Examen final de GIT | OCR Corse

Vincent Sarbach-Pulicani, Violette Saïag, Théophile Miaille, Adrien Escoda

27 janvier 2022

1 Enjeux historiographiques

Avec l'émergence des nationalismes du XIX^e^ siècle se greffent conjointement des mouvements régionalistes d'affirmations et de revendications de particularismes culturels. La Corse s'insère très bien dans cette dynamique et se présente même comme un lieu propice au développement de telles idées. La centralisation de l'État autour d'une capitale forte et les politiques d'assimilation des populations indigènes à la frontière de la France ont poussé certains acteurs à défendre ces particularismes. Tradition jacobine de la « République une et indivisible », parfois largement exagérée, c'est notamment à partir du Second Empire et du règne de Napoléon III que la francisation de la Corse prend tout son sens. Cela s'est traduit par l'apprentissage du français à l'école en lieu et place du corse, des plans de relance économique et industrielle du territoire ou encore par la participation massive des Corses à l'effort colonial. Dès lors naissent les premières brigues d'une lutte régionaliste sur l'île c'est-à-dire par la défense de la langue corse, centrale dans la préservation des identités. C'est le cas de Santu Casanova, poète et rédacteur en chef de la revue *A Tramuntana* (« la Tramontane ») dont il est également le fondateur en 1896. L'un des aboutissements des nationalismes du XIX^e^ siècle est bien connu, c'est la Première Guerre mondiale. Le désastre que cet événement engendre se ressent beaucoup sur l'île que ce soit en termes démographiques ou sociétaux. Encore aujourd'hui le nombre d'insulaires morts lors de ce qu'ils appellent le *scumpientu*, la « catastrophe », est difficile à chiffrer. Par ailleurs, cela a donné lieu à de nombreux débats idéologiques autour du sacrifice des Corses pour la « Grande Patrie » ou au nom d'une guerre qui ne les concernait pas.

Le premier ouvrage s'intitule *Pontenôvu* a été écrit par Petru Rocca et publié par la « Stamparia di a Muvra » en 1927. Il s'agit d'un recueil de poèmes en corse et en français dont les thèmes varient. *A Muvra* est un journal autonomiste corse d'influence maurrassienne qui a existé pendant toute la période de l'entre-deux-guerres. Se revendiquant comme une revue culturelle, la dimension politique de la revue (incarnée par le PCA, ou *Partitu corsu d'azione*), en a fait un mouvement controversé. C'est dans le contexte de lutte politique et d'éveil culturel corse, que nous avons abordé précédemment, que s'inscrit ce recueil. Le second ouvrage s'intitule *A nostra Santa Fede - Catechismu Corsu*, écrit par Ageniu Grimaldi en 1926 sous le pseudonyme de

Saveriu Malaspina. Proche de Petru Rocca, ce dernier est l'un des théoriciens de l'autonomisme corse de l'entre-deux-guerres et fidèle muvriste. Dans l'ouvrage, il est fait mention notamment de la façon dont un vrai corse doit se comporter vis-à-vis de sa foi envers Dieu et son île. Bien qu'il ne s'agisse pas réellement d'un recueil de poèmes, le style d'écriture de cet ouvrage est particulièrement intéressant, reprenant un style se rapprochant des écrits bibliques. Alors que le premier ouvrage a été trouvé sur `gallica`, le second a été téléchargé depuis la plateforme `bucullezzione` de l'Università di Corsica Pasquale Paoli. Il n'est donc disponible qu'en format PDF à l'inverse de *Pontenôvu* disponible en IIIF.

Il existe en Corse une volonté de structurer et d'étudier les évolutions de l'emploi de la langue corse depuis les années 1980. On peut notamment mentionner les travaux de la linguiste Marie-José Dalbera-Stefanaggi avec son *Nouvel atlas linguistique et ethnographique de la Corse* dont le premier volume paraît en 1995 [1]. Dans les rééditions de cette œuvre majeure dans les années 2000, l'autrice incorpore notamment ses travaux sur la création d'un Banque de Données Langue Corse (BDLC). Il s'agit là de la première initiative dans la volonté de lemmatiser dans l'espace et le temps la langue corse. L'apport des humanités numériques dans cette problématique est assez neuf. Les réflexions des chercheurs sur l'outillage des langues régionales en utilisant le TALN (Traitement Automatique du langage Naturel) se sont vraiment accélérées à partir de la deuxième moitié de la décennie 2010. Notre démarche s'inscrit donc dans cette continuité. Face au manque d'outils pour traiter et analyser la langue corse [2], océriser des données textuelles afin qu'elles puissent être exploitées dans des recherches en humanités numériques représente un enjeux de taille dans l'avenir de la discipline en Corse.

