refugees [edit]

The sentence and a small number of hostile agents including Islamic State militants has been removed three times, the first without e was a rumor, though the source does not say that, and the third claims WP:UNDUE, though the Reuters story received ample coverage Insider, Telegraph, etc). If that isn't enough, the same claim that Reuters reported in February has been made again by German auth moments, being picked up by The Wall Street Journal, Politico, etc. I'm undoing the removal one more time, since I believe there is du more sources should be included? Saturnalia0 (talk) 17:43, 23 March 2017 (UTC)

> While this might be factually correct, putting it in the first paragraph like this exaggerates its relevancy and gives the article an anti-i of people were "hiding" among the refugees, and it is obvious that the author wants to emphasis the IS operatives in order to give fue
> — Preceding unsigned comment added by 85.24.238.36 (talk • contribs) 18 August 2017 (UTC)
>
> > The opening paragraph is still not NPOV. The subject of this article is the migrant crisis, not IS terrorism or whether the migrant not. Sourced or not, things like this should be discussed further down in separate sections. I can see the case for reducing Euro encyclopaedia articles to read like opinion pieces. The last sentence of the opening paragraph is no more on-topic than it would on the Great Britain article that the country in question is home to many terrorists and illegal aliens, or to mention in the first p that several of the people killed in the Nazi camps were pedophiles, wife-beaters and murderers. PSjolund (talk) 12:51 19 Augu
> >
> > > The Holocaust claim of yours in wrong - the Holocaust was primarily about trying to wipe out the Jewish people; if some of th "pedophiles, wife-beaters and murderers" (which statisticly is quite likely, since millions of Jews were murdered), this was irr hand, a significant part of the current European migration crisis is the inability to filter out the terrorists; their presence is a עוד מישהו Od Mishehu 09:55, 20 August 2017 (UTC)

no interaction

Numbers outdated, wrong citizenship problem [edit]

According to research in the year 2016, 40% of Moroccans who came via Greece pretended to be syrian. Pretending a wrong citizenship 2015 also many people from Morocco (10.258), Algera (13.883) applied for Asylum in Germany, not regarding those pretending wrong c mention this, numbers mentioned in the article can be considered as outdated or questionable. - Haaklich (talk) 21:16, 30 April 2017 (U

A new statistics report from the United Nations Refugee Agency [edit]

According to a new UNHCR statistics report, less than 3% of the immigrants currently arriving in Europe are actual refugees.

Plan

Une approche Top-Down pour caractériser les discussions

- selon l'acte de parole, selon le topique, selon la posture de l'auteur, ...
- au niveau de la page, du fil, du post, de l'acte, ...
- annoter par des sociologues, des linguistes, des foules, ...
- deux principaux axes : les actes de dialogues et les attaques

Une approche Top-down des actes de dialogues [Fer14]

Annotation de 1864 fils contenant 4923 messages extraits de la E_WP

Scope	Label	Description
Topic	ERROR	Segmentation Errors
	REFOBJ-PART	Comment about specific section of the article
	REFOBJ-WHOLE	Comment about the whole article
	REFOBJ-META	Meta comment not directly referring to article
Turn	***Article Criticism***	
	CRITCOMPL	Information is incomplete or lacks detail
	CRITACC	Lack of accuracy, correctness or neutrality
	CRITLANG	Deficiencies in language and style
	CRITSUIT	Content not suitable for an encyclopedia
	CRITSTRUCT	Deficiencies in structure, organization or visual appearance
	CRITAUTH	Lack of authority
	Self Commitment	
	ACTF	Commitment to action in the future
	ACTP	Report of past action
	Requests	
	REQEDIT	Request for article edit
	REQMAINT	Request for admin or maintenance action
	Interpersonal	

Une approche Top-down des actes de dialogues [Fer14]

- 2729 messages ont pu être associés à au moins un acte (45% ne rentraient pas dans les catégories prédéfinies)

label	nb. msg	% msg	acte
CRIT	1778	49	critique p.r. à l'article (article criticism)
ACT	749	21	"journal d'activité" (self commitment)
REQ	432	12	demande d'edition, de maintenance (requets)
ATT	655	18	sentiment p.r. à un contrib. (interpersonnal)

- Naive Bayse Classifier des Msg (traits : n-grams (1-3) de tokens (-stoplist) dans une fenêtre de messages n-1, n, n+1 + information non lexicale e.g. longueur des messages, position de n dans le fil, durée temporelle entre n-1, n et n+1, niveau d'indentation level
- F1 = 0,78

Une approche Top-down des actes de dialogues [Fer14]

