UNIVERSITÉ PARIS, SCIENCES & LETTRES
ÉCOLE NATIONALE DES CHARTES

Vincent Sarbach-Pulicani
Diplômé de licence d'Histoire
Diplômé de master d'Histoire contemporaine

I sumeri ùn leghjenu micca A Muvra
Lecture distante sur un corpus trilingue

Langue et rubrication dans la presse autonomiste corse de l'entre-deux-guerres

Mémoire de première année du master
« Humanités Numériques »

2022

Résumé

L'emploi de la langue corse dans la presse dialectale insulaire est un enjeu de taille pour les régionalistes de tous bords. Que ce soit à la fin du XIX^e^ siècle ou dans les revues culturelles de notre temps, la préservation du patrimoine linguistique corse est essentielle dans la défense de l'identité insulaire, ce que beaucoup considèrent comme le ferment de l'âme corse. Bien que proche des idéologies fascistes, les activistes de l'entre-deux-guerres réunis autour de Petru Rocca et sa revue *A Muvra* ont posé les bases des revendications de la lutte nationaliste. L'objectif de notre étude est d'identifier le rôle des langues dans une telle publication : en employant des méthodes de *topic modeling*, nous chercherons à déterminer si l'usage d'une langue est propre aux thématiques abordées ou si, à l'inverse, la langue influence directement les thématiques des muvristes.

Mots-clés : études corses, études italiennes ; linguistique computationnelle ; histoire de la presse ; humanités numériques ; topic modeling ; traitement automatique du langage naturel ; languee peu dotées ; plurilinguisme ; histoire culturelle

Informations bibliographiques : Vincent Sarbach-Pulicani, *« I sumeri ùn leghjenu micca A Muvra ». Lecture distante sur un corpus trilingue : langue et rubrication dans la presse autonomiste corse de l'entre-deux-guerres*, mémoire de master 1 « Humanités Numériques », dir. [Jean-Baptiste Camps, Christophe Gauthier], Université Paris, Sciences & Lettres, 2021.

Abstract

The use of the Corsican language in the island's dialectal press is a major issue for regionalists on all sides. The preservation of the Corsican linguistic heritage is essential in the defence of the island's identity, which many consider as the ferment of the Corsican soul. Although close to fascist ideologies, the activists of the inter-war period gathered around Petru Rocca and his review *A Muvra* laid the foundations of the demands of the nationalist struggle. The aim of our study is to identify the role of languages in such a publication : using topic modeling methods, we will try to determine whether the use of a language is specific to the subjects addressed or, conversely, whether the language directly influences the themes.

Keywords : corsican studies ; italian studies ; textometry ; digital humanities ; history of journalism ; natural language processing ; topic modeling ; low-resource languages

Bibliographic Information : Vincent Sarbach-Pulicani, *« I sumeri ùn leghjenu micca A Muvra ». Lecture distante sur un corpus trilingue : langue et rubrication dans la presse autonomiste corse de l'entre-deux-guerres*, M.A. thesis « Digital Humanities », dir. [Jean-Baptiste Camps, Christophe Gauthier], Université Paris, Sciences & Lettres, 2021.

Remerciements

J'aimerais en premier lieu remercier mes directeurs de recherches, Jean-Baptiste Camps et Christophe Gauthier, pour leurs précieux conseils au cours de cette année particulièrement riche.

*

J'aimerais également remercier le directeur du master (par intérim) Chahan Vidal-Gorène pour sa disponibilité durant ces dernières semaines.

*

Je n'oublie pas les personnes m'ayant apporté de l'aide dans la rédaction de ce mémoire ainsi que mes camarades de promotion, notamment Malamatenia Vlachou qui m'a beaucoup appris sur la linguistique et la philologie, domaine que je ne maîtrisais pas beaucoup avant d'arriver à l'École des chartes.

*

Infine, ringraziu di core tutti i studienti corsi chì m'anu aiutatu : pensu à Forteleone Arrighi, Ghjuvan' Santu Olivieri Battestini, Marc'Antone Faure Colonna d'Istria è Lisa Pupponi. (A squadra Tempi).

Introduction

0.1 L'émergence d'une lutte régionaliste et identitaire

0.1.1 Les prémices d'une affirmation culturelle

I sumeri ùn leghjenu micca A Muvra ou « les ânes ne lisent pas le mouflon », cette phrase que l'on trouve notamment sur les bulletins d'abonnements définit bien la ligne éditoriale de la revue autonomiste de l'entre-deux-guerres. Mais d'où vient-elle et quelle est son histoire ? 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 sociaux. 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.

0.1.2 La naissance du muvrisme

Cette défiance grandissante vis-à-vis du gouvernement français s'incarne par la revue *A Muvra*, fondée en 1920 par l'ancien combattant Petru Rocca lui-même assisté de son frère Matteu. Il s'agit d'un hebdomadaire autonomiste corse d'influence maurrassienne qui a perduré tout au long la période de l'entre-deux-guerres. Se revendiquant comme une revue culturelle, la dimension politique de cette dernière (incarnée par le PCA, ou *Partitu corsu d'azione*), en a fait un mouvement controversé. En parallèle des revues, la *Stamparia di A Muvra* (« Imprimerie d'*A Muvra* ») publia tout au long des années de l'entre-deux-guerres ouvrages et almanachs. Écrites en langues française, corse et italienne, cette série de publications s'intègre parfaitement dans l'héritage spirituel de Santu Casanova, dont les auteurs se revendiquent clairement. Si dans les premières années d'existence de cet

hebdomadaire les autorités françaises n'y prêtaient guère attention, la radicalisation des propos dans les années 1930 poussèrent le Gouvernement à censurer le journal en 1939. Les muvristes ont, dans les premières années, davantage des revendications à tendance régionaliste mais une succession de facteurs externes ou internes à la Corse poussèrent ces derniers vers le séparatisme. La similarité de certains arguments ainsi que la promiscuité évidente de certains muvristes avec les autorités italiennes augmentèrent la méfiance des commissaires spéciaux [1] à l'égard des corsistes. En effet, le régime fasciste a rapidement fait de la Corse un objet de convoitise tant stratégique qu'idéologique. S'inscrivant dans une longue tradition doctrinale datant du XIX^e siècle, ceux qu'on appelle les « irrédentistes » [2] ont largement appuyé leur propagande sur les mouvements autonomistes internes à la société insulaire.

0.1.3 Le rôle central de la langue

De fait, la langue devient rapidement un argument de poids dans l'idéologie corsiste. Employer la langue corse dans la presse écrite devient alors essentielle pour la conserver d'une part, mais aussi pour changer les mentalités dans la vie de tous les jours d'autre part. Et ce constat est partagé par tout un ensemble de presses locales que ce soient les journaux autonomistes comme *A Muvra* ou *A Baretta Misgia*, mais aussi la presse régionaliste unioniste comme *L'Annu Corsu*[3]. Ainsi, l'apparition d'*A Muvra* dans le paysage médiatique corse n'est pas une exception : la prise de conscience de la nécessité d'écrire le corse pour le faire vivre est plus générale et est partagée par un ensemble de groupes aux opinions politiques très variées, du moins à leurs débuts. La défense du patrimoine linguistique insulaire passe par diverses revendications comme la reconnaissance de son existence mais aussi son enseignement à l'école et à l'université de Corse [4].

0.2 État de l'art

0.2.1 Un sujet d'étude particulier ?

Les études autour de l'autonomisme corse sont relativement nombreuses et de qualité. Le premier travail académique sur la question est un mémoire de maîtrise de 1974

1. Agents rattachés au ministères de l'intérieur chargés notamment de rapporter aux préfets les activités suspectes du département.

