The WikiDisc corpus: in the backstage of Wikipedia

Lydia-Mai Ho-Dac

The WikiDisc corpus: in the backstage of Wikipedia

Lydia-Mai Ho-Dac

University of Toulouse, CLLE-ERSS, CNRS

June 2019, IDS – Mannheim

Université de Toulouse

CLLE ERSS

Ass. Prof. and Researcher at the University of Toulouse in

- **Linguistics** :
  - The study of discourse organization (*how human build and structure "text worlds" via documents*)
  - Text genres and text types characterization (*The better you know the kind of text you read, the better you understand and process it*)
- **Computational Linguistics – CL** : using computer for studying discourse organization and characterizing text genres and text types
- **Natural Language Processing – NLP** : injecting linguistic knowledge in NLP for improving applications such as information extraction, information retrieval, automatic classification, etc.

Among these areas of research : CMC genres and Wikipedia

- CMC (Computer-Mediated Communications) include a variety of texts written through computers and smartphones such as chatting, texting, writing via online fora, Twiter, Whatsapp, Facebook, etc.
- CMC are producing more and more textual data
- CMC involve new/different linguistic usages (asynchronous communications, between(?) oral and written genres, using new technology devices)
- CMC consist in interactions between users, interactions are structured in threads, CMC may be seen as a "bag" of threads
- Challenge : describe these new kinds of interaction
  - understand how knowledge is sharing and text worlds are building
  - improve NLP when confronting to these kind of texts (e.g. Info. Extraction in Health Fora)
- Focus on text genres and text types characterization of the CMCs : (*The better you know the kind of thread you read, the better you understand and process it*)

Plan

Plan

A chance for sharing and recording humanity knowlegde

1- The *wiki* technology

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

2- Grand projects of free online encyclopedia [Sah15]

Collaborative, free, open copyright, online, written by voluntary authors and checked by voluntary editors

- 1993 : ***Interpedia*** (*The **Inter**net Encyclo**pedia***, R. Gates) free online collaborative encyclopedia, but no collaborative tool for writing
- 1997 : ***Distributed Encyclopedia*** (U. Fuchs, future German Wikipedian) against the "growing importance of the market sphere on the web"
- 1999 : ***GNUpedia*** and *GNE's Not an Encyclopedia* (R. Stallman, GNU project and free software) multilingual encyclopedia aiming at "preserving the human knowledge open and free for everybody"
- 2000 : ***Nupedia*** (J. Wales, a trader), a collaborative advertising-supported encyclopedia written by "experts"

Nupedia, the mother of Wikipedia

the native project Nupedia

- **Jimmy Wales** :
  - @Jimbo
  - founder of the *Bomis* society working on managing advertising-supported *pornographic* web sites
  - planning to build the "free" encyclopedia *Nupedia*, advertising-supported with the help of the Bomis society
- **Larry Sanger** : Dr of Philosophy, hired by Wales as editor-in-chief in the *Nupedia* project
- 2000 : ***Nupedia*** : collaborative advertising-supported encyclopedia written by "experts" (with at least a PhD). "Written by experts" implies "a peer-reviewed process" (as in the academic world)

Nupedia, the mother of Wikipedia

the native project Nupedia

- **Jimmy Wales** :
  - @Jimbo
  - founder of the *Bomis* society working on managing advertising-supported *pornographic* web sites
  - planning to build the "free" encyclopedia *Nupedia*, advertising-supported with the help of the Bomis society
- **Larry Sanger** : Dr of Philosophy, hired by Wales as editor-in-chief in the *Nupedia* project
- 2000 : ***Nupedia*** : collaborative advertising-supported encyclopedia written by "experts" (with at least a PhD). "Written by experts" implies "a peer-reviewed process" (as in the academic world)

Birth of Wikipedia – WP

Birth of English WP – WP[EN]

The ***Nupedia*** peer-review process : two experts per submission ; once the submission reviewed, the two experts have to submit for acceptance a revised version of the submission to all the Nupedia members, for the final revision and acceptance

- This burdensome process → very few articles
- WikiWikiWeb as a solution for facilitating this process
- **15.01.2001, WP[EN] was launched** (Wiki(Nu)pedia), as the "draft side" of *Nupedia*.

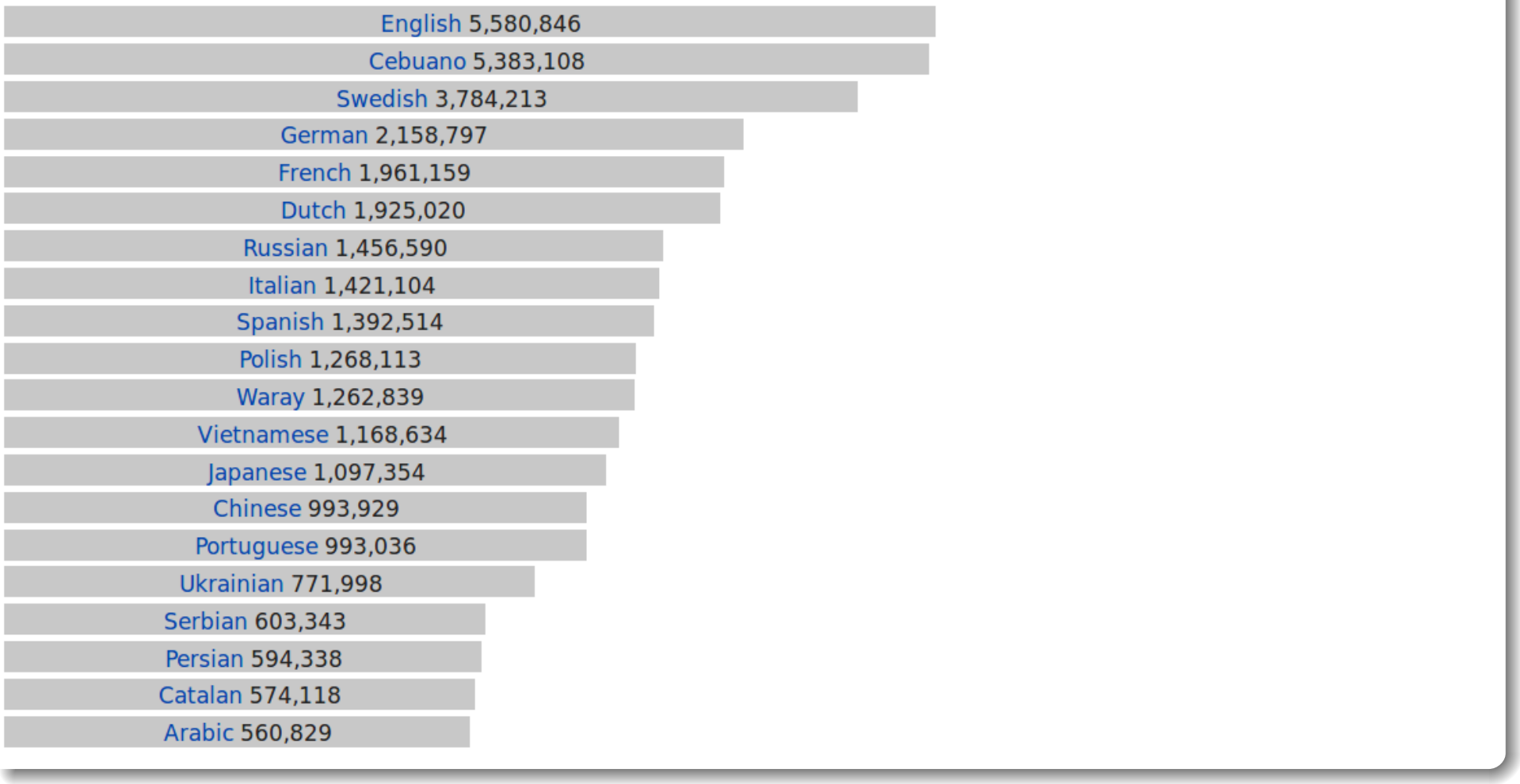
> *"Originally it was the Nupedia Wiki - our idea was to use it as an article incubator for Nupedia. Articles could begin life on this wiki, be developed collaboratively and, when they got to a certain stage of development, be put it into the Nupedia system." (Sanger in 2006* `https://www.theguardian.com/technology/2006/jul/13/media.newmedia` *)*

First steps of Wikipedia – the WP spike

02.2001	600 articles (drafts) in WP[EN]
03.2001	1300
05.2001	3900
01.2002	20,000
09.2003	more than 100,000 articles *Nupedia* was abandoned (with only 24 accepted and published articles)
2018	WP is the 5th most visited web site just after *Google*, *Facebook*, *Youtube*, *Baidu* with 5,184,686 views per hours for WP[EN]. more than 50 million articles, about 300 different languages

Nowadays, a global phenomenon (amount of articles)

https://meta.wikimedia.org/wiki/List_of_Wikipedias



Plan

A research subject of a huge importance

http://wikipapers.referata.com

"a compilation of resources [...] focused on the research of wikis"

Create account Log in

Page Discussion Read Edit View history Search

List of publications

See also: List of authors, List of datasets, List of tools.

