#ww1 and the Centenary of the Great War on Twitter

Frédéric Clavert, unil, Journées suisse d’histoire, 9-10 juin 2016
@inactinique / frederic.clavert@unil.ch
A corpus of (yet) two million tweets

- April 2014-April 2016
- 2,096,968 tweets
  - Few noise but #11Nov #Somme #Verdun
  - 730,111 without retweets
- 542,570 accounts
  - Of all sorts
  - Individuals, memory-based institutions, media, research
- 124,424 hashtags
  - 54,566 just once in the database
  - 107,047 ten times or less in the database
- Harvesting still ongoing

- projects, bots...
Technical infrastructure

- LAMP server
- PHP script: 140dev
  - Connects to Twitter Public Streaming API
  - Collect tweets based on keywords (#ww1, #1gm, #ewk and many others)
  - Parse JSON into MySQL
- Export to CSV for data analyses
  - No «real time» analysis
How to read two millions tweets?
Basic statistics

Number of tweets per day
Clustering

- Text as statistics (text-mining)
- Software: Iramuteq
  - Groups similar tweets in clusters
  - Similarity = co-occurrences of words
  - Tree diagram (dendogram)
  - Words that are shown are the most typical words of each cluster
Network analysis / visualization
As a conclusion

- Understanding the missing parts
  - Where are the German Twitter accounts?
  - Where are prisoners, women, colonial troops, soldiers who did not fall on the battlefield?
  - Where are other war fronts than the West-European one?

- Where are the weak signals?