Textometry on Audiovisual Corpora

Bénédicte Pincemin, Serge Heiden, Matthieu Decorde

Textometry on Audiovisual Corpora
Experiments with TXM software

Bénédicte Pincemin[1], Serge Heiden[2], Matthieu Decorde[3]

[1]Univ. Lyon, CNRS, IHRIM UMR5317 – benedicte.pincemin *at* ens-lyon *dot* fr

[2]Univ. Lyon, ENS Lyon, IHRIM UMR5317 – slh *at* ens-lyon *dot* fr

[3]Univ. Lyon, ENS Lyon, IHRIM UMR5317 – matthieu.decorde *at* ens-lyon *dot* fr

Abstract

Textometry is applied to audiovisual corpora, such as transcripts from semi-directed interviews, or the *Actualités françaises* newsreels archive. A workflow using an assisted or automatic transcription software is efficient to get a rich encoding. New features are added to the TXM software: a specialized import module based on Transcriber XML format, a utility to convert text transcripts to XML, and the MediaPlayer extension to watch the video segment corresponding to a word context selection. Methodological thoughts arise from this experience. It is highly relevant that textometry takes into account internal text structures (such as speech turns) and other meta-information (such as timecodes). Meta-information has to be displayed and available for processing without being mixed with contents. Another challenge is to integrate multiple interrelated representations. A back-to-media feature is as fundamental as the back-to-text one to provide context to interpretation work.

Keywords: manual transcription, speech-to-text software, non-verbal communication, semi-directed interviews, audio recordings, video archive, Actualités françaises newsreels, XML, multimedia, TXM.

1. Background and context

Since its foundations in the 80s, textometry aims at processing corpora of textual documents. More than simply counting words, it is interested in every dimension of language. This has been stimulated by the availability of natural language processing software (NLP) such as lemmatizers, and by the advent of international standards for data and text structuring (XML, TEI).

Few textometric tools still focus on plain text and deal with words as character strings only, sticking to the form they show in the primary data, so as to account for the text “as it stands”, with as little recoding as possible (Lexico). Most tools have integrated a lemmatizer component (usually TreeTagger) to get a three-fold representation for every word: as a graphical word form, as a lemma (to group the various inflections of a same word), and as a part-of-speech category (Hyperbase, DtmVic, Iramuteq). This offered new insights and induced developments towards possible ways to combine these three levels of word description, from multi-level queries in the search engine facility to parameters implied in different steps of a calculus (Pincemin, 2004).

Regarding XML and TEI text encoding, the challenge is to integrate structural information and various other encodings into the textometric workflow (Trameur, TXM), in contrast to discarding tags or extracting plain text only (Hyperbase, DtmVic). A first level of processing consists in managing a tree-view of the text, so that it can be divided according to multiple

partition levels. But XML encoding can also be used, inter alia, to integrate multimedia content: images, time-coding to relate to an audio or video recording, etc. TXM may still be the only one that has developed textometric solutions for such corpus modalities. Specialized software for audiovisual data, such as CLAN (MacWhinney, 2000) for “list of turns” transcription type or ELAN (ELAN, 2018) for “partition of timelines” transcription type, focus first on corpus annotating, then browsing and searching the annotated information, with few analytical functions. The TEITOK web platform (Janssen, 2016) has developed advanced features for creating, viewing and annotating TEI (or XML) corpora, with possible alignment to facsimile images or to audio files, but analytical tools are mainly visualization tools such as frequency graphs and result plots on a geographic map. A textometric approach can then really complement the investigation tool kit for such corpora.

In this paper, we would like to introduce to TXM 10-year experience in re-thinking textometry in the context of multiple textual modalities, and to present software components that have been produced for such corpora. This experience can be divided into two main steps, corresponding to the two main sections of the article.

2. Recording transcripts: data modeling and new textometric features (2010-2015)

2.1. Context and application cases

In 2010, fellow teachers from the EVS geography Lab (UMR5600) invited us to contribute to a Master’s Course in Geography about methodology and tools for fieldwork. Students had to conceive a survey and collect data via semi-directed interviews they recorded. Then, students transcribed the interviews they had conducted. These textual data were all gathered in a corpus, and qualified with socio-demographic information. Then students were introduced to TXM and could systematically analyze the survey data with the help of the textometric tool, in order to address their geographic research questions.

In the first academic year 2010-2011, the survey was conducted in partnership with a local association, CeDRE (Cellule Développement Rural Emploi), that supported women seeking employment or leading a project. The geographic investigation dealt with relationships between territory, work, and personal skills. 33 women were interviewed, which produced a CEDRE corpus of 300,000 words (24 hours of speech). In the following year 2011-2012, the geographic study focused on the French fishermen working on Lac Léman. The geographic question was related to their relationship to their environment and their perception of how it had evolved along various dimensions (economic, natural, social,…). Nearly all fishermen could be interviewed (38 men or women in a total of 46), so that the LEMAN corpus included nearly 400,000 words for 34 hours of speech (Le Lay et al., 2016).

