Transcrire un corpus audio, méthodologie retenue, outils employés

Mounia Illourmane

Mounia Illourmane
Doctorante en Sciences du langage
33, boulevard du Port, 95 011, Cergy-Pontoise cedex
Université de Cergy-Pontoise
EA 7518 LT2D
mounia.illourmane@u-cergy.fr

Colloque jeunes chercheurs de l'École Doctorale CLI les 10 et 11 octobre 2019

Université Paris 8.

"Savoir des mémoires & Mémoire des savoirs"

Titre : *Transcrire un corpus audio, méthodologie retenue, outils employés*

Mots-clés : mémoire, conservation du patrimoine immatériel, transcription, corpus audio.

Introduction

« La mise en mots de la mémoire des chibanis du quartier de la Goutte d'Or à Paris », un sujet qui m'a été proposé par Marie-Madeleine Bertucci, ma directrice de thèse. Ce sujet s'inscrit dans le cadre de la recherche sur les contextes plurilingues de l'espace francophone conduite dans le laboratoire de rattachement. L'idée est de sauvegarder et de mettre en mots la mémoire de ces vieux migrants, en partant d'un corpus d'archives audio, de documents authentiques et d'enregistrements vocaux qui n'ont fait l'objet d'aucune manipulation au préalable.

I. Conservation de la mémoire des chibanis du quartier de la Goutte d'Or à Paris : Patrimoine mémoriel d'une minorité sociale.

La prise en compte de la mémoire de l'immigration en général et plus particulièrement de l'immigration maghrébine est primordiale pour la constitution d'une mémoire collective de la société française, en vue du rôle important de ces oubliés de l'immigration. Une telle démarche aura pour objectif l'étude de la mémoire de ces acteurs sociaux qui ont marqué leur temps et les générations ultérieures en laissant suffisamment d'éléments pertinents pour entreprendre des recherches bien plus approfondies et riches en données brutes et exploitables notamment avec certains éléments saillants du passé comme l'histoire de vie, l'identité et la culture.

La sauvegarde de la mémoire des immigrés est l'objet d'étude de plusieurs associations cultuelles et sociales notamment avec la création de La Cité Nationale de l'Histoire de l'Immigration (2007)[1]. La mémoire devient un vocable fourre-tout, recouvrant des éléments très hétérogènes : facultés, souvenirs, faits historiques, processus, représentations, patrimoine ». (Ribert 2011 : pp 59-78). L'objectif est de mettre en valeur les migrations antérieures et faire connaître leur mémoire. Ils ont vécu seuls et terminent leurs vieux jours dans une solitude bercée par les souvenirs nostalgiques d'un pays qui n'existe plus, qui ne ressemble pas à celui quitté plusieurs décennies plutôt. (Cherfi et al 2016 :15).

Cette étude prend appui sur un corpus d'archives audio constitué d'enregistrements biographiques des chibanis de « La Goutte d'Or » à Paris, réalisés par Jean-Marc Bombeau, un membre de la galerie associative « l'Echomusée », située dans le XVIIIe arrondissement de Paris avec un groupe de chibanis du quartier de 2010 à 2012. Avec la collaboration des membres de « l'Echomusée », Jean-Marc Bombeau a réalisé en 2010 des ateliers de rencontre avec les anciens de « La Goutte d'Or » dont l'objectif est de récolter leurs témoignages et leurs parcours afin de réaliser un film documentaire. Ces témoignages font émerger la mémoire des chibanis en les amenant à pousser leur réflexion sur leur passé, notamment leur arrivée en France et à Paris dans le cadre d'une immigration de travail. Avec ces entretiens l'objectif est de mettre en avant le regard rétrospectif des anciens sur le passé, ces données mémorielles constituent la base d'un patrimoine immatériel. Mémoire et souvenirs : deux termes que tout rapproche dans le vocabulaire commun où la mémoire n'est autre que la faculté de conserver et de rappeler des états de conscience passés (Lavabre 1994 : 15).

[1] www.unesco.org

À partir d'une démarche muséale et ethnographique, il s'agit de transcrire un corpus audio dans une perspective de conservation d'archive et de valorisation de données peu accessibles matériellement. *Quelle méthodologie pour un corpus audio ?* C'est la question à laquelle nous allons tenter d'apporter réponse.

II. Les données

Il s'agit de documents authentiques, séquences vidéo, bandes son et photos, peu accessibles matériellement. Ces données sont brutes et représentent une manne de ressources et de matériaux pour la constitution de notre corpus. La récupération de ces données auprès de Jean-Marc Bombeau a eu lieu dans les locaux de la galerie de « l'Echomusée » dans le XVIII^{ème} arrondissement de Paris.