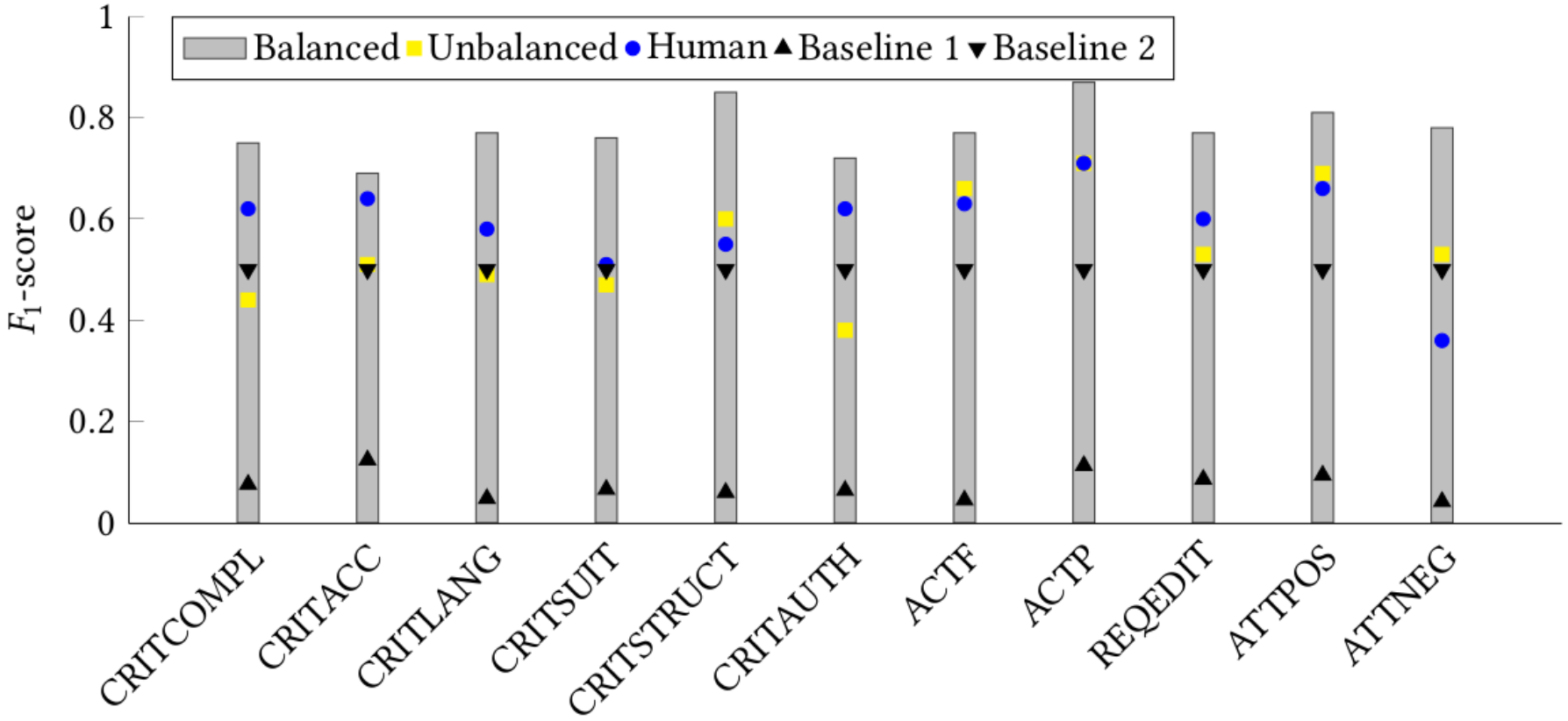


Figure 6.11: F_1-Scores for the classifiers trained on the balanced and unbalanced dataset, the human performance and two random baselines on the EWD corpus. Baseline 1 assigns labels according to their frequency distribution in the unbalanced dataset, while baseline 2 assigns labels at random on the balanced dataset.

Une approche Top-down des messages agressifs [WTD17]

https://meta.wikimedia.org/wiki/Research:Detox/Research

1. Annotation de 1000 posts pris à la fois au hasard et parmi les posts écrits par des contributeurs bannis

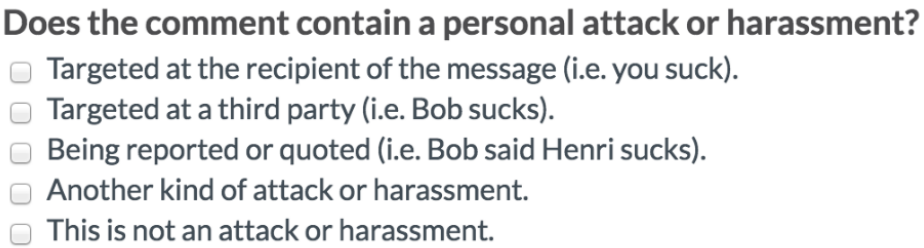


Figure 2: The question posed to our Crowdflower annotators.

2. 115 737 posts annotés (10 codeurs par posts, annotation par les foules *Amazone Mechanical Turk*)
3. 11,7 % étiquetés "attaque" par la majorité des codeurs
4. Entraînement de plusieurs classifieurs

Algo. perceptrons multi-couche, n-gram de caractères, tâche : prédire pour chaque post la classe : attaque oui/non $\Rightarrow$ P 96,59

Les limites de l'approche Top-down

- Quelles catégories pour décrire les interactions ?
  - A quel niveau (messages, fil, page) ?
  - Selon quel point de vue (acte de chaque contributeur, problème dans l'échange, etc.) ?
- Beaucoup d'interactions peu "intéressantes"
  - 45% des messages qui ne correspondent par à un acte de dialogue prédéfinie [Fer14]
  - 88% des messages jugés conflictuels [WTD17]