We met the same kind of corpus in a linguistic and educational research context. Verbal interactions in a class were filmed, then students and teacher communications were transcribed to be analyzed accurately (Blanc and Griggs, 2015; Heiden, 2015).

Video corpora were also used for research in History. The Matrice project organized an exhibition about the liberation from extermination camps in 1945. One of the devices presented was a textometric terminal allowing simple browsing and analysis of the transcripts of 108 video testimonies edited by INA French national archive (https://entretiens.ina.fr/memoires-de-la-shoah) (more than 3 million words).

2.2. Using a transcription software to interactively edit richly-encoded textual data

2.2.1. Basic case: the starting point is an audio recording

In the case of the fieldword done by geographers, based on semi-directed interviews, the primary data are audio recordings. A relevant solution for a workflow is then to rely on the transcription community's experience and tools, by using a specialized software for the transcription task. Thus, one can get a richly encoded corpus through an ergonomically efficient user interface. We chose Transcriber software: it managed the transcription elements we needed in our corpora, it ran on Windows, Mac and Linux, it had a widespread diffusion in the scientific community, it offered a sound XML output format for audio transcription, and it was free and open-source, like TXM (Figure 1).

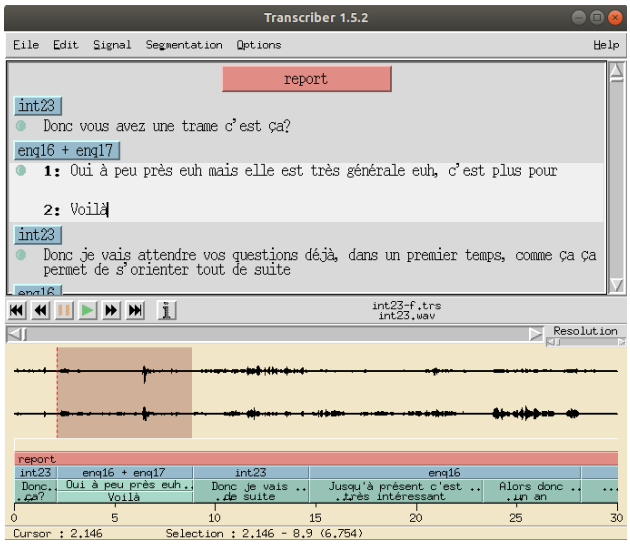

Figure 1. *Transcriber graphical user interface: example of a transcription for the CEDRE corpus.*

The critical importance of applying rules to normalize word representations is well-known for any statistical approach on linguistic data (Muller, 1977). In the same way, interpreting speech as text implies a tremendous number of choices,[1] concerning the way of writing words and managing very common phenomena such as silences, intonation, onomatopoeia, repetition, incomplete or peculiar pronunciation, unknown words, external events, etc. These choices must be directed by explicit and shared conventions in order to get a homogeneous and controllable transcript (one knows how a phenomenon is annotated or if it has been left out of scope). In this respect, we adapted the Transcriber transcription guide to our context (Heiden and Pincemin, 2011), taking into account what annotation could be relevant and available in TXM analysis, next step.

In a transcription software such as Transcriber, the basic transcript unit is the utterance, which corresponds to an audio segment, typically matching a sentence or the part of a sentence between two breaks. Through the graphical interface, the user plays the audio recording and sets its borders: then this audio segment can easily be played and repeated again when the user is typing the written form of what she hears. Utterances are embedded in speech turns, to which a speaker is assigned. If needed, it is possible to encode the fact that several people are speaking at the same time. Speech turns are placed in sections that organize the entire transcript. In a semi-directed interview typically, sections can be used to match the topics of the interview grid.

Facilities are also provided to log non-verbal communication (such as hesitations, laughs, emphasis on a word, gestures), or to describe events which interfere with the current communication and should be known at the analysis stage.

[1] Similar considerations happen in the edition of medieval manuscripts. This is even an overall philologic question: text is not unique and clear, decisions are necessary to establish it. Digital humanities reveal it again through encoding choices, so that a digital philology has to be developed (Guillot et al., 2017).

With such an interface, transcript is easily and systematically synchronized with audio recording, at the utterance level. The text representation is equipped with full timecodes, that embed in the text the alignment information to the media. Furthermore, variables characterizing the interviewee can be associated with each transcript, in the usual way for TXM: they are entered as a spreadsheet, in which a line is an interview, and columns are characteristics for the interviewee.

Nowadays, Transcriber is no longer maintained. Nevertheless, the Transcriber XML format is still an efficient mean to represent transcription data: either other software programs provide a Transcriber XML export of the data, or one can translate another format into the Transcriber one. Even if TXM already converts the Transcriber XML format into XML TEI internally for processing, we would still be interested in a general XML TEI based standard format for transcripts (Schmidt, 2011).

