Crowdsourcing-based Annotation of the Accounting Registers of the Italian Comedy

Adeline Granet, Benjamin Hervy, Geoffrey Roman Jimenez, Marouane Hachicha, Emmanuel Morin, Harold Mouchère, Solen Quiniou, Guillaume Raschia, Françoise Rubellin, Christian Viard-Gaudin

Crowdsourcing-based Annotation of the Accounting Registers of the Italian Comedy

Adeline GRANET[1], Benjamin HERVY[1], Geoffrey ROMAN-JIMENEZ[1], Marouane HACHICHA[1], Emmanuel MORIN[1], Harold MOUCHÈRE[1], Solen QUINIOU[1], Guillaume RASCHIA[1], Françoise RUBELLIN[2], Christian VIARD-GAUDIN[1]

[1]LS2N, UMR CNRS 6004, Université de Nantes, France

[2]L'AMO, EA 4276, Nantes Université de Nantes, France

{firstname.lastname}@univ-nantes.fr

Abstract

In this paper, we present a double annotation system for new handwritten historical documents. We have 25,250 pages of registers of the Italian Comedy of the 18^{th} century containing a great variety and amount of information. A crowdsourcing platform has been set up in order to perform labeling and transcription of the documents. The main purpose is to grasp budget data from the all 18^{th} century and to create a dedicated database for the domain's experts. In order to improve, help and accelerate the process, a parallel system has been designed to automatically process information. We focus on the titles field, segmenting them into lines and checking candidate transcripts. We have collected a base of 971 title lines.

Keywords: handwriting recognition, crowdsourcing, historical data

1. Introduction

The CIRESFI project[1] sets out to reassess a theatrical heritage that has often been considered as inferior to that of the two major, royally-privileged theaters (the Opera and the French Comedy). By studying the theater that was excluded from the system of privileges, the Italian Theater, we analyze the questions of acculturation and institutionalization including the fusion of the Opera-Comique with the Italian Comedy. By employing a mass of untapped and unpublished resources (27,544 pages of registers available at the BnF[2]), this program will take a decidedly fresh look at emerging forms of creation and the changes in the entertainment economy. To this end, CIRESFI takes up a technological challenge such as creating a tool for handwriting recognition and create an interactive database.

Information retrieval is tedious in old handwritten documents for humanities and social science researchers. The digitization of collections facilitates their consultation but it is necessary to extract information to make them fully exploitable. Sometimes, handwriting recognition systems could be set up but they require to be trained on an existing ground truth. Within the CIRESFI project, this assumption is not fulfilled since the corpus is very large, diverse and quite new w.r.t. the information extraction task. Crowdsourcing allows the use of volunteers to annotate historical documents according to our needs: (i) indicate the type of a page; (ii) detect the different areas and identify information type; (iii) transcribe as required; (iv) validate or correct previous transcriptions. In order to support and boost this label-and-transcribe process, we have implemented an automatic annotation enrichment method. The detected areas are segmented into lines. Then, the transcripts associated with these zones, if any, are reworked and validated manually. Figure 1 illustrates this approach.

2. Registers of Italian Comedy

The studied documents are registers of the Italian Comedy which record the daily, monthly and annual receipts from 1716 to 1791. In addition to these valuable information, there are additional annotations concerning the historical context through 63 seasons.

The corpus is made of 25,250 pages over the official 27,544 pages of the accounting registers since the remaining pages are not available in high quality. We have identified several types of page in addition to accounts: cover, blank page, resident statement, and introduction. The most common type is the daily accounts.

Figure 2 shows a daily account with the date and titles of the plays; receipts in the left column, and expenses in the right column; followed by actor names; and sometimes, notes.

The pages have undergone severals changes over the century. Firstly, this layout of information moved in four approaches using more or less than one page for all information. Secondly, at the beginning of the century, the Italian actors drafted the accounts themselves. Later in the century, a cashier was hired to write these accounts. Drafting language has evolved from the various Italian dialects to French.

