

University of Luxembourg

Multilingual. Personalised. Connected.



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

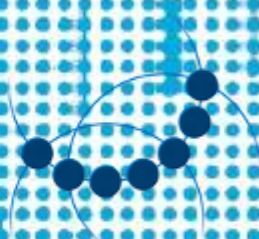
Florentina Armaselu, Elena Danescu, François Klein

Luxembourg Centre for Contemporary and Digital History, University of Luxembourg

florentina.armaselu@uni.lu, elena.danescu@uni.lu, francois.klein@uni.lu



CLARIN



CLARIN Annual Conference 2018

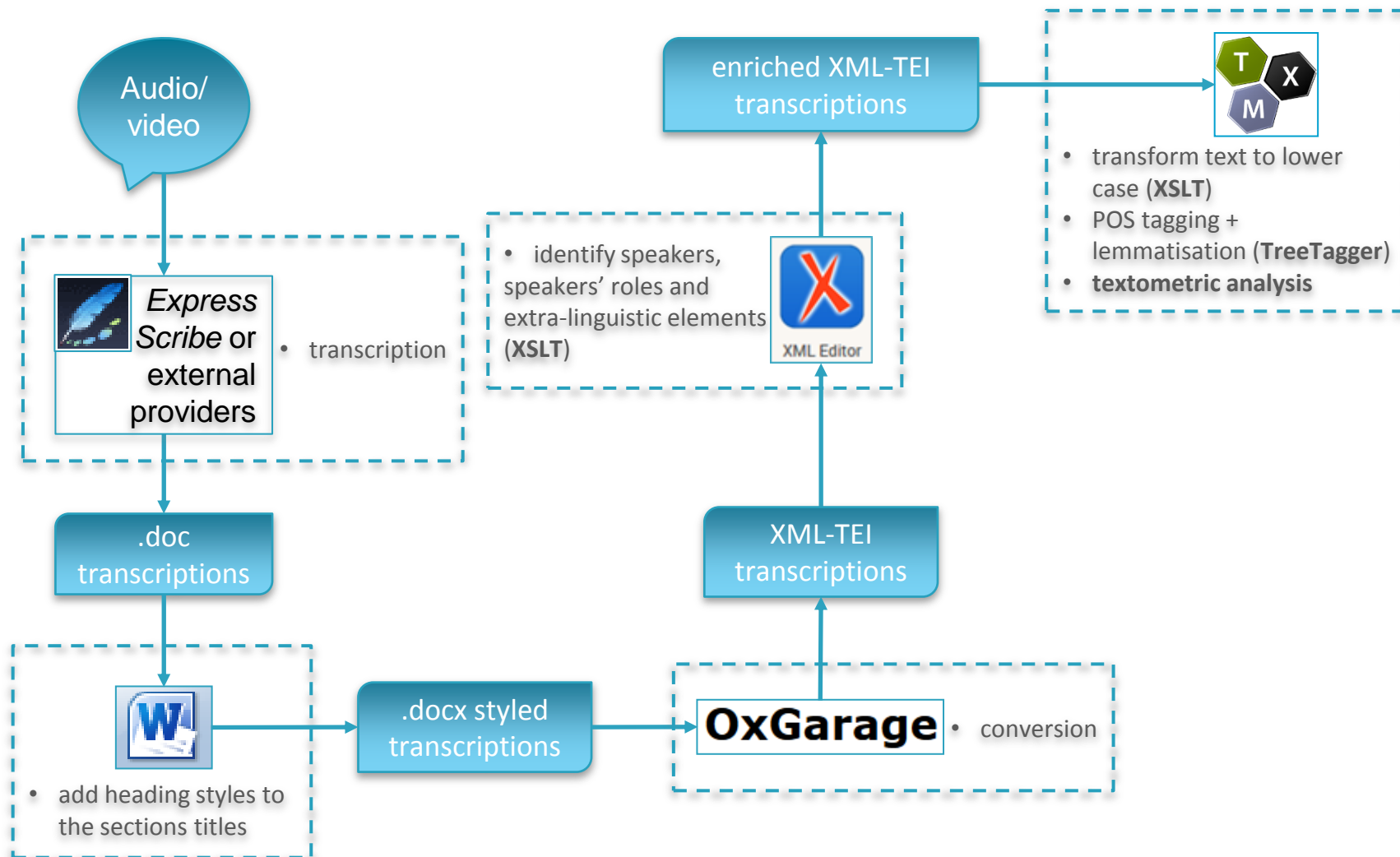
Pisa, Italy, 8 - 10 October 2018



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

- 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*



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/>; 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**.