2 Exploitation du jeu de données

La transcription de ce corpus, si elle fut longue et mouvementée du fait de nos nombreux tâtonnements initiaux, n'a pas été particulièrement problématique. Les principaux soucis rencontrés étaient liés à la richesse de la forme du corpus. Par exemple, les changements de police étaient parfois difficiles à intégrer pour notre modèle. Les mots en italique étaient souvent mal lus, notamment les lettres « u » et « v » qui sont légèrement arrondies lorsqu'elles sont écrites en italique : le modèle les lisait tantôt comme un « u », tantôt comme un « o ». De même, les lettres « z » et « g » étaient souvent confondues. Aussi, l'enjeu d'étudier un texte corse est aussi de faire face à certains accents inhabituels pour un modèle habitué au français ou à l'anglais, tels que les « ì ». Aussi, les tampons rouges en bas de certaines pages cachaient parfois les quelques dernières lignes.

1. Dalbera-Stefanaggi Marie-José, *Nouvel Atlas Linguistique et Ethnographique de la Corse*, Paris, éditions du CNRS, 1995.
2. Kevers Laurent, Gueniot Florian, Tognotti A. Ghjacumina, Retali-Medori Stella, « Outiller une langue peu dotée grâce au TALN : l'exemple du corse et BDLC », *26e Conférence sur le Traitement Automatique des Langues Naturelles*, Toulouse, 2019, p. 371 à 380.

Ainsi, après l'échec de quelques essais avec le « modèle imprimé 16e-18e Fra+Lat » et le « Modèle Manuscrit 19e Lectaurep », nous avons été satisfaits par les transcriptions produites par le modèle « 19th century prints - HTRcatalogs Artlas » que nous avons trouvé assez tardivement, et dont la précision est estimée à 98,7%. Les corrections à effectuer après l'utilisation de ce modèle étaient bien plus légères que celles que nécessitaient les modèles antérieurs.

Ce succès du modèle est en partie dû à l'état relativement correct de notre corpus. En effet, le texte ne comporte pas assez de bruit pour être inintelligible. L'impression est de qualité amplement satisfaisante pour l'époque, bien qu'elle présente tout de même quelques mots légèrement estompés et quelques bavures difficilement lisibles par le modèle. Cependant, une correction manuelle effectuée après chaque transcription automatique a suffi à pallier ces quelques confusions du modèle. La segmentation automatique sur `eScriptorium` se faisait aisément et ne nécessitait que de légères corrections de notre part pour ajouter par exemple un numéro de page oublié, ou encore effacer la ligne tracée automatiquement sous une tache d'encre prise pour un caractère, à l'exception d'un souci rencontré fréquemment : pour quelques pages, le texte était divisé en deux colonnes que le logiciel ne distinguait pas spontanément, il nous fallait donc rectifier cela manuellement.

Enfin, le fait de travailler en groupe a représenté un avantage crucial qui nous a permis de produire une transcription fidèle de manière efficace et relativement peu chronophage. En effet, afin d'optimiser la productivité de notre travail, nous avons réparti les tâches entre les quatre membres du groupe par nos discussions sur `GitHub`. En premier lieu, Théophile s'est chargé de trouver et de tester des modèles sur le corpus d'entraînement afin de déterminer lequel était le plus performant ; ensuite, Violette s'est chargée de la segmentation et de la transcription des textes (quelques premières fois avec les modèles qui furent ensuite jugés obsolètes, ces transcriptions ne furent donc pas gardées, puis entièrement manuellement en l'attente d'un meilleur modèle, et enfin la version finale avec le modèle jugé plus efficace) ; dans la mesure où seul Vincent parle Corse dans le groupe, il fut préposé à la relecture des pages écrites en corse et à la transcription de certaines d'entre elles, ce qui nous fut très bénéfique dans la mesure où il était bien plus difficile d'élucider nos doutes sur quelques lettres ambiguës sans reconnaître les mots, tandis qu'Adrien s'est chargé de relire méticuleusement celles qui étaient rédigées en français et a ensuite effectué la relecture et la correction finales et détaillées de l'intégralité du corpus avant son importation, tant du point de vue de la transcription que de certaines segmentations qui restaient imprécises.

