Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow
Julien Longhi, Ciara R. Wigham

To cite this version:
Julien Longhi, Ciara R. Wigham. Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow. Corpus Linguistics 2015, Jul 2015, Lancaster, United Kingdom. 2015. halshs-01176061

HAL Id: halshs-01176061
https://halshs.archives-ouvertes.fr/halshs-01176061
Submitted on 14 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
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)
Ciera R. Wigham
Université Lumière Lyon 2 – ICAR (ciera.wigham@univ-lyon2.fr)

The CoMeRe project

- Aims to build a thesaural 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 have 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

Corpus structuring in TEI

- Development of the Interaction Space (IS) model to model CMC interaction (Charlier & 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

Openness

- Corpus 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)

Staged quality control process

- Quality control frameworks and processes that are used to define quality and to assess and, if necessary, to guarantee quality
- Staged quality control process, full documentation of the process and the results produced, instance of the process for this corpus

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 (7867 lists), and all accounts that have tweeted at least twice during six months for which the description contained the word ‘politique’ were selected (120 lists in total). Finally, 2394 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 Djamli, 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

- Various fields for the description of Twitter, including the number of retweets, the number of followers, the number of mentions, the number of replies, etc.
- Features that are used to describe the nature of the tweets, such as the type of material they contain.

Ethical issues

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

B. 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 (i) to satisfy any applicable law, regulation, legal process or governmental request, (ii) to 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 Djamli, 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 20460 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 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 23-24 October 2015. See http://ird-cmc-rences-sciencesconf.org/