# Profiling interactions behind the Wikipedia articles

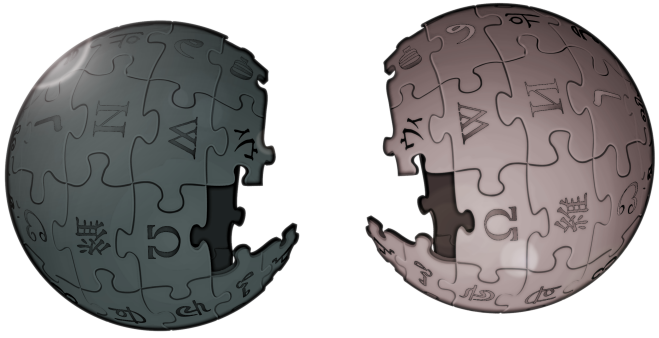
Lydia-Mai Ho-Dac

# Profiling interactions behind the Wikipedia articles

Lydia-Mai Ho-Dac

University of Toulouse, CLLE-ERSS, CNRS

March 2018





## Ass. Prof. and Researcher in

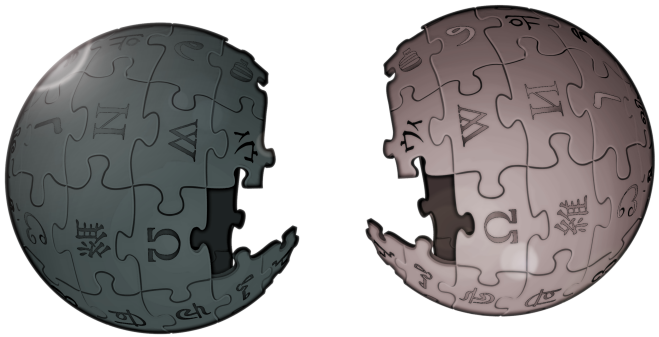
- **Erasmus Teaching mobility** : Partnership between the Department of Linguistics – University of Toulouse – and the School of Languages – University of Turku
- **Linguistics** :
  - The study of discourse organization (*how human build and structure "text worlds" via documents*)
  - Text genres and 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

## Among these areas of research : CMC genres and Wikipedia as a topic research

- Focusing since few years on a range of new genres called CMC (Computer-Mediated Communications)
- 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)
- Challenge : describe these new genres
  - 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)

# Profiling interactions behind the Wikipedia articles

1. Wikipedia as a research topic for human sciences
2. WP Talk Pages, the other side
3. What kind of interactions in the WP talk pages ?
4. Conclusion



# Plan

# Plan

# A chance for sharing and recording humanity knowlegde

## 1- 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***

## 2- The *wiki* technology

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

# A chance for sharing and recording humanity knowlegde

## 1- 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***

## 2- The *wiki* technology

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

# Birth of Wikipedia – WP

## the native project Nupedia

- 2000 : ***Nupedia*** : collaborative advertising-supported encyclopedia written by "experts" (with at least a PhD) who submit an article which must be accepted after a fairly burdensome peer-reviewed process
- **Jimmy Wales** :
  - @Jimbo
  - a trader, 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

# Birth of Wikipedia

## Birth of English WP – WP[EN]

***Nupedia***, a fairly burdensome editorial process : two expert reviewers submit to all the Nupedia members the proposition which will be entirely corrected and then re-reviewed

- 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<br>*Nupedia* was abandoned (with only 24 accepted and published articles) |
| **2018** | WP is the 5th most visited web site (the 6th in Finland) just after *Google*, *Facebook*, *Youtube*, *Baidu* with 5,184,686 views per hours for WP[EN] (36,095 for WP[FI]) |

With a precious help of search engines : in 2011 2/3 of WP visits follow a Search Engine query

# Birth, first steps and influence of non English WPs

## WP[EN] principles https://wikivividly.com/wiki/User:Jimbo_Wales/Statement_of_principles

- 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" speak 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

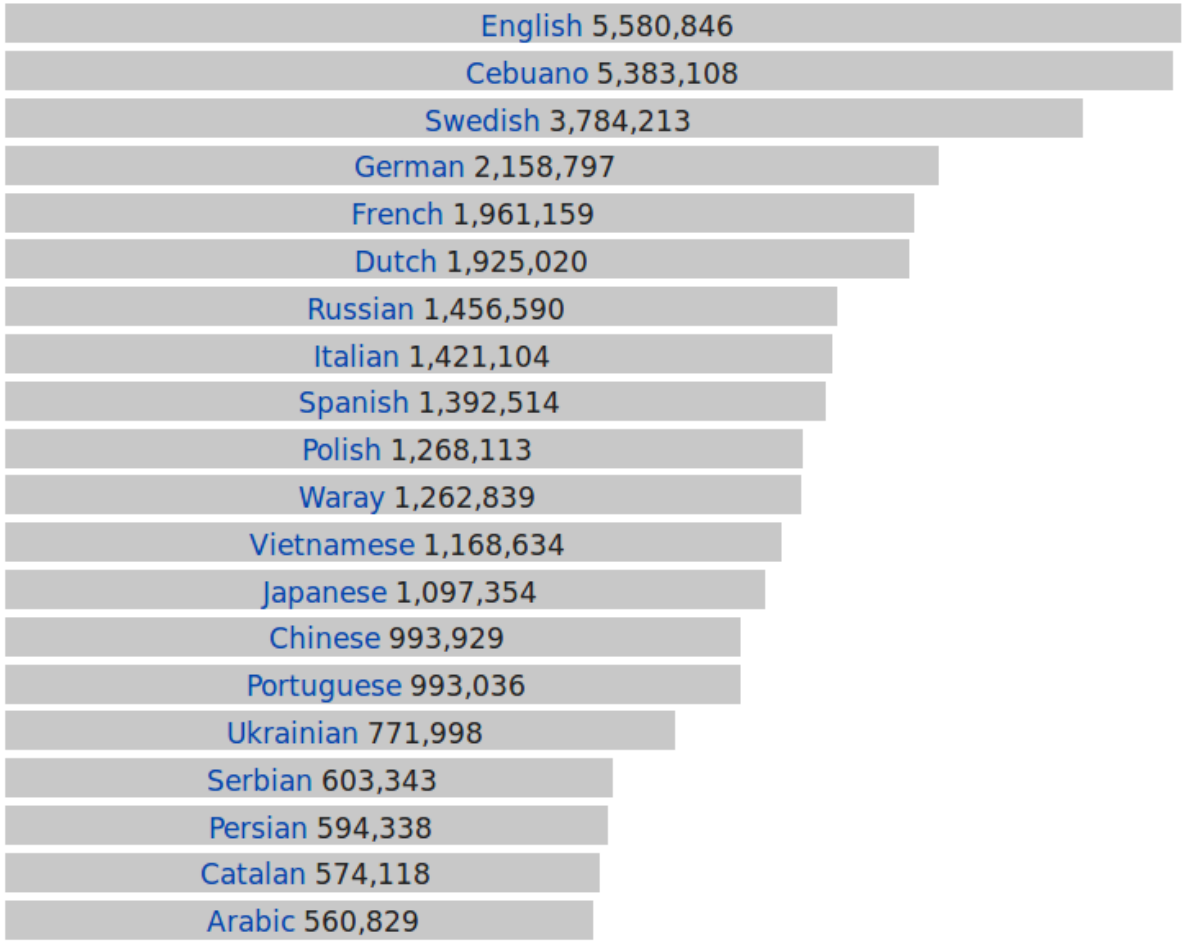
1. WP[DE] (March 2001) use of forums called *Meinungsbilder* (Meinung – opinion) for clarifying issues for which there is no consensus
2. WP[FR] (March 2001) In October 2002, Florence Dévouard (@Anthère) created a page called "decision-making" where "the final choice will depend on a vote instead of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

