The textometric concept of active corpus

Bénédicte Pincemin, Serge Heiden, Franck Mazuet

The Textometric Concept of Active Corpus. Illustration by an Analysis Scenario based on Annotation then Projection.

Bénédicte Pincemin[1], Serge Heiden[2], Franck Mazuet[3]

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

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

[3]University of Paris 1, CHS UMR8058 – fmazuet *at* gmail.com

Abstract

Active corpus provides the possibility to apply searching and statistical computing as if corpus were reduced to selected words, whereas full text still remains visible in context display. This is mainly implemented in paradigmatic processing, yet it may concern syntagmatic processing or text display too. Here we experiment active corpus in syntagmatic processing. A projection generates a new corpus, in which words are semantic tags that were automatically assigned in a first step to the original data. This new corpus makes it easy to explore tag sequences, with any generic textometric tool available, however sparse the original annotation may be. This methodological path was applied to film grammar analysis on 10,000 archival descriptions of news reports. 19 camera shot and angle types were ed through queries and tagged. This annotation became the lexicon of the projected corpus that was used to study shot sequences. The annotation and projection tools we have run are available as utilities in TXM open-source software and should usefully serve many research projects.

Keywords: textometry, active corpus, corpus annotation, corpus projection, film grammar, Les Actualités françaises newsreels, TXM software.

1. Definition and state of the art

1.1. What is active corpus?

We define *active corpus* as the subset of textual data that is taken into account in the textometric processing. It may be called *background* as well, as it provides the total amount of frequencies for contrastive analysis (Pincemin, 2004). This concept is useful to manage two different realities: the full source text and its relevant content for various processing.

Active corpus plays a role at several moments of textometric analysis. *(a)* When the text is *displayed*, parts of the text that do not belong to active corpus may be shown anyway (because the whole text matters in order to achieve the reading and interpretation) but with a distinctive layout (such as a gray print), so as to know what is actually considered. *(b)* For *syntagmatic processing* dealing with the flow of the text and considering words as successive *tokens* (like phrase search or n-gram computing), active corpus is a means for passing over secondary items (insertions, expansions). *(c)* For *paradigmatic processing* listing words out of context and operating on *types* (that unify repeated occurrences), such as statistics on lexical tables, active corpus selects the focus (the items for which results are requested) and the overall set that stands for the frequency distribution (reference corpus and table margins).

1.2. Current implementations

Active corpus is commonly implemented for *paradigmatic processing*, through several ways to adjust the rows of the lexical tables submitted to statistical computing that can be:

- qualitative filtering: stop word list, part-of-speech selection, or analytic categories as assigned by an editable dictionary, like analytic keys in IRaMuTeQ that specify if a word is active, supplementary or excluded (Loubère and Ratinaud, 2014);
- quantitative threshold: either the *n* most frequent words, or words with a frequency higher than a given threshold;
- table advanced building: compiling several selection results, merging or discarding rows, like the *List* function in Hyperbase (Brunet, 2011).

Active corpus is then defined at the word-type level (in a dictionary, in indexes, in table rows) and is used in statistics computing. A preselection at the word-token level can also be combined (restriction to a subcorpus). Most frequently are the selected lexical rows both the statistical reference and the vocabulary under study — the two sets are identical. However, supplementary elements make it possible for a word to be considered in the analysis without weighing in the statistical reference. Conversely, one could display in the results only a subset of the words involved in the statistical calculus.

Concerning the active corpus in *text display*, this is rarely available in text analysis tools. Such a feature is more suitable to TEI-based tools that encode and render complex text internal structures, like TEITOK (Janssen, 2016).

As for active corpus in *syntagmatic processing*, a current illustration may be provided by the possibility, in a concordance view, to consider only some word positions when sorting contexts: for instance, the 2L position meaning “the second word to the left of the target word” (see AntConc for an example of implementation (Anthony, 2022)). Nevertheless, active corpus appears still underused in syntagmatic processing.

1.3. Active corpus in TXM software

TXM implements the usual means to refine a lexical table to fit any active corpus (*paradigmatic processing*): token preselection (subcorpus), qualitative filtering (CQP search engine for rows selection), quantitative thresholds, operations on table rows. Moreover, the lexical table margins can be set either to the current row selection, or to the entire corpus or subcorpus, which allows focusing results on a subset of words while keeping a larger corpus as reference for frequencies. The direct access to specificity score computation through its four parameters (PlotSpecif utility, Fig. 4 below) provides the most accurate settings for the active corpus and the pattern to be studied, for any particular case of specificity evaluation.

