

Analyse de la circulation des textes chantés par similarité multimodale : de la Fronde à la Révolution française

La chanson est un genre qui connaît aujourd’hui des usages politiques renouvelés, dans les médias traditionnels et sur les réseaux sociaux [DB22], notamment par les parodies ou les reprises d’airs connus. Parallèlement, l’exploitation des sources chantées a récemment été réhabilitée pour les recherches historiques et littéraires [Gui08, AR22].

Dans une société largement analphabète comme l’est la France d’Ancien Régime, la chanson est un véritable média : le procédé qui consistait à adapter de nouvelles paroles à une mélodie célèbre est un moyen mnémotechnique efficace pour diffuser des messages (politiques). Les mélodies sont ainsi des canaux de circulation de savoir, de rumeurs, de nouvelles (vraies ou fausses), voire d’opinions, à partir de feuilles volantes relayées par des performances orales. Des moments comme la Fronde ou la Révolution française voient ce phénomène, déjà connu pendant la Ligue, prendre une ampleur considérable, d’une part à la faveur du développement de l’imprimerie et du colportage, d’autre part parce que les pouvoirs (en place ou contestataires) avaient compris l’efficacité de ce média. D’innombrables couplets sont ainsi produits sur la même mélodie (des centaines sur « l’air de la Fronde », des milliers sur l’air de la Marseillaise, par exemple). La chanson satirique peut aussi être répandue par tel ou tel homme politique pour abattre son ennemi du moment en diffusant une image négative de celui-ci : la chanson véhicule ainsi des éléments de langage, et dispose déjà de qualités virales, comparable aux contenus du web aujourd’hui [Abi23]. Ce volet du doctorat permettra de penser des cinétiques de l’information qui ne datent pas, contrairement à ce que l’on croit parfois, de l’ère numérique contemporaine.

Pourtant, cette tradition textuelle et musicale est très mal connue en France à ce jour, en partie parce qu’il est difficile d’accéder à des données regroupées en un corpus combinant taille et homogénéité. Une large gamme de traitements numériques permettrait de rendre disponible un tel corpus, mais aussi de l’exploiter : rendre requêtables les catalogues existants sur le sujet [Pie04], extraire, depuis des bibliothèques numériques, les sources déjà disponibles, enrichir les données et uniformiser les méta-données (voir par exemple le travail mené sur le projet ANTONOMAZ [AL19, ABLT22]). Les techniques, aujourd’hui bien développées, de reconnaissance optique de caractères (OCR) permettent d’obtenir de manière économique de grandes quantités de données, rapidement exploitables. Les textes, les airs, seraient ainsi utilisables au plan statistique : quels airs sont les plus utilisés par tel camp politique, quelles chansons sont le plus passées à la postérité et par quelles manipulations politiques les a-t-on fait servir à des buts différents ?

Outre la constitution de cette ressource, l’outil numérique servira à gérer les variations (textuelle, linguistique et musicale), cruciales pour appréhender ces textes. La chanson est en effet une matière discursive fluide : les documents peuvent compiler des centaines de couplets, ou citer seulement les premiers mots de la chanson, le nom de l’air ... Cette géométrie variable dans la collecte de données bénéficierait de techniques plus avancées de TAL (Traitement Automatique des Langues). En particulier, il faudrait pouvoir améliorer la qualité des paroles extraites en développant des modèles d’apprentissage automatique (par exemple, par réseaux de neurones) pour extraire les paroles des textes bruités, les enrichir tout en questionnant la qualité intrinsèque des données [BL20].

