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

Florentina Armaselu
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
florentina.armaselu@uni.lu

Elena Danescu
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
elena.danescu@uni.lu

François Klein
Luxembourg Centre for Contemporary and Digital History
University of Luxembourg
francois.klein@uni.lu

Abstract

The article presents a workflow for combining oral history and language technology, and for evaluating this combination in the context of European contemporary history research and teaching. Two experiments are devised to analyse how interdisciplinary connections between history and linguistics are built and evaluated within a digital framework. The longer term objective of this type of enquiry is to draw an "inventory" of strengths and weaknesses of language technology applied to the study of history.

1 Introduction

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? The paper will address these questions, within the CLARIN 2018 *Multimodal data (Oral History)* topic, starting from two experiments based on an oral history collection, XML-TEI[1] annotation and textometric analysis.

In her outline of an "oral history à la française", Descamps (2013: 109-110) talks about a "linguistic age" or a "first age of the recorded speech" starting in the 1910s when language scientists manifested their interest in oral sources. Bridging oral history and linguistics in a digital context has made the object of event-oriented initiatives and research, inside and outside CLARIN's framework (CLARIN-PLUS OH, 2016; Oral History meets Linguistics, 2015; Georgetown University Round Table on Languages and Linguistics, 2001). Different tools and perspectives have been approached, such as language technologies for annotating, exploring and analysing spoken data (Drude, 2016; Van Uytvanck, 2016; Van Hessen, 2016), online platforms for Multimodal Oral Corpus Analysis (Pagenstecher and Pfänder, 2017) or the use of oral histories as "data" for discourse analysts (Schiffrin, 2003). However, the question of how oral history and linguistics may impact the historian's exploration and interpretation of data seems less studied so far. This proposal aims to contribute to this topic (in our opinion of potential interest for the CLARIN community, as related to building and evaluating interdisciplinary connections between history and linguistics) and consists in a workflow for: (1) transforming and processing historical spoken data intended to linguistic analysis; (2) evaluating the impact of the use of language technologies in historical research and teaching.

2 Methodology

The study is based on a selection from the oral history collection on European integration published on the CVCE by UniLu Website[2]. The whole collection comprises more than 160 hours of interviews, in

[1] http://www.tei-c.org/index.xml.

[2] https://www.cvce.eu/histoire-orale. CVCE is now part of the Luxembourg Centre for Contemporary and Digital History (C²DH) of the University of Luxembourg, https://www.c2dh.uni.lu/.

French, English, German, Spanish and Portuguese, with some of the actors and observers of the European integration process. The selection included 5-10 hours of audio-video recordings and transcriptions, in French. The selected transcriptions were converted to a structured format, XML-TEI, then imported into the TXM[3] textometry software (Heiden et al., 2010), for linguistic analysis. Two experiments were devised. The first (EUREKA_2017), functioned as a pilot using a shorter corpus and involved a small group of C²DH researchers. The second (MAHEC_2018) was part of a course in *Political and Institutional History* for the Master students in Contemporary European History at the University of Luxembourg. For each experiment, a set of research questions was prepared, and questionnaires were designed to enquire on the role of the language technology in answering the proposed questions (or in discovering and formulating other related questions).

2.1 Corpus selection and research questions

The number of interviewees varied from six (EUREKA) to eight (MAHEC), including personalities such as Jean-Claude Juncker, Viviane Reding, Jacques Delors and Étienne Davignon. The selection criterion focused on important milestones in the construction of the European Union and the interviews had to be in French for homogeneity purposes. One research question was proposed for the pilot experiment and seven for the second. They were either general queries, e.g. discern the multiple dimensions of the European integration process (EUREKA) or more specialised questions related to the topic of the course, e.g. identify the European institutions mentioned in the interviews, their role and interconnections, reconstruct the process of the Economic and Monetary Union (EMU) or determine which of the interviewees is speaking more of the role of Luxembourg in the European integration, which less, and why (MAHEC).

2.2 Corpus preprocessing

The transcriptions were available in Microsoft Word format and contained markers for identifying the interviewer/respondent and, occasionally, timecodes. The transcriptions were first converted from Microsoft Word to XML-TEI[4]. Then, a set of XSLT[5] stylesheets, created for this purpose, were applied to the converted output[6], in order to transform it into specific TEI encoding for the transcription of speech. The extract below shows how the identity and type of speaker were encoded using the <u> tag (*utterance*) and the @who and @corresp attributes. The time points (when present) were encoded using <timeline> and <anchor/> elements, in order to mark the text with respect to time.

```
<u who="#hervé_bribosia" corresp="#interviewer"><anchor synch="#t262"/> Et un siège
   unique pour le Parlement européen, on y arrivera un jour ?</u>
<u who="#wilfried_martens" corresp="#respondent"><anchor synch="#t263"/> Ah, c'est le
   Traité. C'est réglé dans le Traité, il faut l'accord de tous. Même le Parlement
   européen ne peut pas l'imposer. C'est un élément du Traité. Et honnêtement, je
```

