

Correction des données : retour d’expérience sur la plate-forme RECITAL de transcription participative

Benjamin HERVY*, Pierre PÉTILLON†, Hugo PIGEON†, Guillaume RASCHIA*†

*LS2N; †Polytech Nantes

1 Introduction

Ce travail s’inscrit dans le cadre du projet ANR CIRESFI¹. Il 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 BnF², 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 RECITAL³ une plate-forme de production participative (*crowdsourcing*) dédiée 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. L’ampleur, quelques 27544 pages, est également un motif légitime pour justifier l’ouverture au monde de ce travail de transcription. Pour un survol des principaux défis techniques que pose la mise en place d’un système de production participative, le lecteur est invité à se reporter à [CCAY16].

2 Les {RE}gistres de la {C}omédie-{ITAL}ienne

Bien que le recours au *crowdsourcing* soit désormais usuel dans nombre de projets de transcription (*Transcribe Bentham* [MJV11, CGST18], *Transcribathon*⁴, *Testaments de poilus*⁵) en humanités numériques [WWTN12, CMP+13], notamment à des fins d’édition électronique, la plate-forme RECITAL se singularise par (i) la variété et la complexité des tâches à réaliser, (ii) les objectifs poursuivis et (iii) les processus à l’œuvre.

- (i) RECITAL propose 3 grandes classes de tâches, le marquage, la transcription et la vérification, pour un total de 224 micro-tâches différentes encodées dans le système.
- (ii) RECITAL permet, outre la transcription des pages de registres, leur annotation en fonction de la nature des informations recueillies, pour distinguer par exemple un titre de pièce d’une ligne de recette.

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>

4. Site web : <https://transcribathon.com/en/>

5. Site web : <https://testaments-de-poilus.huma-num.fr/>

(iii) RECITAL déploie une stratégie d'affectation des tâches aux participants, un séquençage des tâches et un mécanisme de validation, propres au projet.

D'un peu plus près, 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.*.
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. Il pose néanmoins la question fondamentale de la correction des données ainsi collectées. C'est le thème que nous plaçons au cœur de cette communication.

En date du 10 avril 2018, la plate-forme RECITAL comptabilisait 550 participants totalisant environ 95000 tâches réalisées, pour près de 3800 transcriptions validées.

3 Du bruit dans les données

Il s'agit pour nous de proposer des méthodes et outils d'évaluation de la qualité des données d'une part, et de nettoyage des données d'autre part. Le problème relève à la fois du diagnostic et du remède! Les causes de ce questionnement, qui ont été largement exposées ci-dessus, peuvent être outrageusement réduites à deux facteurs : l'hétérogénéité des contributions et l'hétérogénéité du corpus documentaire. Pour ce qui relève du processus de *crowdsourcing*, il existe des travaux antérieurs, et notamment les expériences pour l'évaluation et la réparation de données menées dans le cadre du projet NSF-ADBC [MMF16].

3.1 À l'origine du problème

Du point de vue du procédé de *crowdsourcing*, et plus particulièrement de la plate-forme RECITAL, nous avons inventorié les mécanismes qui contribuent à « endommager » les données :

- les profils des participants, et notamment l'existence d'utilisateurs inexpérimentés voire malveillants ;
- la recherche de consensus, d'abord par une double transcription, puis par un vote itératif à la majorité des trois quarts ;
- le mode de transcription libre qui n'impose aucun contrôle sur les propositions. Seules les consignes (aide contextuelle, guide de démarrage, tutoriels, forum, *etc.*) offrent des repères normatifs que les participants peuvent suivre ou ignorer ;
- la stratégie d'affectation des tâches aux participants. Chacun choisit son activité parmi Marquer, Transcrire, Vérifier, ensuite le système propose une tâche uniformément aléatoire et indépendante de la séquence déjà réalisée. En outre, un même participant ne peut contribuer plusieurs fois à une transcription.

De façon orthogonale, le corpus documentaire présente, comme nous l'avons détaillé en introduction, des caractéristiques aggravantes pour la qualité des données transcrites :

- le changement de scripteur. Variation de la graphie, de la mise en page, de la morphologie des textes ;
- la nature des comptes, journalier, mensuel ou annuel. Granularité multiple des informations ;
- l'évolution des règles comptables. Variation des modèles de comptes, des catégories de dépenses et de recettes, des modalités de calcul ;

- les erreurs dans le texte original. Calcul, report de valeur, orthographe de noms propres, *etc.*

C'est, de manière évidente, la combinaison de tout ou partie de ces facteurs qui rend compte de la difficulté à traiter le problème de correction des données. L'un des points clés de cette démarche concerne le problème classique de résolution d'entité [KTR10] (ou *record linkage*) induit par l'hétérogénéité morphologique du corpus, et pour lequel chaque manifestation de l'entité, i.e. une marque transcrite, est elle-même incertaine étant donnée les transcriptions multiples et la recherche de consensus !