---

Finnish Wikipedia – WP[FI] opens in Feb 2003

# Nowadays, a global phenomenon (amount of articles)

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




WP[FI] 432,000 articles

# 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:

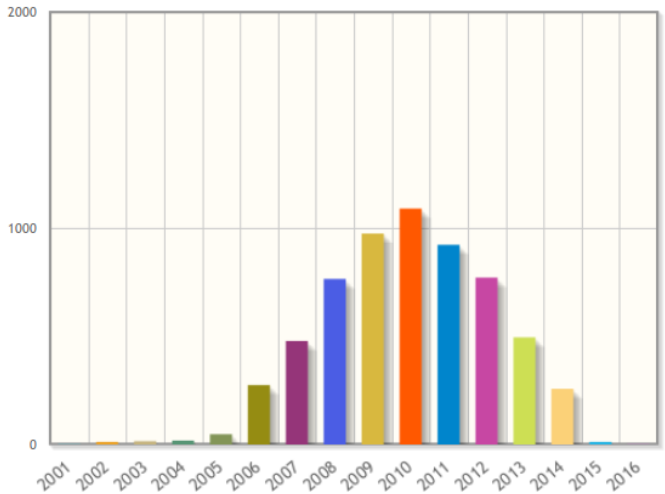
- CLEF, MathWikis, WikiAI, WikiSym, WikiViz

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 human digital artifact

## "What WP is" (https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is)

Wikipedia:Wikipedia is an encyclopedia

From Wikipedia, the free encyclopedia
(Redirected from Wikipedia:What Wikipedia is)

**Wikipedia is an encyclopedia.**

*An encyclopedia is a written compendium of knowledge.*

*Wikipedia is freely available, and incorporates elements of general and specialized encyclopedias, almanacs, and gazetteers.*

## WP :Five pillars

- WP is an encyclopedia
- WP is written from a neutral point of view
- WP is free content that anyone can use, edit, and distribute
- WP's editors should treat each other with respect and civility
- WP has no firm rules

# 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.*

# "WP is not a democracy"

## "What WP is not" (https://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not)

### Community

The above policies are about Wikipedia's content. The following policies apply to Wikipedia's governance and processes.

WP is Encyclopedists' Corner, not Speakers' Corner.

#### Wikipedia is not an anarchy or forum for free speech

Policy shortcuts
WP:NOTANARCHY
WP:NOTFREESPEECH
WP:CHAOS

*"WP:ANARCHY" redirects here. For WikiProject Anarchism, see Wikipedia:WikiProject Philosophy/Anarchism.*
*See also: m:Power structure, WP:User access levels, and WP:Enforcement*
*Main page: Wikipedia:Administration*

Wikipedia is free and open, but restricts both freedom and openness where they interfere with creating an encyclopedia. Accordingly, Wikipedia is not a forum for unregulated free speech. The fact that Wikipedia is an open, self-governing project does not mean that any part of its purpose is to explore the viability of anarchist communities. Our purpose is to build an encyclopedia, not to test the limits of anarchism.

#### Wikipedia is not a democracy

Policy shortcuts
WP:DEM
WP:DEMOCRACY
WP:NOT#DEM
WP:NOTDEMOCRACY
WP:WIKINOTVOTE

*See also: Wikipedia:Polling is not a substitute for discussion and Wikipedia:Elections*

Wikipedia is not an experiment in democracy or any other political system. Its primary (though not exclusive) means of decision making and conflict resolution is editing and discussion leading to consensus—*not* voting (voting is used for certain matters such as electing the Arbitration Committee). Straw polls are sometimes used to test for consensus, but polls or surveys can impede, rather than foster, discussion and should be used with caution.

#### Wikipedia is not a bureaucracy

Policy shortcuts

*See also: Wikipedia:Ignore all rules*

# 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) |

# 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]
  - *diff* between revisions for extracting spelling variants (spell checker), paraphrases (information retrieval), simplifications and summarization (for developing automatic processing)
- 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 personal Attacks detection

# 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]
  - *diff* between revisions for extracting spelling variants (spell checker), paraphrases (information retrieval), simplifications and summarization (for developing automatic processing)
- 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 personal Attacks detection **A new challenge for NLP**

