Éléments d'acquisition du corpus « #Intermittent, constitution d'un corpus lié à un événement discursif controversé »

V2

Version du 8 décembre 2015, Julien Longhi

Pour citer ce document :

Longhi J. (2015). Éléments d'acquisition du corpus « #Intermittent, constitution d'un corpus lié à un événement discursif controversé », In Longhi, J., Borzic, B., Alkhouli, A. (2016). *#Intermittent: constitution d'un corpus lié à un événement discursif controversé*. Banque de corpus CoMeRe. Ortolang.fr : Nancy. [cmr-intermittent-tei-v1]

Coordinateur de la tâche : Julien Longhi

Participants : Julien Longhi, Boris Borzic, Abdulhafiz Alkhouli.

Contenu

1. Objectifs et contraintes de départ

La finalisation de ce corpus a été possible grâce au soutien financier d'Ortolang. La demande de financement portait sur la finalisation de constitution d'un corpus de tweets constitué à partir du mot-dièse (hashtag, #) intermittent, répertoriés dans une base de données de 13 074 tweets avec #intermittent(s) répartis en 4 617 twittos (usagers de Twitter) sur la période de juin à septembre 2014.

Le corpus était répertorié dans une base de données, d'où il a été possible d'extraire tous les champs du format TEI, tel que cela a été fait dans le corpus *Polititweets* du projet CoMeRe (extension proposée par T Chanier et des collègues européens, TEI-CMC).

Après la constitution d'un corpus de tweets politiques, Polititweets, ce second projet avait pour ambition de cerner une moment discursif (controverse sur le statut des intermittents) particulier.

Ce corpus a pour but, du point de vue discursif, de prolonger les travaux de Julien Longhi à propos des intermittents du spectacle autour de la période 2003/2004.

Quelques références :

- 2008 : « Sens communs et dynamiques sémantiques : l'objet discursif intermittent », *Langages*, n°170, p.109-124.
- 2006 : « *Permittent* et *interluttant*, deux néologismes entre lexique et discours », *Cahiers du L.C.P.E.*, n°7, p.95-109.
- 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.

2. Actions des différents participants

Creators: LONGHI Julien; BORZIC Boris
compiler: BORZIC Boris, ALKHOULI Abdulhafiz, LONGHI Julien
depositor: LONGHI Julien
editor: CHANIER Thierry
data_inputter: BORZIC Boris, ALKHOULI Abdulhafiz

Corpus de tweets qui complète la base en ligne CoMeRe dans Ortolang.
Soutien financier de Ortolang : 5000 euros

3. Type de corpus

Twitter est un médium de microbloggage qui permet aux utilisateurs de faire de courtes déclarations publiques, pour partager leur sentiment du moment. Techniquement, un tweet se compose de peu d'éléments, à savoir un utilisateur, un message de 140 caractères ou moins, et un espace de commentaire à la suite pour re-tweeter. Une des particularités de twitter est le "techno-langage" impliqué dans chaque tweet. On y retrouve le hastag (#), permettant de 'tagger' le tweet, et dont le nom est laissé libre de choix à l'utilisateur (ex : #France, #democratie), l'arobase (@) pour adresser son message à un utilisateur particulier, ou le mentionner, et des URLs réduites.

4. Aspects juridiques

Un travail sur cette question avait été fait pour le corpus *Polititweets* (produit dans le cadre du projet CoMeRe), et indiquait notamment les aspects suivants :

Cette licence signifie que vous nous autorisez à mettre vos Tweets à la disposition du reste du monde et que vous permettez aux autres d'en faire de même.

Twitter applique un ensemble évolutif de règles sur la manière dont les partenaires de l'écosystème peuvent interagir avec vos Contenus. Ces règles ont été conçues pour mettre en place un écosystème ouvert, tenant compte de vos droits. Mais ce qui vous appartient vous appartient – vous restez propriétaire de vos Contenus.

Aussi, **la question juridique n'est pas problématique** pour la constitution d'un corpus de tweets.

5. Méthodologie

Le processus de conversion a été manié dans le corpus *Politiweets* (http://comere.ortolang.fr/cmr-polititweets.html), et la base de données a été directement transformée **au format TEI-CMC**.

Concernant la base de données : le laboratoire ETIS impliqué dans la constitution a développé une application sur mesure en 3 étapes : 1) appel une dizaine de fonction de l'API Twitter selon nos besoins, ensuite nous récupérons toutes les informations sous format JSON que nous convertissons 2) ce qui nous permet d'enrichir une base de données Etis avec un design de base qui nous est propre (dizaine de table cinquantaine de champs), 3) ensuite nous pouvons faire un export sur mesure, avec une sous partie des informations stockées dans n'importe quel format de données.

Les données sont directement traitées pour être mises au format TEI du projet polititweets. Le corpus est accessible grâce à la licence *Creative Commons*.

La récupération des tweets a été guidée par le processus suivant :

En 2014 récupération de 13 074 tweets avec #intermittent(s) répartis en 4 617 personnes

En 2015, nous avons créé un seuil correspondant à au moins 10 tweets avec le # #intermitent(s) : nous arrivons à 215 comptes qui avaient donc produit au moins 10 tweets explicitement référencés comme appartenant à cette thématique (afin d'avoir des comptes représentatifs) En récupérant tous les tweets de ces 215 personnes, nous obtenons 586 239 tweets dont 10 876 avec le #intermittent(s) : le corpus #intermittent correspond donc à ces 10 876 tweets.