Mise au point d'un outil de classement, d'indexation et d'arborescence

Les données sont sauvegardées sur un disque dur, toutefois, la matérialité du disque dur impacte la façon dont le classement est effectué et nous amène à modifier nos pratiques. Avec ma directrice de recherche, on a réalisé un pré-test sur un échantillon afin de trouver une méthodologie d'organisation des données reçues. Par la suite, j'ai conçu un outil de mise en ordre des données en utilisant une méthodologie d'inventaire qui consiste à établir une liste exhaustive des éléments figurants dans le disque dur, en élaborant un inventaire physique qui correspond à un comptage manuel de chaque type de document dont les démarches sont le regroupement, le classement, l'indexation et l'arborescence. La matérialité du disque dur impacte la façon dont le classement est effectué et nous amène à modifier nos pratiques.

Présentation des bandes son

Les bandes son sont transcrites en orthographe standard. La partie transcrite comporte 213 pages et repose sur des enregistrements d'archive audio, composé de 18 bandes son. La durée totale des enregistrements est de 6 heures et 58.32 minutes. La durée varie d'une bande son à une autre, la plus courte est de 03.58 minutes et la plus longue est de 1 heure et 2.04 minutes. Dans le corpus, on se situe dans une perspective narrative et d'écriture du réel on distingue au total 33 « quasi personnages » (Ricœur 1983), dont 04 intervieweurs et 29 chibanis. En effet, on se situe dans une perspective narrative et d'écriture du réel en l'occurrence de la mémoire qui est celle de Ricoeur.

Il s'agit d'entretiens biographiques, impulsion narrative sous-jacente aux entretiens qui fait apparaître chez les informateurs une impulsion à se représenter en se racontant (Bertucci, 2006). Les locuteurs mettent en mots des éléments surgissant de leur passé et de leur histoire de migrant. Ces témoignages contiennent des indices identitaires ainsi qu'une expérience de l'immigration. Le détour par le récit de vie a rendu possible, dans cette recherche, l'émergence d'un sujet qui ne pouvait être visible que par un détour par sa vie personnelle. (Bertucci, 2006).

Toutefois, une anonymisation totale du corpus s'avère nécessaire pour la protection des données personnelles des chibanis, elle reste une démarche à forts enjeux scientifiques, juridiques et éthiques.

III. Génération d'un corpus : méthodologie de transcription

Recherche du logiciel

Les logiciels de transcription permettent de gagner du temps et de faciliter la transcription. Après une comparaison de plusieurs logiciels de transcription de corpus, au final le choix s'est porté sur *Express Scribe* dans un premier temps ensuite, on a eu recours à un logiciel de reconnaissances vocales *E-Speaking.*

Prise en main des outils informatiques

Il s'agit d'une transcription mot à mot, c'est-à-dire de mettre par écrit exactement tout ce qui est dit à l'oral incluant des phrases incomplètes, des reformulations et des intonations, on y lira aussi les remarques d'ordre général et les mots répétés à l'oral, des arrêts fréquents sont nécessaires pour ajouter les éléments extralinguistiques comme les rires et les hésitations.

La transcription manuelle reste donc incontournable pour notre corpus. Ce type de transcription a principalement pour but la réalisation d'une retranscription authentique, tout en respectant les enchaînements et les chevauchements de l'enregistrement oral. C'est une technique très utile mais coûteuse en temps. Selon les estimations courantes, un minimum de trente minutes de travail est nécessaire pour transcrire une minute d'enregistrement. (Baude 2006 : 30).

1. Express Scribe

Express Scribe[2] est un logiciel professionnel de lecture audio qui permet de lire des fichiers audio au format Wav ou MP3. Ce programme ne permet pas la conversion automatique des fichiers audio en texte, mais il contient d'autres fonctionnalités pour rendre le processus plus facile. Telle que la transcription, un outil conçu pour faciliter la transcription des enregistrements audio au clavier dans un éditeur de texte. La fonction de contrôle permet de contrôler la vitesse de lecture, qui se gère entièrement à partir des touches d'accès rapides du clavier. On peut sauvegarder, lire ou revenir en arrière en utilisant les touches retour, avance rapide. Le logiciel dispose d'une option de chargement automatique des fichiers audio. Elle est accessible à partir d'un périphérique (lecteur cd ou dvd, clé USB…) ou d'un réseau local.