These data show the evolution of writing which is interesting from a historical point of view but also for the natural handwriting language processing. Table 1 presents some relevant points.

Using the data extracted from the crowdsourcing process, we aim at designing an automatic reading system that focuses on the title field. This contains different levels of information ranging from a short list of the plays performed that day to specific information about a given play. The latter may be the number of acts, if it is a specific piece from the Italian Comedy, how many times it was played, if it was played in special places, or in front of the king's court.

[1]French National ANR-14-CE31-0017 program.

[2]*Bibliothèque Nationale de France*. Digital library: `http://gallica.bnf.fr/accueil/`

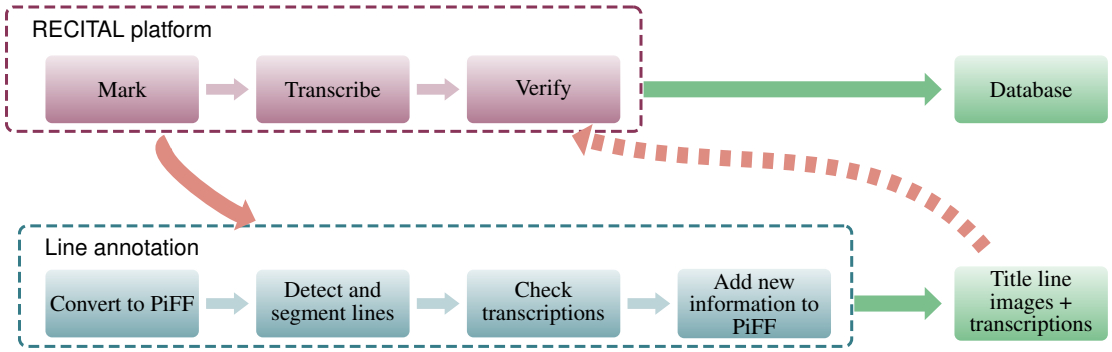


Figure 1: Dataflow of our approach for Italian Comedy documents.

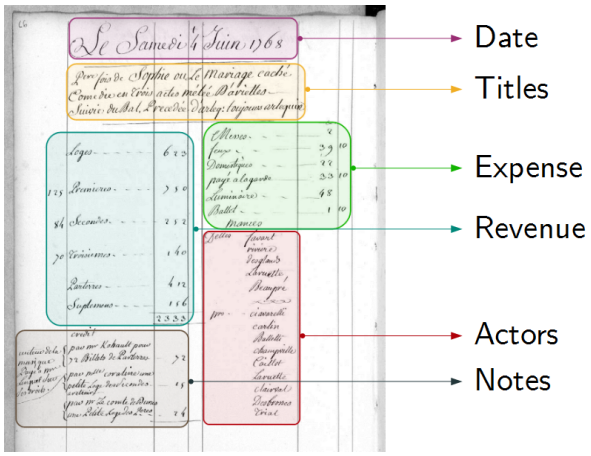
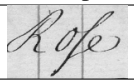
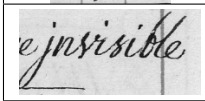
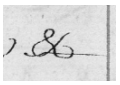
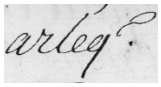


Figure 2: Example of a daily budget record for the Italian Comedy with identification fields.

Table 1: Special characters and abbreviations in the Italian Comedy documents.

Example	Transcription	Relevant Point
	"Rose"	's' has two form : long and short
	"Invisible"	using 'i' or 'j' made no difference
	"etc"	strong abbreviation to replace several words
	"arlequin"	weak abbreviation to cut few characters

3. Related work

The processing steps to be performed are to detect, identify, transcribe and validate the handwritten information contained in these documents. Severals approaches are available : perform all tasks automatically or with a participatory system. The option of having these documents transcribed by at least one specialist was rejected at the outset.The time spent to achieve this, would count in years for one person (see. section 4.3.).

