Twitter data as primary sources for historians: a critical approach

Lessons from two projects: the 2015 Greek referendum and the Centenary of the Great War on Twitter

Sofia Papastamkou (European Center for the Humanities and Social Sciences, Lille)
Frédéric Clavert (C2DH, University of Luxembourg)
Introduction
Introduction: social networks online, a definition

» We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site. «

Introduction: Twitter studies

- Since the beginning of Twitter, its data has been used in Humanities and Social Sciences for different purposes

- Two kinds of historian's work
  - Collective memory
  - Current time events
    Documenting the Now: https://www.docnow.io/
Introduction: why Twitter?

Because we can:

- Relatively (though less and less) open APIs
- Several free and one paying APIs
  - Search API (history)
  - Streaming API (what’s going on, < 1% of )
  - «If you pay you can get whatever you want» API (but we don’t pay, do we?)
- Lots of tools to collect tweets
  - twarc,
  - DMI-TCAT,
  - TAGS,
  - etc.
Introduction

Plan

I. Two projects, two theoretical backgrounds
II. Two projects, many methods and tools for the creation and the analysis of the corpus
III. Twitter hermeneutics
I. Two projects, two theoretical backgrounds
#ww1 - The collective memory of the Great War

- **Context: Centenary of the Great War**
  - First large series of commemorations in the social network online era
  - Multinational(-linguistic) comparisons possible (mainly French and English)

- **Collecting tweets related to the Great War**
  - **mainly** inductive approach

- **Studying collective memory in the digital era**
  - Digital memory studies (Andrew Hoskins)
  - Will collective memory «change» when confronted to information circulation on social networks online?
#greferendum - Studying the 2015 greek referendum

- **Context:** Greek debt crisis, Eurozone crisis,
  - A rich, born-digital (SNS), transnational documentation
  - A personal experimentation: archiving and analysing an event

- **Collecting the #greferendum tweets**
  - An *ad hoc* collect
  - Holistic approach (by hashtag)

- **Studying the event**
  - An important concept for historians (Seignobos 1898; Nora 1972; Le Goff 1999; Sirinelli 2002)
  - Twitter: the medium of the event

- **Studying Twitter as a source for historians**
  - non-institutional; decentralised; wild; born-digital
II. Two projects
Many methods and tools
● Collecting tweets
  ○ 140dev [abandoned] and [the incredible] DMI-TCAT
  ○ a regularly updated tool, that can manage the many and regular changes in Twitter API
  ○ 5 Millions+ tweets, 1 million users (and GPDR headache) stored in a MariaDB database

● Preparing Twitter data for analysis
  ○ spreadsheets, OpenRefine, Dataiku DSS, SQL query, etc.
  ○ «in-between» tools that we don’t always talk about (but we should)

● Analysing tweets
  ○ IRaMuTeQ (iramuteq.org) = data mining
  ○ Gephi = social networks analysis
  ○ Dataiku DSS / spreadsheets for simple stats
#greferendum: collecting tweets

- Dates: 6-16 July 2015 (the "international" phase)
- Holistic collect - main hashtag: #greferendum
- NodeXL: \( \approx 20,000 \) tweets per day

204 714 tweets:
- 139 945 retweets (68.36 %)
- 8 686 replies (4.24 %)
- 56 086 unique tweets (27.39 %)
#greferendum: preparing data for analysis

OpenRefine => data (:hashtag) cleaning (:clustering)

TEI P5 XML => (a very basic) text encoding of data (:tweet text) - subcorpus par date
#greferendum: analysing tweet data

- **Hashtags**
  Qualitative work: a typology of the most frequent hashtags (frq>99, 158 words)
  R (wordcloud package) => textual data visualisation (:hashtags)

- **Tweets**
  text statistical analysis (:cooccurrences) => TXM textométrie

- **Users**
  network metrics and visualisation => Gephi
  Qualitative work on most central accounts

- **Domains**: simple statistics with Voyant tools
Typology 1

Hashtag type

- brand
- common name
- compound
- geoname
- institution
- other
- person
- phrase
- spam
- unknown
Typology 2: hashtag function

Commentary: 14/158 (8,861 %)
Tag: 144/158 (91,139 %)
III. Twitter Hermeneutics
Hermeneutics of APIs

Twitter APIs constraints: choosing an API as the first step to interpretation

- **Search API**: 7 days in the past, around 3000 tweets per hour  
  (some workarounds: [https://github.com/taspinar/twitterscraper](https://github.com/taspinar/twitterscraper))  
  Either sampling / or small corpus

- **Streaming API**: anticipation of what will be the past  
  Limitation: 1% of the tweets that are being published  
  Progressive construction of massive corpus
Hermeneutics of keywords: hashtags

Most research on twitter are based on keywords/hashtags which means that:

- The studied object must be quite well-known by the researcher to find the best keywords
- Hashtags / keywords are not conversation
- Therefore
  - collecting massive data ≠ collecting exhaustive data
  - sampling data can be better than massively collecting data
- Numerous ways to understand what a hashtag is
Hashtags

- tell the big story (quantitative + relational analysis)
- reveal different temporalities related to connected histories of the Eurozone crisis
- a common conversation? a European space? (cf. works of Camille Roth)
Hermeneutics of networks

... some thoughts from the #referendum corpus

- main actors in a graphe - main actors irl?
- what is a Twitter network?
Hermeneutics of tools

Hypothesis: a tool = a method = a theory = a specific way to interpret data

- **Gephi**
  Visualizing social networks => sociology of social networks != sociology of field and *habitus*

- **IRaMuTeQ**
  *Théorie des mondes lexicaux*
  - French School of Data Analysis (yes, there is one)
  - Mondes lexicaux: one point of view = one coherent set of words = social representations (= Émile Durkheim)
… some thoughts from the #greferendum corpus

- Dataviz is useful... metrics are important
- How to be comfortable with the algorithm? (transparency, stability issues)
- Need for tools that behave well with multilingual corpora (TXM is fine)
- Preservation and sharing issues
Hermeneutics of Twitter: Twitter as a primary source

● A primary source in the historian’s point of view
  Traditionally something that is fixed within a set framework (= the Archive)

● Twitter is always moving, is a “source” in the original meaning (source of water): something that is endlessly flowing, that cannot by definition be fixed
  Ex: The “four more years” Obama tweet

● What we do
  ○ transforming something that is not supposed to stay still into an archive, something that is fixed
  ○ What do we lose in this process?
Hermeneutics of Metadata

Information embedded in the metadata are crucial for the interpretation of a / numerous tweet(s)

Ex: timestamps and the interpretation of temporalities

- Timestamps in tweet metadata correspond to the unending (well…) and continuous feed of tweets that is the essence of Twitter
  - Western vision of time
- Many more artifacts of temporalities are embedded in the text of a tweet
  - How to deal with other kind of temporalities whereas collective memory, for instance, is the result of an interlacing of temporalities
Conclusion: what is the allure of born digital archive?
What is a tweet?

```json
{
  "created_at": "Thu Jun 22 21:00:00 +0000 2017",
  "id": 877994604561387500,
  "id_str": "877994604561387520",
  "text": "Creating a Grocery List Manager Using Angular, Part 1: Add & Display Items
https://t.co/xFox78juL1 #Angular",
  "truncated": false,
  "entities": {
    "hashtags": [{
      "text": "Angular",
      "indices": [103, 111]
    }],
    "urls": []
  }
}
```