French Wikipedia Corpus
The WikiTalk Corpus

Lydia-Mai Ho-Dac

University of Toulouse – CLLE-ERSS

June 19th 2017

Wikipedia Talk pages

Online discussion associated with each article where Wikipedian can discuss the ongoing writing process with other Wikipedian

> *"The user discussions on the article Talk pages might shed light on this issue and give an insight into the otherwise hidden processes of collaboration that, until now, could only be analyzed via interviews or group observations in experimental settings." [Ferschke et al., 2012].*

Global Objectives

- Linguistic description of the "online discussion genre" starting from Wikipedia talk pages (relatively well-written in contrast with more popular fora)[Ho-Dac and Laippala, 2015]
- Defining efficient discriminating features for Web classification and more precisely interactions classification i.e. question-answering, debate, co-working, etc.
- Linguistic description of the interactions that take place in the Wikipedia community

WikiTalk Corpus building

Talk pages extraction from the official dump

global backup frwiki-20150512-pages-meta-current#.xml.bz2 available on http ://dumps.wikimedia.org/frwiki/20150512/)

Document structure encoding acc. to *light* TEI-P5

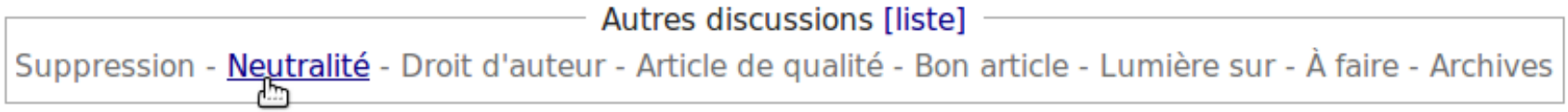
- threads marked up as `<div>`
- threads topic : `<head>`
- posts : `<post who="user" when="timestamp" interactionLevel="#">`

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

Corpus Structure

Meta-Data

- "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)



Profiling for building more homogeneous corpora

First profiles

- 55% single post talks
- 50% under 53 words talks
- Some extremely long talks (up to 1,143 posts and 148,968 words)
- 44% main talk pages (non parallel talk pages)
- 10% talk pages involving between 8 and 228 different writers
- 23% monologue threads (in main talk pages)
- 26% dialogue threads (in main talk pages) i.e. interactions between two identified users
- 2.5% dialogue threads (in main talk pages) with more than 2 posts (i.e. user A posts a message, user B answers)
- 5% debate threads (in main talk pages) i.e. a lot of different users without real exchange

77,536 (19%) relevant threads for a deeper analysis of interaction
(at least 2 users posting more than one single message)

Ferschke, O., Gurevych, I., and Chebotar, Y. (2012).

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.

Ho-Dac, L.-M. and Laippala, V. (2015).

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