# Exploiting edits and forum for discussion for NLP [FDG13]

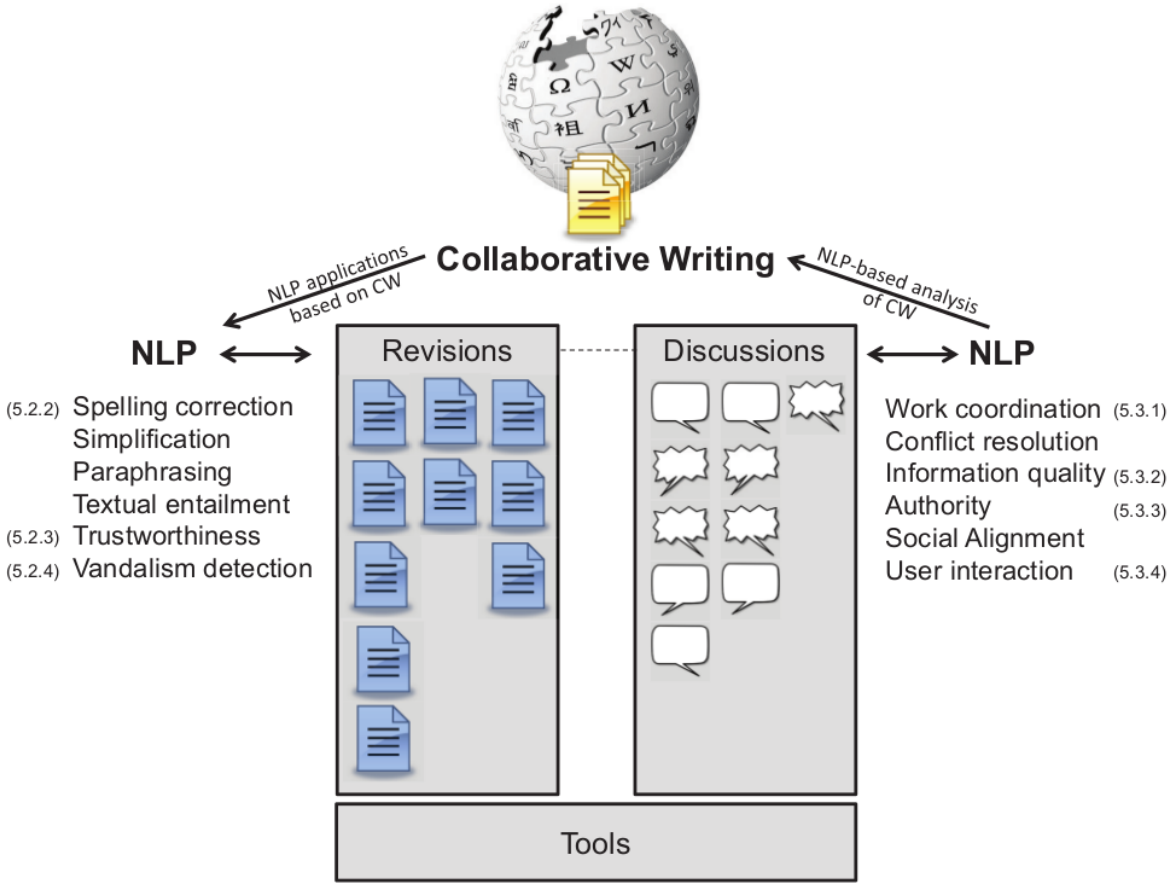



**Fig. 5.1** The role of NLP in collaborative writing (CW): Topics covered in this chapter

# Plan

# WP Talk Pages, the other side

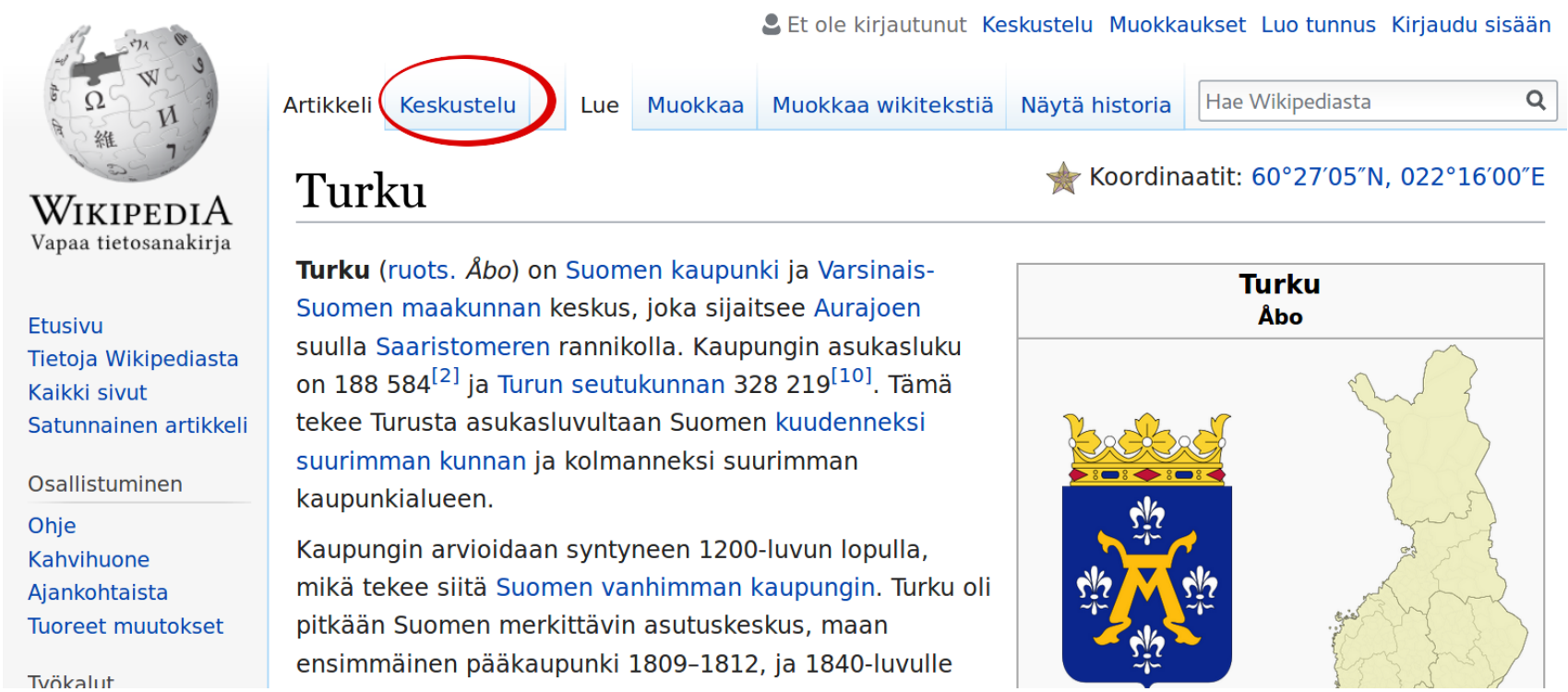
## Exploiting the forum for discussion

- 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]