TXM also provides means to deal with active corpus when *displaying the text*: in the XML-TEI Zero import process (Heiden, 2020: §4.9.5), the textual planes setting labelled “Out of text to edit” lists XML elements whose textual data content will be displayed without being indexed (these words are neither searched nor counted). And a CSS stylesheet can be edited to distinguish the layout for these elements if necessary. An example is given with red section titles in the text displayed in Figure 1.

A recent utility (WordProperty2Word) carries out a projection of a word annotation that creates a corpus whose words are selected annotation values. In doing so, it can retain a subset of original tokens and delete others. We argue this can be a pragmatic way to implement active corpus for subsequent *syntagmatic* analyses, as illustrated in the following section.

2. A use case scenario

2.1. Research context and addressed question

The ANTRACT research project (Carrive *et al*., 2021) studies the audiovisual corpus of *Les Actualités françaises*, composed of about 10,000 news reports that were broadcast as weekly editions in French movie theaters between 1945 and 1969. One of the fields of investigation is an historical description of the production process and its impact on the newsreel's features. From this perspective, the question was raised to capture data about the film grammar of these newsreels. That is, what kind of camera shots and angles do the newsreels actually use, and whether recurring patterns of shots and angle types contribute to film standardization.

We considered that such information might be obtained from the newsreel descriptive forms. In these text documents, the *Sequences* section is a shot-by-shot description of the film footage (see example in Figure 1 and Appendix). We must keep in mind that this is secondary data, as it relies on librarian mediation, which can vary from one librarian to another, and which cannot be totally exhaustive. Nevertheless, we thought these professional descriptions were worth mining and might give insights on film grammar.

2.2. Analysis roadmap within TXM

A TXM formatted AF-NOTICES corpus has been created from the descriptive forms (Pincemin *et al*., 2020). Its size amounts to 2 million words, among which 38% hapax (more information in Appendix). It is a structured corpus, in which *Sequences* sections can be targeted. The idea is to set a typology of shot scales: Close-Up, Medium Shot, etc. For each shot type, CQL search engine queries are elaborated to match the various wording librarians happen to use to mention the shot. An automatic annotation utility applies the queries to the corpus and adds a normalized encoding for shots, available as a new analytic lexical property. Then, we project the corpus along this property, so as to derive a new corpus that retains only sequences of shot types. In this new corpus, shot pattern analysis is no longer hindered, neither by wording variations nor by lexical insertions that break up shot strings in the original text.

2.3. Annotation step

Several annotation modes are available in TXM platform (Heiden, 2018). We opted here for a lexical annotation, assigning a property to a word, for two reasons. First, this annotation is indexed in the CQP model so that it benefits from all TXM usual functionalities. Secondly, a lexical annotation fits the targeted one-shot/one-unit model.

We also chose to systematize the description as every association between a term and a shot value was formally encoded through a CQL query. In addition to the consistency and clarity of the model, this procedure is compatible with corpus updates, and even with description updates (categories can be refined while getting more experienced with the data).

Our model lists 19 camera shot and angle types, plus one miscellaneous type ("divers plans"), and one off-topic type that is used to manage polysemy and mark "plan" or "vue" occurrences that do not denote a shot or an angle. For our needs, 60 queries were designed (see Appendix). We benefited from the CQLList2WordProperties utility to automatically apply these 60 queries to the corpus and encode the new "plan" word property, so that annotating was systematic and fast.