As far as handwritten recognition is concerned, state of the art techniques and methods already exist. OCR softwares such as Abby®FineReader are widely used in production but failed to perform satisfying OCR results on our datasets, even on simple pages, probably due to features listed in table 1. Moreover, handwriting recognition systems could be set up but they require to be trained on an existing ground truth that we do not have for our heterogeneous dataset.

To save and re-use easily all information from the analyzed document, different XML format exits. The best-known format is the Text Encoding Initiative (TEI) dedicated to the representation of the textual components such as the transcription of one play. The main problem is that despite the great complexity of this format, it does not allow us to easily bind spatial information to transcripts as well as their type. Another one, Page Analysis and Ground-truth Elements(Pletschacher and Antonacopoulos, 2010) (Page XML) allows to store image features, layout structure and content page. However, the diversity and complexity of data contained in a daily page of Italian comedy require types and tags to be more specific and hold-back all levels of annotation. This is for those reasons we choose to use a new format called Pivot File Format (PiFF)(Mouchere et al., 2017).

4. Crowdsourcing Platform: RECITAL

Due to the previously described heterogeneity of the documents, a pure OCR-based approach to annotate the accouting registers is doomed to failure. Conversely, labeling and transcription are well-suited HITs (Human Intelligence Tasks) for crowdsourcing (Chittilappilly et al., 2016). The RECITAL crowdsourcing (CS) platform[3] is a fork of ScribeAPI[4], a framework for label-and-transcribe tasks of OCR resistant text-based documents. Although CS approaches to transcription of text-based documents are nowadays usual in digital humanities projects, the RECITAL workflow integrates a pre-labeling step and shows a high level of complexity. Indeed, we need to classify a typology of hundreds of different categories of revenues and expenditures. In addition, we also need to identify dates, titles, and many other information like actors and actresses names, cashiers, etc. Those information are written in different languages, by different people and in different types of documents (daily budgets, annual records, etc.) over the decades covered by the corpus.

[3] http://recital.univ-nantes.fr
[4] https://github.com/zooniverse/scribeAPI

4.1. Overall Workflow for Crowdsourcing

Basically, the workflow (see fig. 3) is composed of 3 activities and follows a sequential implementation:

1. **marking**: this step consists for a worker in classifying the displayed page. 8 different page types can be picked, among which three (covers, blank pages and unclassifiable) close the process and make the page retired. Then, depending on what page type was picked, a sequence of 5 "screens" is suggested. Each screen proposes about 10 marks to identify the different kinds of information (revenues, expenditures, names, etc.). Workers can mark as much elements as possible and will be asked at the end of the sequence if everything in the page has been labeled. They can also stop working whenever they want within the activity.

2. **transcribing**: this step consists in transcribing the text that has been marked in the previous step. Workers have the ability to label the mark as illegible or report the mark as misplaced if necessary.

3. **verifying**: when at least 2 different transcriptions have been proposed, they are submitted as a vote to other workers in order to achieve a consensus.

The CS annotation process offers a very large number of micro tasks, taking all types of documents together (see fig.3): 133 different elements can be classified among 8 types of documents. The marking activity is divided into a sequence of general categories of information to mark (from global to types of revenues, by the way of types of expenditures, names of actors and actresses, etc.).

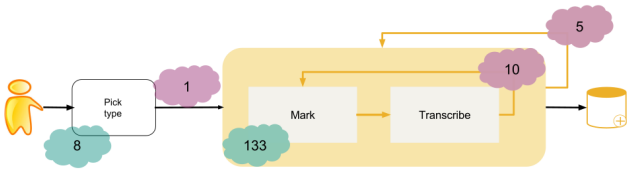


Figure 3: Simplified view of the overall workflow for one page. Voting activity is not represented. Figures in blue represent the numbers of tasks within the activity whereas figures in red represent overall existing categories.

4.2. Crowdsourcing Setup

Task assignment Marking, transcribing and verifying activities can be freely selected by a worker. Within one activity, a random task among available ones is assigned to the worker. But one can also get a direct access to a specific register of accounting records[5] to work on. Moreover, workers (if authenticated) cannot operate the verification on a transcription they made.