The screenshot displays the CVCE.eu website interface. At the top, there are navigation links for FR, EN, and DE. The main header features the CVCE.eu logo and a search bar. Below the header, there are tabs for Accueil, À propos, ePublications, and Recherche & Innovation. A horizontal menu lists various topics: A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z. The main content area is titled 'MARTENS (Wilfried)' and includes a section for 'Photos de l'interview' with four portraits: MARTENS (Wilfried), MAYSTADT (Philippe), MEDEIROS FERREIRA (José Manuel), and MEISCH (Adrien). Below this, there is a section for 'L'interview' with two parts: 'Partie 1 – cofondateur et leader du PPE (1976-2010) (Bruxelles, 19 mai 2010) (durée: 01:45:58)' and 'Partie 2 – Premier ministre et membre du Conseil européen (1979-1992) (Bruxelles, 2 septembre 2010) (durée: 01:49:31)'. The second part includes a list of topics: 'Évolution du rôle et fonctionnement du Conseil européen', 'Des candidats belges pour présider la Commission européenne', 'Nominations belges à la Commission européenne et à d'autres hautes fonctions européennes', 'Les présidences belges de 1982 et 1987', 'L'Acte unique européen', 'La dimension européenne de quelques affaires belges', 'Chute du mur de Berlin et réunification allemande', 'Le traité de Maastricht et l'union politique', 'Organisation et politique belge en matière européenne', and 'L'avenir de la Belgique fédérale dans l'Union européenne'. At the bottom, there is a section for 'Transcription de l'interview'. On the right side, there is a sidebar with a search bar and a list of results. The first result is 'Interview de Wilfried Martens (Bruxelles, 2 septembre 2010) – Extrait: rôle, fonctionnement et évolution du Conseil européen' with a 'Consulter' button. The second result is 'Interview de Wilfried Martens (Bruxelles, 2 septembre 2010) – Extrait: le choix d'Herman van Rompuy à la présidence du Conseil de l'Union européenne' with a 'Consulter' button. Below the results, there is a section for 'Vidéo' with a video player and a 'Partager' button.

- 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

OxGarage



```
<div>
  <head>1. Évolution du rôle et fonctionnement du Conseil
    européen</head>
  <p>[<hi rend="bold">Hervé Bribosia</hi>] [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 ?</p>
  <p>[<hi rend="bold">Wilfried Martens</hi>] [00:01:06] Il
    n'y avait pas de vote. Donc, c'était par le consensus,
    et le Conseil européen donnait des impulsions. Parfois,
```

```
<profileDesc>
  <particDesc>
    <p>Speakers roles: <list xml:id="speakers_roles">
      <item xml:id="interviewer">Interviewer</item>
      <item xml:id="respondent">Respondent</item>
    </list>
    <p>Speakers list: <list xml:id="speakers_list">
      <item xml:id="hervé_bribosia">hervé_bribosia</item>
      <item xml:id="wilfried_martens">wilfried_martens</item>
    </list>
  </particDesc>
</profileDesc>
```

```
<text xml:id="MartensITw" xml:lang="fr" decls="#transcr-video">
  <timeline unit="s" origin="#t0">
    <when xml:id="t0" absolute="00:00:11"/>
    <when xml:id="t1" absolute="00:01:00"/>
    <when xml:id="t2" absolute="00:01:06"/>
    <when xml:id="t3" absolute="00:02:00"/>
  </timeline>
</text>
```

```
<div>
  <head>1. Évolution du rôle et fonctionnement du Conseil
    européen</head>
  <p>[<hi rend="bold">#hervé_bribosia</hi>] <sup>corresp</sup> <sup>#interviewer</sup>
    [...] Comment définiriez-vous le rôle et l'influence du
    Conseil européen à cette époque ? Prenait-il des
    <sup>#t1</sup> décisions ? Recourait-on parfois à la procédure de vote ?</p>
  <p>[<hi rend="bold">#wilfried_martens</hi>] <sup>corresp</sup> <sup>#respondent</sup>
    <sup>#t2</sup> 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 plaignait qu'il
    n'y ait pas d'impulsions ou de conclusions lors d'un
    Conseil européen. En tout cas, je suppose que c'était
    tout à fait différent du Conseil européen d'aujourd'hui qui
    a 27 membres, sauf les ministres des Affaires étrangères.
    À l'époque, les ministres des Affaires étrangères étaient aussi
    présents. Nous étions au début neuf, dans mon expérience, et
    puis <sup>#t3</sup> maximum douze. Et pour moi, les
    débuts étaient caractérisés par le fait que nous sommes entrés
    la première fois, M<hi rend="superscript">me</hi> Thatcher et
    moi-même, et qu'il y avait un mode de joie : 1979, le tandem
```