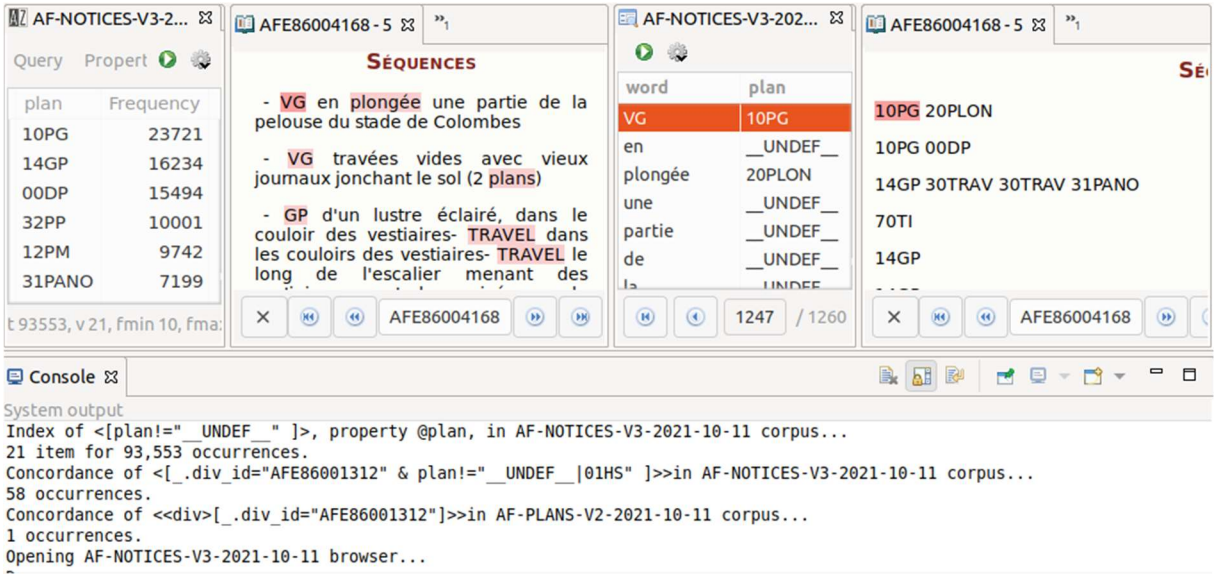

Figure 1. *The* Les Actualités françaises *corpus in TXM, with shot type annotations. From left to right: (i) annotated shot tags and their frequencies in the corpus; (ii) view of a documentary description with all annotated words highlighted; (iii) view of the shot tag assigned to each word of this text; (iv) same text in its projected version. Below: log of executed commands.*

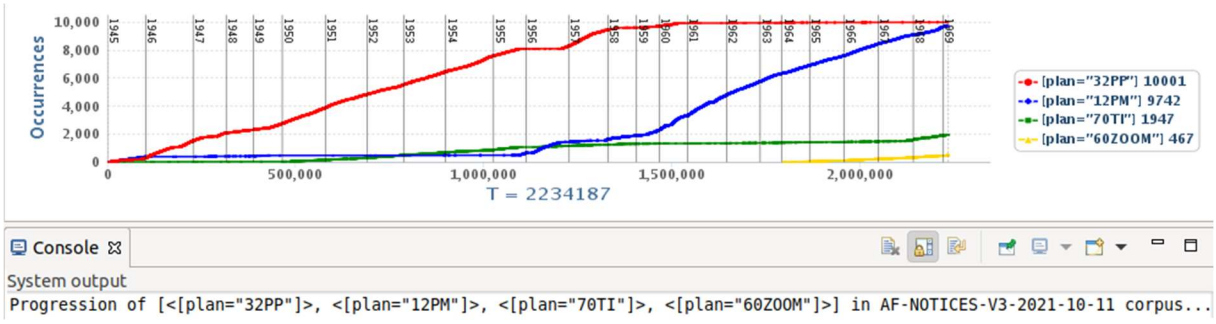

Figure 2. *Progression chart showing the diachronic evolution of 4 shot types. By decreasing frequency: Handheld Shot (red), Medium Shot (blue), Title & Credits (green), Zoom In & Out (yellow)*

Figure 1 shows a summary of the resulting annotation (a developed example is provided in the Appendix). The shot that is the most frequently noted in descriptions is the Extreme Wide Shot (10PG) with 23,721 occurrences, which represents one quarter of all annotated shot mentions (93,553 annotations) in a 2-million-word corpus. For instance, one can examine how shot mentions are distributed along time (Figure 2) and note that Handheld Shots seems to have been replaced by Medium Shots. Another pattern involves the use of Zoom lenses during newsreel shootings. The appearance of this type of camera optics in the progression chart signals its adoption by *Les Actualités françaises* cameramen in 1964 and its increasing use among the company crews from then on.