2. E-Speaking

Un logiciel[3] shareware de contrôle et de reconnaissance vocale, disponible en téléchargement. Il a pour fonction la dictée et l'exécution des tâches sur ordinateur par voix de l'utilisateur, Il assure de meilleurs résultats en s'intégrant à Word (Microsoft office), il contient également des commandes intégrées comme *Voice Dictation* pour la dictée, elle consiste à reformuler sous forme intelligible et grammaticalement correcte les propos de l'utilisateur.

Néanmoins dans le cas de notre corpus la seule utilisation possible est de répéter à voix haute devant un microphone les propos tenus par les chibanis pour que le logiciel les retranscrive en texte. Cependant, il est nécessaire de revoir et corriger les textes retranscrits, étant donné que les logiciels de reconnaissance vocale ne prennent pas en

[2] https://www.nchsoftware.com/fr/index.html
[3] https://e-speaking.com/

compte les différents accents des chibanis et ainsi reproduire fidèlement à l'écrit un discours oral.

Pourquoi l'emploi de ces deux outils ?

Les systèmes de transcription doivent être adaptés en fonction des corpus à transcrire. (Baude 2006 : 31). L'emploi de ces deux logiciels a permis la transcription des enregistrements des chibanis en restant le plus fidèlement possible à l'original tout en tenant compte des nombreuses contraintes d'ordre syntaxique, orthographique et phonologique résultant du caractère oral des données.

CONCLUSION

La transcription des entretiens présente des difficultés d'ordre pratique et technique. La première difficulté est liée au caractère oral des données d'un point de vue syntaxique, orthographique et phonologique. Les logiciels de transcription automatique et de reconnaissance vocale ne correspondent pas à notre corpus, il a donc fallu tirer profit des logiciels que nous avons choisis, cependant nous sommes obligées de recourir à un traitement majoritairement manuel, même si ces logiciels facilitent grandement la tâche de transcription des données. A cette difficulté s'ajoute la qualité des enregistrements, en effet, les bandes son ne sont pas réalisées dans un but scientifique, la prise de son reste donc artisanale.

Il s'agit de difficultés auxquelles tout chercheur peut être confronté dans ce genre de recherches, au travers d'un corpus d'archives orales, nous proposons une approche de traitement : mise en ordre des données, recherche des logiciels et transcription authentique du corpus.

Références

Baude, O. (dir) 2006. *Corpus oraux. Guide des bonnes pratiques*. Paris : CNRS Éditions.
Baude, O. *Corpus oraux : les « bonnes pratiques » des linguistes*. Laboratoire Ligérien de Linguistique UMR 7270. pp. 1-104.
Bertucci, M-M. 2012. *Le récit de vie, un processus réflexif à l'œuvre dans la production des savoirs*, Cahiers internationaux de sociolinguistique, vol. 2, no. 1,pp. 85-102.
Bertucci, M.-M. 2009- 2017. (J. Assier et E. Chemblette, coll. pour la transcription du corpus *Mémoires de l'immigration, vers un processus de patrimonialisation ?* Rapport de recherche universitaire pour le Ministère de la Culture et la Cité nationale de l'Histoire de l'immigration. Université de Cergy-Pontoise. 1 vol. (170 p.).
Cherfi, M., DjemaÏ, N., Labidi, M., Oujdi, M., Quemeneur, T., Sclavis, L. 2016. *Chibanis La question*. Au diable vauvert.
Faure, R., Lemaire, B., Picouleau, C. 2009. Précis de recherche opérationnelle : Méthodes et exercices d'application, Paris, Dunod, 6e éd.
LAVABRE, M-C. 1994. *Le fil rouge : sociologie de la mémoire communiste*. Paris : Presses de la Fondation Nationale des Sciences Politiques.
Matalon, B. 1978. *Les Enquêtes Sociologiques*. Paris : Armand Colin.
Noiriel, G. 2006. *Le creuset français. Histoire de l'immigration XIX^{e}-XX^{e} siècles*. Paris : Seuil.
Ribert, E. 2011. *Forme, supports et usages des mémoires des migrations. Mémoires glorieuses, douloureuses, tues*. pp 59-78.
Ricœur, P. 1983. *Temps et récit*. T.1. Paris : Seuil
https://www.nchsoftware.com/fr/index.html consulté le 05-05-2019.
https://e-speaking.com/ consulté le 06-05-2019.
www.unesco.org consulté le 29-04-2019.