2.2.2. Extension to the case of existing transcripts

The case stated above may not however be the most common one. Researchers often already hold transcripts they made previously. Most of these colleagues are not acquainted with specialized work on verbal interaction, so they used basic word-processing software to get a written version of their interview data. They may have implemented more or less formal transcription conventions, for example with consistent notations to introduce speech turns and speaker identifiers, or to insert timecodes. These notations may be associated with colors or typographic styles for ease of reading.

If these existing transcripts have been annotated following strictly constant rules, an automatic program can be developed to translate this kind of markdown document to a Transcriber XML document (as regards to information that this format can encode). Then, the Transcriber XML document can be imported in TXM (cf. 2.2.1). Such a utility program was released as a TXM macro named TextTranscription2TRS. It processes initial transcripts following specific conventions. A tutorial presents the transcription conventions and the way to call the utility program (Heiden, 2016).

2.3. Textometric enhancements thanks to transcription structuring and encoding

Some pre-processing may be required on Transcriber XML files: concatenation of transcripts when an interview was interrupted, normalization of speaker codes or of topical categories, speaker anonymization. Then, information encoded in the Transcriber XML representation can be used at various stages of the analytical process.

For the import procedure, that inputs the corpus in TXM and makes it available for textometric analysis, one can choose if the interviewer’s words are indexed, and contribute to the corpus, or if only the words of the interviewed people are taken into account when searching, counting, and computing statistics. In both cases anyway, the interviewer’s questions are shown in the transcript edition which is provided by TXM to see words in context: indeed, the full context of both interviewer’s and interviewee’s statements are necessary to interpret textometric outputs.

In the same vein, various contextual information are rendered in the transcript displayed in TXM, such as stage directions, with a characteristic typographical layout (Figure 2). All this information that was considered significant by the transcriber, is made available to contextualize interpretation, however they are not included in corpus words, they don’t weigh in textometric operations. This concerns speakers’ names and timecodes, section indications,

comments about concomitant events, information about pronunciation or intonation, silences, etc. (Heiden and Pincemin, 2011).

Of course, several information are also useful for analytical processes. Some are available as structures defining contexts (topical sections, speech turns, utterances). These TXM structures are qualified with properties that can be queried: the topic for a section, the speaker for a speech turn or an utterance. Every word is fitted with several features: not only its lemma and part-of-speech as provided by TreeTagger software, but also transcription features (word partially pronounced and how, uncertain spelling or unknown word, quoted speech). These information items are available in any analytical processing: for instance, one can partition the corpus into speakers, or kind of speakers (Figure 3); one can look for words that were annotated as truncated.

Timecodes give the opportunity to get an accurate access to the recording. Thus, the transcript does not replace the original media. During textometric analysis, one may need to check not only the textual context, but also the audio or video context which brings complementary information.

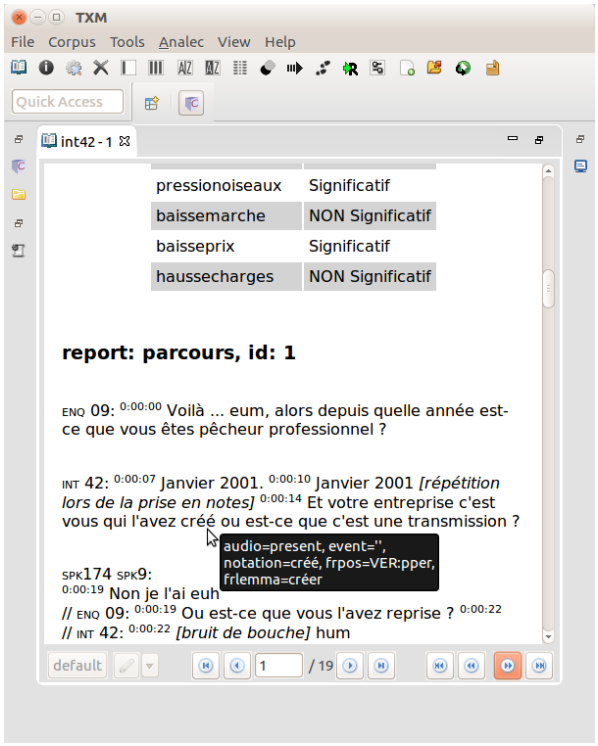

Figure 2. *LEMAN corpus: Rendering of a transcript in TXM after an XML-TRS import: speaker names, timecodes, section titles, comments, etc. are displayed and styled, but they are not searched and counted with corpus words.*

Actually, transcription implies choices and reduction of information, if not errors, because everything cannot be annotated, and on top of that, what is understood depends on the transcriber's perception, interpretation and experience. A transcription can neither be complete, nor neutral. Thanks to the availability of the source media, the necessary reduction applied by the transcription operation becomes more acceptable and less crucial. Nothing from the source media is definitively lost.

