The #intermittent corpus: corpus features, ethics and workflow for a CMC corpus of tweets in TEI

Julien Longhi

The #intermittent corpus: corpus features, ethics and workflow for a CMC corpus of tweets in TEI

Julien Longhi Université de Cergy-Pontoise – Centre de Recherche AGORA (julien.longhi@u-cergy.fr)

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

Corpus structuring in TEI (cf. Longhi & Wigham 2016)

- Development of the Interaction Space (IS) model to model CMC interactions (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-politweets-a392322730999046144" who="#cmr-politweets-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>
```

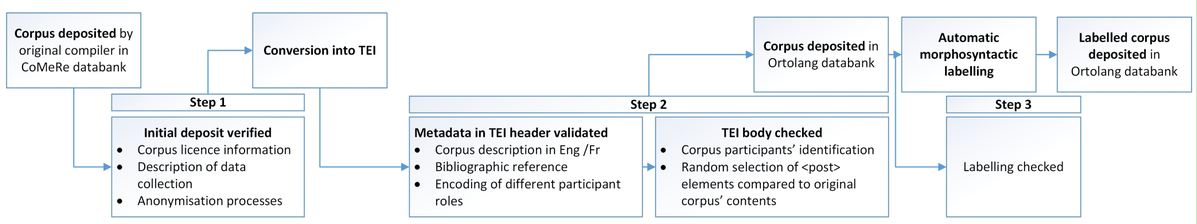
Openess

- 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., Borzic, B. & Alkhouli, A. (2016). #Intermittent: constitution d'un corpus lié à un événement discursif controversé. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. https://repository.ortolang.fr/api/content/comere/v2/cmr-intermittent/cmr-intermittent.html

Staged quality control process



The context of #Intermittent

In March, 2014, a new agreement concerning unemployment benefits for French arts workers was signed by social partners. Protest movements and mass demonstrations took place in Paris and in other French cities for several days. These reactions rapidly invaded social networks, especially Twitter. Millions of tweets were written as soon as the first information about this controversy emerged.

It is hoped that the constitution of the corpus #intermittent will enable researchers to work on this kind of discourse (tweets related to a controversial topic), to characterize it and understand it from different angles in order to extend previous work (Longhi , 2006) concerning the French arts workers around 2003/2004.

The choice of tweets was led by the following steps: 1) In 2015, with the creation of a threshold of at least 10 tweets with the #intermittent(s): we identified 215 accounts , each of which had produced at least 10 tweets explicitly referenced as contributing to this theme (in order to have representative accounts); 2) By gathering all of the tweets sent by those 215 people, we collected 586, 239 tweets; 3) 10,876 of the 586, 239 tweets contained the # : #intermittent(s): The #intermittent corpus corresponds to these 10, 876 tweets.

The corpus can be downloaded at https://repository.ortolang.fr/api/content/comere/v2/cmr-intermittent/cmr-intermittent.html

Including specific features of Twitter

The @

Type of material

Interaction: reply to @ludivineoff

The #

```
<post xml:id="cmr-intermittentstweets-a319968741280583681"
 who="#cmr-intermittents-p212230994" when="2013-04-05T02:24:27.0" xml:lang="fra">
 <p>
  <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
   ref="https://twitter.com/ludivineoff 508264348 "
  >ludivineoff</addressee></addressingTerm> et après on dit que les <distinct
  type="twitter-hashtag"><ident>#</ident><rs
   ref="https://twitter.com/search?q=%23intermittents&amp;src=hash"
   >intermittents</rs></distinct> sont des fainéants ;) (tu l'es encore ? tu l'as
 été, non ?) <distinct type="twitter-hashtag"><ident>#</ident><rs
   ref="https://twitter.com/search?q=%23fausseidéereçue&amp;src=hash"
   >fausseidéereçue</rs></distinct>
 </p>
 <trailer>
  <fs>
   <f name="medium">
    <string>Twitter Web Client</string>
   </f>
   <f name="inReplyToStatusId">
    <numeric value="319967636916150272"/>
   </f>
   <f name="inReplyToUserId">
    <numeric value="508264348"/>
   </f>
   <f name="inReplyToScreenName">
    <string>ludivineoff</string>
   </f>
  </fs>
 </trailer>
</post>
```

Ethical issues

On https://twitter.com/tos?lang=en we can read:

8. Restrictions on Content and Use of the Services

Please review the Twitter Rules (which are part of these Terms) to better understand what is prohibited on the Service. We reserve the right at all times (but will not have an obligation) to remove or refuse to distribute any Content on the Services, to suspend or terminate users, and to reclaim usernames without liability to you. We also reserve the right to access, read, preserve, and disclose any information as we reasonably believe is necessary to (i) satisfy any applicable law, regulation, legal process or governmental request, (ii) enforce the Terms, including investigation of potential violations hereof, (iii) detect, prevent, or otherwise address fraud, security or technical issues, (iv) respond to user support requests, or (v) protect the rights, property or safety of Twitter, its users and the public.

Twitter does not disclose personally identifying information to third parties except in accordance with their Privacy Policy.

Except as permitted through the Services, these Terms, or the terms provided on dev.twitter.com, you have to use the Twitter API if you want to reproduce, modify, create derivative works, distribute, sell, transfer, publicly display, publicly perform, transmit, or otherwise use the Content or Services.

Twitter encourages and allows broad re-use of content. The Twitter API exists to enable this.

Conclusion and perspectives

- Initial textometrical analysis shows differences with previous research abourt art workers represented in *Le Monde* and *Le Figaro*. While the controversy was described in press from different viewpoints, tweets focus on a community that criticise and question the signed agreement. On the one hand, some tweets describe the precariousness of the French arts workers and their various protest movements against the new regime. On the other hand, the tweets denounce the impartiality of the agreement, with links providing information about the act and citing various political personalities involved in the controversy.
- Members of the CoMeRe project, working with other European partners, participate in the TEI CMC Special Interest Group. They are jointly working on a proposal for an extension to the TEI standard adapted to the particularities of a broad range of CMC genres.
- This corpus illustrates the benefits of using social media for analyzing social controversy.

TEI CMC SIG

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/
Longhi (2006). De intermittent du spectacle à intermittent : de la représentation à la nomination d'un objet du discours. *Corela*, n°4 vol. 2, URL : http://corela.revues.org/457.
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., Borzic, B. & Alkhouli, A. (2016). #Intermittent: constitution d'un corpus lié à un événement discursif controversé. In Chanier T. (ed) *Banque de corpus CoMeRe*. Ortolang.fr : Nancy. https://repository.ortolang.fr/api/content/comere/v2/cmr-intermittent/cmr-intermittent.html
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
Reynaert, M., Oostdijk, N., De Clercq, O., van den Heuvel, H., & de Jong, F. (2010). Balancing SoNaR: IPR versus Processing Issues in a 500-million-Word Written Dutch Reference Corpus. In, Seventh conference on International Language Resources and Evaluation, LREC '10, 19-21 May 2010, Malta. http://doc.utwente.nl/72111/1/LREC2010_549_Paper_SoNaR.pdf.

IRD2016