2. Mouvement culturel et politique italien ayant pour doctrine l'annexion de tous les territoires considérés comme italien. Voir : Déborah Paci, « Le mare nostrum fasciste : l'espace politique et culturel en Corse et à Malte à l'époque du fascisme italien », *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines*, 128–2 (2016).

3. Revue fondée par Paul Arrighi et organe principale du cyrnéisme, mouvement régionaliste corse.

4. L'université de Corse est fondée une première fois en 1765 par le *babbu di a Patria* Pasquale Paoli. Elle est cependant dissolue en 1769 lors de l'intégration de la petite île au Royaume de France et il faut attendre 1781 pour que l'université soit rouverte sous l'appellation d'université de Corse-Pascal-Paoli.

réalisé par Daniel Polacci à l'université d'Aix-Marseille [5]. Ce-dernier a fait à une étude statistique centrée sur les muvristes et les procédés de production du périodique. Mais c'est véritablement à partir des années 2000 que fleurissent des travaux plus concrets. Nous pouvons citer différentes thèses de qualité : celle d'Ysée Rogé [6] qui est le premier travail à s'atteler à mettre en perspective corsisme et irrédentisme, celle de Deborah Paci [7] qui compare l'irrédentisme corse avec le maltais et celle d'Ange-Toussaint Pietrera [8] qui étudie le mouvement dans le contexte de l'émergence du nationalisme.

La floraison de thèses et de mémoires sur ce sujet est accompagnée de publications de chercheurs confirmés comme Jean-Paul Pellegrinetti et sa synthèse sur le relations entre la Corse et la République [9]. Mais une telle bibliographie est également teintée, souvent, d'une forte connotation politique a replacer dans le contexte de cette période [10]. Par exemple, dans l'ouvrage de Pellegrinetti cité précédemment, le préfacier Maurice Agulhon précise ceci :

> « Tout le monde n'aimera pas ce livre, républicain, donc pro-français, et certains lui reprocheront sans doute — en polémique, on fait flèche de tout bois — de faire cautionner cette histoire de Corses « républicains » par un préfacier continental. » [11]

L'étude de l'histoire récente de la Corse est donc complexe d'où la nécessité de réactualiser sans cesse l'historiographie.

0.2.2 La Corse et les humanités numériques

Depuis plusieurs années, il existe en Corse une volonté de structurer et d'étudier les évolutions de l'emploi de la langue corse. 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 [12]. Dans les rééditions de cette œuvre majeure dans les années 2000, l'autrice incorpore ses travaux sur la création

5. Daniel Polacci, *Les autonomistes corses de l'entre-deux-guerres*, mém. de mast., Université d'Aix-Marseille, 1974.

6. Ysée Rogé, *Le corsisme et l'irrédentisme 1920-1946 : histoire du premier mouvement autonomiste corse et de sa compromission par l'Italie fasciste*, thèse de doct., Paris 10, 2008.

7. D. Paci, *Il mito del Risorgimento mediterraneo : Corsica e Malta tra politica e cultura nel ventennio fascista*, thèse de doct., Université de Nice Sophia-Antipolis, 2013.

8. Ange-Toussaint Pietrera, *Imaginaires nationaux et mythes fondateurs ; la construction des multiples socles identitaires de la Corse française à la geste nationaliste*, thèse de doct., Université Pascal Paoli, 2015.

9. Jean-Paul Pellegrinetti et Ange Rovere, *La Corse et la République. La vie politique, de la fin du second Empire au début du XXIe siècle*, Paris, Média Diffusion, 2013.

10. Même s'il faut mesurer ce propos, il ne faut pas oublier l'agitation nationaliste de la période avec la fin progressive de la lutte armée du FLNC et l'engagement en politique de partis nationalistes.

11. *Ibid.*, p. 9.

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

d'un Banque de Données Langue Corse (BDLC) [13]. Il s'agit là de la première initiative dans la volonté de lemmatiser la langue corse dans sa diachronie. 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 [14], océriser des données textuelles afin qu'elles puissent être exploitées dans des recherches en humanités numériques représente un enjeu important dans l'avenir de la discipline en Corse.

0.2.3 Travaux individuels

Cette présente étude est la continuité d'un premier mémoire de master réalisé à l'université de Strasbourg. Ce-dernier portait sur une analyse comparée entre la revue *A Muvra* et une revue irrédentiste italienne, *Corsica antica e moderna*, entre 1932 et 1939 [15]. Sans se concentrer sur un aspect particulier de ce corpus, il s'agissait davantage d'une lecture proche d'un corpus limité. L'idée était de faire ressortir les différences idéologiques majeures entre les corsistes et les irrédentistes, malgré une promiscuité évidente. Si les Corses admettaient faire partie d'une entité culturelle et linguistique commune avec l'Italie, ils ne partageaient pas la même volonté d'unification politique même si certains autonomistes se sont rapprochés des idées fascistes juste avant de le début de la Seconde Guerre mondiale.

Pour réaliser cette étude, une large base de données Heurist a été réalisée à partir des archives départementales de Corse du Sud [16]. Elle consiste en 4323 entrées regroupant articles, numéros, auteurs ou encore thématiques. Le but était de faire des études statistiques précises sur cette période afin d'appuyer mon propos. Cette base de données a été faite à la main et peut donc comporter des erreurs, représentant donc l'une des limites à une telle méthodologie. Cette base suit le modèle du schéma 1.

13. https://bdlc.univ-corse.fr/bdlc/corse.php

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

15. Vincent Sarbach-Pulicani, *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*, mém. de mast., Université de Strasbourg, 2021.