$\Rightarrow$ Alternative : se laisser guider par les données en adoptant une approche *bottom-up*

Plan

Approche bottom-up : Vue d'ensemble

Objectif principal

Une typologie des interactions sur la Wikipedia, en identifiant les différents usages et pratiques par la communauté
Basée dans un premier temps sur les caractéristiques génériques externes des interactions (pas sur le contenu des échanges)

Sous-objectif

Identifier les discussions "standards", se rapprochant d'un forum d'échange entre utilisateurs extraire des discussions "normales" permettant des investigations plus fines sur les caractéristiques langagières des interactions

Méthode

Principe généraux

Partir des métadonnées, observer des tendances globales et des sous-classes de discussion
Dégager des cas extrêmes et/ou atypiques pour un examen détaillé montrant des pratiques inattendues

Démarche

Concentration sur les caractéristiques externes des discussions :

- Qui écrit ?
- Quand et sur quelle durée ?
- Quelle est la taille des échanges ?
- Qui répond ?

Vue d'ensemble

Au départ, 10 millions de discussions, 30 millions de messages sur les 3 langues.
Deux sources de bruit identifiées :

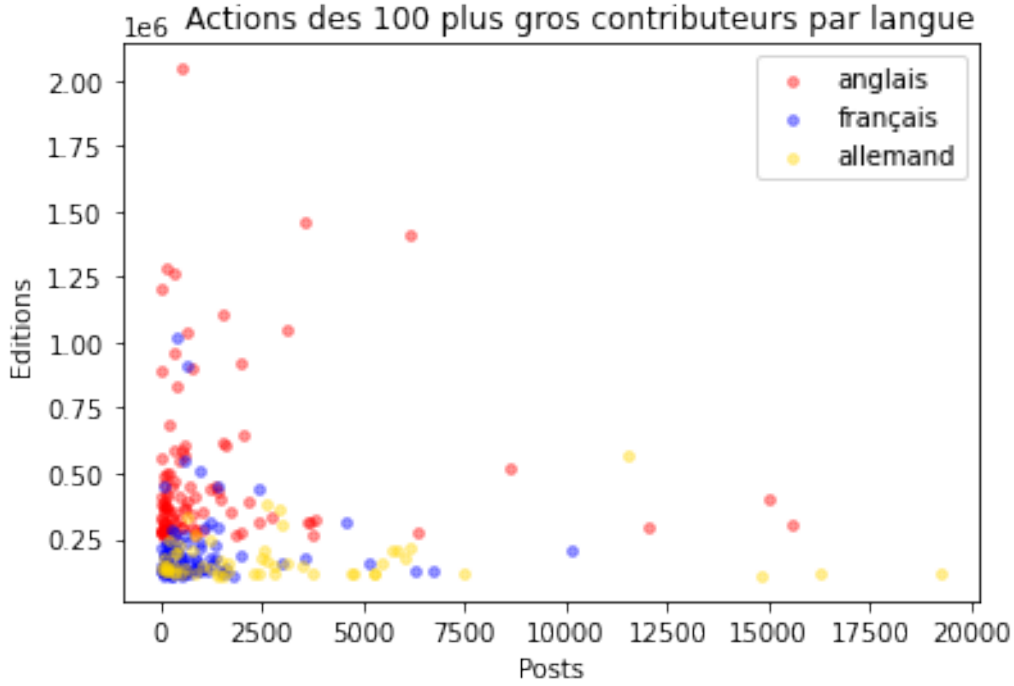
- Les robots (environ 5% des posts)
- Les emboîtages complexes des réponses dus au format Wiki

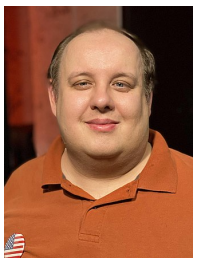
Sélection des seules discussions sans robots impliqués et sans segment isolé (passage indenté sans date ni signature).

	English	French	German	Total
Discussion pages	2 025 888	266 699	713 485	3 006 072
Threads	6 636 783	608 857	2 121 852	9 367 493
Posts/segments	20 025 945	1 832 416	6 938 168	28 796 529
Posts by bots (% of posts)	1 306 737 (6.5%)	55 868 (3.0%)	213 004 (3.1%)	1 575 609 (5.5%)
Stranded segments (% of posts)	3 232 039 (16.1%)	451 613 (24.6%)	824 151 (11.9%)	4 507 803 (15.7%)
Selected threads (% of threads)	3 385 583 (63.6%)	302 475 (49.8%)	1 485 648 (70.0%)	6 009 440 (64.1%)
Posts in selected threads	8 873 620	769 880	3 967 726	15 287 508

Auteurs et contributeurs

Croisement avec le nombre d'éditions réalisées sur les pages de la Wikipédia à partir de la liste des 1000 plus gros contributeurs par langue.
Pas de corrélation : les gros posteurs ne sont pas les gros éditeurs.