Ces problématiques de constitution de corpus se conjuguent avec la question de la recherche de similarités, puisque le principe de cette communication chantée est la réutilisation du même sur fond de variation. Les similarités (et variétés) sont sémiotiquement complexes, relevant de modes (ou canaux) différents : reprises à l’identique de la musique et parodie du texte ou identité du matériau textuel qui permet de connaître la musique programmée. S’y ajoutent diverses variations linguistiques (registres, niveaux de langue) ou matérielles (brochures ou tracts d’actualité, objets livres comme les recueils). Ces dernières recèlent

des traces linguistiques de performances orales comme les vers hypermétriques ou les variations orthotypographiques. La recherche de similarité doit donc fondamentalement être multimodale et nécessite d’encoder des informations purement textuelles, enrichies par de l’étiquetage morpho-syntaxique et de la reconnaissance d’entités nommées, mais aussi des *features* spécifiques à ce genre textuel particulier. Il s’agit notamment de la métrique et du timbre (ou air connu). La capacité de requêter ces corpus en mêlant ces différentes dimensions soulève des questions épistémologiques qui intéressent à la fois les humanités et l’informatique.

La chanson est un média politique qui a une assise sociale extrêmement large, ce qui en fait une donnée exceptionnelle pour la connaissance des cultures politiques du passé. De ce point de vue, les chansons écrites qu’il nous reste des périodes anciennes sont parmi les seules traces qui nous restent de pratiques discursives de peu ou non-lettrés (même si les lettrés et l’élite les pratiquaient tout aussi bien). Une meilleure description des modalités de leur circulation enrichira la connaissance de l’histoire culturelle et dans le même temps les techniques de similarité composite développées bénéficieront au domaine du TAL. Les grands objectifs de cette thèse peuvent être résumés de la manière suivante :

- Établir des liens entre des données textuelles éventuellement entachées d’erreurs de transcription afin d’identifier des séquences répétées avec et sans trous
- Développer un phonétiseur/syllabeur Open Source pour la langue française, en proposer une adaptation pour la période considérée
- Exploiter la mise en forme matérielle pour requêter selon les formes prises par les documents, repérer celles qui sont destinées à l’oralisation ou à la conservation
- Croiser les similarités sur les séquences de graphèmes, de phonèmes, de syllabes avec celles disponibles sur le timbre (ou mélodie connue).
- Permettre des angles d’attaques variés du corpus selon la pondération des différents observables servant au calcul de similarité

Références

- [Abi23] Karine Abiven. Viralité des mazarinades chantées et écrites : tubes et/ou éléments de langage? *Tubes, teasing et viralité : Grand Siècle et pop’culture, Les recettes du succès. Stéréotypes compositionnels et littérarité au XVIIe siècle*, 2023.
- [ABLT22] Karine Abiven, Alexandre Bartz, Gaël Lejeune, and Jean-Baptiste Tanguy. Vers une collection numérique des libelles parus pendant la fronde, ou comment relier des mazarinades. *Le Verger*, 1(23), mai 2022.
- [AL19] Karine Abiven and Gaël Lejeune. Analyse automatique de documents anciens : tirer parti d’un corpus incomplet, hétérogène et bruité. *Recherche d’information, document et web sémantique*, 2(Numéro 1), 2019.
- [AR22] Luce Albert and Mickaël Ribreau. *Polémiques en chanson : IVe-XVIIe siècles*. Interférences. Presses Universitaires de Rennes, Rennes, 2022.
- [BL20] Adrien Barbaresi and Gaël Lejeune. Que recèlent les données textuelles issues du web? In *Conférence TALN*, pages 19–28. ATALA ; AFCP, 2020.
- [DB22] Tatiana Debbagi Baranova. Chanter au temps du coronavirus en france et ailleurs. ... *Écrire l’histoire - Histoire, Littérature, Esthétique*, (22) :121–126, September 2022.
- [Gui08] Eva) Guillorel. *La complainte et la plainte : chansons de tradition orale et archives criminelles : deux regards croisés sur la Bretagne d’Ancien Régime (16e-18e siècles)*. Université européenne de Bretagne, Rennes., 2008.
- [Pie04] Constant Pierre. *Les Hymnes et chansons de la Révolution : aperçu général et catalogue, avec notices historiques, analytiques et bibliographiques*. Imprimerie Nationale, 1904.