# What are the WP talk pages



# An online discussion behind the article...

# ... associated with metadata and containing threads

# headed threads containing dated and signed posts


## Kuvituskränää [ muokkaa wikitekstiä ]

Keskiaikaa käsittelevään lukuun ei pidä sijoittaa 1800-luvun alun kuvaa mansardikattoisine kivitaloineen ja muine ei-keskiaikaisine ilmiöineen. Jos artikkeliin halutaan Turun historiaan liittyvä yleiskuva, sille tarvitaan sijoituspaikka aikakausikohtaisten lukujen ulkopuolelta. --91.156.108.170 30. heinäkuuta 2008 kello 18.53 (UTC)

id User (anonymous)

publication date

> Kuvitus nyt varmaan kunnossa :) -Jontts- 30. heinäkuuta 2008 kello 23.53 (UTC)
>
> id User (Jontts) publication date
>
> > No tuota, eihän tuo piispa Henrik Kupittaan lähteellä liity yhtään mitenkään keskiajan Turkuun, joka perustettiin yli 100 vuotta piispan oletetun kuoleman jälkeen...se on oikeastaan vielä harhaanjohtavampi kuin tuo 1800-luvun alun kuva, jossa sentään näkyy tuomiokirkko taustalla. Kunnioituksesta vaivannäköäsi kohtaan ryhdy enää kuvien poisteluun, mutta fiksaan tuota kuvatekstiä. Toivottavasti jostain vielä löytyy kuva joka oikeasti sopii aiheeseen. Sinulla on selvästi taiteellista silmää kuvien laadun ja asettelun suhteen, ja se näyttää ohjailevan kuvitusvalintojasi. Ja osittain se onkin ihan oikein, mutta minun mielestäni on pohdittava sitäkin, esittääkö kuva oikeasti jotain sijoituspaikkansa läheisyydessä käsiteltävää aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.--130.234.5.137 31. heinäkuuta 2008 kello 09.49 (UTC)
>
> id User (anonymous)
>
> publication date

# Going to the root of Collaborative writing : negotiation process

## WP talk pages, a native need

- Few days after the WP[EN] birth, a Wikipedian raises the following question : what to do with discussions behind the articles ?
- WP talk pages for guaranteeing the quality and the impartiality of WP

## WP talk pages, where problems are resolved but also where conflicts occur

- Conflicts between experts (from *Nupedia*) and non-experts (called "wiki-anarchists")
- Conflicts of interest and self-promotion suspicion (*"WP **is not** a free advertising space"*)
  - 31% of the communications professionals have been contribute to the article about the product they supply [DiS12]
  - *Wikiscanner* (V. Griffith) for identifying the IP addresses of well-known partial writers : *Pepsi*, *CIA*, *Exxon*, *political parties*

# Plan

# WP talk pages as Corpus

## Availability and large amount of data

- (kind of) on-line forum, with the (rare) specificity of being freely available under Creative Commons by-sa
- Multilingual $\rightarrow$ contrastive studies with languages (even if poorly endowed)

## A lot of extra-linguistic characteristics

- thematic (portals, categories)
- specific content (associated article)
- subjectivity (banner)
- writer characteristics (status in WP, activity on WP, etc.)

# What kind or content in the WP talk pages ?

## Collecting WP talk pages for corpus-based linguistic description

The WikiDisc Corpus [HDL15] : talk pages extracted from the WP[FR] Wikipedia snapshot (*dump*) from 12th may 2015 which contains 3,487,480 talk pages ( global backup frwiki-20150512-pages-meta-current#.xml.bz2 available on http ://dumps.wikimedia.org/frwiki/20150512/)

An updated version soon !

## The WikiDisc Corpus : 366,326 talk pages

Among the 3 487 480 /<title>Discussion/ (talk pages) in the 2015.05.12 snapshot

| | | |
|---|---|---|
| User Talk pages | 1,990,927 | 57% |
| **Article Talk pages** | **1,496,553** | **43%** |
| Redirections | 116,432 | 8% |
| Empty talk pages (< 2 words) | 1,013,791 | 68% |
| **Remaining talk pages** | **366,326** | **24%** |

# The WikiDisc Corpus building

## Document structure of a talk page

The 366,326 talk pages were 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. an optional signature including timestamp and eventually user id
  2. a change of indent level indicated with zero, one or more semi-colon ( :) at the beginning of the post