16. https://heurist.huma-num.fr/heurist/?db=vsp_presse_corsiste_irredentiste

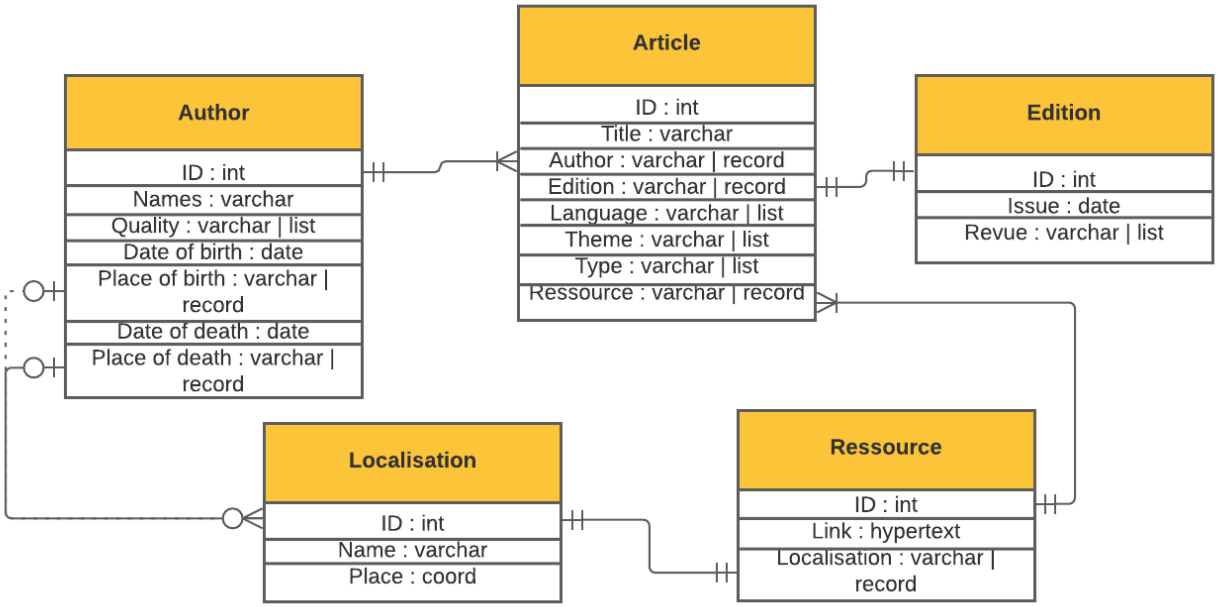


FIGURE 1 – Schéma de la base de données relationnelle

0.3 Objectifs

0.3.1 Approches méthodologiques

Si notre étude s'inscrit dans les sciences historiques, la nature même de notre sujet se doit d'importer des méthodologie de linguistique. Comme nous avons pu le préciser précédemment, nous effectuerons une lecture distante de notre corpus. Notre objectif ici est de prolonger l'observation déjà effectuée en prenant plus de recul sur le corpus et de l'analyser dans son ensemble en utilisant les outils numériques. Ce travail s'inscrit donc dans un double enjeu : celui de mieux comprendre la façon de penser des muvristes en leur temps mais aussi celui d'aider à la pérennisation de l'emploi des humanités numériques dans les études corses. L'une des difficultés est la composition d'un corpus trilingue. Dans la linguistique de corpus, cela n'est pas forcément un problème si les textes sont comparables en nombre, genre et types [17]. Dans notre cas, il est difficile de se retrouver avec un corpus totalement homogène dans la mesure où la répartition des langues et les aspects thématiques et typologiques sont des enjeux d'analyse à part entière. Entre 1932 et 1939, 58% des articles sont écrits en langue corse, contre 31% en français et 11% en italien [18]. De même, dans la mesure où il s'agit d'un corpus de presse culturelle,

17. Isabelle Léglise et Sophie Alby, « Les corpus plurilingues, entre linguistique de corpus et linguistique de contact : réflexions et méthodes issues du projet CLAPOTY », *Faits de langues*, 41–1 (2013), p. 98.

18. V. Sarbach-Pulicani, *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*..., p. 56.

les thématiques sont très variées tout comme les types d'articles. Le *topic modeling* issu de notre corpus devra donc prendre en compte ces métadonnées pour que l'analyse soit pertinente. Néanmoins, si la linguistique multilingue se prête bien à de larges corpus, est-elle efficace sur des corpus plus réduits ? En effet, notre corpus est composé de 12 numéros des années 1924 et 1925 ce qui correspond à 169 articles. cette question est au coeur de l'article d'Inga Hennecke de 2018 :

> « Dans les dernières décennies, le but de la linguistique de corpus a été de compiler des corpus de plus en plus grands afin de pouvoir effectuer des recherches et des analyses quantitatives et qualitatives sur un échantillon le plus représentatif possible. Cette évolution va de pair avec les progrès dans les domaines de l'informatique et de la linguistique quantitative. C'est alors seulement ces dernières années que les corpus de petite taille ont été de nouveau pris en considération par la communauté scientifique. » [19]

Si notre corpus réduit peut paraître une limite dans notre volonté d'effectuer une lecture distante d'*A Muvra*, la chercheuse semble également présenter des avantages non-négligeables. Même si cette-dernière s'attarde sur les petits corpus oraux, certaines considérations de cet article peuvent être assimilables à notre propre étude, comme la possibilité d'effectuer une recherche qualitative en parallèle [20].

Comprendre les muvristes nécessite de prendre du recul sur notre propre manière de penser l'étude de la Corse. Pour se faire, il est primordial d'utiliser le concept d'« insularisme ». Celui-ci est définit dans le *Larousse* par la « tendance d'un peuple insulaire à s'enfermer dans son île et à réduire ses relations internationales » [21]. dans le cadre de la Corse, Deborah Paci complète cette définition en précisant dans un article paru en 2016 :

> « La catégorie d'analyse que constitue l'« insularisme » permet d'observer que la conscience d'appartenir à un territoire commun est la conséquence de l'interaction entre l'espace, la représentation de l'espace et l'organisation sociale. » [22]

Il est donc nécessaire de déporter notre regard « continental » sur une situation bien particulière afin de bien comprendre les motivations des muvristes et les enjeux d'une lutte régionaliste sur l'île méditerranéenne. Néanmoins, « l'insularisme » ne doit pas se limiter qu'aux îles, même si le terme peut le laisser transparaître. Ce concept peut s'appliquer à toutes les sociétés dont l'isolement mental vis-à-vis de l'extérieur s'effectue par des contraintes géographiques autres, à l'image du Pays basque par exemple. Dans un sens,

19. Inga Hennecke, « Petits corpus oraux bilingues et plurilingues–enjeux théoriques et méthodologiques », *Corpus*–18 (2018), p. 2.
20. *Ibid.*, p. 14.
21. https://www.larousse.fr/dictionnaires/francais/insularisme/43489
22. D. Paci, « Le mare nostrum fasciste : l'espace politique et culturel en Corse et à Malte à l'époque du fascisme italien »…, p. 456.

« l'insularisme » est à l'origine des régionalismes car c'est cette volonté de se préserver son identité qui entraîne des initiatives politiques sur le plan local. En poussant cette logique, nous pouvons évoquer une forme de « double insularisme » avec le caractère montagnard du peuple corse, les reliefs servant souvent de refuge aux habitants de l'île et aliment un microcosme particulariste au sein même de la société corse.

0.3.2 Problématisation

En prenant tous ces éléments en compte, nous essaierons de déterminer le cadre dans lequel une langue est employée dans la revue *A Muvra*. Notre étude se portera essentiellement sur la rubrication de la revue. Peut-on déterminer, via le *topic modeling*, s'il existe des champs sémantiques propres à un langage ? D'un point de vue plus technique, nous essaierons de déterminer si notre approche est pertinente sur un corpus de taille réduite en prenant en compte la typologie des articles et les thématiques déjà plus ou moins connues de la revue autonomiste.

FIGURE 2 – *Pipeline* de notre projet de recherche

Notre méthodologie s'appuie sur trois grands axes de travail. Il a fallu dans un premier temps préparer nos données avant l'analyse. Cela consiste en l'extraction d'images en haute résolution, de l'océrisation puis de la structuration des données textuelles pour que celles-ci soient exploitables. Puis nous nous sommes attelés à l'analyse qui consiste en deux volets parallèles, une analyse sur R et une sur Python. Enfin, la dernière étape consistera en l'interprétation des résultats obtenus après avoir visualisé les sujets modélisés. Les scripts produits ainsi que les données sont disponibles sur le GitHub du projet [23].

23. V. Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP*, version 2.0.4, juin 2022, URL : https://github.com/vincentsarbachpulicani/Corsican-Stylometry.

Première partie

Préparation des données

Chapitre 1

Traitement des images

Le traitement des images est essentiel pour notre analyse dans la mesure où nos données ne sont pas disponibles, il faut pouvoir se les procurer grâce à l'océrisation. Pour se faire, il est nécessaire d'avoir des images en haute résolution pour que ce processus soit efficace. Ce chapitre sert à expliquer la démarche effectuée pour obtenir des documents de bonne qualité.

1.1 Récupération des images

Le principal problème auquel nous faisons face en ce début de recherche est l'accessibilité des données et métadonnées inhérentes à nos sources, à savoir la revue corse *A Muvra*. La visionneuse d'archive `THOT` des archives de Corse [1] est difficilement exploitable [2] et la numérisation des journaux à la BNF se termine en 1930. Néanmoins, afin de constituer un premier corpus, nous pouvons nous baser sur les documents disponibles Gallica.

Afin d'exploiter des images de bonne qualité avant d'opérer une phase d'océrisation, nous pouvons utiliser des documents IIIF, une norme d'interopérabilité internationale pour des images en haute résolution [3]. Face à certains problèmes de normalisation du nombre d'images en IIIF disponibles pour chaque numéro du journal et afin d'être certain d'extraire toutes mes sources, nous employons le script `IIIF-Crawler`, développé par Thibault Clérice et Jean-Baptiste Camps. Cela permet aussi, à termes, de pouvoir importer un nombre conséquent d'images de bonne qualité automatiquement.

