Oral History and Linguistic Analysis. A Study in Digital and Contemporary European History

Florentina Armaselu, Elena Danescu, François Klein
Luxembourg Centre for Contemporary and Digital History, University of Luxembourg
florentina.armaselu@uni.lu, elena.danescu@uni.lu, francois.klein@uni.lu

CLARIN Annual Conference 2018

Pisa, Italy, 8 - 10 October 2018
Overview

- Research questions
- Methodology
- The experiments
- Conclusion and future work
Research questions

- To what extent can the combination of digital linguistic tools and oral history assist research and teaching in contemporary history?
  - How can this combination be evaluated?
  - Is there an added-value of using linguistic digital methods and tools in historical research/teaching as compared with traditional means?
  - What are the benefits and limitations of this type of methods?
Methodology. *Data processing workflow*

1. Audio/video
   - *Express Scribe* or external providers
   - transcription

2. .doc transcriptions
   - add heading styles to the sections titles

3. enriched XML-TEI transcriptions
   - identify speakers, speakers’ roles and extra-linguistic elements (XSLT)
   - transform text to lower case (XSLT)
   - POS tagging + lemmatisation (TreeTagger)
   - textometric analysis

4. XML-TEI transcriptions
   - conversion

5. .docx styled transcriptions
   - OxGarage
Methodology. ‘Oral history of European integration’ collection

- **Overview**
  - accounts from people who have **witnessed** and/or been involved in the major events that have shaped the **European integration process**;
  - more than **100 interviews, 160 hours** of material published in a dedicated section on [http://www.cvce.eu/histoire-ORALE/](http://www.cvce.eu/histoire-ORALE/); diversity of languages - French (70%), Spanish, Portuguese, English, German, Dutch, etc.
  - new primary sources for researchers specialising in European studies.

- **Structure**
  - each interview has its own dedicated web page;
  - interviews published in **full** and **indexed by theme**;
  - selected **excerpts** are published to offer easy access to the different topics covered;
  - explanatory **caption** for each selected excerpt;
  - **transcription** of the interview is published, together with a **translation** into English and/or French.
Methodology. ‘Oral history of European integration’ corpus samples for the experiments

- Selection criteria applied for the corpus samples used in the EUREKA and MAHEC experiments:
  - **linguistic** approach:
    - French language
  - **thematic** approach:
    - interviewees involved in the history of Luxembourg in European integration;
    - interviewees involved in the building of the Economic and Monetary Union (EMU).
Methodology. Transcriptions pre-processing

1. Évolution du rôle et fonctionnement du Conseil européen

[Hervé Bribosia] [00:00:11] [...] Comment définiriez-vous le rôle et l’influence du Conseil européen à cette époque ? Prenait-il des [00:01:00] décisions ? Recourait-on parfois à la procédure de vote ?

[Wilfried Martens] [00:01:06] Il n’y avait pas de vote. Donc, c’était par le consensus, et le Conseil européen donnait des impulsions. Parfois, il n’y avait pas d’impulsions. Et donc, à l’époque, quand il était président de la Commission européenne, Jacques Delors se...
What is textometry?

- Methodology allowing **quantitative** and **qualitative analysis** of textual corpora, by combining developments in **lexicometric** and **statistical research** with **corpus technologies** (Unicode, XML, TEI, NLP, CQP, R).

What is TXM?

- **Open-source platform** (Heiden et al., 2010, TXM User Manual 0.7) used for the **analysis** of large bodies of **texts** in various fields of the **humanities** (history, literature, geography, linguistics, sociology, political sciences) and allowing to:
  - **import from different textual sources**, e.g. raw text combined to flat metadata (CSV), raw XML/w+metadata, XML-TEI BFM; **exports** of results in CSV for lists and tables or in graphic format (SVG, JPEG, etc.) for diagrams;
  - manage **NLP tools** for **processing** the input files during the import process (e.g. _Tree Tagger_ for lemmatisation and POS tagging);
  - build a **sub-corpus** or a **partition** based on metadata (date, author, genre, etc.) or structural units (text, section, etc.) of a corpus;
  - **query** for word and word properties patterns (via the CQP search engine);
  - build **frequency lists**, KWIC **concordances** and **co-occurrence** scores for words and words properties;
  - compute **specificity** scores for words/properties in a sub-corpus or a partition, **progression/evolution** of patterns, **correspondence factor analysis** (CFA).
Methodology. **Textometric analysis**