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**

The screenshot displays the C2DH Textometric analysis interface. The top section shows the 'Create subcorpus' dialog with 'Name: u_respondent', 'Structure: u', 'Property: corresp', and 'Value: #interviewer, #respondent'. A bar chart titled 'Dimensions of the u_who partition in the u_respondent corpus' shows the number of words for various categories. The bottom section shows a concordance query for 'européen' and a table of concordances.

Dimensions of the u_who partition in the u_respondent corpus

Category	Number of words
#interviewer	~6000
#respondent	~22000
#interviewer, #respondent	~7000
#interviewer, #respondent, #interviewer	~7000
#interviewer, #respondent, #interviewer, #respondent	~7000
#interviewer, #respondent, #interviewer, #respondent, #interviewer	~18000
#interviewer, #respondent, #interviewer, #respondent, #interviewer, #respondent	~11000
#interviewer, #respondent, #interviewer, #respondent, #interviewer, #respondent, #interviewer	~11000

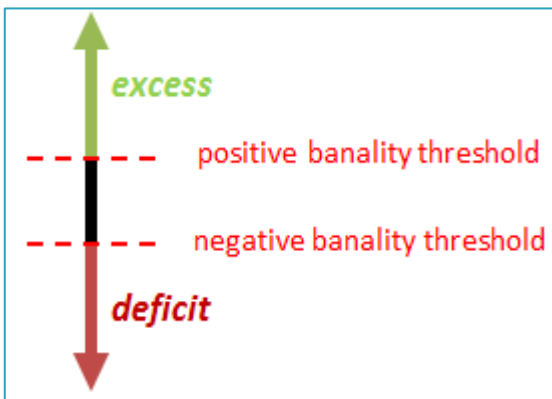
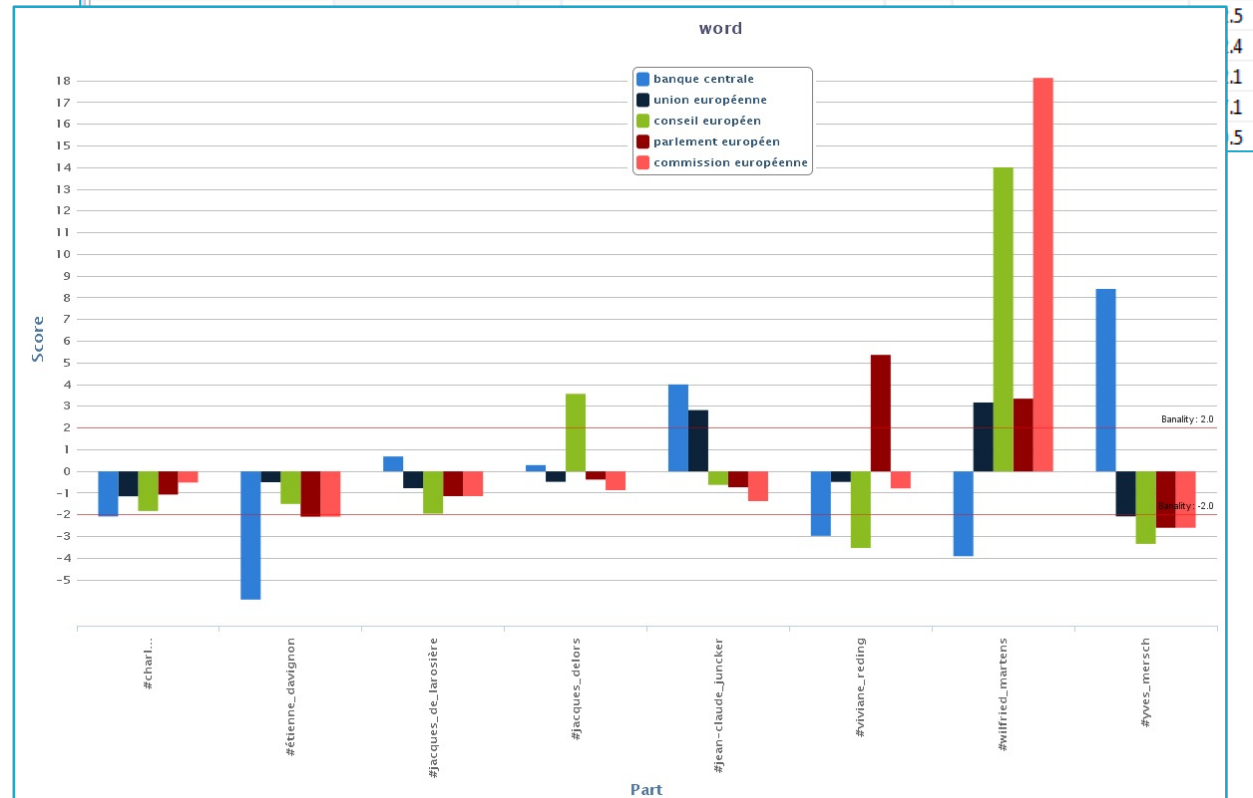
Concordance Query: `[[frlemma="européen"]]* [[words="construction"]]* [[words="construction"]]* [[frlemma="européen"]]` within 11

Concordance Table:

text_id	Left context	Keyword	Right context
MartensITW_FR	où se trouve le cœur de cette union	européenne, ceux qui sont vraiment convaincus que la construction	européenne, l'intégration européenne, après deux guerres