2.3 TXM analysis

The corpus in XML-TEI format was imported into TXM, a textometry software, allowing part of speech tagging and lemmatisation[7], frequency of occurrence counts and statistical analysis of textual corpora. The analysed samples contained a total of 38687 (EUREKA) and respectively 110563 (MAHEC) occurrences. Given the encoding, it was possible to build sub-corpora and partitions corresponding to the type of speaker (respondent/interviewer) and the name of the speakers.

The following TXM features were used by the participants to find answers to the proposed questions: specificities (Lafon, 1980), index, concordances and co-occurrences (TXM manual). Figure 1 illustrates the specificities, i.e. a comparative view on the vocabularies of the speakers (e.g. over-use

[3] http://textometrie.ens-lyon.fr/?lang=en.
[4] Via the OxGarage online service, http://www.tei-c.org/oxgarage/.
[5] https://www.w3.org/TR/xslt/.
[6] Using oXygen XML Editor, https://www.oxygenxml.com/.
[7] Via TreeTagger.

for *banque centrale*[8] in the discourse of Yves Mersch and Jean-Claude Juncker, and respectively deficit in the speech of Étienne Davignon), for the top five European institutions most frequently mentioned in the text. Other features allowed particular queries (index), by a single property or in combination (e.g. *noun + adjective*), detection of forms having a tendency to occur together (co-occurrences, e.g. *banque centrale + européenne*) or a switch from a synthetic, tabular view to mini-contexts (concordances, e.g. *la banque centrale européenne est en charge de la politique monétaire ...*[9]) or document visualisation. Our hypothesis was that this type of linguistic analysis may help the participants in their quest for answers to the proposed questions.