This is a **list of publications** available in WikiPapers. Currently, there are 6246 publications.

Filter by type:

- List of books (27) and List of book chapters (45)
- List of conference papers (4034)
- List of journal articles (1541)
- List of literature reviews (73)
- List of bachelor's theses (11), diploma theses (1), doctoral theses (52), master's theses (26)
- List of essays (11)
- List of peer-reviewed publications (663) and List of non peer-reviewed publications (17)
- List of magazine articles (30)
- List of unpublished works (4)

Filter by year:

- 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014

Filter by language:

- Arabic, Catalan, Chinese, Dutch, English, French, Galician, German, Greek, Hungarian, Italian, Japanese, Polish, Portuguese, Russian, Slovenian, Spanish, Turkish

Filter by conference:

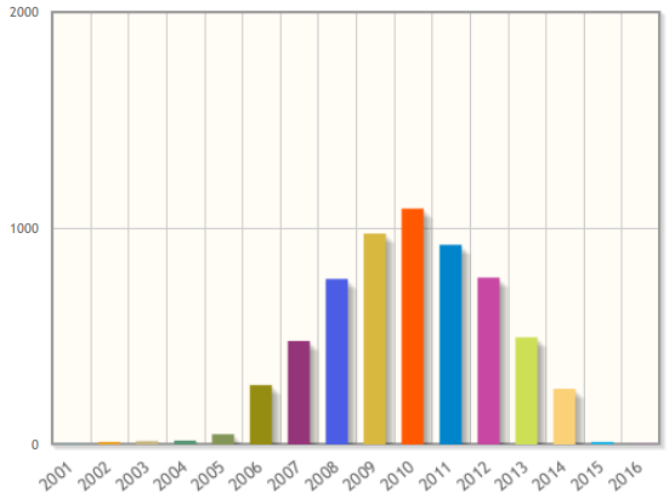
- CLEF, MathWikis, WikiAI, WikiSym, WikiViz

Main page
Publications
Keywords
Authors
Datasets
Tools
Examples
...More lists...
Random page

Create new...
Publication
Author
Event
Keyword
Dataset
Journal
Tool

Activity
Community portal
Recent changes
New pages
RSS feeds
Follow us on Twitter!

Publications distribution by year



(More trends stats)

Plan

A *wikicracy* i.e. "democracy by consensus"

An observatory for human collaboration

- collaborating for free in a democracy by consensus (vs. by the majority)
- the "democracy of the future" : a perpetual rethinking without establishment, a collaborative building with an "open government"...
- a lot of benevolence with some bad (toxic) behavior (vandalism, personal attacks, ..)

Wikimedia (2009). Wikicracy. Retrieved on 4 March 2009 from

`http://meta.wikimedia.org/w/index.php?title=Wikicracy&oldid=1406941`

Wales' mail (13.06.2001) http://lists.wikimedia.org/pipermail/wikipedia-l/2001-June/000187.html

Probably the most astounding fact about Wikipedia is that it is so good without any formal rules or restrictions at all. There are social customs and social pressures that do a really good job of keeping things in line.

Plan

For Natural Language Processing – NLP

Exploiting articles

- Knowledge extraction [ZMG08, MMLW09]
- Multilingual resources building [KRPA10]

Exploiting article history and edits

- Writing process analysis and modeling [FDG13, BWC+15]
  - *diff* between revisions for extracting spelling variants (spell checker), paraphrases (information retrieval), simplifications and summarization (for developing automatic processing) ⇒ `http://contropedia.net/`
- Vandalism detection (about 7% of edits in WP[EN] [PSG08])

Exploiting the forums for discussion

- Negotiating process analysis an modeling [FGC12, FDG13]
- Disagreement, Conflict and Controversial topics detection [HDLPT17]

Plan

A Corpus Factory for Many Languages [KRPA10]

A	Arabic (Arabic web corpus)
B	Basque (basque_WaC) Bengali (bengaliWaC) Bosnian (bosnianWaC14)
C	Cantonese (Cantonese WaC) Chinese (ChineseTaiwanWaC) Croatian (hrWaC, hrWaC_10M)
D	Danish (danishWaC) Dutch (Dutch web corpus, nlWaC, nlWaC_1)
E	English (pukWaC, ukWaC, ukWaC_1, ukWaC_10M, ukWaC_10M_1, ukWaC2, ukWaC2_1, ukWaC3, ukWaC_mcd, ukWaCsst)
F	Filipino (filipinoWaC) Finnish (finnishWaC) Frisian (frisianWaC) French (frWaC, frWaC1_1)
G	Georgian (georgianWaC) German (deWaC, Parsed DeWaC (sDeWaC)) Greek (gkWaC) Gujarati (gujarathiWaC)
H	Hebrew (hebWaC) Hindi (hindiWaC, hindiWaC3)
I	Igbo (igboWaC) Indonesian (indonesianWaC) Italian (itWaC)
J	Japanese (jpWaC, jpWaC_10M, jpWaC2)
K	Korean (koreanWaC) Kannada (Kannada WaC)
L	Latin (latinWaC, latinWaC2) Latvian (latvianWaC, latvianWaC_shallow) Lithuanian (lithuanianWaC, lithuanianWaC_v2, lithuanianWaC_v2_10M)
M	Malay (malayalamWaC, malaysianWaC2) Maltese (malteseWaC, malteseWaC2, malteseWaC2_sample) Maori (maoriWaC)
N	Nepali (nepaliWaC) Norwegian (norwegianWaC)
P	Persian (WBC-Per) Polish (Polish Web Corpus)
R	Romanian (romanian_WaC) Russian (Russian Web Corpus)
S	Samoan (SamoanWaC) Serbian (serbianWaC, serbianWaC14, srWaC, srWaC22M) Setswana (setswanaWaC, setswanaWaC2) Spanish (Spanish wen corpus) Swahili (swahiliWaC, swahiliWaC_1) Swedish (swedishWaC, swedish_WaC, swedish_WaC_10M)
T	Tamil (tamilWaC) Tatar (Tatar Sample) Telugu (teluguWaC, teluguWaC2) Thai (thaiWaC) Turkish (turkishWaC, turkishWaC2, turkishWaC2_1, turkishWaC2_1_s, turkishWaC2_1_uniattr)
U	Urdu
V	vietnameseWaC2 (Viatnamese)
W	Welsh (welshWaC)
Y	Yoruba (Yoruba web corpus)

WP for Linguistic Corpus-Studies

- Enclyclopedia discourse as a genre, from the Enlightenment to Wikipedia
- CreativeCommons long exporsitory texts available for linguistic annotation (document structure, headings, enumeration, complex referential chaining, etc... e.g. [AMB+17])
- Topic evolution through a thread, across languages, since almost 20 years !
- Linguistic features of interactions (politeness, implicature, appraisal, argumentation, etc.)