Management of outliers During the workflow, workers can report unclassifiable pages (for the page type label), but also mistakes (misplaced mark for example) done by other workers. Finally, illegible parts can be reported. After two reports from two distinct workers, the element (page/mark) is retired and flagged accordingly in the database.

Consensus achievement Each mark is submitted for transcription to at least 2 distinct workers. Approximate matching offered by lossy algorithms (case and punctuation insensitive comparison, ignoring whitespaces) is used to increase the ability to achieve consensus as suggested in (Matsunaga et al., 2016). If it yields to the same transcript, the proposition is accepted and the mark is retired from the workflow. Otherwise, a new vote is submitted. During the vote task, a worker can either choose one of the already existing transcripts or rather propose a new one. In case of distinct transcriptions, we want to reach a consensus based on majority voting. The threshold for majority is 75% of voters, having a number of voters between 3 and 10. When 10 distinct workers have voted without reaching a majority, the vote is closed and the annotation is considered as a dissensus.

4.3. Monitoring

Working hours We estimate that there is an average of 30 information to be transcribed per page. Based on existing data (timestamps of users' actions) on the platform, we can approximate the average time spent by one of the experts on each information: 23.5 seconds for marking, and 13 seconds for transcription. Thus, an expert would spent approximately 26 minutes per page. Given the number of 1540 hours worked per year, it would take almost 3.5 years for one full-time expert to complete the transcription of the 25,250 pages !

Users and answers The RECITAL platform is hosting 25,250 out of the 27,544 total pages. At the date of the 23^{th} of January 2018, 68 540 tasks (see Fig. 4) have been performed by 314 workers. Though this number is an indicator of the activity on the platform and reveals the worker's engagement, it is not a good overview of the overall progress in labeling and annotating the corpus. So we computed (fig. 5) the number of marks, transcriptions and pending (or closed) votes per page for each page of each register (one register per year).

Annotations and consensus On a total of 46,504 marks created (defining an area to be transcribed), 43.6% were transcribed by at least one worker. Among the marks that have at least 1 transcription, 76.3% have been transcribed by only one worker, 11% have been completed (case when 2 distinct workers make the same transcription successively), 10.1% are pending votes (at least 2 different transcriptions), and 2.6% (536 cases) have yield a consensus (we only have 7 cases of dissensus, e.g. consensus failures, in our dataset).

4.4. Limits and Perspectives

At this step of the CS process, we are able to identify some limitations and perspectives both in the methodology and the underlying framework.

Document ordering Figure 5 illustrates the issue caused by displaying direct access to documents. Therefore, we added a condition on the workflow based on an arbitrary

[5] Accounting records are grouped by year in registers. The volume of pages per register varies between 192 and 590.

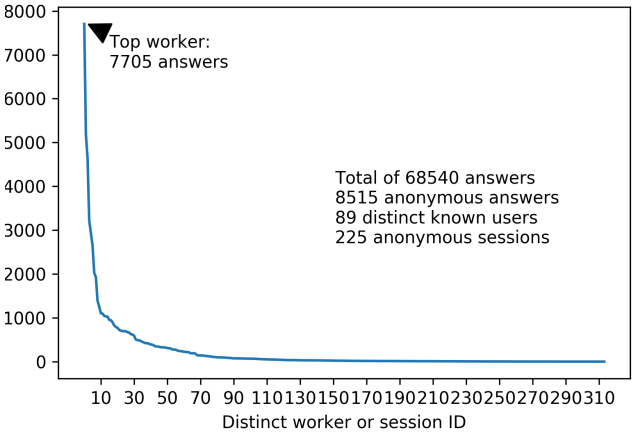


Figure 4: Distribution of answers (taking all activities together) per worker at the date of 2018-01-23. Only the workers that completed at least one task are considered.