Units	Freq	gros t=14468	index
équiper	36	26	8.4
étudier	10	9	4.5
prévoir	23	15	4.2
surgeler	24	15	3.9
suivre	68	31	3.7
placer	12	9	3.4
HAPAX	580	180	3.0
rééquilibrer	5	5	3.0
confondre	5	5	3.0
créer	33	17	3.0

Units	Freq	moyen t=31468	index
pêcher	718	463	7.0
respecter	51	43	5.1
concerner	19	19	5.0
associer	22	21	4.5
vendre	527	332	4.0
lier	31	27	3.8
savoir	887	539	3.7
parler	188	126	3.3
transformer	68	50	2.9
payer	208	136	2.9

Units	Freq	petit t=11389	index
apprendre	138	58	8.6
sauver	24	16	6.1
espérer	29	17	5.3
lever	239	75	4.8
subir	11	9	4.7
travailler	416	118	4.7
former	53	24	4.6
dépêcher	20	12	4.0
permettre	112	38	3.5
progresser	7	6	3.4
appréhender	7	6	3.4

Figure 3. *LEMAN Corpus (fishermen interviews): 10 top-specific verbs according to the size of the fishing company (from left to right: large, medium-sized, small) (cf. Le Lay et al., 2016)*

To facilitate direct access to the source media, a special feature was implemented in TXM: as the "back to text" feature is necessary to correctly interpret textometric outputs, a "back to

media" feature was developed. It provides hyperlinks from any concordance line to an embedded media player. An advanced user, Bertrand Gaiffe, extended this feature to access the media not only from concordance view but also from the full text edition within TXM. He shared[2] a groovy script for TXM, linking utterances in the transcript text edition to the media. This states the centrality of a back-to-media feature, as fundamental as the back-to-text feature in any textometric approach.

2.4. Main outcomes

These first experiments with audio and video corpora, which recorded communication interactions or interviews, emphasized several important features for the textometric method. In every corpus, we had to deal with several speakers, so it was relevant to use a software program that could manage categorized divisions inside a text, and model speech turns with an assigned speaker. A second feature shows highly relevant, whereas it is still rarely implemented in textometric software: dealing separately with what is displayed for text reading, and what is processed in analytical operations. Speakers' names, timecodes, events interfering with speech or non-verbal communication should be available at some steps of the analysis (corpus partitioning, word contextualization and text reading) without being mixed to the lexical content submitted to statistical computations. As a third important feature, we note the link to the source media, so as to gain a fuller context and to overcome the unavoidable lacks and warps of the transcript.

Over the same period, we observed that these features, which were obvious in the case of multimedia corpora, turned out to be fundamental and should be generalized from various kinds of corpora. For instance, for a linguistic and philological study on medieval manuscripts too, one wants to separate critical notes from the original text, to display the text with notes but process the text without note contents; and a view to the source media (the manuscript scan) restores a more complete perception of the original data. We developed the XML TEI zero import module (XML-XTZ) to meet these new requirements in a generic manner. We enlarged the text edition feature to a synoptic edition feature, allowing aligned views of textometric data on the one hand, and a digital copy of the original source on the other hand. Yet, the Transcriber XML import still manages specialized corpora with speech turns, timecodes, and all specific information conveyed by the Transcriber format.

As regards software innovation, several new components augmented TXM software and were made publicly available: the XML-TRS import module, the TextTranscription2TRS utility with an associated tutorial, and a beta version of the MediaPlayer extension.

3. Exploration of a large corpus of videos (since 2018)

3.1. Context: historical analysis of French newsreels 1945-1969 (ANTRACT ANR project)

The ANTRACT project is dedicated to the transdisciplinary analysis of the *Actualités Françaises* newsreels. From 1945 to 1969, a national weekly edition was broadcast in movie theaters as a preamble to films. Within 10 to 20 minutes, it exposed a dozen of topics, as various as political, economic, social or cultural subjects as well as sports events. As a textual corpus, this corresponds to a few million words (we will see that diverse textual representations are possible). Sponsored by the ANR research agency, this collaborative project brings together the Center of social history of contemporary worlds (CHS, Paris 1 & CNRS), the French National Audiovisual Institute (INA), Eurecom Graduate school and

[2] Mail sent on txm-users list, November 2, 2017, 2:42 PM.

research center in digital sciences (specialized in image processing), the computer science laboratory from Le Mans University (LIUM) (specialized in speech-to-text technologies), and our team in IHRIM for the text analysis component (Carrive et al., in press).

In a previous research project in 2014-2015, we had met an even bigger video corpus, composed of the automatic transcription of news bulletins and reports about World War II (WWII) that were broadcast on French national television from 1980 to 2010. In the context of the Matrice project, dedicated to interaction between individual and social memory, these news were used to account for collective representations of WWII events, in comparison with outputs from neurolinguistic investigations (Gagnepain et al., 2019). The transcripts were encoded in the Vocapia XML format, synchronized at the word level. A dedicated TXM import module was developed. The resulting TXM corpus of 3,766 texts containing 125 million of lemmatized words was optimized to only build lexical tables, that were then processed by external topic modeling software such as MALLET (McCallum, 2002). At that time, no specific processing was applied to this corpus in relation with its audiovisual source.