WP for Linguistic Corpus-Studies

- Enclyclopedia discourse as a genre, from the Enlightenment to Wikipedia
- CreativeCommons long exporsitory texts available for linguistic annotation (document structure, headings, enumeration, complex referential chaining, etc... e.g. [AMB+17])
- Topic evolution through a thread, across languages, since almost 20 years !
- Linguistic features of interactions (politeness, implicature, appraisal, argumentation, etc.)

$\Rightarrow$ exploring the WP Talk Pages

Plan

What are the WP talk pages

Nicht angemeldet Diskussionsseite Beiträge Benutzerkonto erstellen Anmelden

Artikel | Diskussion | Lesen | Bearbeiten | Quelltext bearbeiten | Versionsgeschichte | Wikipedia durchsuchen

Koordinaten: 49° 29′ N, 8° 28′ O

Mannheim

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter Mannheim (Begriffsklärung) aufgeführt.

Die Universitätsstadt **Mannheim** (kurpfälzisch *Mannem* [manəm],[2] auch *Monnem*) ist ein Stadtkreis mit 307.997 Einwohnern (31. Dezember 2017)[3] im Regierungsbezirk Karlsruhe in Baden-Württemberg. Sie ist nach Stuttgart und Karlsruhe die drittgrößte Stadt des Landes. Die ehemalige Residenzstadt (1720–1778) der Kurpfalz mit ihrem stadtprägenden Barockschloss, einer der größten Schlossanlagen der Welt, bildet das wirtschaftliche und kulturelle Zentrum der Metropolregion Rhein-Neckar mit 2,35 Millionen Einwohnern. Mannheim liegt unmittelbar im Dreiländereck mit Baden-Württemberg, Rheinland-Pfalz und Hessen. Von seiner rheinland-pfälzischen Schwesterstadt Ludwigshafen am Rhein, mit der es ein zusammenhängendes Stadtgebiet bildet, ist Mannheim durch den Rhein getrennt.

Erstmals 766 im Lorscher Codex urkundlich erwähnt, erhielt Mannheim 1607 die Stadtprivilegien, nachdem Kurfürst Friedrich IV. von der Pfalz den Grundstein zum Bau der Festung Friedrichsburg gelegt hatte. Das damals für die mit der Festung verbundene Bürgerstadt Mannheim angelegte gitterförmige Straßennetz mit Häuserblöcken statt Straßenzügen ist in der Innenstadt bis heute erhalten geblieben. Darauf ist die bis heute geläufige Bezeichnung *Quadratestadt* zurückzuführen.

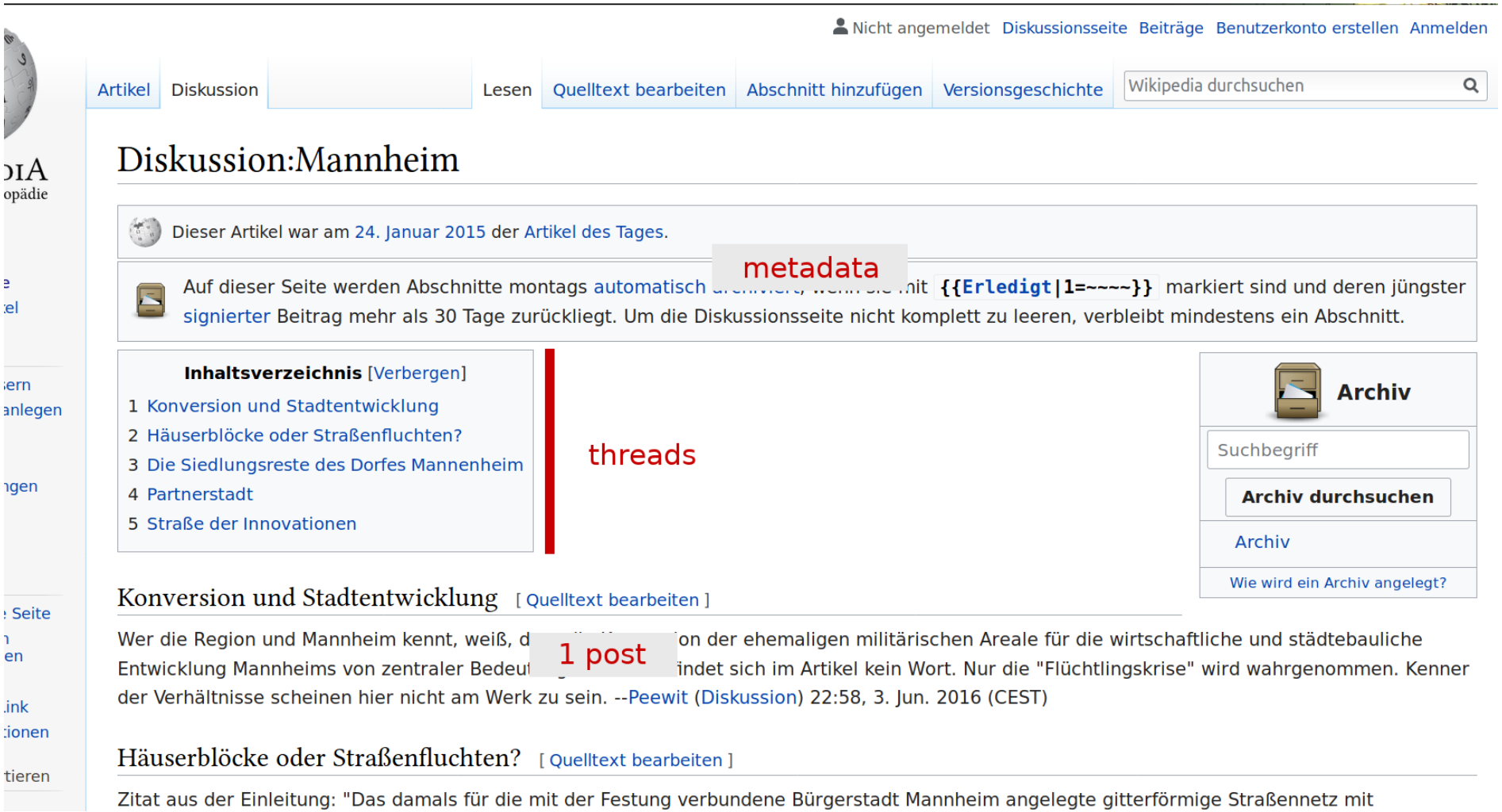
Seit 1896 Großstadt, ist Mannheim heute eine bedeutende Industrie- und Handelsstadt, Universitätsstadt und wichtiger Verkehrsknotenpunkt zwischen Frankfurt am Main und Stuttgart, unter anderem mit einem ICE-Knotenpunkt, dem zweitgrößten Rangierbahnhof Deutschlands und einem der bedeutendsten Binnenhäfen Europas.

Wappen	Deutschlandkarte

Basisdaten	
Bundesland:	Baden-Württemberg
Regierungsbezirk:	Karlsruhe
Höhe:	97 m ü. NHN
Fläche:	144,96 km²
Einwohner:	307.997 *(31. Dez. 2017)*[1]
Bevölkerungsdichte:	2125 Einwohner je

An online discussion behind the article...

... associated with metadata and containing threads and (dates and signed) posts



Birth, first steps and influence of non English WPs

Few days after the WP[EN] birth, a Wikipedian raises the following question : what to do with discussions behind the articles ?