MartensITW_FR	, ceux qui sont vraiment convaincus que la	construction européenne	, l'intégration européenne, après deux guerres mondiales
MartensITW_FR	l'intégration européenne prévient ces tendances ? la	construction européenne	ou l'intégration européenne ou la création d'une union eur
MerschITW_FR	voir l'aboutissement de ce bout de la	construction européenne	qui lui tenait tant à cœur, et auquel il avait tant
MerschITW_FR	niveau pour des échanges au niveau de la	construction monétaire européenne	, où il y avait un échange entre personnes qui restaient activ
MerschITW_FR	était surtout la dimension du progrès de la	construction européenne	qui était à l'avant-plan, de par mon éducation à la
MerschITW_FR	qu'il ait attaché le développement de la	construction européenne	comme quelque chose attaché à une personne, et que le fl.
MerschITW_FR	reconfirmé la relance de l'idée d'une	construction monétaire européenne	. et c'est là, suite à un premier paragraphe,
MerschITW_FR	cette absence de réponse a menacé toute la	construction européenne	. c'est la raison principale pour laquelle les plus grands pa

- 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.

Units	Frequency T 2769	#charles-ferdinand_nothomb t=217	score	#étienne_davignon t=573	score
banque centrale	58	0	-2.1	0	-5.9
union européenne	53	1	-1.1	9	-0.5
conseil européen	51	0	-1.8	5	-1.5
affaires étrangères	38	5	0.8	18	3.7
commission européenne	30	1	-0.5	1	-2.1
parlement européen	30	0	-1.1	1	-2.1
union économique	25	5	1.4	1	-1.6



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**.