With these corpora, we shifted from manual transcription of audio or video recordings to automatic transcription of video recordings, scaling up the size and multiplicity of data representation.

3.2. Structural considerations: a multi-layered representation with various segmentations, various modalities; and archival material redundancy and overlaps

The *Actualités françaises* case presented a complex set of data: various modalities are available in archives, with non homothetic physical units, and new secondary representations can be added through automatic audio or image processing (Carrive et al., in press). The corpus can be thought of as a compound and multimodal view on a common object to be observed. Data derive from three main materials: (a) the video recordings themselves, (b) their archival description in INA documentation system, and (c) paper documents related to the production process.

a) Video recordings associate both audio and video modalities; the primary unit is the film reel, which does not match a logical unit such as the newsreel program or a topical news report. These data are digitized so as to provide files which can be handled by software. Speech-to-text applied to the audio track provides text transcripts of the voice-over commentary (LIUM). Automatic image analysis was run to detect and track the presence of some famous people (Eurecom) or to extract written data displayed on the screen (INA). Manual annotation from historian researchers can be added within the INA Okapi web platform, and modeled as layers aligned with the video timeline.

b) A documentary database provides for every news report a scope and content note written by a librarian. The information unit here is the report. Its archival description is composed of numerous fields: some are categorical data (such as the broadcast date), some are sets of descriptors (for instance keywords about topics, places, people shown), and there are also several textual data (title, abstract, sequence description). These data were exported from the INA database as spreadsheets, one report on a line and one descriptive field in a column. The full documentary scheme includes 69 fields, among which we selected the ones that could be relevant for our corpus and questions, which made a focus on about twenty fields.

c) A third source of data was the original paper typescripts which archive a written version of the voice-over commentary. These may not reflect exactly what is said, they may have been prepared before the report capture, or finalized later. INA scanned these papers, then ran an

optical character recognition (OCR) application to obtain a digital text representation. The information unit here is neither the film reel (as for videos), nor the news report (as for archival description), but the page, whose boundaries don't match any of the previous units. In particular, a news report can run on several pages, and a typescript page can also include several news reports. INA works on the automatic pairing of image zones and news reports, based on textual similarity between OCR output and speech-to-text transcription.

From these data we have generated two different TXM corpora so far.

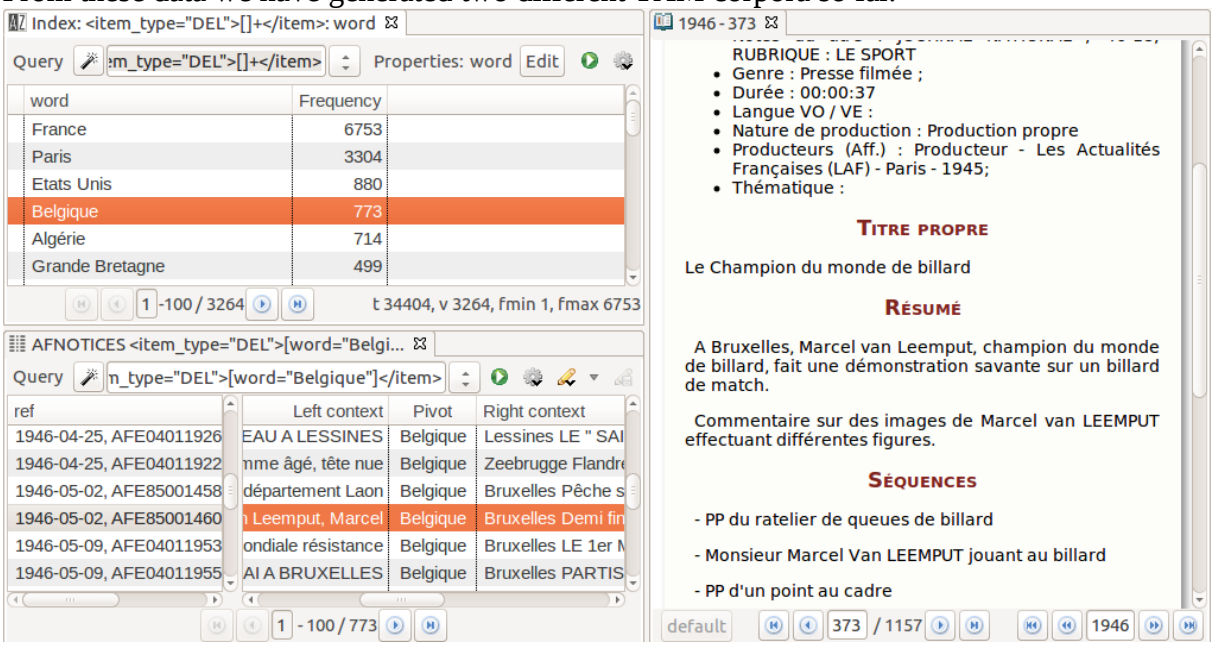

Figure 4. *AFNOTICES Corpus: an index (upper left) lists the location descriptors; one of them is selected and viewed in a concordance (lower left); a double-click on a concordance line displays the corresponding text notice.*

The first one, AFNOTICES, is based on documentary descriptions. As these are professional and natively digital productions, textual data are very clear and reliable. They can be represented in TXM in a linguistically-aware and structured manner (words are identified and lemmatized, lists are encoded as such, keyword types are distinguished, etc.) so that queries can be very accurate (Figure 4). However, from an historian point of view, one works here on secondary data: librarians are mediators between the researcher and the original video data. Moreover, librarians themselves cannot escape time passing, what they notice in films and the words they use in free description fields are time dependent.