Lors d'un stage effectué en février 2021 auprès de Guillaume Porte, ingénieur d'étude en humanités numériques rattaché à l'UR3400 ARCHE, j'ai eu l'occasion de développer un

1. http://archives.isula.corsica/Internet_THOT/FrmSommaireFrame.asp
2. Une API interne est exploitable mais l'extraction de ressources depuis la plateforme prendra un peu plus de temps afin de ne pas surcharger les serveurs avec des requêtes anormalement conséquentes.
3. https://iiif.io/

script permettant de récupérer automatiquement des métadonnées de revues de sociétés d'histoire locale alsacienne grâce aux identifiants `ark` [4]. La réalisation de celui-ci rentrait dans le cadre du projet *Alsatia Numerica* visant à la création d'un portail permettant d'avoir accès facilement à toutes les ressources numérisées concernant l'histoire médiévale dans la région du Rhin supérieur comme les thèses, les archives ou encore les articles de revues [5]. L'idée est de réutiliser ce script [6] et de l'adapter afin de récupérer les codes `ark` des articles d'*A Muvra* puis de les importer dans un fichier `.tsv` compatible avec le `IIIF-Crawler` :

```
python3 iiifcrawler.py ID -source SOURCE -start 1 -end 2
```

En effet, cette commande ci-dessus permet de lancer le script `IIIF-Crawler`. La documentation du script est disponible sur le dépôt du projet [7]. L'ID correspond dans notre cas à l'identifiant `ark`, source est Gallica, le *start* et l'*end* sont les pages 1 et 4 dans la mesure où le format d'édition de la revue corse respecte ce nombre de pages à travers ses deux décennies d'existence avec des pages au format « berlinois » [8].

```
python3 iiifcrawler.py example.tsv
```

La commande ci-dessus permet de lire un fichier `.tsv` qui contient 4 colonnes : l'ID, la source, le *start* et l'*end*. Ainsi, générer automatiquement un fichier de ce type grâce à un script Python qui exploite l'API de Gallica permet d'automatiser la démarche de récupération des images en haute résolution.

Il a donc fallu adapter le script réalisé dans le cadre de mon stage pour qu'il puisse récupérer le code `ark` de la revue, puis de chaque numéros. Le script s'occupe ensuite de créer un fichier `.tsv` qui peut être lu par le script `IIIF-Crawler`. La question s'est posée de savoir s'il fallait intégrer ce-dernier au script `IIIF-crawler-periodic` [9] mais il était plus aisé d'utiliser directement le code initial à l'aide d'une commande de la librairie `os` [10].

Avec ce script, un corpus de 238 numéros de la revue *A Muvra* a été récupéré, soit environ 730 pages, ce qui correspond à la totalité des exemplaires disponibles sur la plateforme Gallica.

4. Code d'identification pérenne utilisé par la BNF pour ses documents numérisés.
5. Voir : Guillaume Porte, « Alsatia Numerica », *Source(s). Cahiers de l'équipe de recherche Arts, Civilisation et Histoire de l'Europe*–2 (2013).
6. Dans le dossier projets_ext. V. Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP…*
7. https://github.com/Jean-Baptiste-Camps/IIIF-Crawler
8. Un format approximatif de 470 × 320 mm. Les premiers numéros sont néanmoins au format « demi-tabloïd » (290 × 210 mm), c'est au cours de l'année 1921 que le format change au « berlinois ».
9. Dans le dossier ressources. *Ibid.*
10. Une des perspective d'amélioration du script est d'unifier les deux scripts pour éviter des dépendances mais ce n'est pas essentiel pour l'instant.

1.2 Phase de nettoyage

Avoir des images de bonne qualité n'est pas suffisant pour que l'océrisation de celles-ci soit complètement efficace, il s'agit désormais de les nettoyer, c'est-à-dire les rendre plus lisibles pour le moteur d'OCR. Pour cela, nous pouvons utiliser le logiciel libre de droit ScanTailor [11]. Après la première étape de récupération des images, ces dernières étaient classées dans des dossiers distincts en fonction de leur identifiant `ark`). Pour régler ce problème, il a fallu écrire un script [12] qui permet de trier les images obtenues par l'étape précédente en fonction de leur numérotation, c'est-à-dire réunir les pages 1, 2 et 3 entre elles, en les renommant en fonction de leur identifiant `ark`. Cette étape est nécessaire car l'outil que nous voulons utiliser peut automatiser l'édition d'images depuis un dossier.

Une fois cette première étape réalisée, intéressons nous au logiciel ScanTailor. L'idée est donc de binariser [13] avec la méthode d'Otsu (paramétrée en 30-Thicker). En sortie, nous avons des images en noir et blanc avec un contraste plus net pour faciliter la détection optique des caractères mais aussi les séparateurs de colonne des journaux. La manchette des revues ne nous intéresse pas donc nous pouvons dès le début rogner toutes les entêtes afin de ne conserver que les rubriques.

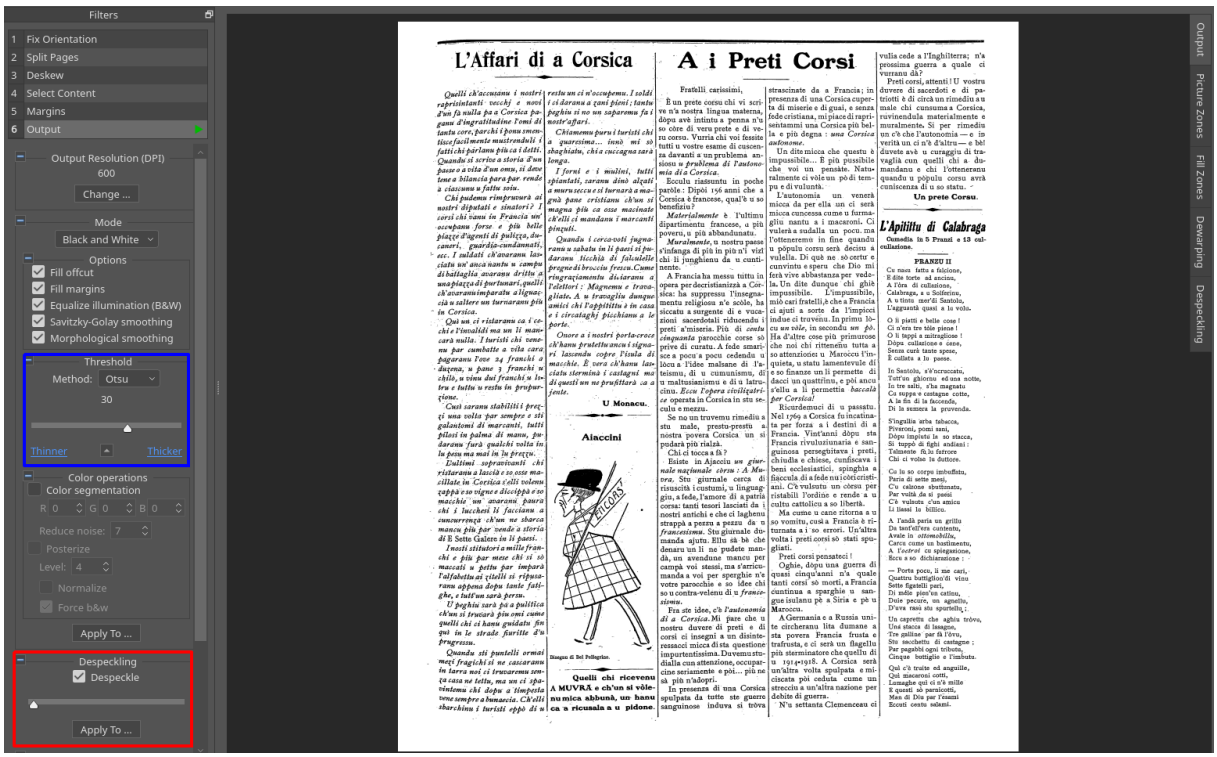


FIGURE 1.1 – Exemple d'utilisation du logiciel ScanTailor

11. https://scantailor.org/
12. Dans le dossier ressources. *Ibid.*
13. Mettre l'image en noir et blanc afin d'avoir un contraste plus clair entre le texte d'une image et le fond.