- WP[EN] : moderation by consensus with the help of the "Benevolent Dictator" : Wales (as a super editor)
- Wales : WP as a topic must be discussed in another place (e.g. on mailing lists) :

  > *Wikipedia is an encyclopedia. The topic of Wikipedia articles should always look outward, not inward at Wikipedia itself.*

- No place for discussion

In German, French and Italian first, a need for discussion

Mostly because the "Benevolent Dictator" speaks only English, developing WP in other Languages require a forum for discussion and negotiation (vs. consensus and decision)[Lan14]

In German, French and Italian, another way to moderate

1. WP[DE] (March 2001) use of forums called *Meinungsbilder* (Meinung – opinion) for clarifying issues for which there is no consensus
2. WP[FR] (March 2001) In October 2002, Florence Dévouard (@Anthère) created a page called "decision-making" where "the final choice will depend on a vote instead of a simple consensus"
3. WP[IT] (May 2001) introduction of forums called *Sondaggio* "easy, quick and simple solution for resolving problems"

WP Talk Pages, the other side of WP

Exploiting the forum for discussion

- The WP talk pages : Online discussions associated with each article where Wikipedian can discuss the ongoing writing process with other Wikipedian
- Computer-Mediated Communications – CMC

> *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.*

See [Fer14, p. 111]

Plan

The WikiDisc Corpus building

Collecting WP talk pages

The WikiDisc Corpus [HDL15] : talk pages extracted from the WP[FR] Wikipedia snapshot (*dump*) global backup file

frwiki-20181020-pages-meta-current.xml.bz2 available on

https ://dumps.wikimedia.org/enwiki/20181020/)

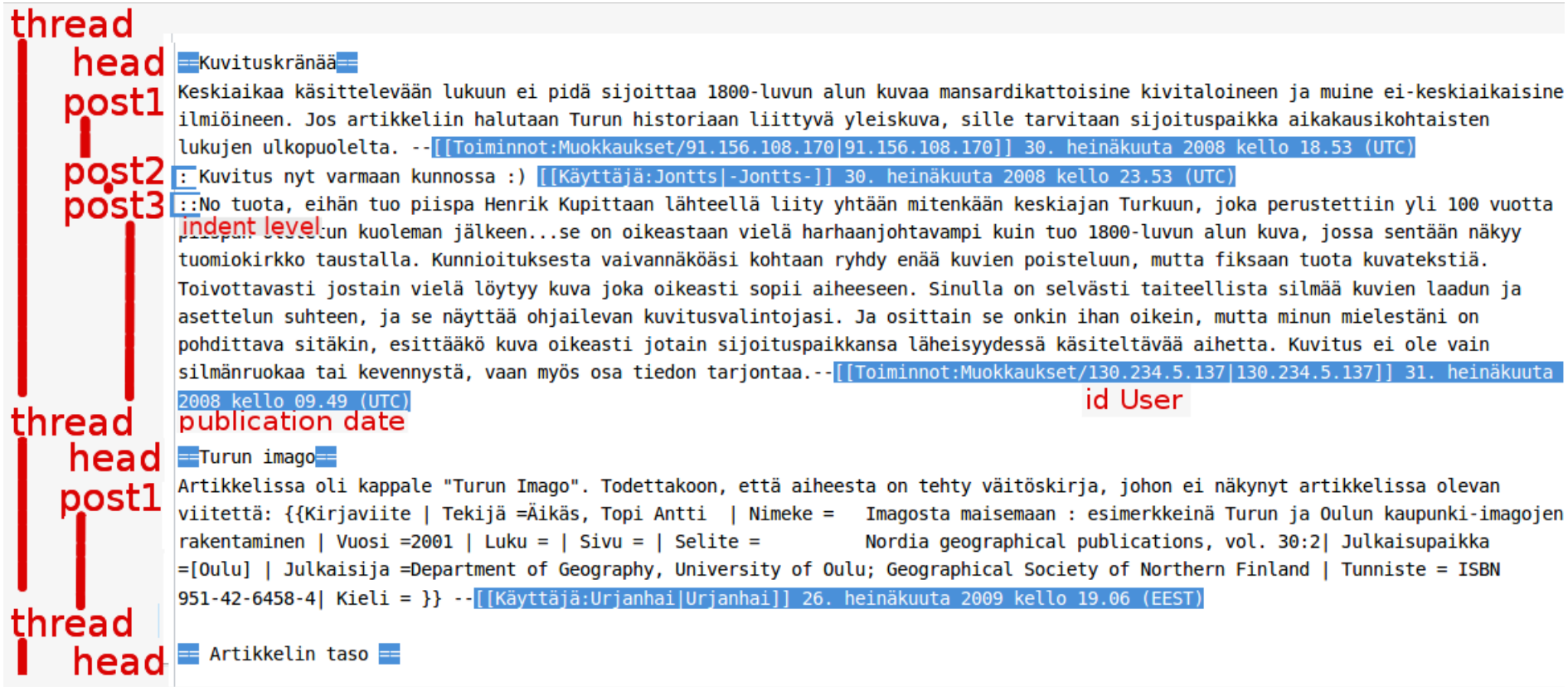
Document structure of a talk page

Talk pages are structured into threads and posts delimiting more or less explicitly in the *wikicode* (the wiki traditional syntax)

- Threads correspond to division delimited by (sub)headings signaled with /==.*?==/ in the wikicode
- Posts are delimited by
  1. timestamp and eventually user signature
  2. a change of indent level indicated with zero, one or more semi-colon (:) at the beginning of the post.

Talk Page behind a talk page

Corresponding Wikicode



The WikiDisc Corpus building

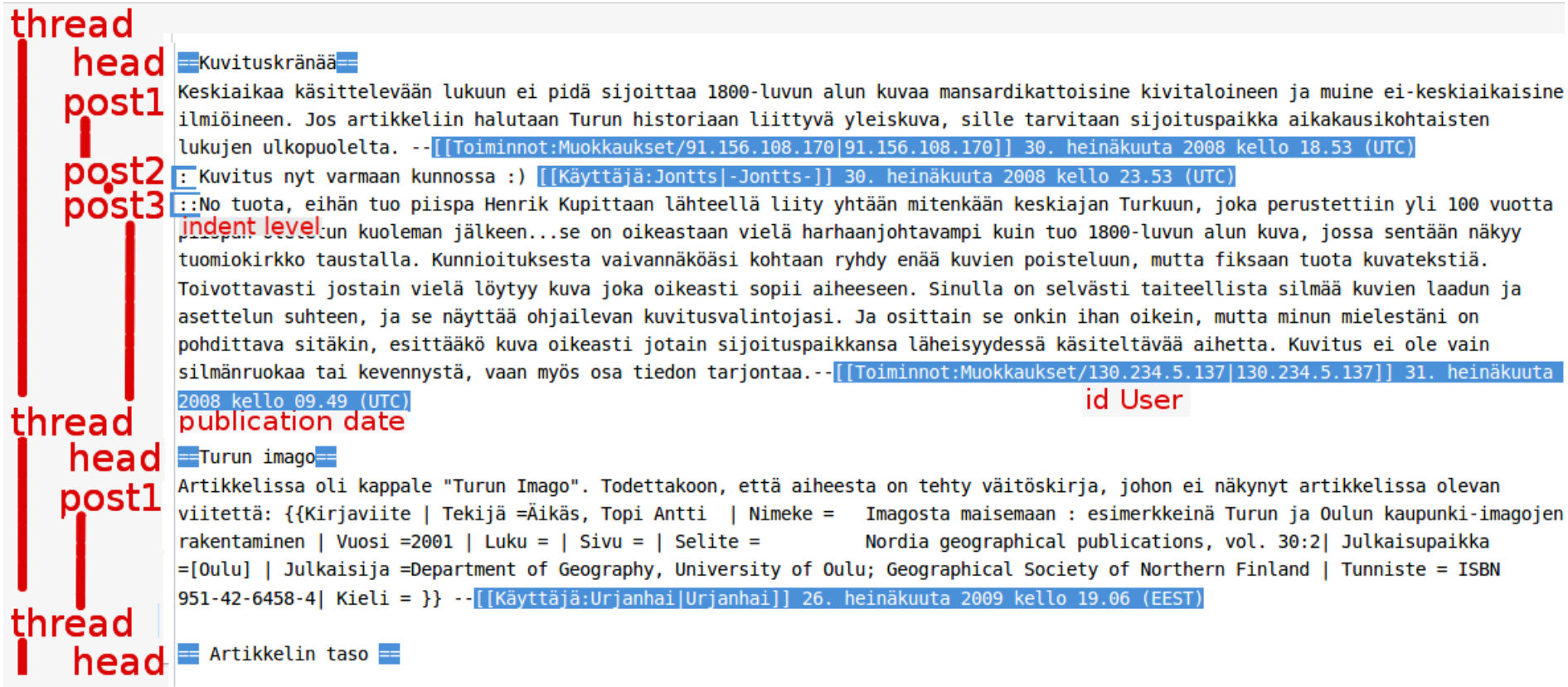
Document structure encoding acc. to TEI-P5