A second corpus was generated, AFVOIXOFF, that was based on the automatic transcription of the voice-over commentary recorded in the videos. As a first step, INA achieved an automatic recomposition of film reel digitized files so as to get one video file per news report (instead of one reel): this was necessary to connect to each news report film its date and other relevant metadata. A second step consisted in selecting a consistent set of news reports: if all the archive was taken as a corpus, concordance views showed significant overlapping in data. Focusing on national broadcast editions (and removing regional or international editions as well as unused supplementary material) appeared as the best way to get an overview of *Actualités françaises* without the redundancy that was difficult to manage in textometric analysis.

In its current version, the AFVOIXOFF corpus is more comprehensive than the AFNOTICES one, because it associates both the video transcription output as main text and the documentary description as metadata. However, querying textual documentary fields is less efficient in AFVOIXOFF, because these fields are processed as character strings, not as textual content. A new corpus design is to be elaborated, to integrate multiple full text representations. TXM already deals with aligned corpora, but the current dedicated import module only reads corpora in TMX format. The new textometric corpus modeling will have to be designed to allow the analysis of aligned, structured and multimedia representations. In such a corpus, new modalities could be added, for instance text and images from typescripts.

3.3. Strengthening and expanding the back-to-media feature

As the *Actualités françaises* videos are at the heart of the ANTRACT project, we invested in bringing to maturity the MediaPlayer TXM extension. Moreover, when working on automatic transcriptions which had to accommodate archival material, we could expect more errors than in manual transcriptions, so a convenient means to watch the source video is essential.

Considering the digital data scale of such a video corpus, four different access strategies were developed: media files can be stored within the TXM corpus, or in a local directory; they can alternately be accessed on a remote repository, with or without authentication control.

The definition of the timespan to be played for a given word focus was refined: either it is a *window* of *n* words before and *m* words after the selected words; or the segment played corresponds to the words of a transcript *structure* in which the word focus occurs (speech turn, or whole section for instance).

In addition to linking concordance lines to the media player, a new feature was developed to access the media from any text selection in the full-text display window.

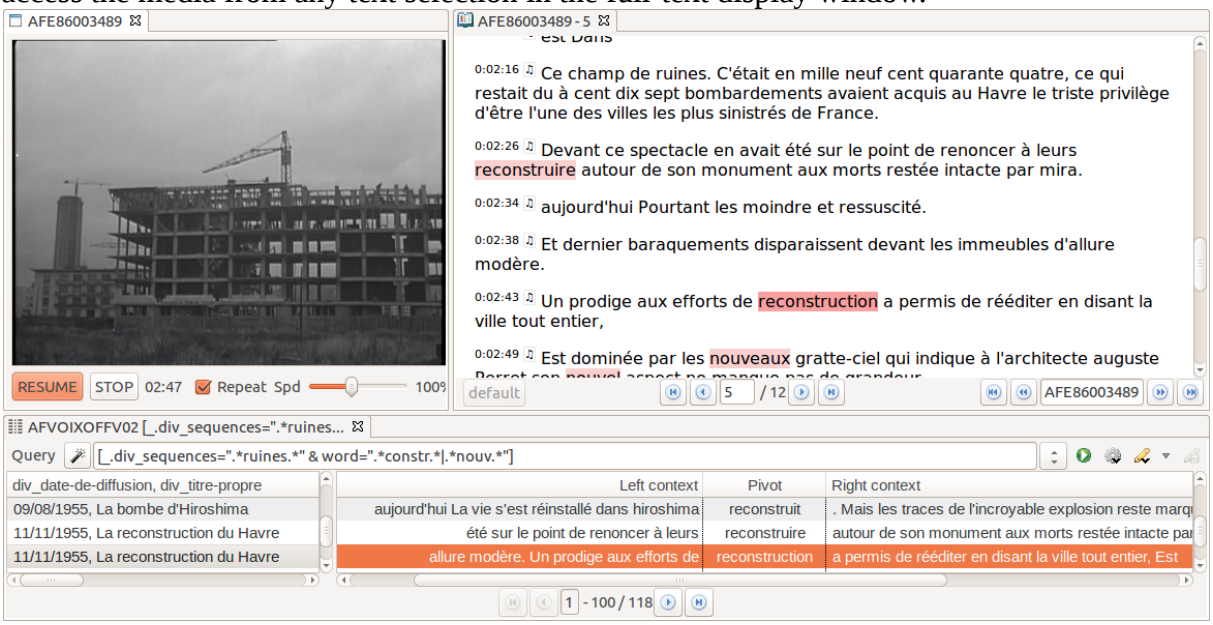

Figure 5. *Corpus AFVOIXOFFV02: concordance with back-to-text and back-to-media hyperlinks and synoptic view.*