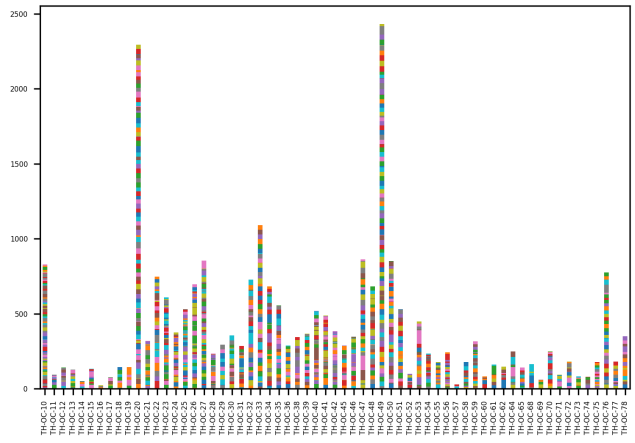


Figure 5: Overview of crowdsourcing progress at the date of 2018-01-23 by accounting registers (time-based ordering on x-axis). Colored bars are single pages and their size depends on the number of marks, transcriptions and consensus existing for this page.

order. This order is defined as follows: accounting records after 1747 are much more easier to mark and transcribe so should be prioritized. This is expected to drastically reduce time-to-complete document annotation and improve user engagement.

Free *vs.* controlled transcription Annotation may yield to slightly different transcripts from distinct workers. It depends of course on the worker's expertise, but also her own judgment of the expected result. For instance, a date like "*Du Mardy 12 Juillet 1768*" (in French) can be transcribed as a facsimile, or can be interpreted and transcribed to a canonical form like "*mardi 12/07/1768*". Normalizing is good for data post-processing and undesirable for pattern recognition. Abbreviations, approximate handwriting quality, multiple writers and multilingual documents highlight the importance and the difficulty to set up instructions, help content and input controls in such crowdsourcing workflow.

Convergence to consensus Lossy algorithms to merge transcripts by similarity are expected to reduce the effort towards reaching consensus by majority voting. Besides, it has been shown (Little et al., 2010) that an iterative updating process can be quite efficient to achieve the consensus in a transcription task.

User engagement Online platforms can improve user engagement through rewards, challenges, community management with discussion forum, etc. Such simple feedback and motivational techniques have been proved useful when you face a small crowd of volunteers inherently interested in the task (Clematide et al., 2016).

User profiling There exist state-of-the-art methods to insert fake tasks with known results to classify users (*ELICE for Expert Label Injected Crowd Estimation*) or learn from disagreement.
Finally, an interesting perspective is related to the combination of CS results and supervised segmentation and transcription. This idea is described in the next section.

5. Line Transcription

CS platforms requires a lot of workers to reach a consensus. In order to accelerate the annotation process, an automatic recognition would be very helpful. Such a system requires a training stage with annotated text lines. Figure 1 illustrates how the first collected data (even if they weren't validate yet) are used to simplify this huge annotation task. As a first step we focused on the title field, but we plan to extend to other types of field.

Converting to PiFF Firstly, we select and convert all data related to title field into a Pivot File Format (PiFF)[6]. This format is a solution promoting the exchange of information between different systems. At each step, the document can be enriched by the new associated results. The PiFF file format allows to store locations of polygons (text areas in our case) and one or several annotations attached to each of them.

Detecting and segmenting lines The coordinates specified by the users are used to crop the full title area. Then, the text line extraction algorithm (TLA) proposed by (Arvanitopoulos and Süsstrunk, 2014) is applied on it. It provides a new polygon for each line in this area. At this step, all of them are added to the PiFF without to sort error line. We associate those new polygons with all candidate transcriptions given by user for this area.

Checking transcriptions Our aim is to create a ground truth for handwriting recognition system focused on title line. So, we build a simple user interface (see Figure 6) to check if:

- the current polygon is correct, i.e it's not void line or just a thread;
- the transcription is a perfect match with the current polygon, i.e sort facsimiles and standardized transcripts.

In case several transcripts have been proposed for a title area, they are all proposed in order to associate the best one

[6] https://gitlab.univ-nantes.fr/mouchere-h/PiFFgroup

or to correct the closest. When the user valids her transcriptions and polygons for one page, once again, the associated PiFF is upgraded with new information. We validated 971 title line images and their transcriptions.