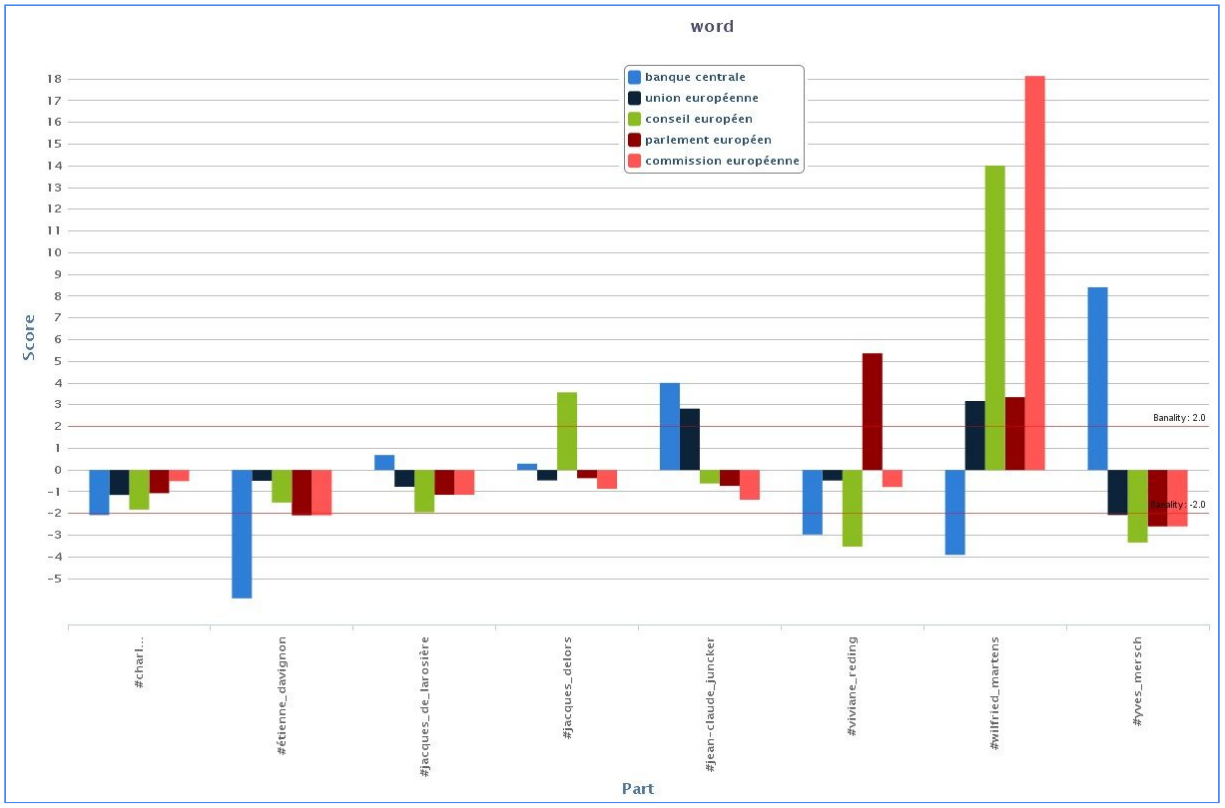

Figure 1. Specificities for European institutions within the respondents' partition (MAHEC_2018)

2.4 Evaluation

The evaluation was intended to confirm/disconfirm this hypothesis and to "measure" the impact of the linguistic technology, its innovative aspects and limitations, when applied to the study of history. The online (anonymised) questionnaires have included: *Yes/No* questions (e.g. *Have you found answers to the research questions?*), Likert-scale queries (e.g. *How do you appreciate the role played by the textometric analysis in the discovery of the answers?* with five possible answers from *Very weak* to *Essential*), open questions (e.g. *Can you shortly describe the added value of this approach, if any?*).

3 The experiments

The pilot experiment EUREKA_2017[10], took place from 11 to 15 and 18 to 22 September 2017 and implied the study of: (1) online audio-video interview sequences and transcriptions; (2) transcriptions using TXM analysis. Evaluation questionnaires were filled-in at the end of each phase. The participants were four C^{2}DH researchers specialised in *European integration*, *Contemporary history*, *Historical and political studies*. Their knowledge varied on a five values scale from *Not at all* to

[8] Eng. *Central Bank*.
[9] Eng. *the European Central Bank is in charge of the monetary policy … .*
[10] Enquiring on the "Eureka effect" of the use of linguistic technology in historical research. The experiment was presented at *Les rendez-vous de l'histoire. Eurêka-inventer, découvrir, innover*, Blois, France, 4-8 October, 2017.

Expert in the fields of: *European integration history*, *Multimedia and oral history*, and *Textometric analysis*. While the data showed specialisation in European integration history with medium knowledge in multimedia and oral history, the self-evaluation of the textometry skills was placed at the lower end of the scale. The second experiment, MAHEC_2018, involved five Master students in *Contemporary European History*, and took place from 16 April to 14 May 2018. The assignment consisted of seven research questions and the evaluation of the added-value/limitations of the language technology in completing the task. The students' background varied from *History* and *Contemporary European history* to *Medieval history*, with medium and good knowledge of *European integration history* reported. Compared with the previous experiment, the self-evaluation of the *Textometric analysis* skills covered a larger spectrum from *Not at all* to *Good*.

The results of the first experiment (Figure 2) indicate moderate valuation by the participants concerning the role of textometric analysis in finding the answers (left) and the response to the question whether there is a discovery, "Eureka" effect determined by the use of this technology (right), on a scale from -2 to +2, *Very weak* to *Essential* and *Not at all agree* to *Fully agree,* respectively.

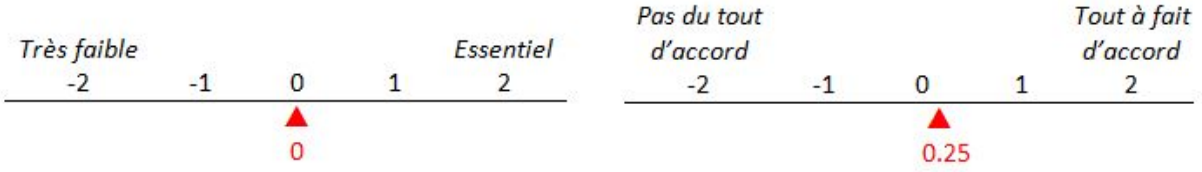

Figure 2. Average scores for textometric analysis (EUREKA_2017): role; "Eureka" effect

As an added value of the method, the participants mentioned: usefulness for analysing large corpora, allowing both local and global observation, rapid identification of the main themes, graphical representation of results. It was also observed that the textometric analysis alone is not sufficient in research. Less positive points were: the interface could have been more intuitive[11], the graphics more attractive and the selected sample larger, in order to fully exploit the potential of the method.