# Talk Page behind the Turku WP[FI] article

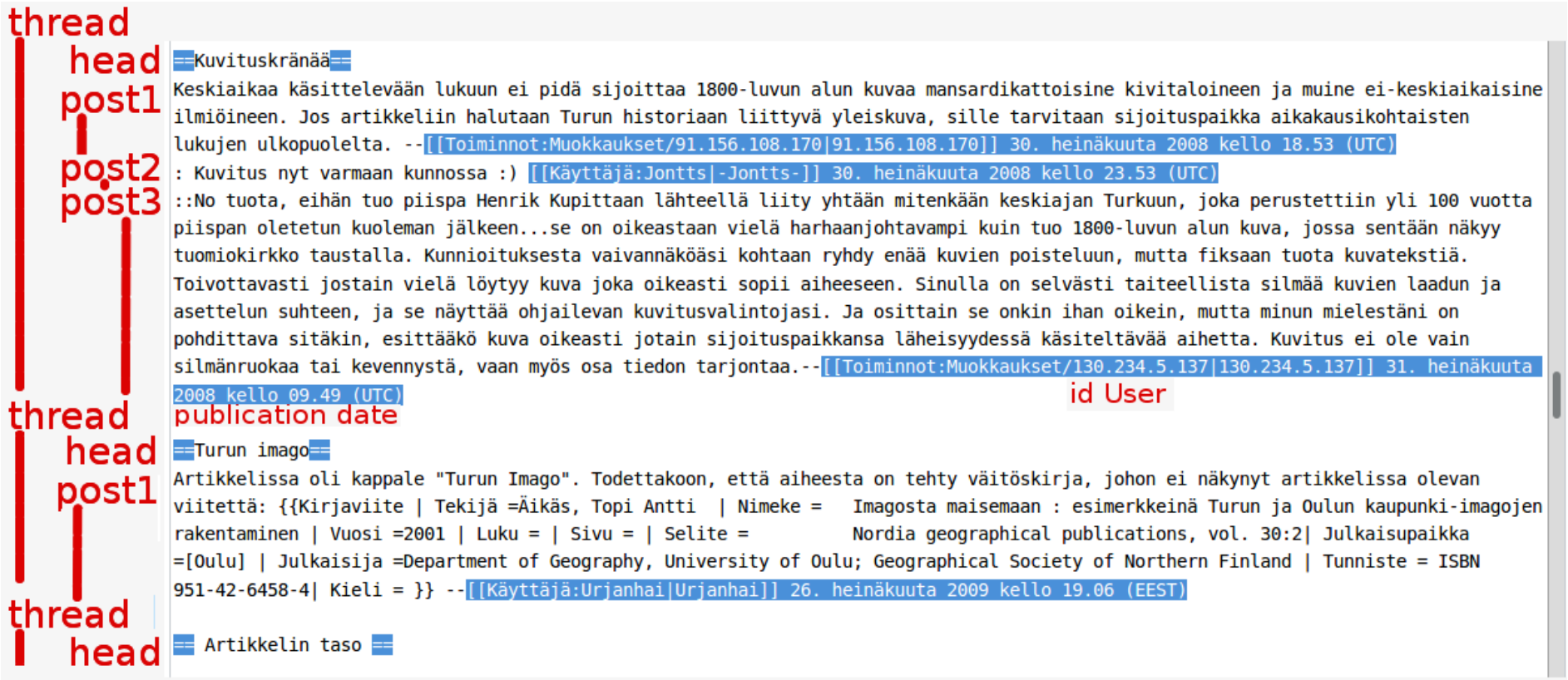

## Kuvituskränää [ muokkaa wikitekstiä ]

Keskiaikaa käsittelevään lukuun ei pidä sijoittaa 1800-luvun alun kuvaa mansardikattoisine kivitaloineen ja muine ei-keskiaikaisine ilmiöineen. Jos artikkeliin halutaan Turun historiaan liittyvä yleiskuva, sille tarvitaan sijoituspaikka aikakausikohtaisten lukujen ulkopuolelta. --91.156.108.170 30. heinäkuuta 2008 kello 18.53 (UTC)
id User (anonymous)
publication date

Kuvitus nyt varmaan kunnossa :) -Jontts- 30. heinäkuuta 2008 kello 23.53 (UTC)
id User (Jontts) publication date

No tuota, eihän tuo piispa Henrik Kupittaan lähteellä liity yhtään mitenkään keskiajan Turkuun, joka perustettiin yli 100 vuotta piispan oletetun kuoleman jälkeen...se on oikeastaan vielä harhaanjohtavampi kuin tuo 1800-luvun alun kuva, jossa sentään näkyy tuomiokirkko taustalla. Kunnioituksesta vaivannäköäsi kohtaan ryhdy enää kuvien poisteluun, mutta fiksaan tuota kuvatekstiä. Toivottavasti jostain vielä löytyy kuva joka oikeasti sopii aiheeseen. Sinulla on selvästi taiteellista silmää kuvien laadun ja asettelun suhteen, ja se näyttää ohjailevan kuvitusvalintojasi. Ja osittain se onkin ihan oikein, mutta minun mielestäni on pohdittava sitäkin, esittääkö kuva oikeasti jotain sijoituspaikkansa läheisyydessä käsiteltävää aihetta. Kuvitus ei ole vain silmänruokaa tai kevennystä, vaan myös osa tiedon tarjontaa.--130.234.5.137 31. heinäkuuta 2008 kello 09.49 (UTC)
id User (anonymous)
publication date

# Wikicode behind the talk page



# The WikiDisc Corpus building

## Document structure of a talk page

The 366,326 talk pages were structured into threads and posts on the basis of *wikicode* (the wiki traditional syntax)

- Threads correspond to division delimited by (sub)headings signaled with /==.*?==/ in the wikicode
- Posts are delimited according to
  1. timestamp and eventually user signature such as : *Viking59 10 Mai 2009 at 17 :16 (CEST)*
  2. a change of indent level indicated with zero, one or more semi-colon ( :) at the beginning of the post.

| talk pages | threads | posts | words |
|---|---|---|---|
| 366,326 | 1,024,351 | 3,022,240 | 159,578,279 |

# The WikiDisc Corpus building

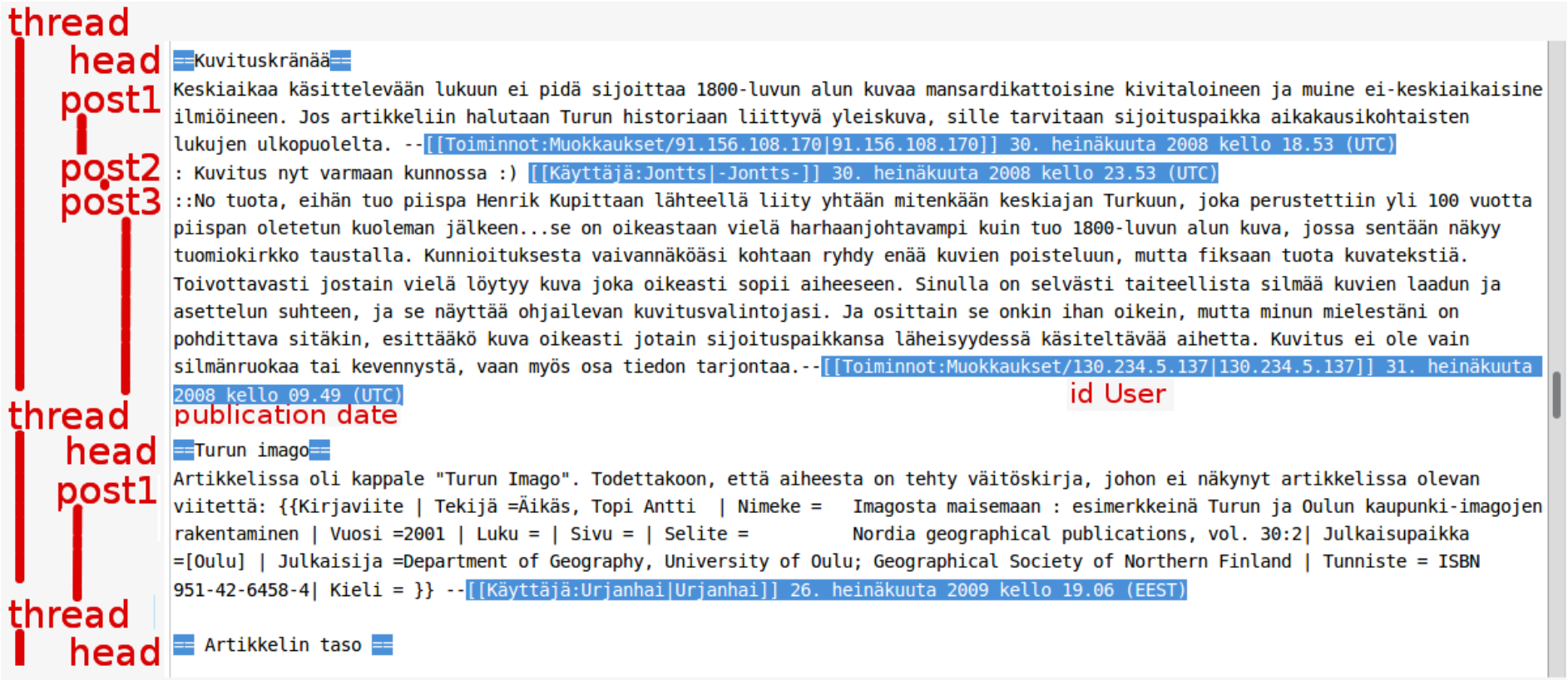
## Document structure encoding acc. to TEI-P5