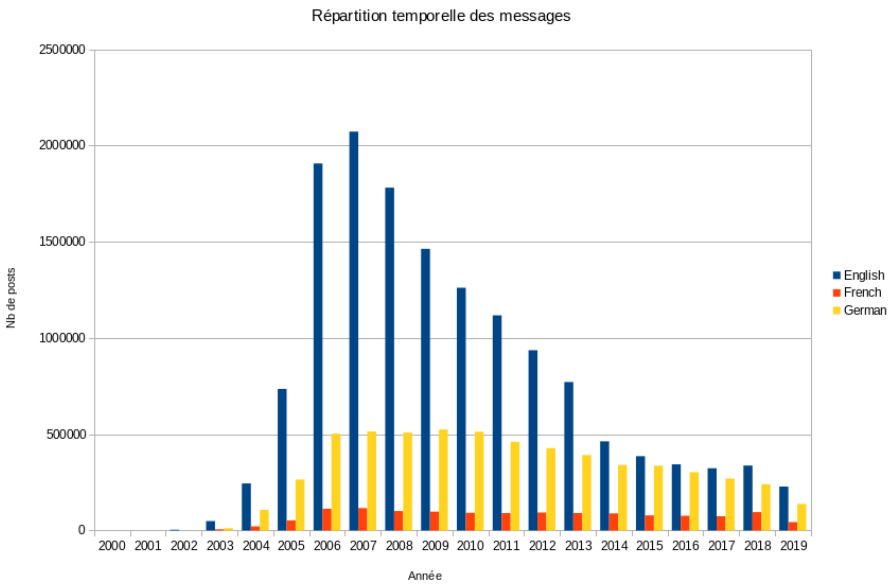
Steven Pruit (aka *Ser Amantio di Nicolao*) : 3 millions d'éditions en 2019 (5 millions aujourd'hui), aucun post sur les pages de discussion (quelques-uns sur les pages des utilisateurs).

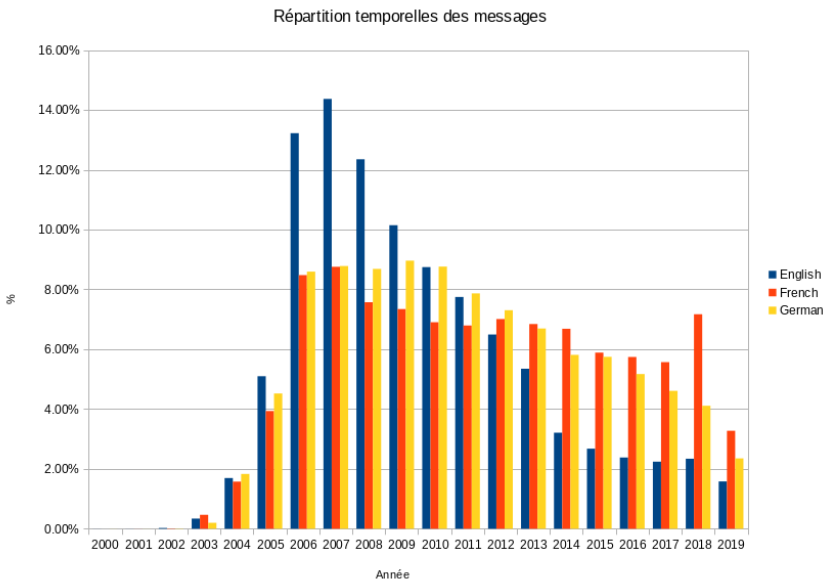
Zoom sur les principaux auteurs

English		French		German	
Will Beback	25078	Jean-Jacques Georges	10119	Kopilot	27126
Jayjg	24191	Racconish	8905	Phi	26348
Pmanderson	18143	Panam2014	6730	Jesusfreund	19255

- Anglais : WillBeback est un administrateur respecté, Jayjg est accusé d'être un pro-israëlien
- Français : JJ Georges "dictateur de Wikipedia", banni à jamais en 2019, Racconish souvent cité pour des modifications suite à des scandales
- Allemand : Kopilot et Phi font l'objet de plusieurs documentaires sur les pratiques problématiques de la Wikipedia (notamment autour du 11 septembre)

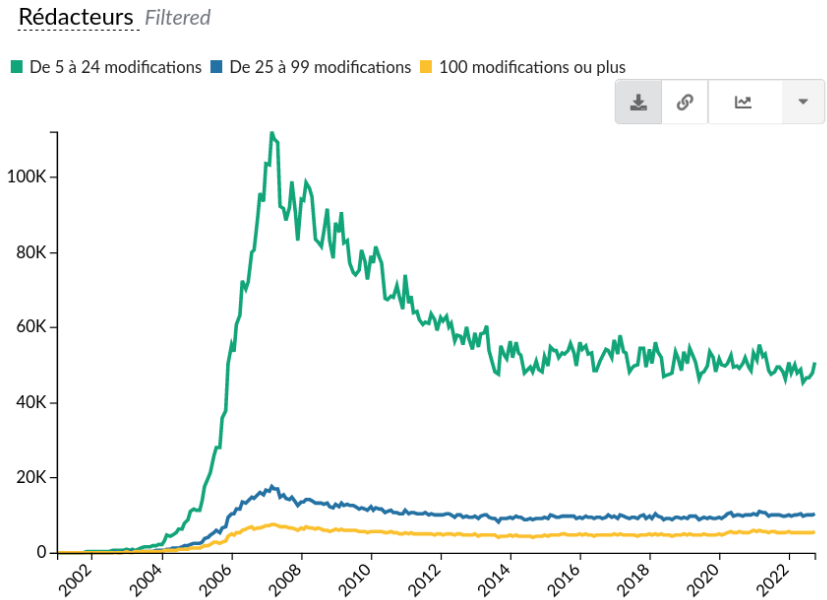
Répartition temporelle des messages





Une désaffection croissante depuis 2007, mais une bonne stabilité pour le français

Répartition temporelle des messages (2)



Une évolution qui suit celle de la croissance de Wikipedia, et du nombre de rédacteurs actifs.