- Text Encoding Initiative, a norm for encoding all the properties of a document (content structure and metadata)
- An international consortium, towards a universal document representation
- Ensuring the sustainability and interoperability of the resource

A *light* TEI-P5

- all available metadata in the teiHeader (genre, thematic portal, etc.)
- threads marked up as `<div>`
- threads topic indicated in the `<head>` element, a part of the first post
- posts : `<post who="id User" when="publication date" indentLevel="#">`
- signature : `<signed><name>xxxx</name><date>xxxx</date></signed>`

Wikicode behind the talk page



Text TEI-P5 Structure

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
<teiHeader/>
  <text>
    <front/>
    <body>
      <div id="1" level="1">
        <head>Kuvituskränää</head>
        <post id="1" who="anonymous" bot="no" when="2008-07-30T18:53" indentLevel="0">
          <p id="1">Keskiaikaa käsittelevään lukuun ei pidä sijoittaa 1800-luvun alun kuvaa mansardikattoisine kivitaloineen ja mu
          Jos artikkeliin halutaan Turun historiaan liittyvä yleiskuva, sille tarvitaan sijoituspaikka aikakausikohtaisten lukujen
          heinäkuuta 2008 kello 18.53 (UTC)</date></signed></p>
        </post>
        <post id="2" who="Jontts" bot="no" when="2008-07-30T23:53" indentLevel="1">
          <p id="1">Kuvitus nyt varmaan kunnossa :) <signed><name>Jontts</name> <date>30. heinäkuuta 2008 kello 23.53 (UTC)</date>
        </post>
        <post id="3" who="anonymous" bot="no" when="2008-07-31T09:49" indentLevel="2">
          <p id="1">No tuota, eihän tuo piispa Henrik Kupittaan lähteellä liity yhtään mitenkään keskiajan Turkuun, joka perustett
          oletetun kuoleman jälkeen...se on oikeastaan vielä harhaanjohtavampi kuin tuo 1800-luvun alun kuva, jossa sentään näkyy
          Kunnioituksesta vaivannäköäsi kohtaan ryhdy enää kuvien poisteluun, mutta fiksaan tuota kuvatekstiä. Toivottavasti josta
          oikeasti sopii aiheeseen. Sinulla on selvästi taiteellista silmää kuvien laadun ja asettelun suhteen, ja se näyttää ohja
          osittain se onkin ihan oikein, mutta minun mielestäni on pohdittava sitäkin, esittääkö kuva oikeasti jotain sijoituspaik
          aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.-- <signed><date>31. heinäkuuta
          </date></signed></p>
        </post>
      </div>
      <div id="1" level="1">
        <head>Turun imago</head>
        <post id="1" who="Urjanhai" bot="no" when="2009-07-26T19:06" indentLevel="0">
          <p id="1">Artikkelissa oli kappale "Turun Imago". Todettakoon, että aiheesta on tehty väitöskirja, johon ei näkynyt arti
          {{Kirjaviite | Tekijä =Äikäs, Topi Antti  | Nimeke =     Imagosta maisemaan : esimerkkeinä Turun ja Oulun kaupunki-imago
          Luku = | Sivu = | Selite =      Nordia geographical publications, vol. 30:2| Julkaisupaikka =[Oulu] | Julkaisija =Depart
          Oulu; Geographical Society of Northern Finland | Tunniste = ISBN 951-42-6458-4| Kieli = }} --va oikeasti jotain sijoitus
          käsiteltävää aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.-- <signed><name>U
          heinäkuuta 2009 kello 19.06 (EEST)</date></signed></p>
        </post>
      </div>
      <div id="1" level="1">
        <head>Artikkelin taso</head>
          [...]
  </text>
</TEI>
```

WikiDisc Corpus Structure – Metadata

Metadata associated to a talk page

- "portal" i.e. associated portal sections e.g. *History*, *Art*, *Sport*, etc. (up to 7 sections associated with a same article). 11 sections
- "grade" i.e. article's quality assessments (from "draft" to "A-class")
- informations about the article (if it has been partly translated, if it is part of the Wikipedia 1.0 project, if its status has been discussed, if a problem happened, etc.) and the discussion (if it could be sensitive, out of range, or temporary blocked)

teiHeader TEI-P5 Structure

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI>
  <teiHeader>
    <fileDesc>
    <encodingDesc>
      <projectDesc/>
      <classDecl>
        <taxonomy>
          <bibl>Wikipedia</bibl>
          <category type="genre">
            <catDesc type="main">discussion</catDesc>
            <catDesc type="sub">Wikipedia talk page</catDesc>
          </category>
          <category type="Wikipedia article portal">
            <catDesc>,geographie,histoire,,,,religion,,,,,</catDesc>
          </category>
          <category type="discipline">
            <catDesc>Seconde Guerre mondiale</catDesc>
            <catDesc>Israël</catDesc>
            <catDesc>Paix</catDesc>
          </category>
          <category type="avancement">
            <catDesc>BD</catDesc>
          </category>
          <category type="interaction">
            <catDesc>{{Appel au calme}}</catDesc>
          </category>
        </taxonomy>
      </classDecl>
    </encodingDesc>
    <profileDesc>
  </teiHeader>
  <text>
</TEI>
```