2.4. Projection step

The projection (WordProperty2Word) of the annotated corpus along its shot-type annotation ("plan" word property) creates a new corpus whose words are the sequence of shot types (see Figure 1, right panel). Thus, shot mentions are normalized, and other words are ignored. This new corpus is 91,898-word long. It can be explored with all usual textometric tools.

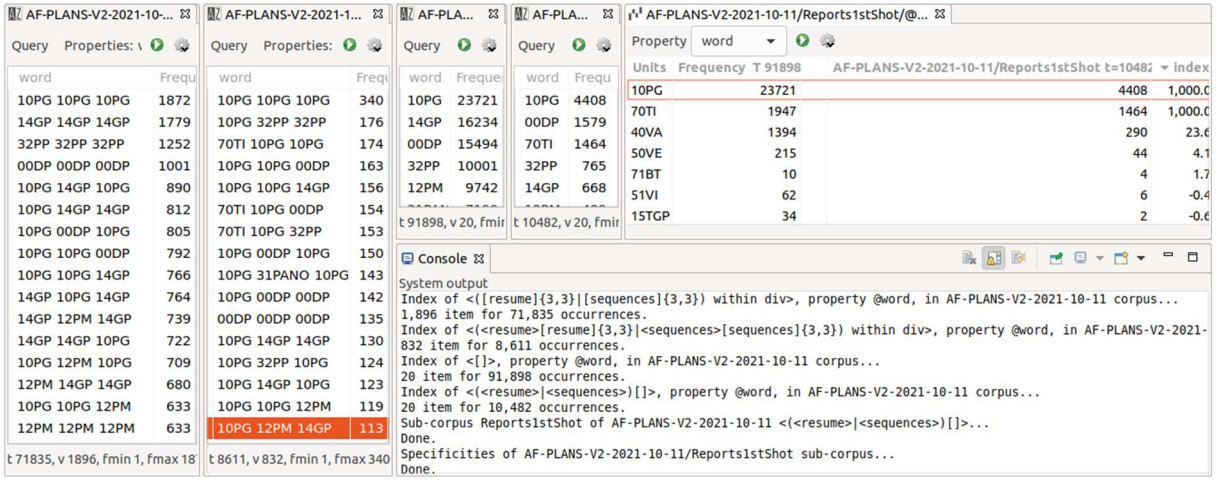

Figure 3. *Study of the film grammar in* Les Actualités françaises.
From left to right: (i) frequency list of 3-shot patterns and (ii) same thing for initial 3-shot patterns only; (iii) frequency list of shots and (iv) same thing for first shot in a report only; (iv) specificity score of shot types for the first position in reports. Bottom right panel: log of executed commands.

An illustration of the usefulness of this new representation is the study of the patterns of shots beginning a report (Fig. 3 & Appendix): we can search and count for 3-shot patterns, whether they begin the report (ii) or occur anywhere within a report (i); we can formulate a basic query matching the first shot in a report (iv), and measure specificities of shot types in this leading position (v). Specificity can be computed for patterns too (Figure 4).

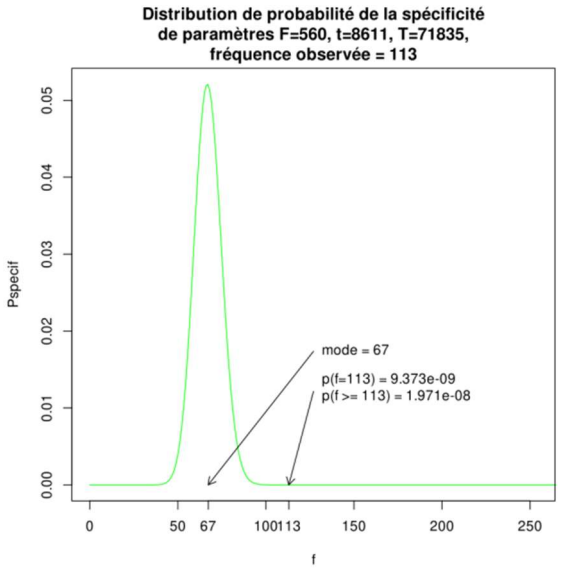

Figure 4. *Specificity result +7.7 for the funnel pattern10GP 12PM 14GP (Extreme Wide Shot, Medium Shot, close-Up) as first shot sequence in a report (TXM PlotSpecif utility output).*