Taille et durée des discussions

	English	French	German
Nombre de posts par discussion			
min, max	1-651	1-149	1-305
médiane, moyenne	2, 2.50	1, 2.54	2, 2.67
Nombre de participants par discussion			
min, max	1-97	1-43	1-120
médiane, moyenne	2, 1.88	1, 1.72	1, 1.88
Durée (au moins 2 posts)			
min, max	1 mn, 20 years,	1 mn, 16 years,	1 mn, 19 years
médiane, moyenne	5.3 days, 260 days	2.1 days, 184 days	6.1 days, 349 days

Discussions avec de nombreux messages

En anglais

Des listes de points à discuter un par un, par exemple :

- garder ou pas une section pour chacun des 134 personnages d'un manga (650 messages, 12 participants) URL
- critique méthodique des éditions d'un utilisateur : *Chris Tomic's edits : Recent edits to this article have introduced unnecessarily flowery language. I will address each problem I have with it below, point by point* (177 messages, 3 participants, football) URL Note : se termine par *"I hope you don't mind that I have reformatted them slightly to make the conversation easier to read."*

En français

Des discussions moins structurées, toutes très tendues et partant de sujets variés (lieu de naissance d'un homme politique, événement de la guerre en Syrie, taille d'un footballeur, etc.). Point Godwin souvent atteint.

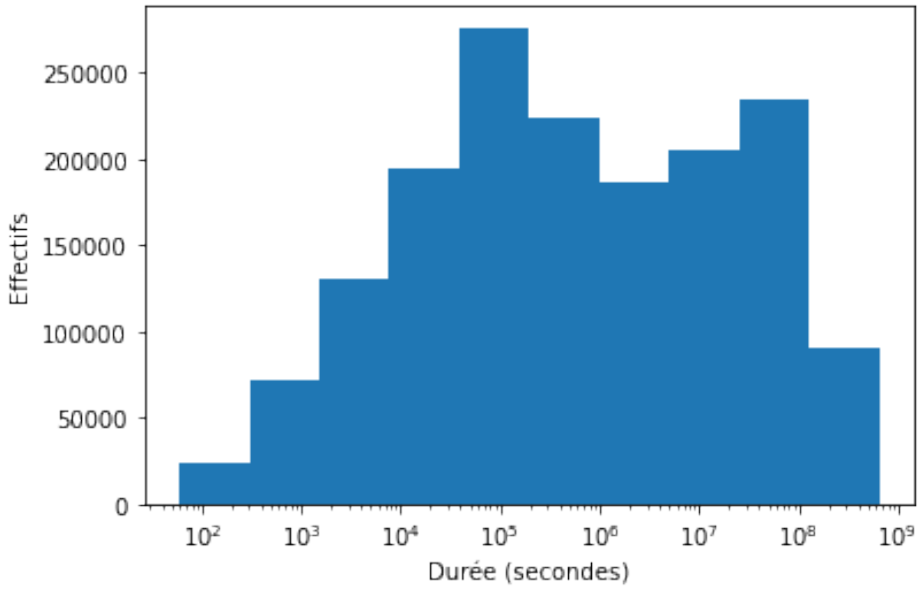
Discussions avec de nombreux participants

Toutes langues confondues

Essentiellement des *votes* sur des choix à faire pour la page : la supprimer, la fusionner, quel titre lui donner, quelle image choisir, etc.

- Donc des messages courts (pour/contre, A/B/C), mais dans certains cas les participants argumentent leur réponse et cela déclenche des discussions locales
- Des sujets variés mais généralement polémiques : statut de planète ou pas d'un corps céleste, désambiguïsation de *Georgia*, titre de la page sur les attaques du 11 septembre, quelles images mettre sur la page des caricatures de Mahomet, faut-il une page pour l'expression "E... de ta r..." etc.

Discussions qui durent dans le temps - méthodologie



Des milliers de discussions étalées sur plus de 10 ans (3×10^8 secondes)
Pas de caractéristiques génériques particulières (nombre de messages, nombre d'utilisateurs)
Variation importante des pauses entre les messages : certaines discussions sont "continues" avec des pauses courtes (2-3 ans)

Discussions qui durent dans le temps - en anglais

Une dizaine de cas de discussions régulières sur plus de 10 ans :

- URL (Discussion sur le titre d'une chanson, étalée de 2002 à 2019 avec des messages tous les 2-3 ans par des utilisateurs différents.)

