The #Idéo2017 platform

Julien Longhi

The #Idéo2017 platform

Julien Longhi Université de Cergy-Pontoise – Centre de Recherche AGORA (julien.longhi@u-cergy.fr)
Members of the project: Hassine N., Marinica C., Borzic B., Alkhouli A

General context: the CoMeRe project

- Aims to build a kernal corpus of computer-mediated communication (CMC) genres in French
- Mono and multimodal interactions stemming from networks including the Internet and telecommunications that may be synchronous or asynchronous
- Members had previously collected and structured different types of CMC corpora within their local teams (in a variety of formats with disparities in corpus compilation choices)
- Corpora are structured and referred to in a *uniform* way in order that they may form part of the forthcoming *French National Reference Corpus*

Project website

First project: the *Polititweet* corpora

- Development of the Interaction Space (IS) model to model CMC interaction (Chanier & Jin, 2013).
- Includes descriptions of time, set of participants, online location(s) defined by the properties of the set of environments used by participants.
- Description of the IS within the TEI header and messages and turns encoded in the TEI body using a common <post> element

```
<post xml:id="cmr-polititweets-a392322730999046144" who="#cmr-polititweets-p131211166"
 when="2013-10-21T18:13:22" xml:lang="fra">
 <p>Une préfiguration de la CTAP sur le thème du sport et des <distinct
    type="twitter-hashtag"><ident>#</ident><rs
     ref="https://twitter.com/search?q=%23CREPS&amp;src=hash">CREPS</rs></distinct>
  pourrait voir le jour en Poitou-Charentes.</p>
 <trailer>
  <fs>
   <f name="medium">
    <string>web</string>
   </f>
   <f name="retweetcount">
    <numeric value="4"/>
   </f>
  </fs>
 </trailer>
</post>
```

Openess

- Corpora released as open-data – paves way for scientific examination, replication and cumulative analyses
- Released on ORTOLANG (French equivalent of DARIAH the European infrastructure for Humanities)
- Bibliographic reference created for each corpus and given in <*titleSmt*> of TEI header. e.g.

Corpus depository

Longhi J., Marinica C., Borzic B., Alkhouli A., 2014: Polititweets, corpus de tweets provenant de comptes politiques influents. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets- tei-v1]. https://repository.ortolang.fr/api/content/comere/v3.3/cmr-polititweets.html