3.2 Analyse et résolution

Parmi les expérimentations menées pour l'évaluation de la qualité des données produites, nous présenterons dans cette communication les modalités et une interprétation de :

- statistiques générales sur les données recueillies, avec une mise en lumière des zones présentant le plus d'indécision ;
- trois études particulières sur les dates, les noms d'acteurs et les valeurs monétaires ;

Dans un second temps, nous commenterons quelques résultats au sujet du nettoyage des données, résultats obtenus par les approches suivantes :

- trois mécanismes de fusion, ou résolution de conflits, appliqués aux dates, aux noms propres et aux valeurs monétaires. Nous détaillerons les heuristiques mises en œuvre pour la réconciliation automatique à partir des propositions de transcription et des votes non clôturés ;
- alignement de séquences pour établir une correspondance entre la transcription des titres de pièces jouées lors d'une même soirée avec une liste de titres de référence ;

Parallèlement, nous présenterons le modèle retenu pour représenter ces données historiques et leurs méta/para-données. Pour cela, nous avons partiellement intégré l'incertitude et la provenance des informations recueillies, comme la trace des traitements opérés à partir d'une source d'information (une archive, une donnée tierce) jusqu'à la donnée. Ce mécanisme a pour vocation de (i) favoriser l'adhésion des experts aux thèses issues des données quantitatives que nous produisons et (ii) s'engager dans une démarche scientifique de transparence et de reproductibilité des expériences et des jeux de données.

4 À terme...

...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.

Par ailleurs, nous travaillons conjointement avec une équipe de recherche qui traite le problème de la reconnaissance automatique de l'écriture dans les registres, étant entendu que la transcription automatique intégrale des registres est inenvisageable à l'aune des technologies de l'état-de-l'art. Une thèse de doctorat est en cours sur la reconnaissance des titres de pièces et leur transcription à l'aide de modèles d'apprentissage adaptés (réseaux de neurones). Cette démarche « concurrente » de RECITAL offre un double avantage, sur le modèle de la fertilisation croisée : les résultats issus de l'apprentissage automatique peuvent être réintroduits comme propositions d'un « robot-participant » dans la plate-forme de *crowdsourcing* et soumis au processus de validation par le nombre. Inversement, les données collectées par RECITAL peuvent servir de jeu d'entraînement pour l'apprentissage des modèles statistiques.

5 Conclusion

Il est établi, grâce à une batterie d'évaluations quantitatives, que la nature du corpus, la communauté de volontaires, la complexité des micro-tâches et l'ensemble des choix de conception pour RECITAL (pas de modélisation des participants, pas de validation par l'expert, transcription libre) sont autant de facteurs qui participent conjointement à dégrader la qualité des données recueillies.

Par le biais d’algorithmes d’alignement et de nettoyage des données, il est cependant possible de mettre en œuvre des correctifs automatiques pour améliorer la qualité des propositions.

Parmi les transcriptions ne permettant pas d’aboutir à une proposition majoritaire, on retrouve des situations résultant des choix de conception de la plate-forme de *crowdsourcing* : transcription littérale vs. diplomatique, gestion des abréviations, mauvaise utilisation des séparateurs choisis pour les valeurs monétaires. On retrouve également bien évidemment des divergences provenant du corpus lui-même principalement liées au déchiffrement de graphies difficilement lisibles.

Les résultats obtenus sont plutôt encourageants sur les annotations induisant le plus de divergences. Par exemple, 85% des transcriptions des marques de type "date du jour" aboutissent à une proposition. Ce constat est sensiblement équivalent pour le cas des transcriptions des titres de pièces.

Pour la suite du projet, plusieurs pistes sont envisagées pour améliorer la qualité des données, notamment la modification du processus de Vérification par l’intégration d’une approche itérative [AKK14] plutôt que cumulative. Par ailleurs, la conjonction des informations de provenance et des instruments de restitution, augmentée d’une capacité d’édition, doit permettre la validation résiduelle par les experts qui pourront s’appuyer sur les para-données.

Enfin, une des difficultés majeures dans ce type de projet de *crowdsourcing* repose sur l’animation d’une communauté de volontaires [SRA⁺16].

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.
- [KTR10] Hanna Köpcke, Andreas Thor, and Erhard Rahm. Evaluation of entity resolution approaches on real-world match problems. *Proc. VLDB Endow.*, 3(1-2) :484–493, September 2010.
- [MJV11] Moyle Martin, Tonra Justin, and Wallace Valerie. Manuscript transcription by crowdsourcing : Transcribe bentham. *LIBER Quarterly*, 20 :347–356, 2011.
- [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.
- [SRA⁺16] J. Schönböck, M. Raab, J. Altmann, E. Kapsammer, A. Kusel, B. Pröll, W. Retschitzegger, and W. Schwinger. A survey on volunteer management systems. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, pages 767–776, Jan 2016.
- [WWTN12] Claire Warwick, Claire Warwick, Melissa Terras, and Julianne Nyhan. *Digital Humanities in Practice*. Facet Publishing, 2012.