- 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="#">`

# 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

- "discipline" i.e. associated portal sections e.g. *History*, *Art*, *Sport*, etc. (up to 7 sections associated with a same article). 11 sections
- "avancement" i.e. article's quality assessments
- "conflictness" i.e. information manually inserted by Wikipedians via the template/banner {{keep calm}}
- "talk type" (a specific characteristic of the French Wikipedia)

# 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

| **On the 366,326 talk pages** | # | % |
|---|---|---|
| Single post talks | 202,856 | 55 |
| Talks under 53 words talks | 181,503 | 50 |
| Few extremely long talks (up to 1,143 posts and 148,968 words) | | |
| Talks involving 8 to 228 different writers | 40,413 | 10 |
| **On the 1,024,351 threads (in main talk pages)** | | % |
| "monologue" | | 35.8 |
| discussion | | 44.7 |
| "dialogue" between two writers | | 26 |
| between 3 and 5 different writers | | 16.5 |
| "debate" i.e. more than 5 different writers | | 2.2 |
| **On the 3,022,240 posts** | | % |
| anonymous posts | | 80 |

# Very active users

In 2006 90% of WP[FR] has been written by 5% of the wikipedians

In 2015 3/4 of WP[FR] by 0.5% [Sah15, 263]

| Edits >= | Wikipedians | | Edits Total | |
|---|---|---|---|---|
| 1 | 617,271 | 100.0% | 61,504,184 | 100.0% |
| 3 | 249,901 | 40.5% | 60,924,528 | 99.1% |
| 10 | 128,919 | 20.9% | 60,227,944 | 97.9% |
| 32 | 54,416 | 8.8% | 58,958,965 | 95.9% |
| 100 | 23,184 | 3.8% | 57,260,629 | 93.1% |
| 316 | 10,705 | 1.7% | 55,107,911 | 89.6% |
| 1000 | 5,303 | 0.9% | 52,114,334 | 84.7% |
| 3162 | 2,621 | 0.4% | 47,397,591 | 77.1% |
| 10000 | 1,164 | 0.2% | 39,108,206 | 63.6% |
| 31623 | 352 | 0.1% | 25,225,891 | 41.0% |
| 100000 | 60 | 0.0% | 10,170,150 | 16.5% |
| 316228 | 5 | 0.0% | 2,266,015 | 3.7% |

Wikipedian edits in WP[FR]

`https://stats.wikimedia.org/EN/TablesWikipediaFR.htm`

# Very active users

In 2006 90% of WP[FR] has been written by 5% of the wikipedians

In 2015 3/4 of WP[FR] by 0.5% [Sah15, 263]

| Edits >= | Wikipedians | | Edits Total | |
|---|---|---|---|---|
| 1 | 88,123 | 100.0% | 7,217,170 | 100.0% |
| 3 | 34,073 | 38.7% | 7,132,202 | 98.8% |
| 10 | 17,059 | 19.4% | 7,034,263 | 97.5% |
| 32 | 7,005 | 7.9% | 6,864,273 | 95.1% |
| 100 | 3,055 | 3.5% | 6,651,358 | 92.2% |
| 316 | 1,470 | 1.7% | 6,379,489 | 88.4% |
| 1000 | 755 | 0.9% | 5,981,372 | 82.9% |
| 3162 | 377 | 0.4% | 5,321,465 | 73.7% |
| 10000 | 162 | 0.2% | 4,121,290 | 57.1% |
| 31623 | 43 | 0.0% | 2,136,281 | 29.6% |

Wikipedian edits in WP[FI]

`https://stats.wikimedia.org/EN/TablesWikipediaFI.htm`

# The top ten of "benevolent activists users" in WP[FR]

| used ID | nb. talk pages | % | nb. posts | % |
|---|---|---|---|---|
| total | 59,593 | 16.3 | 86,595 | 2.9 |
| Chris a liege | 12,511 | 3.4 | 15,254 | 0.5 |
| schlum | 8,107 | 2.2 | 13,706 | 0.5 |
| Patrick Rogel | 7,255 | 2.0 | 12,021 | 0.4 |
| Azurfrog | 3,733 | 1.0 | 8,601 | 0.3 |
| Hégésippe Cormier | 4,804 | 1.3 | 8,088 | 0.3 |
| McLushFR | 5,371 | 1.5 | 6,613 | 0.2 |
| Rosier | 5,260 | 1.4 | 5,964 | 0.2 |
| Axou | 3,911 | 1.1 | 5,540 | 0.2 |
| Taguelmoust | 4,434 | 1.2 | 5,459 | 0.2 |
| Lomita | 4,207 | 1.1 | 5,349 | 0.2 |

# Plan

# Preliminary linguistic description of the talk pages

## Three aspects of linguistic characteristics

- writing level and readability by measuring the amount of unknown words and the sentences length
- subjectivity by counting the nb. of occurrences of first personal pronouns and by projecting an affect lexicon ([ABHB+08])
- discourse structures by extracting posts' opening

## In contrast with two other genres

| **Corpora** | **tokens** | **genre** |
|---|---|---|
| WP article (2013) | 226,207,672 | encyclopedia articles |
| Health Fora* | 236,368,151 | online discussion |
| | | 2,585,188 posts |
| *WikiDisc* | *132,406,816* | *online discussion* |
| | | *3,022,240 posts* |

*for representing classic on-line discussions

# First results

| | #Words | unknown Words | mean length | | Affect | Pro1 |
|---|---|---|---|---|---|---|
| | | (%) | phrases | mots | % | % |
| Health fora | 236 368 151 | 22 | 10.3 | 5.2 | 4.7 | 3.48 |
| WikiDisc | 161 833 298 | 5 | 18.2 | 5.4 | 2.1 | 0.05 |
| WP Articles | 226 207 672 | 5 | 14.6 | 5.5 | 1.8 | 1.34 |
| | 622 154 102 | 12 | 13.4 | 5.4 | 2.8 | 1.53 |

- On the 560,841 unknown types of words, 19% also occur in the WP articles (i.e. not a spelling error)
- The "Affect" feature must be interpret very carefully because of the "quick and dirty" method of extraction (e.g. top "affect words" are *expect* and *beautifull* in health fora, *believe* and *ask* in WikiDisc, *play* et *considered* in WP articles.

