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

Data correction for transcription in crowdsourcing. A feedback from RECITAL platform.

Benjamin HERVY[1], Pierre PÉTILLON[2], Hugo PIGEON[2], Guillaume RASCHIA[1,2]

[1] LS2N - Polytech Nantes

[2] Polytech Nantes School of Engineering

Rue Christian Pauc, BP50609, 44306 Nantes Cédex 3, France

benjamin.hervy@univ-nantes.fr, guillaume.raschia@univ-nantes.fr

pierre.petillon@etu.univ-nantes.fr, hugo.pigeon@etu.univ-nantes.fr

RÉSUMÉ. Les sciences participatives trouvent une place de choix dans les projets d'humanités numériques. En effet, le recours à la foule, notamment dans le cas de la transcription de documents anciens, permet de pallier en partie les contraintes des techniques de reconnaissance automatique d'écriture. Cette approche apporte également des outils complémentaires à la validation de l'expert pour des tâches fastidieuses comme la classification ou l'extraction d'information à partir d'un texte. Cependant, ce type d'approche génère des problématiques inhérentes à la méthodologie employée et à la nature du corpus. Dans cet article, nous proposons des pistes d'évaluation et de résolution du problème de la qualité des données et de leur correction. Pour cela, nous nous appuyons sur le projet CIRESFI et la plateforme RECITAL de transcription des registres comptables de la Comédie Italienne.

ABSTRACT. Crowdsourcing have been widely deployed to cover some challenges in digital humanities, like in the transcription of old handwritten documents. Such approach is especially useful to tackle existing limits in automatic handwriting recognition techniques. Crowdsourcing allows workers to help experts in extraction and classification of information, when the workload is daunting. Yet, it yields some specific challenges related to the quality of produced data. In this paper, we discuss data quality in a research project called CIRESFI which aims at transcribing Italian Comedy financial archives through the RECITAL web platform. We finally propose some leads to tackle these issues.

MOTS-CLÉS : Sciences participatives, Humanités numériques, Manuscrits anciens, Transcription, Qualité des données, Comédie Italienne.

KEYWORDS: Citizen sciences, Digital humanities, Old handwritten documents, Transcription, Data quality, Italian Comedy.

1 Introduction

Ce travail s'inscrit dans le cadre du projet ANR CIRESFI [1]. 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 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 ;

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.

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*) dévolue à la transcription de ces archives. RECITAL s'appuie sur le framework de transcription participative *ScribeAPI* [4], à l'origine de la plateforme *Zooniverse* [5]. 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 (Chittilappilly *et al.*, 2016).

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* (Martin *et al.*, 2011; Causer *et al.*, 2018), *Transcribathon* [6], *Testaments de poilus* [7])) en humanités numériques (Warwick *et al.*, 2012; Carletti *et al.*, 2013), 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.
(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.

On retrouve notamment ce découpage de tâches dans certains projets comme *Old Weather* présent sur la plateforme Zooniverse qui propose de marquer et de transcrire des éléments présents dans des documents d'archives

3. Site web : `http://recital.univ-nantes.fr`
4. Site web : `https://github.com/zooniverse/scribeAPI`
5. Site web : `https://www.zooniverse.org/`
6. Site web : `https://transcribathon.com/en/`
7. Site web : `https://testaments-de-poilus.huma-num.fr/`

de baleiniers. Cependant, le nombre de catégories utilisées pour le découpage dans le cadre du projet *Old Weather* est de l'ordre de 5 à 10 contre 133 pour RECITAL, et la tâche de vérification est également ajoutée dans le processus participatif.

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 28 octobre 2018, la plate-forme RECITAL comptabilisait 730 participants totalisant environ 122000 tâches réalisées, pour 4350 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 (Matsunaga *et al.*, 2016).

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 (Köpcke *et al.*, 2010) (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 avons :

— établi des statistiques générales sur les données recueillies, avec une mise en lumière des zones présentant le plus d'indécision ;

— conduit trois études particulières sur les dates, les noms d'acteurs et les valeurs monétaires ;

Dans un second temps, les approches suivantes ont été mises en œuvre pour le nettoyage des données :

1. trois mécanismes de fusion, ou résolution de conflits, appliqués aux dates, aux noms propres et aux valeurs monétaires. Nous décrivons succinctement ci-dessous les heuristiques mises en œuvre pour la réconciliation automatique à partir des propositions de transcription et des votes non clôturés ;
2. un 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 ;

Les dates du jour, telles qu'elles sont inscrites sur les registres quotidiens, apparaissent sous la forme : `<Jour de la semaine> <Jour> <Mois> <Année>` (cf. Figure 3.1). Quelques variations peuvent subvenir, notamment dans la présence ou non du `jour de la semaine`, ou de l'`année`. En outre, les dérivations lexicales du `mois` occasionnent de nombreux conflits. En effet, il n'est pas rare d'observer un mois écrit sous la forme « 9bre », « 10bre », *etc.* De fait, pour des participants non-experts, il est aisé de confondre septembre et novembre pour l'abréviation « 9bre » (la première option étant incorrecte).