Age range	Genre	Expertise domain	Knowledge
20 – 34	1	F M	3 1
		European Construction	1
			History of European Integration
			Not at all Expert
			1 3
35 – 44	2		
		Contemporary History	2
			Multimedia + Oral History
			Not at all Expert
			1 2 1
45 - 54	1		
		History and Political Sciences	1
			Textometry
			Not at all Expert
			3 1

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**.

Age range	Genre	Expertise domain	Knowledge
18 – 34	5	F M	1 4
		History	2
			History of European Integration
			Not at all Expert
			3 2
		Contemporary History	2
			Textometry
		Medieval History	1
			Not at all Expert
			1 1 2 1

▪ 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).

- 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**.

The experiments. *Results (excerpts)*

There is an “Eureka” effect created by the use of this technology in this study. [EUREKA, textometry]

Not at all agree

Fully agree

1

2

1

- 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)

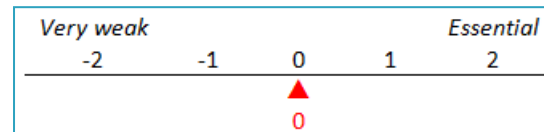
The experiments. *Results (excerpts)*

- Average scores by participants' answers

- EUREKA_2017

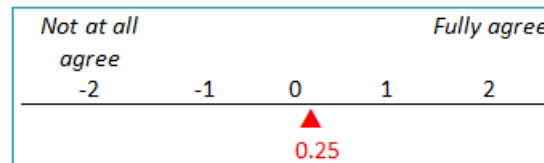
- Role of the **textometric analysis** in **discovering** the **answers** to the question

$$(-1) \times 1 + (0) \times 2 + (1) \times 1 = 0$$



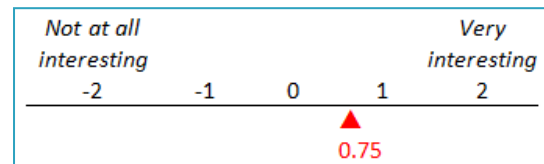
- There is an “**Eureka**” **effect** created by the use of this technology

$$[(-1) \times 1 + (0) \times 2 + (2) \times 1] / 4 = 0.25$$



- Proposed **experimental scenario**

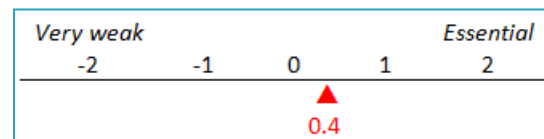
$$[(0) \times 1 + (1) \times 3] / 4 = 0.75$$



- MAHEC_2018

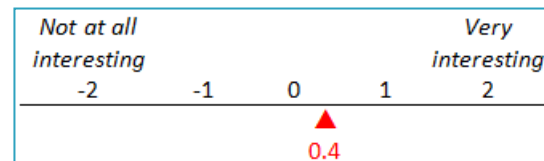
- Role of the **textometric analysis** in **discovering** the **answers** to the questions

$$[(0) \times 3 + (1) \times 2] / 5 = 0.4$$



- Proposed **experimental scenario**

$$[(-1) \times 1 + (0) \times 1 + (1) \times 3] / 5 = 0.4$$



- 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**.

- Guyard M.-R. « Spécificités d'auteurs dans *Le Surréalisme au service de la Révolution* ». In: *Mots*, n°2, mars 1981. Qu'est-ce que le vocabulaire spécifique d'un texte politique? pp. 95-122. DOI : [10.3406/mots.1981.1023](https://doi.org/10.3406/mots.1981.1023). www.persee.fr/doc/mots_0243-6450_1981_num_2_1_1023.
- Heiden, S., Magué, J-P., Pincemin, B. (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/>. TXM Website: <http://textometrie.ens-lyon.fr>.
- Lafon P. (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.
- TEI: Text Encoding Initiative. <http://www.tei-c.org/>.
- *TXM User Manual 0.7* - June 2015. <http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf>.
- XML: Extensible Markup Language. <https://www.w3.org/XML/>.
- XSLT: Extensible Stylesheet Language Transformations. <https://www.w3.org/TR/xslt/all/>.

Stay connected with us!



@C2DH_LU



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

www.c2dh.uni.lu