Cette répartition du travail au sein du groupe s'est dessinée naturellement au fil du projet sans concertation préalable. Si nous avons tous travaillé sur tous les différents pôles du projet, tant sur `GitHub` que sur `eScriptorium`, certains rôles se sont tout de même définis. Vincent a trouvé et proposé le corpus sur lequel nous avons travaillé, car il s'agit d'une partie du corpus qu'il compte utiliser pour son mémoire de cette année. Le reste du groupe s'est ensuite fami-

liarisé avec le corpus et a été très intéressé par ce thème, et a donc approuvé ce choix. Il s'est donc spontanément chargé de la mise en place du dépôt. Théophile, pour qui `GitHub` était une plateforme familière, a grandement participé à la mise en place du dépôt et à l'import de la documentation nécessaire pour qu'un œil extérieur puisse plus aisément prendre connaissance du corpus en ajoutant les versions PDF et IIIF des documents. Adrien et Violette ont quant à eux pris en charge la transcription du texte et la correction de celle-ci en communiquant leur avancée par de nombreuses issues sur `GitHub`.

3 Organisation du dépôt

Le dépôt est composé dans la section « Code » de différents dossiers : le fichier `README.md` qui présente le projet et ses participants ; le dossier « Ressources » qui rassemble les PDF des pages de l'ouvrage *Pontenôvu* (qui est également en IIIF) et *A nostra Santa Fede - Catechismu Corsu* à transcrire, chacun dans un dossier différent ; ainsi que le dossier « Transcriptions » qui est quant à lui composé des transcriptions réalisées (pages 4 à 20 pour le *Pontenôvu*, soit 16 pages ; pages 9 à 38 pour *A nostra Santa Fede - Catechismu Corsu*, 30 pages, soit 46 pages en tout). Chacun des ces dossiers contient les transcriptions au format TXT, ainsi qu'un sous-dossier contenant les transcriptions au format XML/ALTO agrémentées des images des pages numérisées (au format PNG). Chaque dossier est également composé d'un fichier `README.md` qui explique ce qu'il contient.

A l'origine, seul l'ouvrage *Pontenôvu* avait été sélectionné pour la transcription, mais étant donné qu'il ne contenait pas assez de pages, il a été décidé d'ajouter un second ouvrage, *A nostra Santa Fede - Catechismu Corsu*, ce qui a eu pour effet d'ajouter deux sous-dossiers dans le dossier « Transcriptions », dont chacun d'eux correspond à un ouvrage, comme décrit ci-dessus. Une traduction avait initialement été envisagée, d'où l'existence d'un dossier « Traduction » dans l'historique, mais l'idée a finalement été abandonnée car cela ne représentait pas d'utilité particulière pour notre travail (bien qu'elle n'eût cependant pas été dénuée d'intérêt).

Les issues soumises par les différents membres du groupe nous ont permis de mieux identifier les problèmes rencontrés. On peut distinguer trois grands types d'issues : celles relatives aux problèmes de transcriptions (concernant donc principalement Adrien et Violette) ; celles signalant des fin de tâches, à savoir la fin d'une transcription ou de sa vérification pour Violette, Vincent et Adrien, ou celles concernant les réorganisations du dépôt réalisées par Vincent ou Théophile ; et celles relatives à des propositions émises par les membres du groupe. Bien que la plupart des labels de base furent suffisants pour décrire nos différentes issues, nous en avons ajouté quelques uns plus spécifiques, tels que « bonne idée » ou « suggestion », ainsi que « bad transcription », bien que ce label n'ait finalement pas été utilisé. Notre travail a naturellement été émaillé au début du projet de nombreux tâtonnements relatifs à la prise en main progressive

d'`eScriptorium` ou de `GitHub`, sur lesquels nous aurions pu communiquer plus explicitement. Cependant, les problèmes de chevauchement de transcriptions récurrents ont rapidement entraîné des interactions plus importantes et plus suivies, permettant ainsi l'amélioration de notre rapport à ces logiciels.