L'image 1.1 est un exemple de sortie d'une page de une de la revue *A Muvra* après avoir été traitée par le logiciel ScanTailor. Ainsi, comme évoqué précédemment, la méthode d'Otsu permet de binariser l'image (voir l'encadré bleu sur la figure 1.1). Elle consiste en le calcul d'un seuil de variance des pixels à parti de l'histogramme de l'image, d'où le terme de *thresholding* (qui signifie « seuillage ») sur l'interface du logiciel. Sans s'attarder sur les opérations mathématiques derrière cet algorithme [14], il faut retenir que ce-dernier déduit alors si tel pixel est noir ou blanc à partir du seuil sélectionné. Dans notre cas, le seuil choisi est de 30 pour avoir des caractères bien lisibles [15].

La deuxième phase de nettoyage est l'étape de *despeckling*, qui signifie « dépoussiérage ». Dans un article paru en 2021, des chercheurs italiens définisent le « speckling » par ces termes :

> « Unlike the Gaussian noise typically affecting optical images, speckle noise is spatially correlated and signal-dependent, and appears as a grainy texture superimposed to the image, which greatly affects its interpretability and scientific exploitation. » [16]

Concrètement, la « poussière » est un type de bruit qui correspondent à des clusters aléatoires de pixels noirs nuisent à la qualité intrinsèque d'une image binarisée. l'étape de *despeckling* est donc nécessaire pour supprimer ces tâches [17]. Sur la figure 1.1, ce paramètre correspond à l'encadré rouge qui a été réglé au maximum selon nos propres besoins.

Pour résumer cette étape, si nous mettons en entrée des images `.jpg` issues du *crawling* de Gallica, nous obtenons en sortie des images au format `.tif` [18] après les avoir nettoyées et recalibrées. Le nettoyage a consisté en la binarisation des images et l'ajustement du contraste, puis en le dépoussiérage de *cluster* aléatoires de pixels. Pour nos besoins futurs, il a fallu replacer les images dans leur dossier d'origine, triées selon leur identifiant `ark`.

14. L'article est un peu ancien, mais la justification mathématique reste très pertinente : Nobuyuki Otsu, « A threshold selection method from gray-level histograms », *IEEE transactions on systems, man, and cybernetics*, 9–1 (1979).

15. Le seuil est peut-être un peu élevé mais, à mon sens, le rendu des images est plus cohérent. Tester un OCR avec un seuillage plus faible pourrait être pertinent à l'avenir.

16. Giulia Fracastoro, Enrico Magli, Giovanni Poggi, Giuseppe Scarpa, Diego Valsesia et Luisa Verdoliva, « Deep learning methods for synthetic aperture radar image despeckling : An overview of trends and perspectives », *IEEE Geoscience and Remote Sensing Magazine*, 9–2 (2021).

17. Il est intéressant de noter que les auteurs de cet article font remonter les premières méthodes de *despeckling* aux années 1970, à peu près à la même époque des premières méthodes de *thresholding*.

18. Format d'image en haute qualité. ici, les images sont en 600dpi.

Chapitre 2

Mise à disposition des données textuelles

Une fois la récupération des images en bonne qualité effectuée, il faut dorénavant en extraire les données textuelles. Même si nous pouvons compenser ce fait par la masse de données, sans entrer dans le *big data*[1], il est nécessaire d'avoir une corpus relativement propre. En effet, cette étape est centrale pour que les analyses de données soient pertinentes.

2.1 Le défi de l'océrisation

Malgré le nettoyage des images, l'étape d'extraction des données textuelles s'est révélée rapidement un réel un problème. Que ce soit avec le moteur Kraken[2] ou avec Tesseract[3], la grande difficulté a été de passer outre la segmentation complexe des revues.

2.1.1 Un cas d'analyse de segmentation complexe

Les sources de presse sont caractérisées par des mises en pages complexes (*Complex Layout Anaylisis*) qui diffèrent des ouvrages traditionnels, entendez ici des pages de textes avec une colonne unique ou deux au maximum. Le modèle de segmentation de base Kraken se trouve être très peu performant pour mes sources ce qui nécessiteraient l'entraînement d'un nouveau modèle adapté. Pour se faire, l'utilisation de la version de l'Inria de l'interface eScriptorium[4] s'est révélée utile mais a vite montrée ses limites. L'entraînement d'un modèle est chronophage et les documents que nous utilisons sont particulièrement longs à segmenter manuellement. Il semblerait que ce problème d'analyse de segmenta-

1. Jeux de données extrêmement complexes et volumineux.
2. https://kraken.re/master/index.html
3. https://tesseract-ocr.github.io/
4. https://gitlab.inria.fr/scripta/escriptorium

tions complexes ait déjà été traité [5] mais il est difficile de trouver des jeux de données de modèles disponibles facilement en ligne. On peut notamment noter l'initiative de la *Library of Congress* et du *Newspaper Navigator* [6], mais rien de réellement adapté pour nous pour le moment. L'image 2.1 est un exemple de segmentation complexe d'un numéro d'*A Muvra* :