More details in [HDL15]

# How to start a post in a Talk Page ?

## Extraction of N-grams (n < 4) occurring at the beginning of a post

| | Post/Sentence Initial | %of Msg |
|---|---|---|
| Discussions | 99.9 | 18 |
| Avis | 99.7 | 17 |
| supprimer | 97.1 | 6 |
| conserver | 95.7 | 5 |
| Neutre | 98.9 | 4 |
| Votes | 99.8 | 3 |
| Signalé_par | 98.8 | 3 |
| des_articles_admissibles | 99.2 | 3 |
| Si_vous_êtes | 15.6 | 2 |
| Il_me_semble | 32.8 | 2 |
| Bilan | 97.9 | 2 |
| Il_y_a | 23.4 | 2 |
| Merci | 28.2 | 2 |
| Ce_n'_est | 22.8 | 2 |
| pourBA | 98.7 | 2 |
| Bon | 53.6 | 2 |
| Je_viens_de | 60.8 | 2 |
| Je_ne_vois | 32.3 | 1 |
| Je_suis_d'accord | 59.7 | 1 |
| En_effet | 23.0 | 1 |
| Je_pense_qu' | 30.1 | 1 |
| Effectivement | 57.4 | 1 |
| pour | 88.1 | 1 |
| Je_ne_sais | 29.9 | 1 |

- Post/Sentence Initial : N-grams specificity for starting a post (and not simply starting a sentence, either or not the first one of the post)
- %of Msg : % of posts starting with this N-gram

# How to start a post in a Talk Page ?

## Extraction of N-grams (n < 4) occurring at the beginning of a post

| | Post/Sentence Initial | %of Msg |
|---|---|---|
| **Discussions** | **99.9** | 18 |
| **Avis** | **99.7** | 17 |
| **supprimer** | **97.1** | 6 |
| **conserver** | **95.7** | 5 |
| **Neutre** | **98.9** | 4 |
| **Votes** | **99.8** | 3 |
| **Signalé_par** | **98.8** | 3 |
| **des_articles_admissibles** | **99.2** | 3 |
| Si_vous_êtes | 15.6 | 2 |
| Il_me_semble | 32.8 | 2 |
| **Bilan** | **97.9** | 2 |
| Il_y_a | 23.4 | 2 |
| Merci | 28.2 | 2 |
| Ce_n'_est | 22.8 | 2 |
| **pourBA** | **98.7** | 2 |
| Bon | 53.6 | 2 |
| Je_viens_de | 60.8 | 2 |
| Je_ne_vois | 32.3 | 1 |
| Je_suis_d'accord | 59.7 | 1 |
| En_effet | 23.0 | 1 |
| Je_pense_qu' | 30.1 | 1 |
| Effectivement | 57.4 | 1 |
| pour | 88.1 | 1 |
| Je_ne_sais | 29.9 | 1 |

- if Post/Sentence Initial $> 85$

$\Rightarrow$ language specific to WP (e.g. for voting pro or against the article removing or the article ranking as good article)

- if Post/Sentence Initial $< 85$

$\Rightarrow$ Argumentative marker, towards a consensus

# 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*

# Plan

# 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)

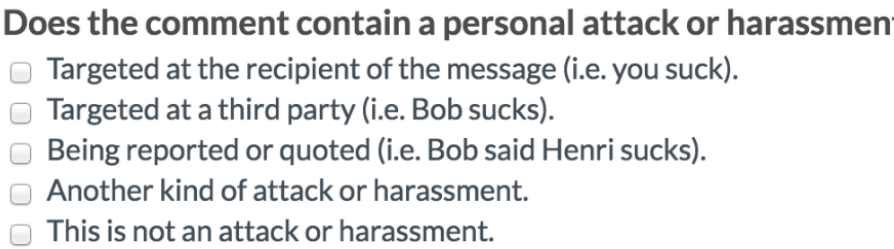



Figure 2: The question posed to our Crowdflower annotators.

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)

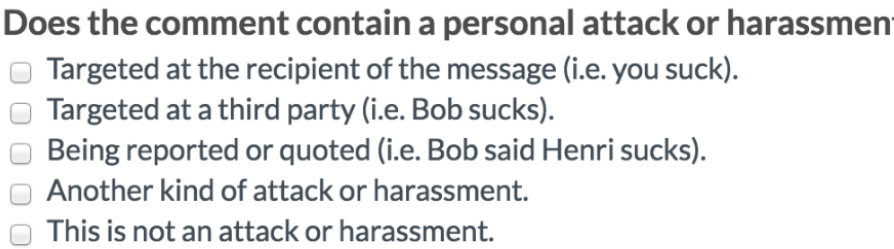



Figure 2: The question posed to our Crowdflower annotators.