4 Conclusion

Après un début marqué par des tâtonnements incertains, l'ensemble du groupe a pu peu à peu prendre en main l'outil pour améliorer le travail collaboratif autour de ces transcriptions. L'ajout de nouveaux dossiers et les différentes réorganisations du dépôt ont permis de mieux organiser notre travail, ce qui nous a menés à envisager petit à petit des pistes de perfectionnement du projet. Le choix que nous avons fait de de mettre à disposition sur le repository à la fois les images, le texte en XML/ALTO et le texte brut permet aux utilisateurs de se réapproprier ce corpus transcrit et de pouvoir l'utiliser de manière plus libre.

D'un point de vue purement historiographique, ces transcriptions sont intéressantes car il y existe plusieurs types de textes différents : certains textes en français, d'autres en corse, sous forme de poésie ou de récit. Effectuer une analyse textométrique de ces données donnerait déjà un aperçu intéressant des termes employés et des sujets abordés en fonction de la langue du texte. Le modèle « 19th century prints - HTRcatalogs Artlas » est particulièrement satisfaisant mais il serait intéressant de comparer ces résultats avec ceux d'un modèle entraîné spécifiquement pour ce projet. Les opportunités à envisager sont donc intéressantes, et, en utilisant la lemmatisation effectuée par la *Banque de Données Langue Corse* (BDLC), nous pourrions donc effectuer des analyses encore plus approfondies en étudiant les évolutions linguistiques dans la presse et la littérature corse aux XIX^{e} et XX^{e} siècles, par exemple en s'intéressant plus spécifiquement aux variations grammaticales et orthographiques du langage.

5 Bibliographie

5.1 Sciences humaines

Dalbera-Stefanaggi Marie-José, *Nouvel Atlas Linguistique et Ethnographique de la Corse*, Paris, éditions du CNRS, 1995.

Delporte Christian, Blandin Claire et Robinet François, *Histoire de la presse en France,* XX^{e} - XXI^{e} *siècles*, Paris, Armand Colin, 2016 (coll. « Collection U »).

Paci Deborah, *Il mito del Risorgimento mediterraneo : Corsica e Malta tra politica e cultura nelventennio fascista*, Thèse de doctorat en Histoire contemporaine, Nice, Université de Nice-Sophia- Antipolis (cotutelle avec l'université de Padoue), sous la direction de PELLEGRINETTI

Jean-Paul, 2013.

Pellegrinetti Jean-Paul, *La Corse et la République : la vie politique de la fin du second Empire au début du* xxi^{e} *siècle*, Paris, Seuil, 2004. (coll. « xx^{e} siècle »).

Rogé Ysée, *Le corsisme et l'irrédentisme 1920-1946 : histoire du premier mouvement autonomiste corse et de sa compromission par l'Italie fasciste*, Thèse de doctorat en Histoire contemporaine, Paris, Université Paris 10 Nanterre, sous la direction de MUSIELDAK Didier, 2008.

Sarbach-Pulicani Vincent, *La presse corsiste et irrédentiste des années 1930 : étude comparative et quantitative des revues « A Muvra » et « Corsica antica e moderna » entre 1932 et 1939*, Mémoire d'histoire contemporaine, Université de Strasbourg, sous la direction de BOURGUINAT Nicolas, 2021.

5.2 Humanités numériques

Dalbera-Stefanaggi M.-J. et Retali-Medori S., « Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse. », *Actes du colloque Tribune des chercheurs, études en linguistique*, Bastia, Société des Sciences Historiques et Naturelles de la Corse, 2015, p. 17 à 25 (coll. « Corse d'hier et de main – Nouvelle série »).

Kevers Laurent, Gueniot Florian, Tognotti A. Ghjacumina, Retali-Medori Stella, « Outiller une langue peu dotée grâce au TALN : l'exemple du corse et BDLC », *26e Conférence sur le Traitement Automatique des Langues Naturelles*, Toulouse, 2019, p. 371 à 380.

Leixa J., Mapelli V. et Choukri K., *Inventaire des ressources linguistiques des langues de France*, ELDA, 2014. Accessible à cette adresse : `http://www.elda.org/media/filer_public/2014/12/17/rapport_dglflf_05112014-1.pdf`.

Vergez-Couret M., Bernhard D., Ligozat A.-L., Eloy J.-M. et Rey C., *TALaRE 2015 - Traitement Automatique des Langues Régionales de France et d'Europe. Atelier de TALN 2015*, Caen, 2015.