For the second experiment, the average value regarding the role played by the textometric analysis in finding the answers was a bit higher than above (0.4 instead of 0 on the -2 to +2 scale). The aspects evoked as added value were similar to those mentioned in the first case, e.g. allowing the analysis of a large corpus of documents instead of reading them one by one, "fast reading", speed and rigour. As strong points, it was noted the use of part of speech based queries and the suitability of textometric analysis for assisting interpretation. As weak points were mentioned the results window that should have been larger and the heterogeneity of the questions proposed to the interviewees instead of a common set that would have allowed a better basis for comparing their responses. Concerning the innovative side of the studied technology, it was pointed out that it often served just to prove the position or the role of a given personality within the European integration process, rather than providing new information. An aspect that ought to be further examined in future experiments.

4 Conclusion and future work

The project combined oral history and digital linguistic analysis, and evaluated the use of language technology in history research and teaching. Two experiments have been devised. Although rapidity in processing and visualising linguistic features in large amounts of texts were mainly valued, the results showed a 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. For comparison purposes, more evaluation results, from different groups of participants with different degrees of knowledge about the proposed topic, are needed. The longer term objective of this type of evaluations would be to draw an "inventory" of strengths and weaknesses of language technology applied to the study of history.

[11] For EUREKA, no initial TXM training was provided, just a tutorial and assistance with the tool during the experiment. For MAHEC, a short TXM training was provided, as well as a tutorial and assistance.

References

[CLARIN 2016] CLARIN. 2016. CLARIN-PLUS OH workshop: "Exploring Spoken Word Data in Oral History Archives", University of Oxford, United Kingdom. https://www.clarin.eu/event/2016/clarin-plus-workshop-exploring-spoken-word-data-oral-history-archives

[Freiburg Institute for Advanced Studies 2015] Freiburg Institute for Advanced Studies. 2015. Conference "Oral History meets Linguistics", Freiburg, Germany. https://www.frias.uni-freiburg.de/en/events/frias-conferences/conference-oral-history-and-linguistics.

[Descamps 2013] Florence Descamps. 2013. "Histoire orale et perspectives. Les évolutions de la pratique de l'histoire orale en France". In F. d'Almeida et D. Maréchal (dir.), *L'histoire orale en questions*, p. 105-138. INA, Paris.

[Drude 2016] Sebastian Drude. 2016. "ELAN as a tool for oral history", CLARIN-PLUS OH workshop.

[Georgetown University 2001] Georgetown University. 2001. Georgetown University Round Table on Languages and Linguistics (GURT), Washington, DC, USA.

[Heiden et al. 2010] Serge Heiden, Jean-Philippe Magué and Bénédicte Pincemin. 2010. "TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement". In Sergio Bolasco, Isabella Chiari, Luca Giuliano (Ed.), *Proc. of 10th International Conference on the Statistical Analysis of Textual Data - JADT 2010*, Vol. 2, p. 1021-1032. Edizioni Universitarie di Lettere Economia Diritto, Roma, Italy. https://halshs.archives-ouvertes.fr/halshs-00549779/fr/.

[Lafon 1980] Pierre Lafon. 1980. "Sur la variabilité de la fréquence des formes dans un corpus". *Mots*, N°1 , p 127-165. http://www.persee.fr/doc/mots_0243-6450_1980_num_1_1_1008.

[ENS de Lyon & Université de Franche-Comté 2017] ENS de Lyon & Université de Franche-Comté. 2017. *Manuel de TXM 0.7.8*, http://txm.sourceforge.net/doc/manual/manual.xhtml.

[Pagenstecher and Pfänder 2017] Cord Pagenstecher and Stefan Pfänder. 2017. "Hidden Dialogues:Towards an Interactional Understanding of Oral History in Interviews". In *Oral History Meets Linguistics*, edited by Erich Kasten, Katja Roller, and Joshua Wilbur, pp. 185–207. Fürstenberg/Havel: Kulturstiftung Sibirien, Electronic Edition. http://www.siberian-studies.org/publications/PDF/orhili_pagenstecher_pfaender.pdf.

[Schiffrin 2003] Deborah Schiffrin. 2003. "Linguistics and History: Oral History as Discourse". *Georgetown University Round Table on Languages and Linguistics (GURT) 2001: Linguistics, Language, and the Real World: Discourse and Beyond*, edited by Deborah Tannen and James E., pp. 84–113. Alatis, Georgetown University Press, Washington, D.C. http://faculty.georgetown.edu/schiffrd/index_files/Linguistics_and_oral_history.pdf.

[Van Hessen 2016] Arjan van Hessen. 2016. "Increasing the Impact of Oral History Data with Human Language Technologies, How CLARIN is already helping researchers". CLARIN-PLUS OH workshop.

[Van Uytvanck 2016] Dieter van Uytvanck. 2016. "CLARIN Data, Services and Tools: What language technologies are available that might help process, analyse and explore oral history collections?". CLARIN-PLUS OH workshop.