Recherche de consensus : les enseignements de la plate-forme RECITAL de transcription participative

Guillaume Raschia, Benjamin Hervy

Recherche de consensus : les enseignements de la plate-forme RECITAL de transcription participative

Guillaume Raschia*†, Benjamin Hervy*

*LS2N; †Polytech Nantes

1 Introduction

Le projet ANR CIRESFI [1] propose de mettre en relation tous les éléments et événements qui concernent les spectacles des théâtres de la Foire et ceux de la Comédie-Italienne au XVIIIe siècle, depuis le coût de production, les accessoires utilisés, les acteurs employés, jusqu'à la composition sociale du public, les instruments de l'orchestre, les danses et les textes.

Pour ce faire, le projet exploite un corpus de registres comptables du théâtre de la Comédie-Italienne, couvrant la période de 1716 à 1791. Ces 27544 pages de documents manuscrits, numérisés et mis à disposition par la BnF [2], renferment un gisement d'informations inédites mais difficilement accessibles étant donnée l'hétérogénéité à la fois formelle et morphologique du contenu. En effet, bien que la source soit unique, si l'on considère les 63 registres saisonniers comme un seul corpus, le contenu de ces documents d'archive varie d'une page à l'autre ; tantôt sont consignés des comptes journaliers, tantôt des synthèses mensuelles voire annuelles. En outre, les registres sont rédigés en dialecte vénitien au début du siècle, puis en vieux français jusqu'à la fin du siècle. La nomenclature comptable évolue également au fil du temps. À cela s'ajoutent des changements de scripteurs et donc de graphie, de mise en page, de style, *etc.* Par conséquent, les mêmes informations se trouvent formulées et présentées de manières très différentes au cours du siècle.

L'hétérogénéité formelle met en échec les techniques de reconnaissance automatique d'écriture qui sont incapables de produire des résultats fiables pour l'extraction d'information à partir de notre corpus. C'est pourquoi, nous avons mis en place RECITAL [3] une plate-forme de production participative (*crowdsourcing* [CCAY16]) dévolue à la transcription de ces archives. La fragmentation de ce travail titanesque en une myriade de micro-tâches pseudo-indépendantes est un processus particulièrement bien adapté aux projets de transcription documentaire [CGST18, CMP+13]. L'ampleur, quelques 27544 pages, est également un motif légitime pour justifier l'ouverture au monde de ce travail de transcription.

L'approche proposée vise à disposer d'une base de données historiques fiable, documentée, et fidèle au corpus original, pour laquelle nous réfléchissons au développement de mécanismes exploratoires, et également à des modes d'analyse quantitative qui peuvent se révéler précieux pour les spécialistes en histoire culturelle. Il s'agit ainsi de proposer, entre autre, des modalités de visualisation des données au moyen de statistiques brutes ou agrégées, des représentations graphiques, une exploration calendaire ou encore une recherche par facettes.

RECITAL offre les 3 activités suivantes :

1. **Marquage** : il s'agit de définir le type de page présentée, puis en fonction de ce choix, de marquer la position et la nature des informations contenues dans la page. Il est ainsi possible de catégoriser une information parmi 133 types disponibles : dépenses, recettes, calculs budgétaires, informations générales (date, titre de pièce, *etc.*), *etc.*.

1. Site web du projet : http://cethefi.org/ciresfi/doku.php
2. Bibliothèque nationale de France. Une convention de coopération documentaire avec la BnF régit l'exploitation de ce corpus dans le cadre du projet CIRESFI.
3. Site web : http://recital.univ-nantes.fr

2. **Transcription** : cette activité permet de transcrire le contenu de chaque marque réalisée à l'étape précédente. La production à l'issue de cette étape est donc une séquence de caractères, annotée par l'une des catégories disponibles, associée à une marque dans l'une des pages du corpus, et éventuellement bruitée. C'est pourquoi la même marque est systématiquement soumise pour transcription à deux participants. Si les deux transcriptions sont identiques, alors la donnée produite est validée, sinon, elle entre dans une procédure de vote.
3. **Vérification** : la fin du processus consiste à « élire » la transcription définitive d'une marque, dès lors qu'une divergence a été observée. Il est en outre possible de suggérer une nouvelle transcription, si aucune des propositions ne convient. Le consensus est obtenu par un vote à la majorité des trois quarts, ou alors un contentieux est déclaré à l'issue de 10 votes.

Ce processus contourne la nécessité d'une coûteuse transcription et/ou validation d'experts, remplacée par un mécanisme de **recherche de consensus** au sein d'une communauté nombreuse mais non fiable.

De nombreux facteurs tels que les choix de conception de la plate-forme, la nature du corpus, le profil des participants, entravent la recherche d'un consensus lors d'une transcription. Nous proposons dans cette communication, de détailler les causes de dissensus, de présenter un bref état-de-l'art [AKK14, MMF16, SL13] sur la recherche de consensus dans les systèmes de production participative et enfin, de commenter les expérimentations menées au sein du projet CIRESFI dans le but d'atteindre ou de corriger le consensus.

En date du 25 mai 2018, la plate-forme RECITAL comptabilisait 600 participants totalisant environ 100K tâches réalisées, pour près de 4000 transcriptions validées.

Références

[AKK14] Paul André, Robert E. Kraut, and Aniket Kittur. Effects of simultaneous and sequential work structures on distributed collaborative interdependent tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 139–148, New York, NY, USA, 2014. ACM.

[CCAY16] Anand Inasu Chittilappilly, Lei Chen, and Sihem Amer-Yahia. A survey of general-purpose crowdsourcing techniques. *IEEE Trans. on Knowl. and Data Eng.*, 28(9) :2246–2266, September 2016.

[CGST18] Tim Causer, Kris Grint, Anna-Maria Sichani, and Melissa Terras. Making such bargain : Transcribe bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities*, 2018.

[CMP+13] Laura Carletti, Derek McAuley, Dominic Price, Gabriella Giannachi, and Steve Benford. Digital humanities and crowdsourcing : An exploration. In *Proceedings of MW2013 : Museums and the Web 2013*. Museums and the web, April 2013.

[MMF16] Andréa Matsunaga, Austin Mast, and José A.B. Fortes. Workforce-efficient consensus in crowdsourced transcription of biocollections information. *Future Generation Computer Systems*, 56 :526 – 536, 2016.

[SL13] Aashish Sheshadri and Matthew Lease. SQUARE : A benchmark for research on computing crowd consensus. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2013, November 7-9, 2013, Palm Springs, CA, USA*, 2013.