- Create **sub-corpus** and **partition** using structural properties.
- Build **queries** and look for **co-occurrences** of words/properties.
- Build **concordances** and visualise contexts at the **document** level.
Methodology. **Textometric analysis**

- Compute **specificities** - probabilistic model (Lafon, 1980) allowing to:
  - study the **frequency distribution** of words/properties in a (sub-)corpus divided on several parts;
  - compare the parts, in terms of **specific** (excess/deficit) or **basic** use of words/properties.
The experiments. Layout

- **EUREKA_2017** (pilot)
  - time frame: 11 to 15 and 18 to 22 September 2017;
  - target group: four C²DH researchers;
  - data sample:
    - online audio-video interview sequences (5 hours, 6 interviewees) and transcriptions;
    - interviews transcriptions in XML-TEI format (38687 words);
  - assignment:
    - answering one research question using online multimedia recordings of interviews and TXM (tutorial + assistance);
  - evaluation.

- **MAHEC_2018**
  - time frame: 16 April to 14 May 2018;
  - target group:
    - five Master students in Contemporary European History at the University of Luxembourg, as part of a course in Political and Institutional History;
  - data sample:
    - interviews (10 hours, 8 interviewees) transcriptions in XML-TEI format (110563 words);
  - assignment:
    - answering seven research questions using TXM (1 hour training + tutorial + assistance);
  - evaluation.
The experiments. *Proposed questions (excerpts)*

- **EUREKA_2017**
  - What “dimensions” of the European integration process can be discerned from the discourse of the different interviewees?

- **MAHEC_2018**
  - Can you identify the *European institutions* mentioned in the interviews, their role and *interconnections*?
  - Reconstitute the process of the *creation of Economic and Monetary Union* (EMU), with these testimonies, while describing the role played by the different actors of these developments (countries, personalities, principles).
  - With these testimonies, describe the specific role that *Luxembourg* has played in the *European Integration* process? Which of the *interviewees* is speaking more of the role of Luxembourg in the European integration, which less, and why?
  - Draw the “*lexical profile*”\(^1\) (Guyard, 1981:110) of the personalities interviewed. What conclusions do you draw?

---

\(^1\) List of words/properties with the highest positive specificities scores for a respondent, e.g. by category (noun, verb, adjective, adverb).
The experiments. *Evaluation*

**Hypothesis**
- **linguistic analysis** may help the participants in their *quest for answers* to the proposed questions and eventually in formulating other questions.

**Evaluation**
- EUREKA_2017 -> at the end of each *phase*;
- MAHEC_2018 -> at the end of the *assignment* period in the course.

**Questionnaires - Sections**
- Participant:
  - ID, gender, expertise, knowledge.
- Evaluation of:
  - **multimedia** technology + **oral history** collection (EUREKA);
  - **textometric analysis**.
- Evaluation of:
  - proposed *experimental scenario*.

**Questionnaires - Questions**
- **Yes/No:**
  - Have you *found answers* to the research questions?
  - Would you like to *formulate* other language-related *questions* for the studied sample?
- **Likert-scale** queries (five possible answers from *Not at all agree* to *Fully agree* or *Very weak* to *Essential*):
  - There is an *“Eureka” effect* created by the use of this technology in this study. (EUREKA)
  - How do you appreciate the *role* played by the **textometric analysis** in the discovery of the answers?
- **Open** questions:
  - Can you formulate a short *description* of the *“Eureka” effect*, or of its absence, observed during the experiment? (EUREKA)
  - Can you shortly describe the *added value* of this type of analysis?
  - Other reflections on the *innovative* character of the considered technology and/or its *limitations, bias*, etc. for the studied case.
  - Please, enumerate some *strong/weak points* of the proposed *scenario*. 
There is an “Eureka” effect created by the use of this technology in this study. [EUREKA, textometry]

<table>
<thead>
<tr>
<th>Not at all agree</th>
<th>Fully agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

**Can you formulate a short description of the “Eureka” effect, or of its absence, observed during the experiment? [EUREKA, textometry]**