Interests and limits

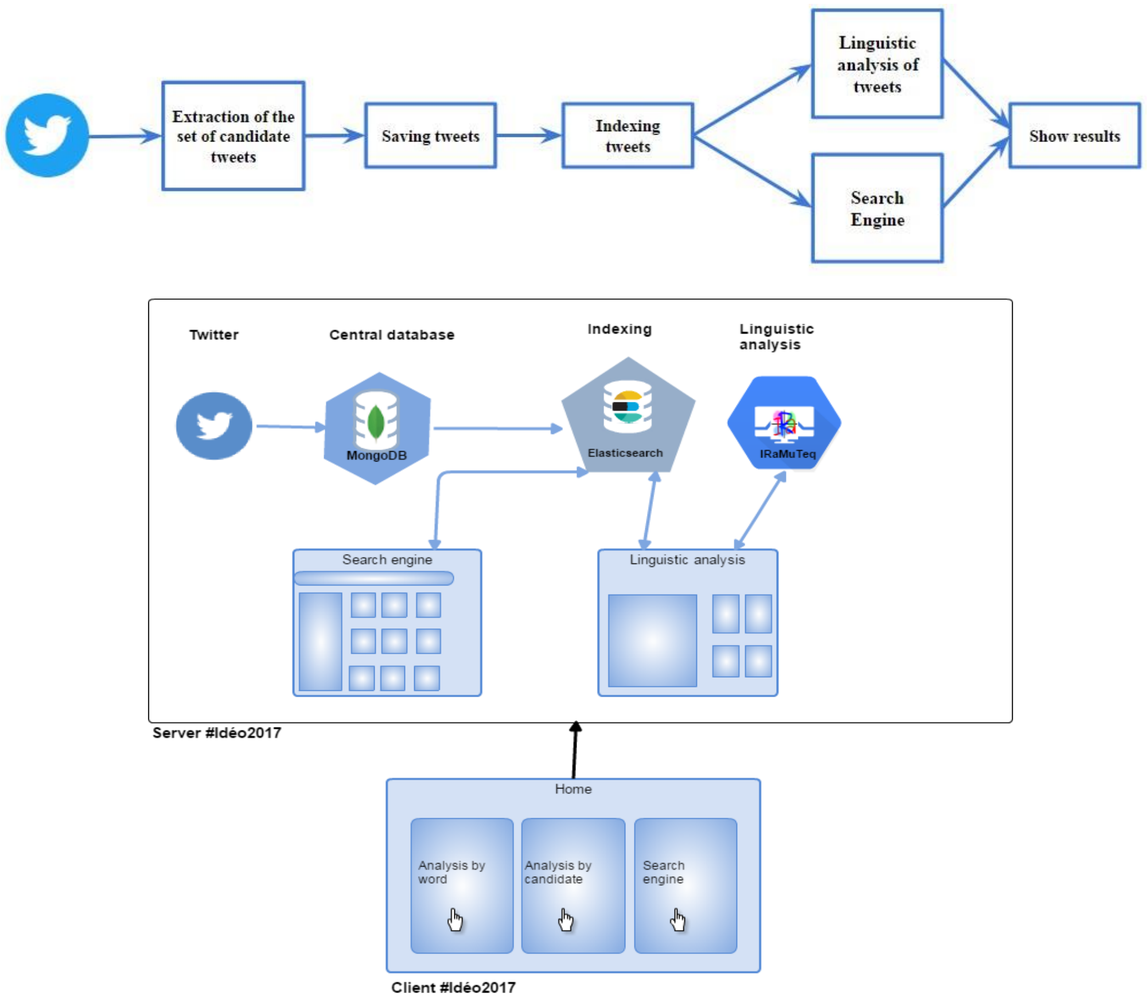
- The analysis of political tweets during the election campaigns, or specific events, is increasing and can be seen as a specific type of political discourse (Longhi, 2013)
- Extensive literature on the analysis of political tweets: but these works are difficult to gather because they come either from the computer sciences, either from the humanities and social sciences.
- Despite the unquestionable interest in outlining political facts, citizens do not have access to these results.

Challenges of #Idéo2017

- #Idéo2017 is a web platform available online allowing the analysis of the messages, posted on Twitter, related to political news.
- Its objective is to make available on the web for average citizens a set of statistical analyses and data visualization tools applied to the Twitter messages. #Idéo2017 allowed French electors to analyze the discourse of the candidates by means of their tweets.
- Citizens can make their own queries (based on linguistic and textometric criteria, more precisely, the most commonly used words by political personalities, analyses of similarities, ALCESTE algorithm, etc.) and obtain comprehensible results.