Mais généralement un long silence (jusqu'à 15 ans). Des discussions "normales" sans marques spécifiques

- *URL Is it T-shirt or T-Shirt ? What's the correct capitalisation ? (2002)*
  I think T-shirt, or t-shirt. (2003)
  The whole virtue in using the letter T in the name is because the garment resembles a capital T (albeit with a rather fat middle). Whereas t-shirt suggests it ought to have a sideways tail-vent. (2019)
- *URL The original text said "fluid resistance". I replaced this with "viscosity". Is this correct ? (2003)*
  No. The fluid property called viscosity is [...] (2019)
- *URL* ***[13 years later...]*** *. o O (How is "turning a log file into an HTML report" not exactly text manipulation ? I have to wonder what this IP user* ***imagined*** *HTML documents are made of ?)*

Discussions qui durent dans le temps - français

Là aussi des réponses "normales" mais qui actent généralement le décalage temporel :

- *Mon commentaire vient 14 ans après ceux qui précèdent... Les modifications que je viens d'opérer sont censées répondre aux "objections" faites à l'époque.*
- *Je réponds 14 ans plus tard... Oui, il faut donner à Aron la place qui lui revient.*
- *À peine 850 résultats sur Google, c'est un peu léger(...) (2004)*
  Certes mais 15 ans plus tard certaines de leurs vidéos ont plus de 16 millions de vues (2019)

Des décalages justifiés :

- *(...) mais je pense que nous devons nous en tenir à la terminologie de la commission de toponymie de l'ONU (...) On peut changer de référence, mais pas avant un vrai débat général – 2003*
  Le nom officiel est maintenant Tchéquie ! Ce n'est pas ouvert au débat, ça ne nous appartient pas (...) – 2018

Plan

Principe

Représenter la structure d'une discussion

- Considérer la suite des messages suivant leur auteur
- Chaque participant différent est représenté par une lettre arbitraire : A pour celui qui initie la discussion, B pour le premier qui répond, etc.
- Pas de prise en compte des structures complexes, on se base sur la linéarité observée sur la page

Principe

Représenter la structure d'une discussion

- Considérer la suite des messages suivant leur auteur
- Chaque participant différent est représenté par une lettre arbitraire : A pour celui qui initie la discussion, B pour le premier qui répond, etc.
- Pas de prise en compte des structures complexes, on se base sur la linéarité observée sur la page

Exemples :

- ABAB :

Principe

Représenter la structure d'une discussion

- Considérer la suite des messages suivant leur auteur
- Chaque participant différent est représenté par une lettre arbitraire : A pour celui qui initie la discussion, B pour le premier qui répond, etc.
- Pas de prise en compte des structures complexes, on se base sur la linéarité observée sur la page

Exemples :

- ABAB : dialogue
- AAA :

Principe

Représenter la structure d'une discussion

- Considérer la suite des messages suivant leur auteur
- Chaque participant différent est représenté par une lettre arbitraire : A pour celui qui initie la discussion, B pour le premier qui répond, etc.
- Pas de prise en compte des structures complexes, on se base sur la linéarité observée sur la page

Exemples :

- ABAB : dialogue
- AAA : monologue
- ABCDEFGHIJKL :

Principe

Représenter la structure d'une discussion

- Considérer la suite des messages suivant leur auteur
- Chaque participant différent est représenté par une lettre arbitraire : A pour celui qui initie la discussion, B pour le premier qui répond, etc.
- Pas de prise en compte des structures complexes, on se base sur la linéarité observée sur la page

Exemples :

- ABAB : dialogue
- AAA : monologue
- ABCDEFGHIJKL : vote/consultation
- ABCBCDEFA :

Principe

Représenter la structure d'une discussion

- Considérer la suite des messages suivant leur auteur
- Chaque participant différent est représenté par une lettre arbitraire : A pour celui qui initie la discussion, B pour le premier qui répond, etc.
- Pas de prise en compte des structures complexes, on se base sur la linéarité observée sur la page

Exemples :

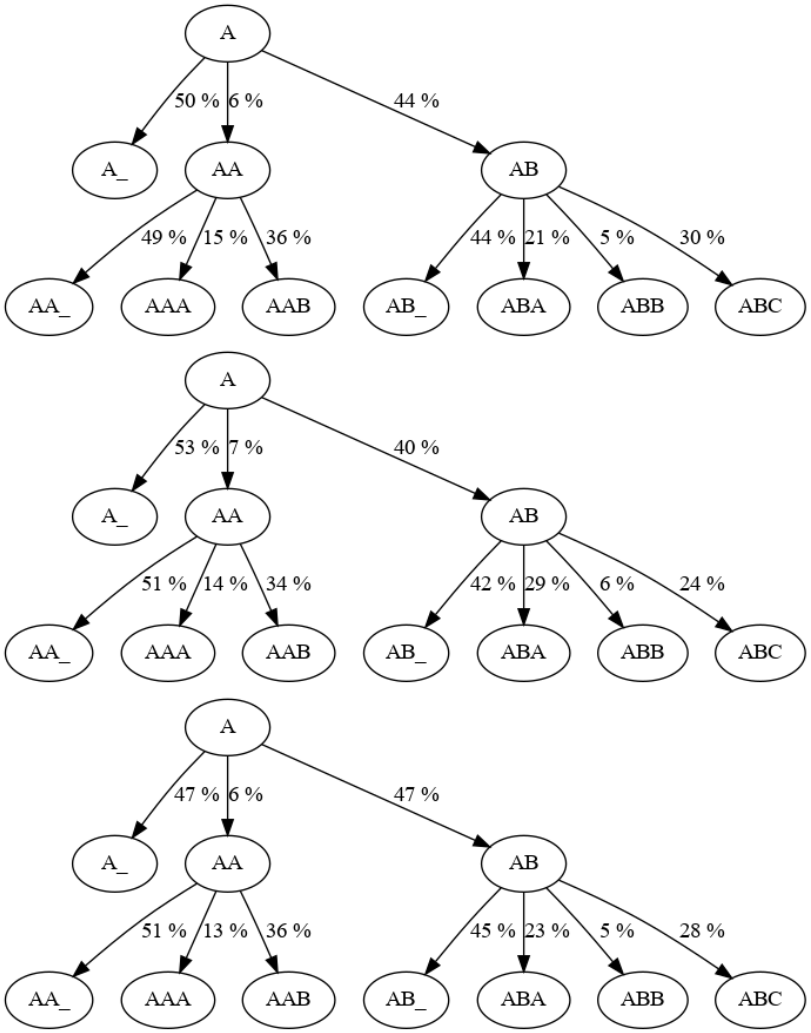
- ABAB : dialogue
- AAA : monologue
- ABCDEFGHIJKL : vote/consultation
- ABCBCDEFA : c'est plus compliqué