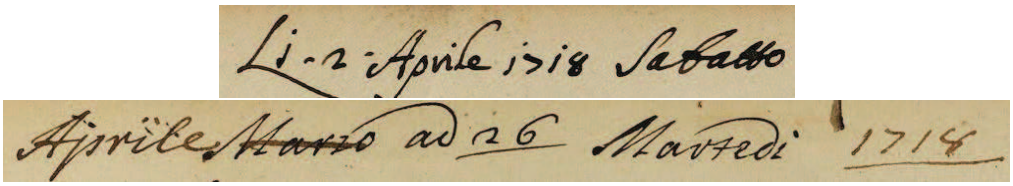

FIGURE 3.1. *Deux exemples de zones de transcription identifiées pour le type "Date du jour".*

L'heuristique choisie pour évaluer la résolution automatique des conflits concernant les dates est la suivante :

1. séparer les différents éléments constitutifs de la date du jour pour une transcription de type texte ;
2. évaluer la répartition des votes sur chacun des éléments constitutifs pour chaque proposition ;
3. pour chaque élément constitutif, garder la proposition majoritaire avec un indice de confiance de 80%
4. reconstituer la date complète à partir des résultats de l'étape 3.

Une vérification supplémentaire à partir d'un calendrier permet de s'assurer que le jour de la semaine, s'il est bien renseigné, correspond à la date du jour identifiée.

Cette heuristique a été adaptée pour les valeurs monétaires en exploitant le séparateur d'unités (« : ») pour extraire `<Livres>:<Sous>:<Deniers>`. Une vérification supplémentaire a également été intégrée sur base des règles de conversion de la monnaie de l'Ancien Régime : 1 livre = 20 sous et 1 sou = 12 deniers. Une analyse plus fine pour les valeurs monétaires met en évidence un conflit bien marqué sur la distinction entre les chiffres « 1 » et « 2 ». Mais il semble que le système de vote permette de résoudre rapidement ces cas particuliers. Le tableau 3.2 propose un exemple de propositions conflictuelles pour lesquelles la méthode proposée aboutit au consensus `12:34:56`.

Enfin, concernant les titres de pièces, une modélisation des transcriptions en N-Grams ($N = 3$) permet d'évaluer la similarité des propositions deux à deux et de les catégoriser par groupes. Si une proposition candidate cumule un nombre de votes, parmi toutes les propositions similaires, supérieur à 75% du nombre total de votes, elle est retenue. Le tableau 3.1 illustre un cas simple de transcriptions conflictuelles présentant une différence d'un mot.

Une analyse plus fine des conflits pour les titres de pièces indique qu'ils sont majoritairement issus de caractères mal transcrits (confusion entre deux lettres, ajout ou omission d'un caractère).

Le tableau 3.3 détaille la répartition du nombre de marques conflictuelles par rapport au nombre de marques ayant fait l'objet d'au moins deux transcriptions, et, pour ces marques conflictuelles, le pourcentage de cas résolus à partir de l'approche décrite précédemment.

Proposition	Nombre de votes
Gioco di Fortuna	2
Arlecchino gioco di fortuna	3

Tableau 3.1. *Exemple de transcriptions de titres de pièces conflictuelles.*

Proposition	Nombre de votes
12 :3 :56	2
22 :34 :56	2
12 :34 :	1

Tableau 3.2. *Exemple de propositions conflictuelles sur des valeurs monétaires.*

Type	*2+ trans.*	*Cons.*	*Conflits*	***Conflits résolus***
Dates	799	324	475	**382** (80%)
Titres	869	339	530	**352** (67%)
Val. monétaires	1739	1234	505	**327** (65%)

Tableau 3.3. *Pourcentage de conflits résolus par type de marque pour un indice de confiance de 80% (voir détail de la procédure de résolution de conflit). La colonne «2+ trans» correspond au nombre de marques ayant fait l'objet d'au moins 2 transcriptions, «Cons.» correspond au nombre de cas résolus soit à l'étape de transcription soit à l'étape de vérification (i.e après un consensus avec un vote à la majorité des 3/4).*

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. Parallèlement, un modèle conceptuel spécifique a été 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.

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 déchiffrage de graphies difficilement lisibles.

Les résultats obtenus sont plutôt encourageants sur les annotations induisant le plus de divergences. Par exemple, 80% 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 (André *et al.*, 2014) 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 (Schönböck *et al.*, 2016).

Références

André P., Kraut R. E. & Kittur A. (2014). 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, p. 139–148, New York, NY, USA : ACM.

Carletti L., McAuley D., Price D., Giannachi G. & Benford S. (2013). Digital humanities and crowdsourcing : An exploration. In *Proceedings of MW2013 : Museums and the Web 2013* : Museums and the web.

Causer T., Grint K., Sichani A.-M. & Terras M. (2018). Making such bargain : Transcribe bentham and the quality and cost-effectiveness of crowdsourced transcription. *Digital Scholarship in the Humanities*.

Chittilappilly A. I., Chen L. & Amer-Yahia S. (2016). A survey of general-purpose crowdsourcing techniques. *IEEE Trans. on Knowl. and Data Eng.*, 28(9), 2246–2266.

Köpcke H., Thor A. & Rahm E. (2010). Evaluation of entity resolution approaches on real-world match problems. *Proc. VLDB Endow.*, 3(1-2), 484–493.

Martin M., Justin T. & Valerie W. (2011). Manuscript transcription by crowdsourcing : Transcribe bentham. *LIBER Quarterly*, 20, 347–356.

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

Schönböck J., Raab M., Altmann J., Kapsammer E., Kusel A., Pröll B., Retschitzegger W. & Schwinger W. (2016). A survey on volunteer management systems. In *2016 49th Hawaii International Conference on System Sciences (HICSS)*, p. 767–776.

Warwick C., Warwick C., Terras M. & Nyhan J. (2012). *Digital Humanities in Practice*. Facet Publishing.