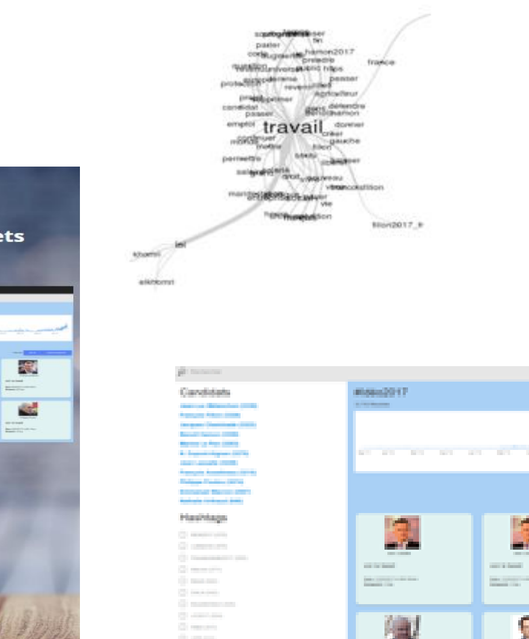
Tools and analysis

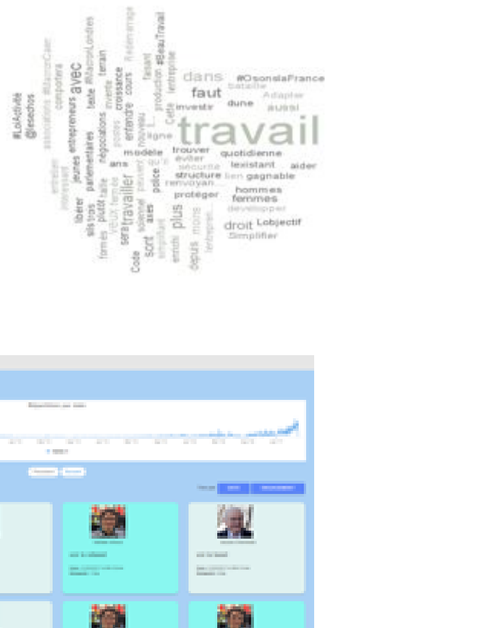
Tool Development

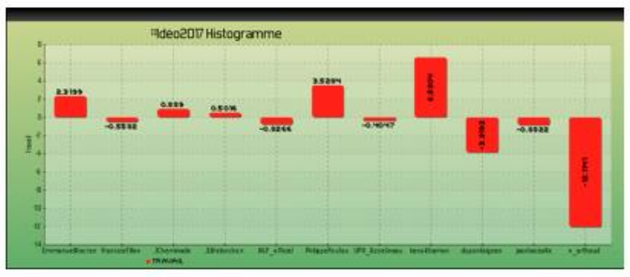


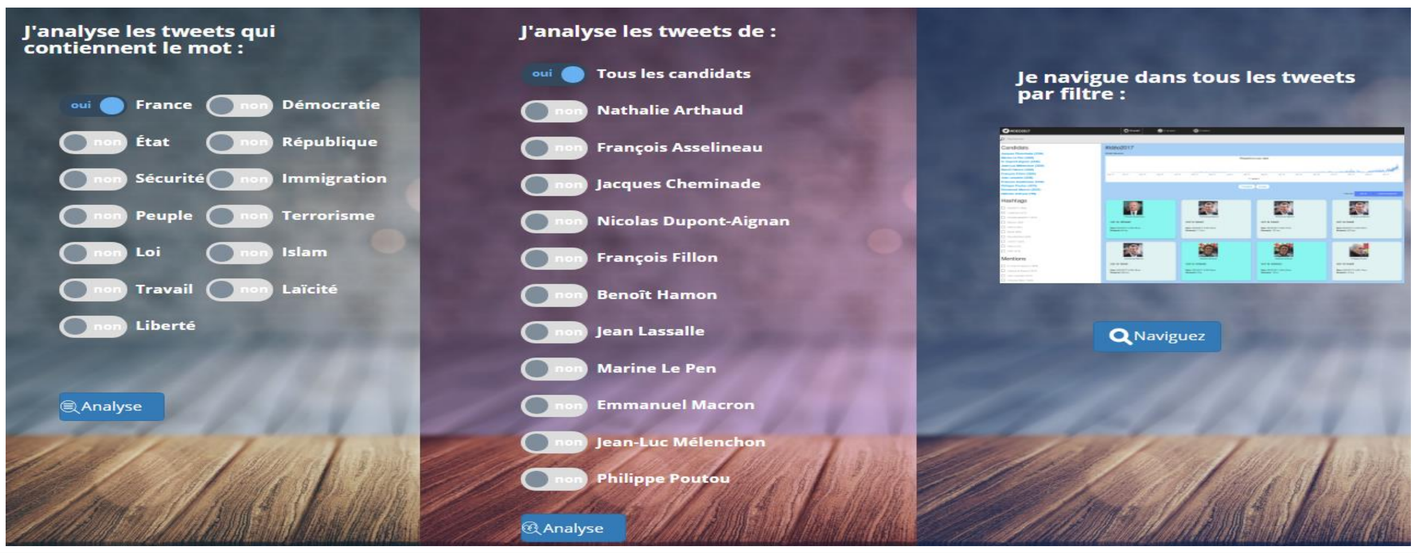
Analysis

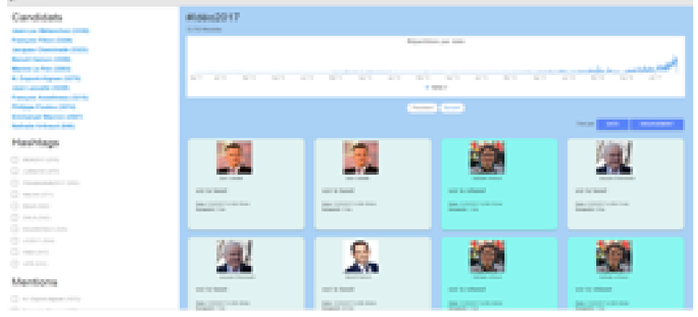
The #Ideo2017 platform proposes:
The Analysis “I Analyze the Tweets that Contain the Word [Word]”
The Analysis “I Analyze the Tweets of [Candidate]”
The Search Engine











Conclusions

- #Idéo2017 combines different technologies and inputs, which give citizens the opportunity to grasp a part of the discursive issues of the election.
- This development, which can be enriched, allows citizen to easily use a set of features usually accessible by software requiring different transformations of the data.
- For the period from September 1st 2016 to May 7th 2017, 42290 tweets were extracted for the 11 candidates. These tweets were gathered in a collection published in a TEI corpus in the standards of the Ortolang platform, thanks to the Corli Consortium.

CoMeRe Repository (2014). Repository for the CoMeRe corpora [website], http://hdl.handle.net/11403/comere
Burnard, L. & Bauman, S. (2013). TEI P5: Guidelines for electronic text encoding and interchange. TEI consortium, *tei-c.org*. http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf
Chanier, T., Poudat, C., Sagot, B., Antoniadis, G., Wigham, C.R., Hriba, L., Longhi, J. & Seddah, D. (2014). The CoMeRe corpus for French: structuring and annotation heterogeneous CMC genres, in Beißwenger, M., Oostdijk, N., Storrer, A & van den Heuvel, H. Building