Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow

Julien Longhi, Ciara R. Wigham

Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow

Julien Longhi Université de Cergy-Pontoise – Centre de Recherche AGORA (julien.longhi@u-cergy.fr)
Ciara R. Wigham Université Lumière Lyon 2 – ICAR (ciara.wigham@univ-lyon2.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

- 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-politweets-a393227309990461 44" 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>
```

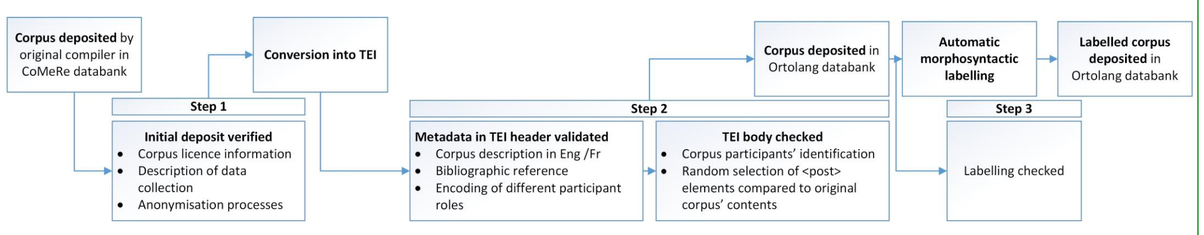
Openess

- 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., 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. http://hdl.handle.net/11403/comere/cmr-polititweets

Staged quality control process



The context of politweets: HumaNum_DJ

This work is part of the "Digital Humanities and Data Journalism" transdisciplinary project (funded by the Foundation of the Cergy-Pontoise University, France).
The purpose of this research was to take advantage of discourses produced on social media to leverage the semantics of discourse in relation to social issues.
The corpus was built starting from seven French politicians from six different political parties. In order to generate political tweets, a set of lists citing these politicians was generated (7087 lists), and lists that have tweeted at least six times and for which the description contained the word 'politics' were selected (120 lists in total). Finally, 2934 tweets were recovered. In order to be sure that we selected politicians' tweets (and not, for example, those of journalists), only the accounts cited in more than 12 lists were considered; 205 politicians were tweeting. We took the last 200 tweets of each of the 205 accounts on 27 March 2014 (34,273 tweets). This allowed us to recover data that focused on the period between the two rounds of the 2014 municipal elections in France
Analyses have started to be carried out: some ideas have been launched in Djemili, Longhi *et al.* (2014) but further analyses must adhere rigorously to methodologies stemming from the natural language processing (NLP) field.

Including specific features of Twitter


```
<tweet id="448491380948885504">
    <author>
        <user_id>
            80820758
        </user_id>
        <name>
            Jean-Luc Mélenchon
        </name>
        <screen_name>
            JLMelenchon
        </screen_name>
    </author>
    <creation_date>
        2014-03-25 17:07:32.0
    </creation_date>
    <tweet_text>
        RT @LePG: A #Bruxelles, @JLMelenchon conclue la rencontre-débat sur le #GMT. Nous live-tweetons - #UE #USA #Europe - http://t.co/hNGmnAagIM
    </tweet_text>
    <entities_hashtags>
        Bruxelles,GMT,UE,USA,Europe
    </entities_hashtags>
    <entities_urls>

    </entities_urls>
    <entities_user_mentions>
        35022597,80820758
    </entities_user_mentions>
    <geo_lat>
        0.0
    </geo_lat>
    <geo_long>
        0.0
    </geo_long>
    <source>
        <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
    </source>
```

The @

The #

Type of material

```
<post xml:id="cmr-politweets-a448491380948885504" who="#cmr-politweets-p80820758"
 when="2014-03-25T17:07:32" xml:lang="fra">
  <p><distinct type="twitter-retweet"><ident>RT</ident>
      <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
          ref="https://twitter.com/LePG 35022597"
        >LePG</addressee></addressingTerm>:</distinct> A <distinct type="twitter-hashtag"
        ><ident>#</ident><rs ref="https://twitter.com/search?q=%23Bruxelles&amp;src=hash"
        >Bruxelles</rs></distinct>,
        <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
        ref="#cmr-politweets-p80820758">JLMelenchon</addressee></addressingTerm> conclue
    la rencontre-débat sur le <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23GMT&amp;src=hash">GMT</rs></distinct>. Nous
    live-tweetons - <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23UE&amp;src=hash">UE</rs></distinct>
    <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23USA&amp;src=hash">USA</rs></distinct>
    <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23Europe&amp;src=hash">Europe</rs></distinct> -
      <ref target="http://t.co/hNGmnAagIM">http://t.co/hNGmnAagIM</ref></p>
  <trailer>
    <fs>
      <f name="medium">
        <string>Twitter for iPhone</string>
      </f>
      <f name="retweetcount">
        <numeric value="22"/>
      </f>
      <f name="isRetweet">
        <binary value="true"/>
      </f>
      <f name="retweetedstatus_id">
        <numeric value="448490939338588160"/>
      </f>
    </fs>
  </trailer>
</post>
```

The #

The @

Type of material

Is retweeted/ number of retweets

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

❖In Djemili, Longhi *et al.* 2014, the main objective was to detect whether or not a tweet is an ideology tweet. We tested a system against a set of 20400 tweets of French politicians in order to experiment rules' implementation and their accuracy. The evaluation of the rules and their implementation gave us good results for the system's accuracy since 66.66% of tweets identified as ideological were indeed so and 96.64% of tweets identified as non-ideological (after sampling) were validated as non-ideological by the expert.

❖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.

❖ Members of the CoMeRe project are organising the international research days (IRDs) on Social Media and CMC Corpora for the eHumanities to be held in Rennes, France on 23-24th October 2015. See http://ird-cmc-rennes.sciencesconf.org/

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
Djemili S., Longhi J., Marinica C., Kotzinos D. & Sarfati G.-E. (2014). « What does Twitter have to say about ideology », Konvens 2014 - *Workshop proceedings vol. 1* (NLP 4 CMC: Natural Language Processing for Computer-Mediated Communication / Social Media – Pre-conference workshop at Konvens2014) , Germany (2014), p.16-25.
DCMI (2014). Dublin Core Metadata Initiative. http://dublincore.org/
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. http://hdl.handle.net/11403/comere/cmr-polititweets
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.