- “… possibility to visually transform results as tables or graphics …” (EKA-PIL_P01); “no new elements as compared with the first phase but quicker identification of the main themes” (EKA-PIL_P02); “Sample not representative enough, since too consensual, for a real Eureka effect. Difficulty in using the tool …” (EKA-PIL_P03); “… Eureka effect … to be taken with care since the only use of textometric analysis is insufficient in research. However, textometric analysis … good tool for ‘mind mapping’.” (EKA-PIL_P04)

**Other reflections on the innovative character of the considered technology and/or its limitations, bias, etc. for the studied case. [EUREKA, textometry]**

- “… without previous knowledge in linguistics and discourse analysis, I don’t see how to interpret the deficit in the usage of a term …” (EKA-PIL_P01); “The interface could be more intuitive and the visualisations and graphics more appealing.” (EKA-PIL_P02); “This technology has great potential but more time is needed and a larger sample in order to fully exploit the potential of the tool.” (EKA-PIL_P03); The selection of the interviews and excerpts is subjective; which may produce bias in the critical analysis of the research question (EKA-PIL_P04).

Can we speak of an "added value" in using this type of analysis as compared with a "traditional" study in (oral) history? [MAHEC, textometry]

| Yes | 4 |
| No  | 1 |

**Can you shortly describe the added value of this type of analysis? [MAHEC, textometry]**

- “The textometric analysis allows the study of a large text corpus and saves a lot of time to the historian. Especially, the analysis of the vocabulary is greatly facilitated.” (TXM-HO_P01); “Possibility to analyse several documents instead of reading them one by one.” (TXM-HO_P02); “Speed, rigorous analysis.” (TXM-HO_P06); “Efficiency in ‘fast reading’ …” (TXM-HO_P10)

**Other reflections on the innovative character of the considered technology and/or its limitations, bias, etc. for the studied case. [MAHEC, textometry]**

- “A problem of the textometric analysis is the question if there is a real gain of new information. In most cases the textometric analysis proved the position and role already known of a character, but did not really bring new information. (TXM-HO_P01)
The experiments. *Results (excerpts)*

- Average scores by participants’ answers
  - **EUREKA_2017**
    - **Role** of the **textometric analysis** in **discovering** the **answers** to the question
      \[(-1) \times 1 + (0) \times 2 + (1) \times 1 = 0\]
  - There is an “Eureka” **effect** created by the use of this technology
    \[
    \left[\frac{(-1) \times 1 + (0) \times 2 + (2) \times 1}{4}\right] = 0.25
    \]
  - Proposed **experimental scenario**
    \[
    \left[\frac{(0) \times 1 + (1) \times 3}{4}\right] = 0.75
    \]
  - **MAHEC_2018**
    - **Role** of the **textometric analysis** in **discovering** the **answers** to the questions
      \[
      \left[\frac{(0) \times 3 + (1) \times 2}{5}\right] = 0.4
      \]
    - Proposed **experimental scenario**
      \[
      \left[\frac{(-1) \times 1 + (0) \times 1 + (1) \times 3}{5}\right] = 0.4
      \]
Conclusion and future work

- Project combining:
  - oral history data;
  - digital linguistic analysis;
  - evaluation of the use of language technology.

- Experiments results:
  - valuation of rapidity in processing and visualising linguistic features in textual corpora;
  - certain reserve concerning the innovative added value of the analysis tool (perhaps, since, as specialists or students in the field, the topic of European integration was, to a certain extent, already known to the participants?).

- Experiments limitations:
  - small number of participants;
  - relatively small samples (~ 5% and ~ 9% of the total hours of interview in French from the Oral History collection).

- Prospects:
  - more evaluation results, from various, larger groups of participants with different degrees of knowledge about the proposed topic and larger samples will be needed.
  - longer term objective: to draw an “inventory” of strengths and weaknesses of language technology applied to the study of (oral) history.


XSLT: Extensible Stylesheet Language Transformations. https://www.w3.org/TR/xslt/all/.
Stay connected with us!

www.c2dh.uni.lu

@C2DH_LU

https://www.facebook.com/c2dh.lu/

www.c2dh.uni.lu