Preference parameters were defined to manage the display layout when several windows are to be shown together (Figure 5) : typically, a concordance window with the output of a word search, a text window providing the transcript context of a given concordance line, and a

videoplayer window providing a targeted access to the media to get a full context for any word or text segment. The user sets her default preference according to the data characteristics, to her digital equipment, and to her analytical patterns: does the mediaplayer window open above, under, or on the right- or left-hand side of the text to which it is linked? Another parameter manages the position of the text opened with a back-to-text hyperlink from a concordance line. Afterwards, every window position may be adjusted by drag-and-drop if needed.

Furthermore, the setup process was simplified, and its compatibility was extended, thanks to the use of JavaFX technology instead of relying on the external VLC component.

3.4. New analytical possibilities with an aligned multimodal corpus

Innovation comes from dealing with several representations together. A first benefit from this multiple representation is that one representation contextualizes another. In a search query, one can cross criteria from several representations: for instance, one can select news reports which show ruins (according to the documentary description) and whose comments mention any word related to building and novelty (in the voice-over commentary) (Figure 5).[3] The synoptic view of video and transcript, or of words in context and dates, are invaluable for an informed interpretation.

A second benefit is related to getting a better control on observations: combining representations allows cross-checking, which is especially useful in a framework where some data are secondary descriptions (here, documentary descriptions, that are mediated by librarians) or automatic productions (here, speech-to-text transcriptions). For instance, if one finds an amazingly high frequency of a given word in the voice-over comments, one can check if it was actually pronounced in the videos, or if it is due to a transcription error. This checking activity leads to correction needs. In the Antract framework, edition of transcripts or annotation data can be provided either within TXM through its new annotation tools, or with the help of the INA Okapi platform, which is dedicated to video corpora creation and annotation.

A third benefit is that some statistical induction can be made from the aligned data, applying the resonance principle (Salem, 2004). For instance, when a crowd is shown, without any mention of a president, what is it told about?[4] In TXM, this kind of calculus is achieved via specificity analysis on a subcorpus; and new corpus operators (intersection, union) have been implemented to allow complex definition of subcorpora. The resonance approach is quite promising for our *Actualités françaises* data, to study implicit relationships, or unexpected deviations, between what is told and what is shown; or between what has been noted in documentary descriptions and what occurs in the film itself; or even between the written archived commentary and what was actually recorded and displayed. This could be a inspirational entry point for considerations about textual genres and semiotic genres.

3.5. Main outcomes

The context of a large and complex video archive raised new questions about the implementation of a textometric approach. We had to develop and precise the definition of textual units, the various kinds of contexts involved, and the alignment possibilities, thanks to which various representations could contextualize one another. The textometric corpus could

[3] Query for AFVOIXOFFV02 corpus: [_.div_sequences=".*ruines.*" & word=".*constr.*|.*nouv.*"]

[4] This example is illustrated in (Carrive et al., in press).

be modeled as multiple views on an object which has to be defined, because it is not given as a unique and obvious reality. Relevant data include both source data and derived data, especially textual representations automatically computed from audio data or image data (analysis of the video part of the films as well as scans of paper documents). This experience emphasizes the corpus design phase, the need for accessing various types of source data (searching, visualizing), and the relevance of a powerful query language and of the resonance statistical feature in order to analyze such multi-representation corpora.

These thoughts about fundamental aspects of data modeling and processing were implemented in different parts of TXM. The Transcriber XML import module was extended in order to manage synchronization at the word level, and to organize metadata with a better rendering for textual metadata conveying parallel textual representations. The MediaPlayer extension matured and gained robustness, flexibility and usability, thanks to renewed technologies, parameters, and configuration options. In particular, parameters were introduced to manage an optimal arrangement of windows for complex synoptic analyses.

4. Conclusion

These experiments and innovations in the textometric processing of audiovisual corpora show that the special requirements of this kind of corpora are an opportunity to both extend software capabilities and generalize a few elements of textometric methodology.

Our first experiments with manual transcription of audio or video recordings, were extended to large multimedia data with automatic processing for text extraction. In every case, data structuring and structure-aware processing were key features. Textual data may be divided into speech turns; speakers' names or timecodes have to be encoded and available for reading and analysis without being mixed with speakers' statements. Several textual or audiovisual representations may be relevant, and have to be organized so that cross-searches are available for augmented context, for better control of errors in data, and for new transmedia investigation possibilities.