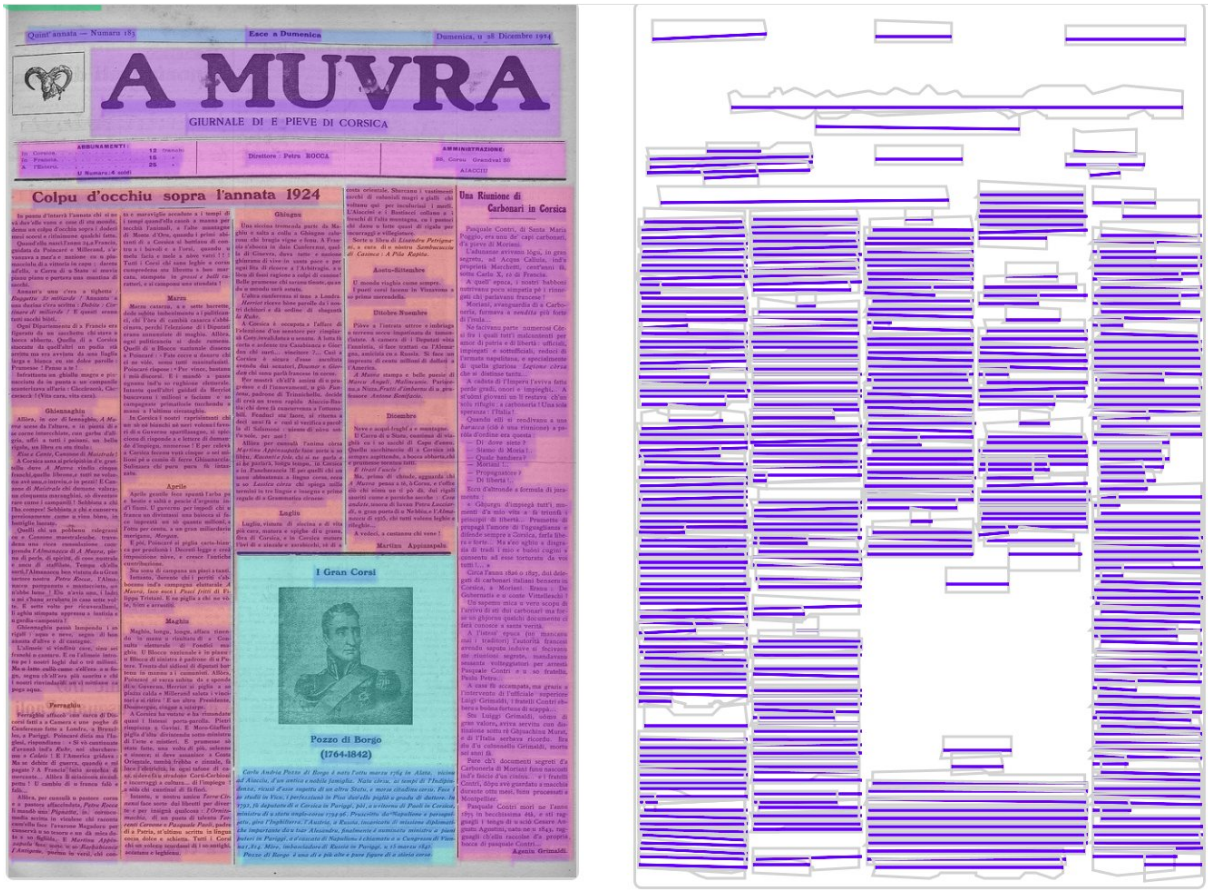


FIGURE 2.1 – Exemple de segmentation d'une page de une d'un numéro d'*A Muvra*

Ce visuel est assez évocateur de la difficulté que peut rencontrer un moteur d'OCR pour détecter la mise en page d'un journal. Dans le cas de la figure 2.1, la segmentation a été faite manuellement, mais la détection automatique se base sur le même principe avec la définition de régions et de lignes.

Ce problème est le même pour Tesseract-OCR, qui suit un modèle de segmentation basique de détection des lignes successives : très efficace pour les textes ayant qu'une seule colonne mais rapidement dépassé lorsque la segmentation devient trop difficile. Il semblerait néanmoins qu'il existe des paramètres de modifications des PSM (*Page Segmentation Modes*) afin de prendre en compte différentes mises en page mais, après plusieurs essais, ces paramètres ne semblent pas très efficaces pour notre corpus.

5. De nombreux projets de numérisation de presse existent comme le projet ANR Numapresse : http://www.numapresse.org/presentation-du-projet/

6. https://news-navigator.labs.loc.gov/

L'analyse de segmentation complexe est, nous l'aurons compris, un enjeu de taille dans l'utilisation de moteurs d'OCR ou de HTR. À terme, l'objectif pour notre recherche est d'avoir résolu ce problème. L'une des possibilités, si la segmentation ne peut pas être faite rapidement, est d'utiliser les fichiers XML/ALTO proposé par Gallica de leur propre OCR. L'idée serait de proposer ces fichiers à Tesseract pour que le moteur n'ait qu'à s'occuper de la phase de reconnaissance des caractères et non pas à la phase de segmentation.

2.1.2 Le choix du moteur d'OCR

Le choix du moteur d'OCR s'est d'abord porté sur Kraken, via l'interface eScriptorium à l'aide du modèle *manuscrit DAHN NFC*[7] développé par Floriane Chiffoleau et trouvé sur le dépôt GitHub HTR-United[8]. Néanmoins, se révélant finalement peu efficace dans le cadre de presses écrites en langue corse, je me suis orienté vers le modèle *19th century prints - HTRcatalogs Artlas*. Bien plus efficace que le précédent, celui-ci semblait être une bonne source de travail pour entraîner un nouveau modèle propre aux presses corses afin d'avoir un taux d'efficacité optimal.

Néanmoins, face à la difficulté de la segmentation évoquée précédemment, le basculement sur Tesseract-OCR a semblé évident. Ce moteur, en plus de posséder un module de transcription corse, permet de compiler différentes langues dans le cadre de corpus plurilingues. S'il existe en effet une librairie dédiée pour Python, les commandes `bash` sont largement suffisantes pour notre utilisation du moteur.

```
for d in */ ; do
        (cd "$d"
        pdftoppm -r 300 -tiff *.pdf document
        for f in *.tif; do
                tesseract $f $(basename $f .tiff) -l fra+cos+ita
                tesseract $f $(basename $f .tiff) -l fra+cos+ita alto
        done
        cat *.txt > $d.txt)
done
```

Cette suite de commandes permet de lancer la transcription de tous les fichiers `.tif` et d'avoir en sortie un fichier `.txt` et un fichier `.xml` au format ALTO. Enfin, on concatène tous les fichiers texte en un seul pour obtenir un fichier `.txt` qui contient le texte brut de chacune des revues.

7. Floriane Chiffoleau, *dahncorpus*, version 1.0.0, mars 2021, DOI : 10.5281/zenodo.5911868.
8. Alix Chagué et Thibault Clérice, *HTR-United : Ground Truth Resources for the HTR and OCR of patrimonial documents*.

Il nous reste à comparer la qualité de la transcription issus de deux moteurs différents pour décider duquel utiliser. Si l'idée de Kraken a été écartée, il reste pertinent de comparer les résultats entre l'OCR proposé par Gallica et celui effectué grâce à Tesseract. Pour cela, il a fallu retranscrire entièrement une page de une en respectant la rubrication [9] et d'effectuer un calcule de distance entre la vérité de terrain et les résultats d'OCR proposés par Tesseract et Gallica. Ainsi, en calculant la distance Levenshtein [10], nous obtenons les résultats suivants un taux d'erreur de 3,79% avec l'OCR de Gallica alors que tesseract atteint péniblement 82% d'erreurs. Mais comment expliquer ce résultat ? Le texte brut proposé par Gallica est déjà nettoyé et « prêt à l'emploi » alors que Tesseract enregistre beaucoup de bruit tout en ayant une segmentation approximative. Ceci explique pourquoi, avec une métrique comme celle de Levenshtein qui s'occupe à comparer chaque caractères entre eux, le taux d'erreur s'envole avec Tesseract. Quelque soit la métrique utilisée, il y aura de toute façon un décalage important entre les différents résultats.

Il faut donc remettre ce résultat en perspective et essayer de comparer un peu plus empiriquement les résultats. Si la qualité intrinsèque de Tesseract semble inférieure, ce moteur présente l'avantage d'être malléable et d'avoir un potentiel non négligeable une fois la segmentation corrigée. En revanche, l'OCR de Gallica présente quand même des erreurs de transcriptions qui ne sont pas homogènes, entendez ici que chaque pages possède son lot d'erreurs qui varient en fonction des exemplaires de la revue. Il est donc plus difficile de corriger ces erreurs alors que celles de Tesseract sont plus aisément identifiables et modifiables, c'est ce que nous verrons dans la partie suivante de ce chapitre. Cet aspect de notre recherche reste encore un domaine à approfondir. Si le choix de Tesseract a été fait pour ce mémoire de 1re année, les prochains essais que nous effectuerons seront essentiels dans notre choix final d'OCR.

2.2 Exploitation des données brutes

Maintenant que nous possédons un texte brut, il s'agit de le rendre exploitable pour l'analyser. Comme nous avons vu précédemment, un OCR n'est jamais fiable à 100%. Le tout est de déterminer, via une approche dans un premier temps manuelle, des erreurs récurrentes afin de les corriger automatiquement. Mais la normalisation du jeu de données est également un enjeu méthodologique notamment pour ce qui est des formes contractées des mots outils.

9. L'exemplaire choisi est le numéro 183 de la cinquième année, datant du 28 décembre 1924 : https://gallica.bnf.fr/ark:/12148/bpt6k1330777p

10. En se fondant sur le script développé par Louis-Fiacre Franchet d'Esperey disponible sur son dépôt GitHub : https://github.com/Louis-Fiacre/OCR_corpus

2.2.1 Structuration en XML

Avant de procéder, il était nécessaire de structurer les textes en XML. En effet, si le nettoyage des textes est essentiel, une fonction `clean_text` ne s'applique qu'à un langage particulier et il faut d'abord différencier la langue d'expression des différentes rubriques.

