The #Intermittent Corpus: Corpus Features, Ethics and Workflow for a CMC Corpus of Tweets in TEI

Julien Longhi
Cergy-Pontoise University, AGORA
E-mail: julien.longhi@u-cergy.fr

Abstract

This poster aims to describe issues encountered whilst structuring a corpus of tweets compiled from the key word intermittent (arts worker) in order to analyse a discursive topic related to the controversy surrounding the status of French arts workers. This corpus is part of the CoMeRe project (CoMeRe, 2014): it aims to build a kernel corpus of computer-mediated communication (CMC) genres with interactions in the French language. Three key words characterize the project: variety, standards and openness. A variety of interactions was sought: public or private interactions as well as interactions from informal, learning and professional situations. The CoMeRe project structured the corpora in a uniform way using the Text Encoding Initiative format (TEI, Burnard & Bauman, 2013) and described each corpus using Dublin Core and OLAC standards for metadata (DCMI, 2014; OLAC, 2008). The TEI model was extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse (Chanier et al., 2014). The term 'openness' also characterizes the project: The corpora have been released as open data on the French national platform of linguistic resources (ORTOLANG, 2013) in order to pave the way for scientific examination by partners not involved in the project as well as replicative and cumulative research.

This poster presentation aims to give an overview of the corpus building process using, as a case study, a corpus of tweets cmr-intermittent (Longhi et al., 2016). The following steps led to the choice of tweets:

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 poster will focus, firstly, on how features that are specific to Twitter were included and structured in the interaction space TEI model. We will exemplify how certain features are accounted for in TEI. These include hashtags that label tweets in order that other users can see tweets on the same topic and at signs that allow users to mention or reply to other users. Secondly, the poster will evoke some of the ethical and rights issues that had to be considered before publishing this corpus of tweets. Finally, the workflow and multi-stage quality control procedure adopted during the corpus building process will be illustrated.

Keywords: tweets, corpus, TEI, CMC corpora