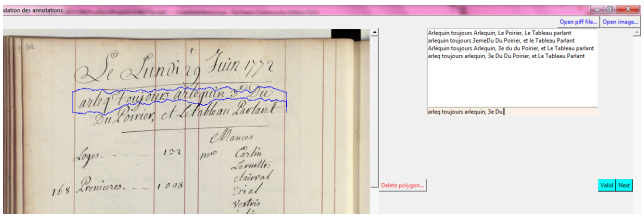


Figure 6: User interface to check manually transcriptions for title line image. The left part shows the polygons and area. The blue polygon is the result from TLA. The right part provides all candidate transcriptions. The user can validate her choice, pass to next polygon (without validation) or delete one polygon with its candidate transcriptions.

This assisted annotation system allowed us to create new resource of labeled title line images and formalize collected information in a dynamic format. The automatic handwriting recognition of our project can start the training phase with title line images.
Currently, the interaction is unilateral but in the future we are going to make it bilateral by feeding the RECITAL platform with the polygons obtained automatically with DMOS (Couasnon, 2001) for the mark information. Furthermore, we have a recognition system consisting of a Convolutional Neural Network (CNN) for the automatic features extraction; and a BLSTM-CTC neural network for the transcription part as described in (Granet et al., 2018). This deep neural network for handwriting recognition is going to provide more transcriptions for the database. This is illustrated in Figure 1 by the red dotted arrow. We can use those new data at two locations in the workflow. Firstly, the neural network results giving sufficient confidence are going to be proposed for the verifying task to accelerate crowdsourcing process. Another part of the data are going to be used to compare with crowdsourcing final data for checking data quality.

6. Conclusion

We presented two types of production to study the economy of Italian Comedy through the accounting registers. The hybrid system combining crowdsourcing platform and annotation system allows to identify and annotate unknown documents. This allowed us to collect information in several forms: 25,250 digitized pages from accounting registers; a database of crowdsourced transcriptions, and a new handwriting database focusing on the title plays, that includes 971 images. All PiFF documents produced are going to be distributed with Creative Commons Attribution-Non Commercial-Share Alike 3.0 Unported License.

7. Bibliographical References

Arvanitopoulos, N. and Süsstrunk, S. (2014). Seam carving for text line extraction on color and grayscale historical manuscripts. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, pages 726–731. IEEE.

Chittilappilly, A. I., Chen, L., and Amer-Yahia, S. (2016). A Survey of General-Purpose Crowdsourcing Techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28(9):2246–2266, sep.

Clematide, S., Furrer, L., and Volk, M. (2016). Crowdsourcing an OCR Gold Standard for a German and French Heritage Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 975–982, Portorož. European Language Resources Association (ELRA).

Couasnon, B. (2001). Dmos: A generic document recognition method, application to an automatic generator of musical scores, mathematical formulae and table structures recognition systems. In *Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*, pages 215–220. IEEE.

Granet, A., Morin, E., Mouchere, H., Quiniou, S., and Viard-Gaudin, C. (2018). Transfer learning for handwriting recognition on historical documents. In *7nd International Conference on Pattern Recognition Applications and Methods, ICPRAM 2018*.

Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010). Exploring iterative and parallel human computation processes. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '10, pages 68–76, New York, NY, USA. ACM.

Matsunaga, A., Mast, A., and Fortes, J. A. (2016). Workforce-efficient consensus in crowdsourced transcription of biocollections information. *Future Generation Computer Systems*, 56(C):526–536, mar.

Mouchere, H., Kermorvant, C., Rojas, A., Coustaty, M., Chazalon, J., and Couasnon, B. (2017). Piff: a pivot file format. In *Proceedings of the 1st International Workshop on Open Services and Tools for Document Analysis*. IEEE.

Pletschacher, S. and Antonacopoulos, A. (2010). The page (page analysis and ground-truth elements) format framework. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 257–260. IEEE.