La question du format de balisage s'est alors posée. Le choix de la TEI [11] pouvait sembler évident mais un format personnalisé s'est révélé dans un premier temps plus pertinent. Dans notre méthodologie d'analyse, seule certaines informations paraissaient pertinentes à conservées, à savoir :

— Le titre de la rubrique

— La langue d'écriture

— L'auteur

— La typologie

— Le contenu

L'utilisation du XML/TEI semblait n'être pas forcément idéale étant donné l'exploitation que l'on voulait en faire. En effet, dans la mesure où notre corpus est relativement réduit et qu'il ne nécessite pas une granularité fine, encoder manuellement avec de la TEI de tels documents semblait être une perte de temps pour un gain méthodologique incertain. Néanmoins, l'objectif est de publier ces données en XML/TEI une fois l'acquisition d'un modèle pour automatiser un minimum l'encodage et de convertir en TEI à l'aide d'une fiche XSLT.

Une fois ces considérations prises en compte, voici un exemple d'encodage d'une rubrique d'*A Muvra* :

```
<metadonnees>
    <numero>187</numero>
    <date>1925-01-25</date>
    <ark>bpt6k1330781k</ark>
</metadonnees>
<rubrique>
    <type>article</type>
    <langue lan="cos">cos</langue>
    <titre>Riflessioni d'un veru corsu</titre>
    <texte>« A Corsica un è mica stata lampata a mezzu ...</texte>
    <auteur>Sambucucciu di Casinca</auteur>
</rubrique>
```

11. https://tei-c.org/

L'entête `<metadonnees>` est unique et contient quelques informations sur l'exemplaire d'*A Muvra* : son identifiant `ark`, sa date de parution et le numéro d'édition. Puis chaque rubrique doit contenir au moins une balise `<type>`, `<langue>`, `<titre>`, `<texte>` et `<auteur>`. Dans le cas de rubriques un peu particulières, il y a la possibilité d'ajouter d'autres balises comme `<chronique>` ou `<compositeur>`. La balise `<langue>` contient un attribut qui prend en valeur la langue de l'article : `cos` pour le corse, `ita` pour l'italien et `fra` pour le français [12].

Afin de normaliser l'encodage et être sûr qu'aucune erreur n'a été faite, il a fallu traduire dans une règle DTD afin de normaliser les observations précédentes :

```
<!DOCTYPE revue[
<!ELEMENT revue (metadonnees,rubrique*)>
<!ELEMENT metadonnees (numero,date,ark) >
<!ELEMENT numero (#PCDATA)>
<!ELEMENT date (#PCDATA)>
<!ELEMENT ark (#PCDATA)>
<!ELEMENT rubrique (type+,langue+,chronique*,titre+,texte+,auteur+,compositeur*,tradu
<!ELEMENT type (#PCDATA)>
<!ELEMENT langue (#PCDATA)>
<!ELEMENT auteur (#PCDATA)>
<!ATTLIST langue lan (cos|fra|ita) #REQUIRED>
<!ELEMENT titre (#PCDATA)>
<!ELEMENT texte (#PCDATA)>
<!ELEMENT compositeur (#PCDATA)>
<!ELEMENT traducteur (#PCDATA)>
<!ELEMENT chronique (#PCDATA)>
<!ELEMENT origine (#PCDATA)>
]>
```

La limite d'un balisage XML personnalisé se trouve dans la rigueur à avoir lors de l'encodage des fichiers, il faut faire attention à ne pas faire d'erreurs. De plus, dans la mesure où les règles DTD ne s'appliquent qu'à un fichier unique, cela nécessite de vérifier chaque fichier pour être sûrs qu'ils soient valides et cohérents. Comme évoqué précédemment, cela suffit amplement pour ce mémoire de 1re année mais il faudra améliorer l'encodage pour la suite de notre recherche.

12. Une perspective d'amélioration : définir des attributs pour les balises `<type>` et `<auteur>` pour être sûr d'avoir des identifiants normalisés.

2.2.2 Nettoyage automatique du texte

Une fois avoir séparé les différents textes et de les avoir caractérisés selon leur langue, nous pouvions appliquer une fonction de nettoyage de texte [13]. L'idée est d'utiliser la librairie Python `lxml` et de parser les différents fichiers `.xml` créés durant l'étape précédente. On doit extraire le contenu qui se situe entre les balises textes et appliquer la fonction de nettoyage adéquate à la langue utilisée. Une fois cela fait, il faut renvoyer en sortie un fichier `.xml` avec le contenu nettoyé. Pour identifier le langage du contenu, il faut se fonder sur la valeur de l'attribut `@lan` de la balise `<langue>` de la rubrique en question, d'où l'importance d'avoir des fichiers `.xml` bien normalisés. Il faut donc utiliser des expressions `xpath` telles que celle-ci : `//langue[@lan="cos"]/following-sibling::texte`. Le nettoyage s'effectue ensuite en quatre étapes principales :

1. La suppression de la ponctuation
2. La réduction de la casse
3. La normalisation de la syntaxe
4. L'élimination des accents

Pour la ponctuation, il a fallu faire une liste personnalisée plutôt que d'utiliser une liste déjà mise à disposition par des librairies Python comme `string`. En effet, ce terme de « ponctuation » englobe aussi certains caractères spéciaux. La suppression des caractères est essentielle pour avoir un jeu de données propre et une tokenisation qui prend en considération uniquement les mots. De même, les étapes de réduction de la casse et d'élimination de l'accentuation sont assez évidentes lorsqu'on compte des occurrences de mots.

La phase la plus délicate dans notre méthodologie est celle de la normalisation de la syntaxe. Elle est importante car pour des raison euphoniques, des contractions s'opèrent à l'écrit sous forme d'élisions qui reflètent les pratiques discursives des locuteurs du corse. Par exemple, l'expression *s'è ellu hè* (« si il est ») devient *s'ell'è* à l'écrit [14]. Se pose également la question de la règle de normalisation, doit-on se baser sur la syntaxe du XX^{e} ou sur l'actuelle ? De plus, un certain nombre d'ambiguïté peuvent se glisser dans une telle correction comme le mot *e* qui peut dire, selon le contexte, soit « les » ou soit « et ». De plus, il ne faut pas oublier de prendre en compte que la langue corse est une « langue par élaboration » [15] et que, par conséquent, la syntaxe présente une complexité du faite

13. dans le dossier ressources. V. Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP...*

14. Inversement, restituer la forme originale de l'élision nécessite de prendre en compte le contexte du genre et du nombre : *ell'* peut donner *ellu*, *ella*, *elli* ou *elle*.

15. En sociolinguistique, une « langue par élaboration » ou langue *Ausbau* est une variante d'un langage structuré (comme l'italien) et érigée comme langue élaborée distincte. Pour en savoir plus, voir : Alain Viaut, « Marge linguistique territoriale et langues minoritaires », *Lengas. Revue de sociolinguistique*–71 (2012).

des instanciations distinctes suivant les auteurs.

Ces questions sont abordées dans le chapitre sur les statistiques exploratoires car elle n'est pas forcément pertinente pour le moment. En effet, nous verrons plus tard que nous supprimons les mots outils pour le *topic modeling*. Néanmoins, pour l'analyse stylométrique prévue pour la suite de la recherche, ces mots sont essentiels et il nécessitera une phase de désambiguïsation de la langue. Mais pour le moment le choix est fait de suivre la syntaxe moderne du corse et de ne pas désambiguiser, la liste de tous les changements sont disponibles dans le script `clean_text.py` disponible sur le GitHub du projet [16].