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*

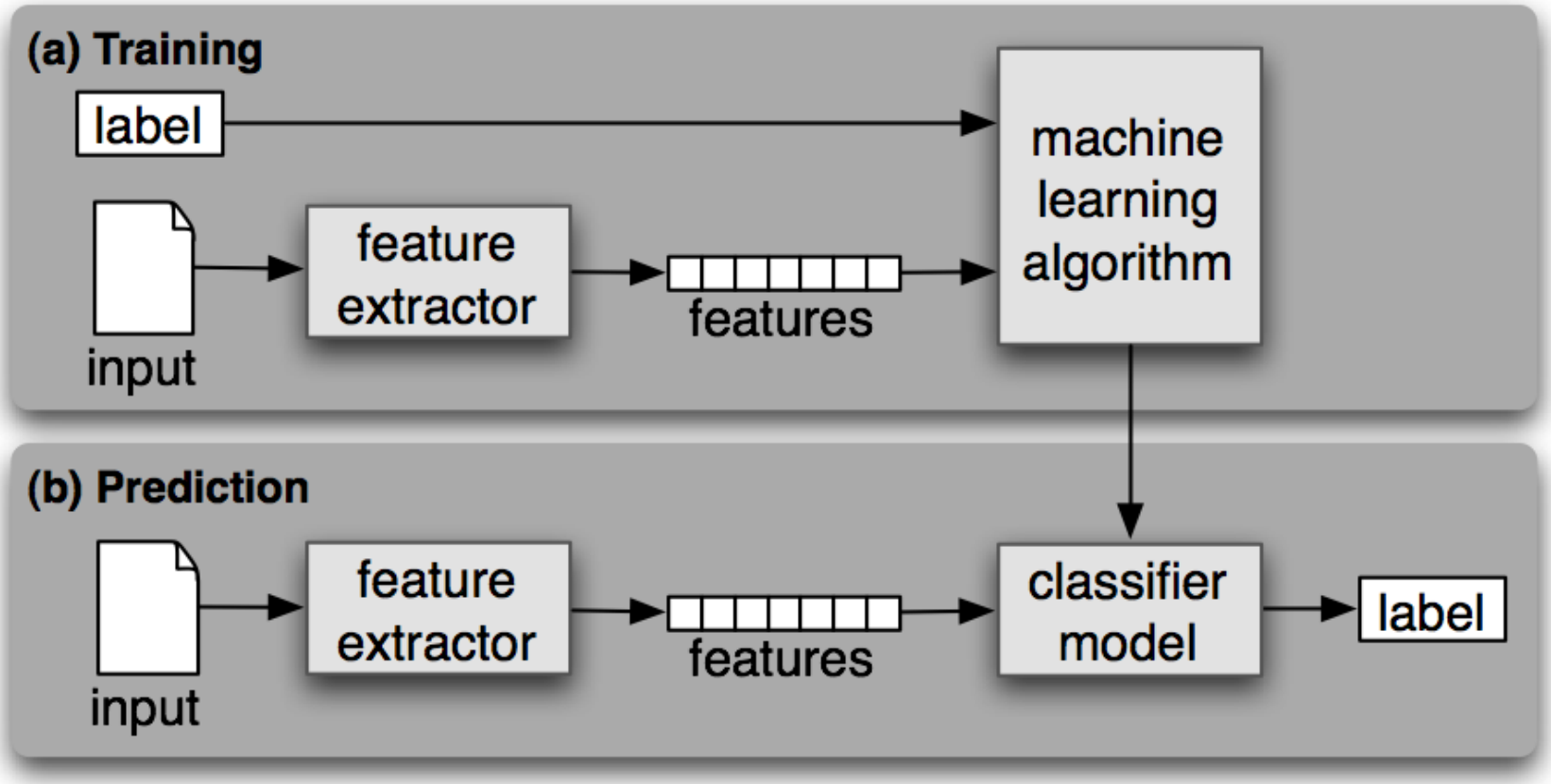



Figure extracted from [BKL09]

# 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 Talk Page 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 (119,359 talk pages)
- 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

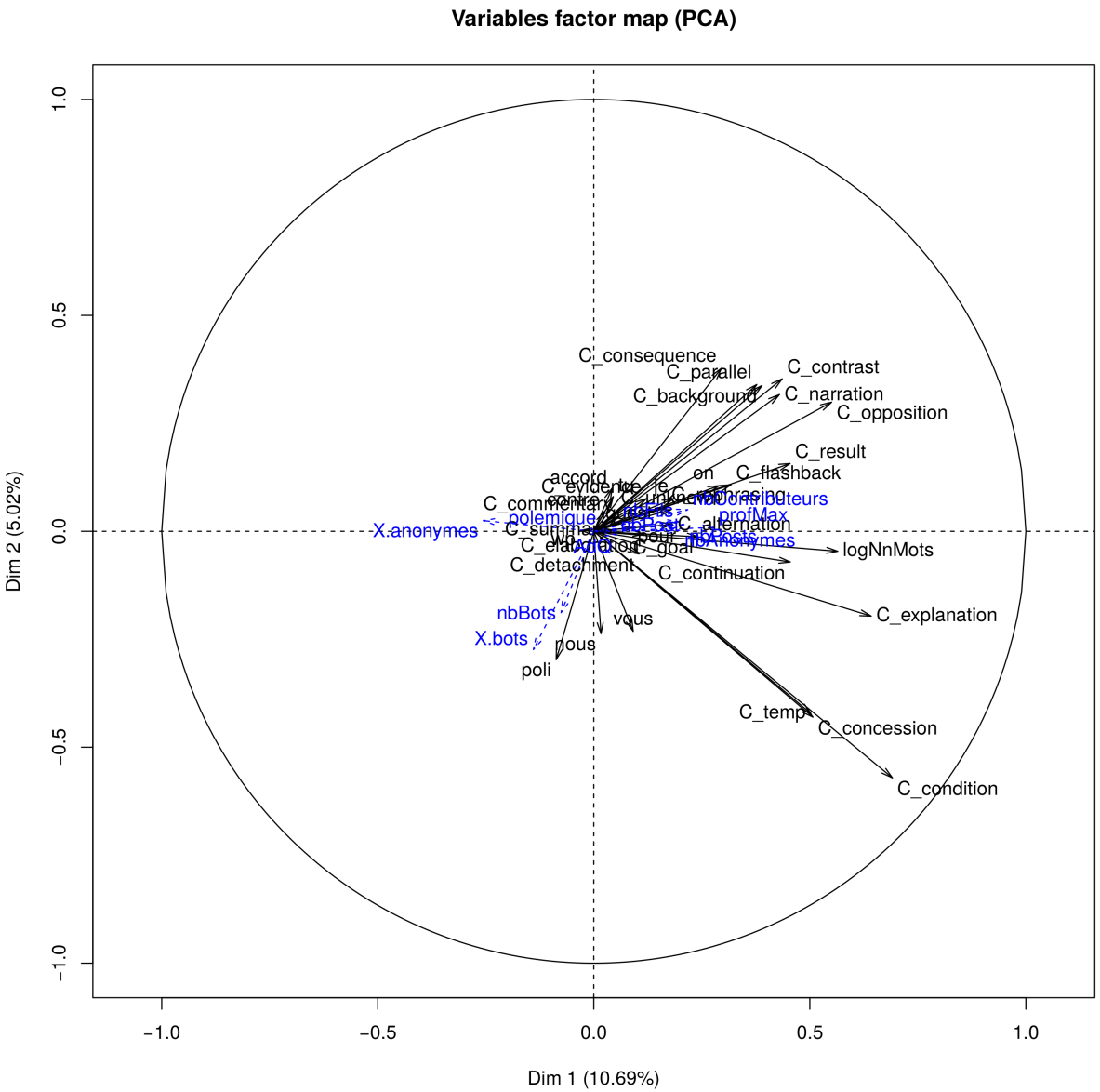
# Results : ACP on linguistic features

- 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

- A first annotation on two WP[FR] talk pages shows that only 50% of threads in two *a priori* conflict talk pages are conflict threads
- 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 complex and wide research area

- headings study
- portals specificities
- link between articles and talk pages
- 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 ?

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.

Steven Bird, Ewan Klein, and Edward Loper.

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

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.