and Annotating Corpora of Computer-Mediated Discourse: Issues and Challenges at the Interface of Corpus and Computational Linguistics, *Journal of Language Technology and Computational Linguistics* (special issue). pp1-31. http://www.jlcl.org/2014_Heft2/Heft2-2014.pdf
DCMI (2014). Dublin Core Metadata Initiative. http://dublincore.org/
Laboratoire AGORA (AGORA), Equipes Traitement de l'Information et Systèmes (ETIS) (2017). *Présidentielle2017: corpus des tweets de la #présidentielle2017* [Corpus]. ORTOLANG (Open Resources and TOols for LANGuage) - www.ortolang.fr, https://hdl.handle.net/11403/corpus-presidentielle2017/v1.
Longhi J., Marinica C., Borzic B., Alkhouli A., 2014: Polititweets, corpus de tweets provenant de comptes politiques influents. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-polititweets- tei-v1]: https://repository.ortolang.fr/api/content/comere/v3.3/cmr-polititweets.html
Longhi, J., Wigham, C. (2015). Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow. Poster presented in Corpus Linguistics 2015, Jul 2015, Lancaster, United Kingdom, available in: https://halshs.archives-ouvertes.fr/halshs-01176061.
Longhi J. (dir.), 2017 : #Idéo2017, plateforme d'analyse de tweets politiques en campagne électorale.
OLAC. (2008). Best Practice Recommendations for Language Resource Description. *Open Language Archives Community*. University of Pennsylvania. http://www.languagearchives.org/REC/bpr.html
ORTOLANG (2013). Open Resources and TOols for LANGuage [website]. ATILF / CNRS - Université de Lorraine: Nancy, http://www.ortolang.fr