Plan

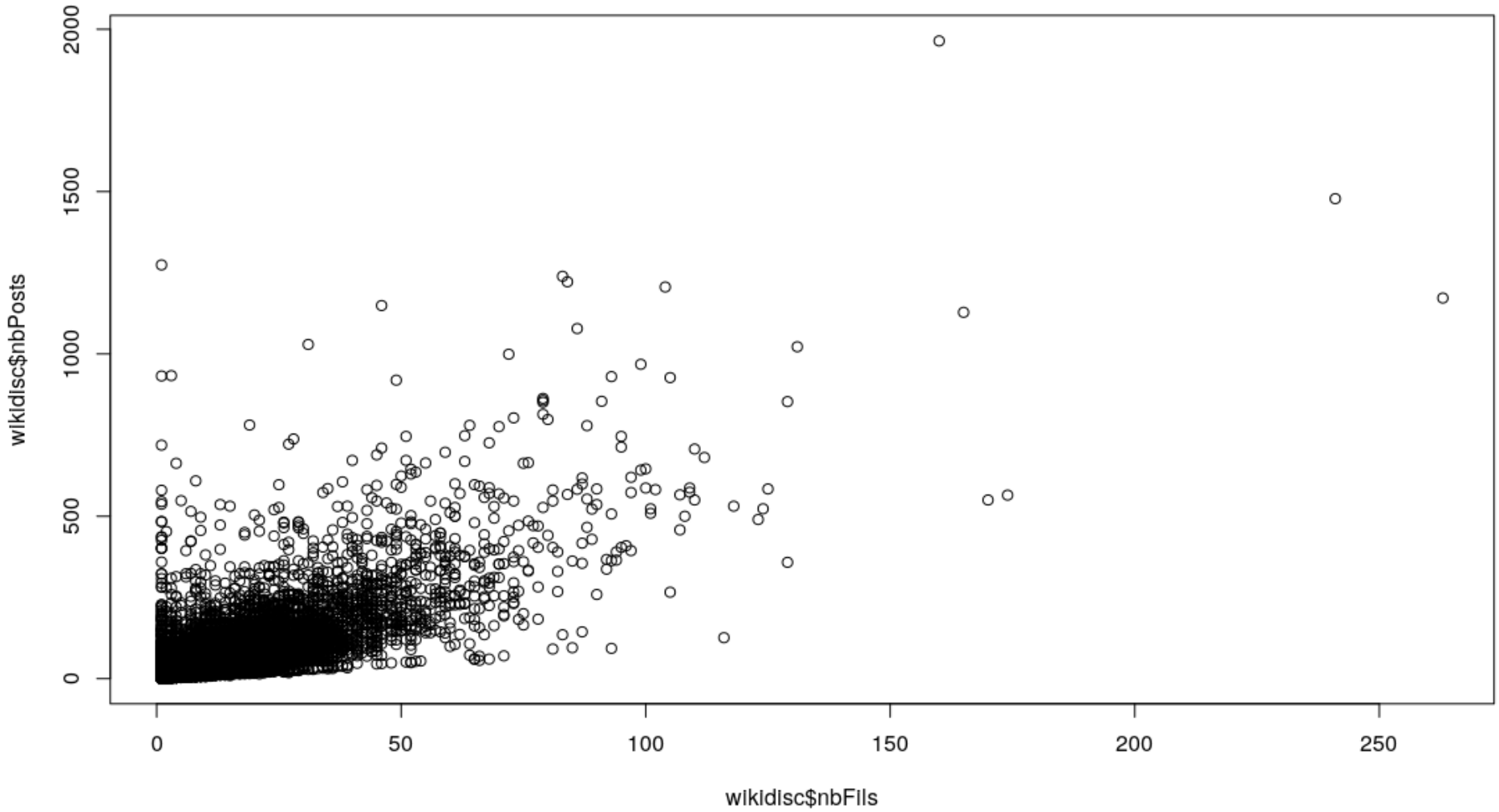
A wide variety of talk pages

version	talk pages	threads	posts	words
2018	439,638	1,243,676	3,145,943	213,286,460

A lot of talk pages with few interactions

On the 439,638 talk pages	#	%
Single thread talks	263,103	60
Single post talks	213,721	49
Talks under 50 words talks	166,108	38
Talks involving one single contrib. (anonymous as one single)	281,519	64
Talks involving 8 up to 228 different contrib.	44,578	10

A lot of talk pages with few interactions



A need for cleaning and structuring

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.*

See [Fer14, p. 111]

A "structured access" could be...

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 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]

A "structured access" could be...

content and contrary points of view

no interaction

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)

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

Plan

What kind of interactions in the WP talk pages ?

WP talk pages, where conflicts occur

- Conflicts between experts and non-experts
- Conflicts of interest and self-promotion suspicion
- Conflicts between opposite points of view

Conflict management is absolutely necessary

From Wikipedia point of view, conflicts must be regulated as it affects productivity

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

Disagreements, conflicts, harassment and personal attacks

An obstacle for the wikicracy

- Disagreements may turn to conflicts when the editing process and/or the discussion process are deadlocked
- When a conflict grows in intensity, discussions may turn to verbal abuse and personal attacks
- In on-line discussions, the article and talk page may be blocked and some users may be banished
- In WP such talk pages are tagged with specific labels signaling that a conflict is ongoing (e.g. NPOV or relevance disputes, "Calm talk" template) $\rightarrow$ MetaData
- Examples of pages with such labels are quite numerous : *Abortion in Iran*, *Bengali cuisine*, *Religion and sexuality*

A Top-Down Approach to conflict and personal attacks

Classifying interactions acc. to their degree of toxicity

Ex-Machina [WTD17] : Detecting conflict toxic posts and toxic writers

1. First experiment on WP talk pages : the "**Wikipedia DeTox**", an automatic detector of toxic comments.
2. The "Wikipedia DeTox" is currently adapted to other CMC under the name of "**Perspective API**"

A "toxic" post is

> *a rude, disrespectful or unreasonable comment that is likely to make you leave the discussion*

Different level of investigations

- Verbal violence and toxicity are generally detected at the post level [WTD17]
- Conflicts are better observed and detected at the thread level

A Top-Down Approach to conflict and personal attacks

Classifying interactions acc. to their degree of toxicity

Ex-Machina [WTD17] : Detecting conflict toxic posts and toxic writers

1. First experiment on WP talk pages : the "**Wikipedia DeTox**", an automatic detector of toxic comments.
2. The "Wikipedia DeTox" is currently adapted to other CMC under the name of "**Perspective API**"

A "toxic" post is

> *a rude, disrespectful or unreasonable comment that is likely to make you leave the discussion*

Different level of investigations

- Verbal violence and toxicity are generally detected at the post level [WTD17]
- Conflicts are better observed and detected at the thread level

A Top-Down Approach to conflict and personal attacks

Classifying interactions acc. to their degree of toxicity

1. Annotation of 1000 posts by using a crowd sourcing platform (posts selected randomly and also written by users who where blocked for violating Wikipedia's policy on personal attacks)

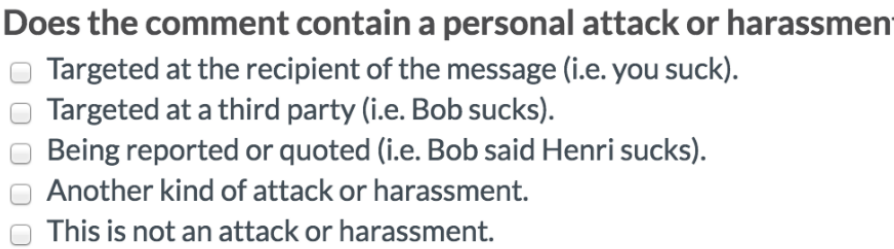


Figure 2: The question posed to our Crowdflower annotators.

2. Database : 115,737 annotated posts (10 coders per post) among which 11.7 % was labeled by the majority as an attack
3. Training a classifier with different configurations
4. The best is using a multi-layer perceptrons algorithm based on n-gram of characters for predicting the percentage of coders who consider the post as an attack

A Top-Down Approach to conflict and personal attacks

Classifying interactions acc. to their degree of toxicity

1. Annotation of 1000 posts by using a crowd sourcing platform (posts selected randomly and also written by users who where blocked for violating Wikipedia's policy on personal attacks)

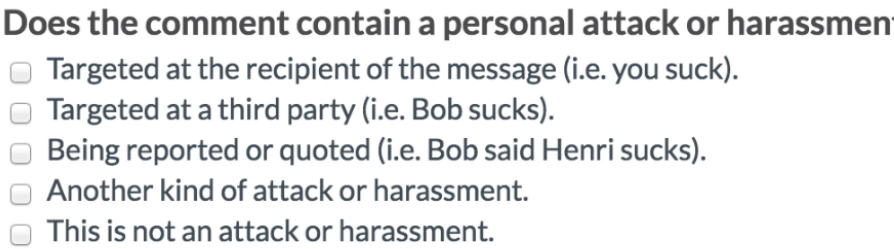


Figure 2: The question posed to our Crowdflower annotators.

2. Database : 115,737 annotated posts (10 coders per post) among which 11.7 % was labeled by the majority as an attack
3. Training a classifier with different configurations
4. The best is using a multi-layer perceptrons algorithm based on n-gram of characters for predicting the percentage of coders who consider the post as an attack

Automatic Classification (reminder)

Multi-layer perceptrons algorithm based on n-gram of characters for predicting the class of a post : attack or not