Schémas de discussions - statistiques

	English	French	German
Nombre de discussions	3 385 583	302 890	1 485 648
Nombre de schémas différents	116 893	13 103	54 834

English (%)		French (%)		German (%)	
A	40.2	A	53.4	A	47.2
AB	35.1	AB	16.8	AB	21.1
ABC	4.2	ABA	3.8	ABC	5.1
AA	2.3	AA	3.4	ABA	3.3
ABA	2.2	ABC	3.2	AA	2.8
ABCD	1.2	ABAB	1.3	ABCD	1.4
ABAB	0.7	ABCD	0.9	ABAB	1.2
ABB	0.6	ABB	0.8	ABB	1.1
AAB	0.6	AAB	0.7	AAB	0.7
ABAC	0.5	ABABA	0.6	ABAC	0.7

Début de discussion



Quelques spécificités du français :

- Plus faible probabilité d'avoir une réponse
- Plus forte probabilité de répliquer après une réponse
- Plus faible probabilité d'avoir un troisième participant

Les monologues : ampleur du phénomène

	English	French	German
Nb de discussions	3 385 583	302 475	1 485 648
Nb de monologues	1 812 457	174 009	749 456
Part de monologues	53.5 %	57.4 %	50.4 %
Dont monologues avec plus de 2 messages	6.4 %	6.9 %	6.5 %
Part de monologues dans les discussions avec plus de 2 messages	6.9 %	8.5 %	6.2 %

Un phénomène fréquent, stable et étonnant.

Typologie des monologues

En se basant sur les cas extrêmes (longueur et durée), plusieurs cas de figure identifiés :

- Cas le plus fréquent : liste des modifications à apporter, et indication que c'est fait (nouveau message). Dans certains cas, il semble que l'auteur attende une réponse, mais finit par faire les modifications et les acter
- Sur la durée : l'utilisateur acte l'absence de réponse :
  Voyez-vous une opposition au changement de l'infobox ?
  Sans réponse depuis trois ans, j'ai changé l'infobox.
- Les "posteurs en série". Par exemple, 84 discussions monologales avec comme titre "voix française" : un utilisateur liste les voix françaises doublant un acteur, film par film, au fil du temps. Un autre utilisateur finira par lui suggérer un peu sèchement d'arrêter de faire ça...
- Des questions ou appels répétés comme ici : une série de questions adressées à un même utilisateur (et des remarques sur son refus de discuter)

Les dialogues

Environ 27 % des discussions, mais 55% des discussions avec plus d'un message

English (%)		French (%)		German (%)	
AB	70.6	AB	61.0	AB	68.9
ABA	10.1	ABA	14.0	ABA	10.8
ABAB	3.3	ABAB	4.9	ABAB	3.9
ABB	2.9	ABB	3.0	ABB	3.6
AAB	2.7	AAB	2.6	AAB	2.3
ABABA	1.5	ABABA	2.3	ABABA	1.8
ABAA	0.9	ABAA	1.2	ABAA	0.9

Des schémas stables sur les trois langues

	English	French	German
Taux de dialogues classiques (alternance A/B)	87.1 %	84.1 %	87.2 %

Exploration des divergences de l'alternance A/B

- A+B : Série de commentaires faits par A, en plusieurs messages, simultanés ou non, puis réponse de B ou ajout d'un commentaire à la suite
- A+B : Série de demandes (changements à faire, appel à aide sur des questions), ajout d'une demande supplémentaire par B
- AB+ : Un commentaire global de A, puis un ensemble de réponses/critiques par B, point par point
- AB+ : Une demande globale de A (traduire, sourcer, vérifier), puis un ensemble d'actions par le même B
- AB+ : Une question de A, puis plusieurs réponses (arguments, sources) par B, au fur et à mesure qu'il les trouve
- A+B+ : A fait plusieurs remarques/demandes. B répond par autant de messages, à la suite. URL
- A+B+A+B+ : cas complexes d'interaction par séries de messages, par exemple une réponse + une citation ou une traduction à chaque échange. URL