As regards to software, TXM was augmented with a new import module based on the Transcriber XML format, the TextTranscription2TRS TXM utility was provided and documented to help the automatic conversion of existing transcripts towards the Transcriber format for the XML-TRS import module, and the MediaPlayer extension was developed to implement a "back-to-media" feature, providing a hypertextual efficient and targeted access to the source recording. At first glance, playing the video corresponding to a textometric corpus might seem like a technology gimmick, something nice and brilliant but simply added aside. Quite the contrary, this "back-to-media" feature is as essential for analysis and interpretation as the undisputed "back-to-text" function.

The textometric perspective on audiovisual corpora still gives a central position to textual data. Several textual representations, following rules from different text genres, compose a complex representation of a research object and connect media sources. Contextualization is also still confirmed as a key concept for semantic analysis, as its encoding get diversified: situational information may be recorded not only with metadata, but also with comments annotated in the text and with a multiplicity of parallel representations.

This research has benefited from the ANTRACT ANR project (ANR-17-CE38-0010) and the Matrice Equipex project (IA 10-EQPX-0021).

References

Blanc N. and Griggs P. (2015). Tracer la procéduralisation dans le contexte interactionnel et multimodal d'une classe d'immersion. *Recherches en didactique des langues et des cultures* [online], 12 (3). URL: http://journals.openedition.org/rdlc/1004 ; DOI: 10.4000/rdlc.1004.

Carrive J., Beloued A., Goetschel P., Heiden S., Laurent A., Mazuet F., Meignier S., Pincemin B., Poels G., Troncy R. (in press). Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *Digital Humanities Quaterly*, Special Issue on AudioVisual DH.

ELAN (Version 5.2) [Computer software]. (2018, April 04). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan

Gagnepain P., Vallée T., Heiden S., Decorde M., Gauvain J.-L., Laurent A., Klein-Peschanski C., Viader F., Peschanski D., Eustache F. (2019). Collective memory shapes the organization of individual memories in the medial prefrontal cortex. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-019-0779-z

Guillot C., Lavrentiev A., Rainsford T., Marchello-Nizia C., Heiden S. (2017). La "philologie numérique": tentative de définition d'un nouvel objet éditorial. In Trachsler R., Duval F., Leonardi L., *Actes du XXVIIe Congrès international de linguistique et de philologie romanes*, pp. 143-154.

Heiden S. (2010). The TXM Platform: Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme. In Otoguro R. et al., eds, PACLIC24, Waseda Univ., pp. 389-398.

Heiden S. (2015). *P1S8 Corpus* (April 4th, 2014; transcript and recording of a high school physics course). Data collected within a project that was coordinated by Andrée Tiberghien (CNRS, Lyon, ICAR). https://sourceforge.net/projects/txm/files/corpora/p1s8-course-transcription.

Heiden S. (2016). *Tutoriel d'import de transcriptions d'enregistrements dans TXM*. Retrieved from https://groupes.renater.fr/wiki/txm-users/public/tutoriel_import_transcriptions

Heiden S., Decorde M., Jacquot S. (2018). *TXM User Manual. Version 0.7*. ENS Lyon and Univ. Franche-Comté. http://textometrie.ens-lyon.fr/files/documentation/TXM%20Manual%200.7.pdf.

Heiden S., Magué J.-Ph., Pincemin B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In Bolasco S. et al., eds, Statistical Analysis of Textual Data -Proceedings of JADT 2010, Edizioni Univ. di Lettere Economia Diritto, Rome : 1021-1031.

Heiden S. and Pincemin B. (2011) - *Guide de transcription d'entretien avec Transcriber pour TXM*, 2e édition, Lyon : Laboratoire ICAR, Projet Textométrie, 25 pages.

Janssen M. (2016). TEITOK: Text-Faithful Annotated Corpora. In Calzolari N. et al., editors, *Proc. of the 10th International Conf. on Language Resources and Evaluation*, Portorož, p. 4037-4043.

Le Lay Y., Heiden S., Merchez L., Pincemin B. (2016) - Retour de pêche. Le métier de pêcheur à travers le discours des professionnels français du lac Léman. In Comby É., Mosset Y., de Carrara S., eds, Corpus de textes : composer, mesurer, interpréter. Lyon: ENS éditions, pp. 117-133.

McCallum A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. http://www.cs.umass.edu/~mccallum/mallet.

McWhinney B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. L. Erlbaum Associates.

Muller Ch. (1977 [1992]). *Principes et méthodes de statistique lexicale*. 1992 reprint:Champion, Paris.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. In Purnelle G. et al., eds, 7e Journées internationales d'Analyse statistique des Données Textuelles, Presses univ. Louvain, pp. 865-873.

Salem A. (2004). Introduction à la résonance textuelle. In Purnelle G. et al., eds, 7e Journées internationales d'Analyse statistique des Données Textuelles, Presses univ. Louvain, pp. 986-992.

Schmidt T. (2011). A TEI-based Approach to Standardising Spoken Language Transcription. *Journal of the Text Encoding Initiative*, 1. https://doi.org/10.4000/jtei.142