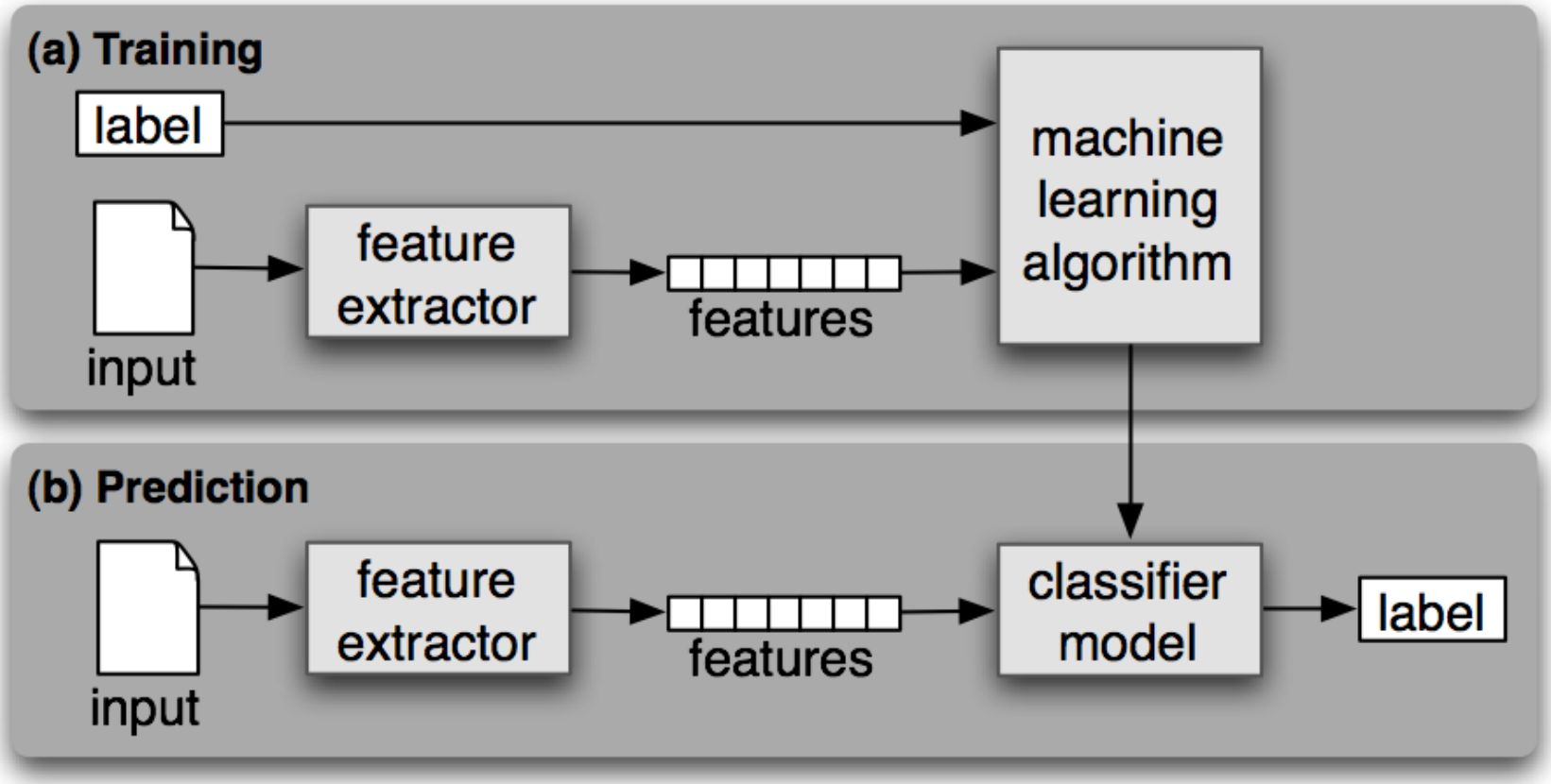


Figure extracted from [BKL09]

Ex-Machina, automatic classification of toxic comments

Multi-layer perceptrons algorithm based on n-gram of characters for predicting the class of a post : attack or not

- Evaluation metrics : accuracy of 96.59 ("the score between the models' predicted probability of being an attack and the majority class label in the set of annotations for each comment" [WTD17])
- Resulting resource : a full corpus of machine-labeled discussions in Wikipedia
- From 115,737 (human-)labeled posts to more than 63,400,000 (machine-)labeled posts
- Enough data for statistics

A Top-Down Approach to conflict and personal attacks

Profiling toxic writers

- What is the impact of anonymity ?
- How do attacks vary with the quantity of a user's contributions ?
- Are attacks concentrated among a few highly toxic users ?
- When do attacks result in moderation ?
- Is there a pattern to the timing of attacks ?

Answers and new insights in the paper [WTD17]...

Plan

A Bottom-Up Approach to Pages and Threads Profiling [HDLPT17]

- Exploring Rich Linguistic Features usually associated with interactional and rhetorical structures
- Mining talk pages and threads for discovering classes (without *a priori* i.e. unsupervized approach)
- Identifying relevant classes that we could linguistically interpret ad describe (not the case with n-grams)

Data mining method

Data mining tool

- R package FactoMineR dedicated to multivariate exploratory data analysis
- Principal Components Analysis (PCA) on talk pages and threads

Rich Linguistic Features

- **Global** : general quantitative characteristics of texts (talk pages and threads) e.g. number of words, number of contributors, presence of a "keep calm" banner ;
- **Thema** : portal sections of the associated article
- **Interact** : the frequency per texts of a wide range of interaction and politeness cues e.g. social deixis, marks of (dis)agreement, etc.
- **DiscRel** : the frequency per texts of connectives for each discourse relations as defined in the LEXCONN[RDM12].

Global features

Information extracted from the talk page itself

logNnMots	number of words
nbFils	number of threads*
nbPosts	number of posts
profMax	"interactional depth"
nbContributeurs	number of different participants
nbAnonymes	number of anonymous posts
X.anonymes	% of anonymous posts
nbBots	number of posts written by bots
X.bots	% of posts written by bots
AdQ	"1" if the talk page is linked to a A-class article
polemique	"1" if the talk page has the banner "keep calm"

Thema features

WP section of the associated article

- 11 WP sections : art, geography, history, leisure, medicine, politics, religion, sciences, society, sport, technology
- Some articles are simultaneously in 7 sections !
- *Geography* is the most frequent section (170,246 talk pages, 39%)
- 11 features binarized (e.g. geography = 1/0)
- The same feature for talk pages and threads

Interact features

11 features automatically identified with simple regular expressions

Politeness	*thanks*, *hello*, *goodbye*, *hi*, *sincerely*, *cheers*, *please*, *would you*, etc.
Agreement	*OK*, *agree*, *yes*, *no*, *actually*, etc.
Question	*?*
Je	1st singular person pronouns + *personally*
Tu	2nd sing. pers. pronouns, informal "you"
Vous	2nd plur. pers. and formal "you" pronouns
Nous	1st plur. pers. pronouns
On	Informal "We"
WP	*Wikipedia* or *WP*
pour	Sentence-initial *For* or *I'm for*
contre	Sentence-initial *Against* or *I'm against*

DiscRel features

22 discourse relations with the number of identified connectives as value

- discourse relations as defined in the LEXCONN[RDM12] : "a French lexicon of 328 discourse connectives, collected with their syntactic categories and the discourse relations they convey"
- When a connective is polysemious, all possible relations are considered
- alternation ; background ; commentary ; concession ; condition ; consequence ; continuation ; contrast ; detachment ; elaboration ; evidence ; explanation ; flashback ; goal ; narration ; opposition ; parallel ; rephrasing ; result ; summary ; temporality ; unknown relation

ACP parameters

- Considering only discussions with more than 100 words
- Only Interact and DiscRel features are taken into account (normalized on the number of words)
- The other features are just indicated (in blue) for permitting a global overview