Quelques remarques sur les polylogues

Si on ne considère que les discussions à plus de 2 messages :

	English (%)		French (%)		German (%)	
Taux de polylogues		69.3		57.2		65.5
3 premiers messages	ABC	47.6	ABA	43.4	ABC	44.9
	ABA	33.9	ABC	36.0	ABA	37.3
	AAB	7.8	AAB	8.6	ABB	8.3
	ABB	7.6	ABB	8.5	AAB	6.9
	AAA	3.1	AAA	3.6	AAA	2.6

Différence de comportement en français : moins de chance de voir un troisième locuteur intervenir

Le cas des ABC

Un schéma relativement improbable dans une conversation normale.
Proposition de typologie du rôle de C :

- apporter une réponse alternative à celle de B
- supporter A ou B en cas de désaccord (avec ou sans argument)
- répondre à une question posée par B
- acter un changement fait à la suite du consensus de A et B
- faire un commentaire général, mettre en perspective, rappeler la pratique de WP
- quelques cas complexes identifiés :
  - A et C sont le même utilisateur sous deux noms
  - A est posté après B et C et comment la discussion qui suit
  - le message de C n'a pas de rapport identifiable avec ce qui précède

Plan

Conclusion

- Les pages de discussion WP offrent une fenêtre inédite sur les coulisses des si populaires articles WP
- Les pages de discussion WP restent des objets complexes qui remettent en cause les modèles et méthodes traditionnels utilisés en linguistique et TAL, en commençant par leur acquisition et filtrage
- Interroger les pages de discussion WP nécessite la prise en compte de différents niveaux : l'identité des contributeurs, la temporalité, la non-linéarité
- Nous avons choisi de privilégier les caractéristiques extra-linguistiques comme point d'entrée dans des données massives et multidimensionnelles
- Les techniques de fouille permettent d'identifier des phénomènes pour lesquels il est nécessaire de recourir à des analyses locales, qualitatives et manuelles
- Des hypothèses vont pouvoir ensuite être testées qui impliqueront le contenu langagier

Conclusion

Quelques premières découvertes néanmoins :

- Les comportements globaux semblent stables d'une langue à l'autre et dans le temps
- La forme des pages de discussion permet une temporalité très longue
- ... mais est une source de déstructuration problématique pour l'analyse
- Les pages de discussion sont le lieu d'usages variés :
  - Des discussions "classiques" : consultation, demande d'information, remarques, critiques
  - Un tableau blanc avec la liste des choses faites ou à faire
  - Pas toujours dans l'idée d'une interaction (plutôt un pense-bête)

Nos pistes d'investigation actuelles

- Caractérisation linguistique des schémas emergeants
- Étude des titres de fil (mémoire M1 LITL)
- Étude des fin de fil (mémoire M1 LITL)
- Annotation des actes de langage (e.g. [FGC12])
- Focus sur des types spécifiques d'interaction et des thématiques particulières
  - Débats terminologiques, e.g. gravissimes vs. mortels, Maladie Multigénique)
  - À propos de la neutralité et du conflit : quels sont les sujets les plus controversés ? contropedia
- Comment sont invoqués les principes WP, e.g. la neutralité (mémoire M1 LITL)

Marcia W DiStaso.

Measuring public relations wikipedia engagement : How bright is the rule.
Public Relations Journal, 6(2) :1–22, 2012.

Oliver Ferschke, Johannes Daxenberger, and Iryna Gurevych.

A survey of nlp methods and resources for analyzing the collaborative writing process in Wikipedia.
In *The People's Web Meets NLP : Collaboratively Constructed Language Resources*. Springer, 2013.

Oliver Ferschke.

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

Oliver Ferschke, Iryna Gurevych, and Yevgen Chebotar.

Behind the article : Recognizing dialog acts in wikipedia talk pages.
In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 777–786. Association for Computational Linguistics, 2012.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and PVS Avinesh.

A corpus factory for many languages.
In *LREC*, 2010.

Pierre-Carl Langlais.

La négociation contre la démocratie : le cas wikipedia.
Négociations, (1) :21–34, 2014.

Olena Medelyan, David Milne, Catherine Legg, and Ian H Witten.

Mining meaning from wikipedia.
International Journal of Human-Computer Studies, 67(9) :716–754, 2009.

Martin Potthast, Benno Stein, and Robert Gerling.

Automatic vandalism detection in wikipedia.
In *Advances in Information Retrieval*, pages 663–668. Springer, 2008.

Gilles Sahut.

Wikipédia, une encyclopédie collaborative en quête de crédibilité : le référencement en questions.
PhD thesis, Université Toulouse Jean Jaurès ; Université de Toulouse, 2015.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon.

Ex machina : Personal attacks seen at scale.
In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399. International World Wide Web Conferences Steering Committee, 2017.

Torsten Zesch, Christof Müller, and Iryna Gurevych.

Extracting lexical semantic knowledge from wikipedia and wiktionary.
In *LREC*, volume 8, pages 1646–1652, 2008.