We then were able to conduct a comprehensive analysis (see Appendix) of how scale evolves in the very first shots. 9,551 reports have got at least two shots mentioned in their text description. Among these, 5,582 reports (58 %) have a second shot type that is at a closer scale than the first one, which we call a funnel pattern. 2,065 other reports (22 %) use a unique and constant type of shot in their description. So all other patterns (either a second shot type wider than the first one, or a funnel pattern occurring later than in the very first shots) happen in the few remaining 1,904 reports (20 %). We interpret this prominence of the funnel shot pattern as a kind of standardization of film grammar. This approach has a dual goal. It offers the public a didactical presentation of the news based upon a simple editing where the gradual narrowing of successive shot frames enables a fluid visual exposition of any given subject. It also uses the standardization and reiteration of the *ad hoc* sequences to accelerate the shooting and editing process, thus providing an economical production method to the newsreel companies.

A drawback of this projection-based approach for active corpus is some loss of context, since original words are no longer present. However, one can work simultaneously with both corpora (the one before and the one after projection) and view both representations of the text

(see example Fig. 1 with panels *(ii)* and *(iv)*). Textual structures (sections, paragraphs) and their metadata are preserved, and each shot-type word in the projected corpus has got a property that records the lexical form of the original tagged word in the shot mention.

3. Summary and prospects

Active corpus is a critical parameter in textometric analysis as it allows finetuning of textual data for searching domain and for statistical computing. On the one hand, active corpus is quite fully implemented in textometric software as regards *paradigmatic* processing (on lexical tables); on the other hand, it is much less advanced concerning text *display* and *syntagmatic* processing. For this last case, we devised and tested a new analysis scenario based on utilities that were recently released in TXM open-source software. First step is annotating through queries, which enriches the corpus with consistent and relevant analytic categories, even on evolving data with periodical releases. Second step is a projection that derives a new corpus focused on the chosen analytic categories. Both corpora are useful to the analysis: the original one provides the lexical and textual contexts when looking back at text content; and the projected one is efficient to address textometric questions about analytic patterns. Such a strategy is worth disseminating since it provides an effective solution for many similar research scenarios and needs. A more powerful implementation of active corpus for syntagmatic processing may still be achieved with a full integration of analytic representation and textual context into a unique corpus.

This research was supported by the ANTRACT ANR project (ANR-17-CE38-0010).

References

Anthony L. (2022). *AntConc (Windows, MacOS, Linux) Build 4.0.5. Help file*. Waseda University.

Brunet É. (2011). *Hyperbase, Manuel de référence*. Université de Nice.

Carrive J., Beloued A., Goetschel P., Heiden S., Laurent A., Lisena P., Mazuet F., Meignier S., Pincemin B., Poels G., Troncy R. (2021). Transdiciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *Digital Humanities Quarterly*, 15 (1).

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

Heiden, S. (2018). Annotation-based Digital Text Corpora Analysis within the TXM Platform. In Iezzi D. F. et al. editors, *Proc. of JADT 2018*. Roma: UniversItalia, pp.367-374.

Heiden, S. (2020). *Manuel de TXM, v. 0.8*. ENS de Lyon. https://pages.textometrie.org/txm-manual/

Janssen, M. (2016). TEITOK: Text-Faithful Annotated Corpora. In Calzolari, N. *et al*., editors, *Proc. of LREC'2016*. Portorož, Slovenia, pp. 4037-4043.

Loubère L. and Ratinaud P. (2014). *Documentation IRaMuTeQ 0.6 alpha 3 v. 0.1*. Univ. Toulouse.

Pincemin B. (2004). Lexicométrie sur corpus étiquetés. In Purnelle G. et al., editors, *Proc. of JADT 2004*. Presses universitaires de Louvain, vol. II, pp. 865-873.

Pincemin B., Heiden S., Decorde M. (2020). Textometry on Audiovisual Corpora: Experiments with TXM software. In Ratinaud P. and Marchand P., editors, *Proc. of JADT 2020*. Univ. Toulouse 3.

Appendix

Full resolution figures and supplementary materials are available in the HAL-SHS archive: https://halshs.archives-ouvertes.fr/halshs-03667319