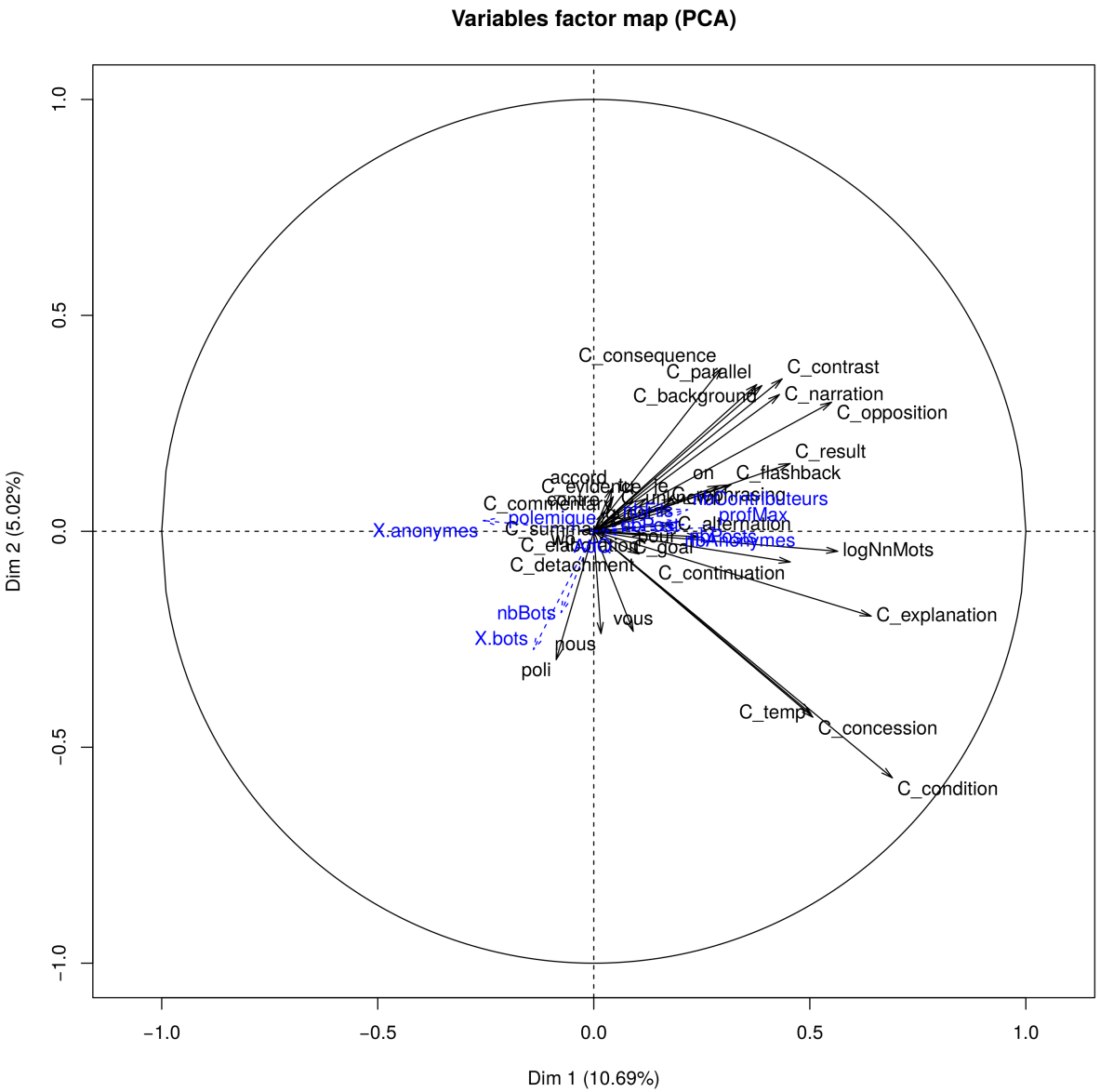
Results : ACP on linguistic features

- 5 dimensions that explain around 30% of the total variance
- A first dimension simply related to the number of words (the more words, the more features)
- A second dimension that differs acc. to the unit taken into account (thread or talk page)
- For **threads**, it opposes :
  - Dimension 2+ : more *I*, informal *we* (*on*) and discourse relations expressing **contrast**
  - Dimension 2- : agreement cues, formal *you* and discourse relations expressing alternation, consequence, goal and temporal relations

Results : ACP on linguistic features

- 5 dimensions that explain around 30% of the total variance
- A first dimension simply related to the number of words (the more words, the more features)
- A second dimension that differs acc. to the unit taken into account (thread or talk page)
- For **talk pages**, it opposes :
  - Dimension 2+ : more discourse relations expressing **contrast**, background/narration and causality
  - Dimension 2- : politeness cues, formal you and we and discourse relations expressing concession, condition and temporal relations

Results : ACP on linguistic features for talk pages



- Dimension 1 : the more words the more features
- Dimension 2+ : discourse relations expressing contrast, background/narration and causality
- Dimension 2- : politeness cues, formal you and we and discourse relations expressing concession, condition and temporal relations

Difficulties to go from these results to examples we may interpret

"few politeness cues, formal you and more discourse relations expressing contrast"

- Few politeness cues because few words or no real Interaction (one post per threads)
- Potentially only one formal you (perhaps included in a specific locution as "s'il vous plait" *please*)
- 17 connectives associated with contrast in the LexConn including the two very polysemous "but" and "while"

Plan

Conclusion

- WP talk pages shed light on the other side of the well-known WP articles : collaboration and writing processes
- WP talk pages remain complex objects that challenge the traditional models and methods used for linguistic characterization
- WP talk pages genres require different levels of investigation : toxicity on the post level, conflict on the thread level, "controversity" on the talk page level...
- Data mining techniques may give us some leads but...

Qualitative analyses and manual annotation are crucial

- We must improve features that describe the thread level as for example by looking at the headings, the first post of the section and the context (ex : `https://fr.wikipedia.org/wiki/Discussion:Psychanalyse/arch1#choqu.C3.A9`)
- A lot of work in progress....

A complex and wide research area

- headings study
- expression of (dis)agreement
- annotation of the speech acts (e.g. [FGC12])
- focus on more detailed interactions and special topics
  - discussion about terminology issues : what is the right (layout) word that must be used in an article about a technical domain ?
  - about neutrality and conflict : what are the most controversial topics (plus, timeline) and what are the pros and the cons ?
  - about negotiation : is there (linguistic) cues for detecting threads that will bring about a consensual solution vs. threads that will sink into chaos ?
- Still at the first stage : bleaching threads for identifying prototypical interactions :
  - A opens a thread by asking a question, B answers to A who thanks B
  - A opens a thread by lauching a vote, B votes, followed by C, D, ... Z
  - A puts a cat among the pigeons (toxic post ?), B reacts strongly, C too... fight

M. Augustyn, S. Ben Hamou, G. Bloquet, V. Goossens, M. Loiseau, and F. Rynck.

Autour Des Langues Et Du Langage : Perspective Pluridisciplinaire, chapter Constitution de ressources pédagogiques numériques : le lexique des affects, page 407–414.
Grenoble : Presses Universitaires de Grenoble, 2008.

Nicholas Asher, Philippe Muller, Myriam Bras, Lydia Mai Ho-Dac, Farah Benamara, Stergos Afantenos, and Laure Vieu.

ANNODIS and Related Projects : Case Studies on the Annotation of Discourse Structure, pages 1241–1264.
Springer Netherlands, Dordrecht, 2017.

Steven Bird, Ewan Klein, and Edward Loper.

Natural language processing with Python : analyzing text with the natural language toolkit.
O'Reilly Media, Inc., 2009.

Erik Borra, Esther Weltevrede, Paolo Ciuccarelli, Andreas Kaltenbrunner, David Laniado, Giovanni Magni, Michele Mauri, Richard Rogers, and Tommaso Venturini.

Societal controversies in wikipedia articles.
In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 193–196, New York, NY, USA, 2015. ACM.

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.

Lydia-Mai Ho-Dac and Veronika Laippala.

Les discussions wikipedia : un corpus pour caractériser le genre "discussion".
In *International Research Days Social Media and CMC Corpora for the eHumanities*, Rennes, France, october 2015.

Lydia-Mai Ho-Dac, Veronika Laippala, Céline Poudat, and Ludovic Tanguy.

Exploring Wikipedia talk pages for conflict detection.
In Darja Fišer and Michael Beißwenger, editors, *Investigating Computer-Mediated Communication : Corpus-Based Approaches to Language in the Digital World*, Translation Studies and Applied Linguistics, pages 146–168. Ljubljana University Press, Faculty of Arts, 2017.

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.


Charlotte Roze, Laurence Danlos, and Philippe Muller.
Lexconn : A french lexicon of discourse connectives.
Discours, 10, 2012.


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, International World Wide Web Conferences Steering Committee*, page 1391–1399, 2017.


Torsten Zesch, Christof Müller, and Iryna Gurevych.
Extracting lexical semantic knowledge from wikipedia and wiktionary.
In *LREC*, volume 8, pages 1646–1652, 2008.