16. Dans le dossier ressource. V. Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP...*

Deuxième partie

Valorisation des données

Chapitre 3

Analyse lexicométrique d'un corpus en langue corse

La première partie de ce mémoire a mis en évidence des problématiques de fond et de forme propres au traitement automatique de la langue corse. Ce chapitre constitue un essai d'approche lexicométrique [1] comparative entre un corpus en langue corse datant des années 1920 et d'un corpus écrit avec la syntaxe moderne. L'intérêt de cette étude exploratoire est double pour notre recherche : il s'agit d'une part de mettre en évidence les évolutions grammaticales les plus visibles en un siècle d'écriture de la langue et, d'autre part, de voir si la transition d'une syntaxe ancienne à la syntaxe moderne est efficace dans le traitement du corse [2].

3.1 Méthodologie et limites

Pour cette étude, nous avons donc comparé deux corpus différents. Le premier est composé de plusieurs textes variés [3] publiés dans les années 1920 par la *Stamparia di a Muvra*, organe de publication des muvristes [4]. Le second corpus est une traduction contemporaine du *Comte de Monte Cristo* disponible sur le site de la revue dialectale *Tempi Rivista* [5]. Les textes sont variés et abordent des thèmes et styles différents, mais notre travail se concentrera essentiellement sur les mots outils donc cela n'est, en principe, pas un problème pour nous.

Dans un premier temps, il était nécessaire de nettoyer correctement nos textes. Pour cela, il a fallu créer une fonction `clean_text` qui, de la même manière que dans

1. Étude quantitative d'un lexique à l'aide de méthodes statistiques.
2. Tout le processus se trouve sur le dépôt, dans le dossier statistiques_exploratoires *Ibid.*
3. *Almanaccu di A Muvra per 1932* (1932) ; *In giru eletorale* de Francescu Piazzoli (1932) ; *U partitu corsu autonomista. Statuti e duttrina* (1935) ; et *Pontenôvu 9 maghiu 1769* (1923).
4. Les autonomistes corses étaient les principaux producteurs de textes dialectaux et les règles grammaticales qui structurent le langage écrit n'en sont qu'à leurs balbutiements.
5. https://tempicorsica.com/

le nettoyage des numéros de revues transcrits, se débarrassait de la ponctuation et de la casse. Cette fonction sert également à normaliser les textes de deux manières :

— Une normalisation de l'encodage, les textes provenant pour certains d'eScriptorium, l'unicode n'est pas forcément respecté, particulièrement au niveau des accents, faussant ainsi les résultats.

— Une normalisation de l'écriture, les textes anciens (OLD) ayant une graphie différente des textes contemporains (NEW). Par exemple, le verbe être à la 3^{e} personne du singulier s'écrit soit *è* ou soit *hè*.

Les 8 textes sont ensuite divisés en deux dossiers pour former les deux corpus différents, qui seront par la suite importés dans RStudio. Séparer les deux corpus est très important car nous sommes dans le cadre d'une étude comparative : si le script peut en effet être plus long, cela reste la manière la plus efficace de comparer deux jeux de données entre eux.

Dans la première phase de notre analyse, un *dataframe* a été créé qui contient le nombre de caractères total de chacun des corpus puis le nombre de mots une fois la tokénisation effectuée. Le choix a ensuite été fait de calculer le nombre de caractères moyen par mots, élément intéressant pour déterminer s'il y a un phénomène de littérarisation [6] de la langue corse ou non. Nous observerons les résultats obtenus dans notre deuxième sous-partie.

Une fois ce premier tableau effectué, qui compare des données relativement basiques des deux corpus, il a fallu calculer les fréquences de mots, ce qu'on appelle les MFW (*Most Frequent Words*). C'est à cette étape que la constitution de deux corpus différents prend tout son sens. Il convenait donc de créer deux objets *dataframes* comportant les 15 MFW de chaque corpus avec leur fréquence absolue et leur taux en pourcentage donnant en sortie deux histogrammes.

Enfin, dans une dernière étape, il était nécessaire d'effectuer une réduction de dimension en ne conservant que le calcul de la fréquence des mots afin de produire une matrice de correspondance du taux de fréquence de chacun desdits mots dans les textes du corpus. Cela permet ensuite de visualiser ces résultats à travers une carte de chaleur classifiée par des dendrogrammes (voir la figure 3.2) permettant de repérer des groupes de mots plus ou moins utiliser dans les corpus. La métrique utilisée pour réaliser ces dendrogramme est la distance euclidienne, la métrique de base proposée par RStudio.

Durant cette étude, plusieurs problèmes se sont posés mettant en exergue les limites d'une telle méthodologie. D'abord, le corpus a présenté deux problèmes qui tendent à

6. Processus de constitution d'une langue écrite principalement orale à l'origine.

légèrement fausser les résultats : déjà vis-à-vis des différences de graphies évoquées précédemment, comme la négation *ùn* où l'accentuation n'est pas présente dans les textes anciens, causant une ambiguïté par homonymie syntaxique avec le déterminant cardinal *un* ; d'où la limite d'avoir un texte non lemmatisé car on décontextualise le propos. Puis d'un point de vue sémantique, les textes anciens sont variés, allant de la monographie à la poésie, alors que les textes contemporains sont une traduction. L'usage fait des mots vides n'est pas forcément le même et donc leur quantité peut varier suivant le type de texte.

3.2 Des premiers résultats à saisir

TABLE 3.1 – Tableau représentant les caractéristiques des deux corpus

Corpus	**Nb de caractères**	**Nb de mots**	**Caractères par mots**
Old	84634	14894	5.682422
New	77733	14978	5.189812

Le tableau 3.1 permet de nous rendre compte que, pour deux corpus de taille relativement équivalente, les mots sont en moyenne 10% plus long dans les textes anciens que dans les textes contemporains. Cela peut s'expliquer peut-être par le genre des textes, comme nous avons vu précédemment, ou alors par la littérarisation d'une langue principalement orale dans les années 1920, en cherchant à complexifier le langage pour le légitimer. Ce phénomène correspond à l'un des trois projets de régionalisme linguistique selon le sémioticien Jean-Marie Klinkenberg, celui du moment défensif et culturel :

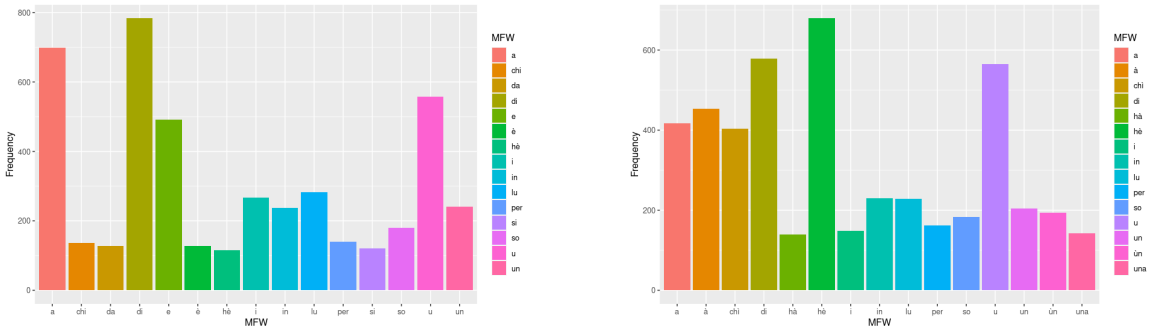
> « Le travail linguistique sur les langues régionales reproduit ainsi le travail qui se fait alors sur la culture nationale : publier et faire connaitre les auteurs du passé, en illustrant la langue dans le présent. Le mouvement est le fait d'une catégorie d'acteurs qui sont en même temps des écrivains et des érudits. Ces écrivains associent, en les distinguant soigneusement, leur travail de créateur et leur quête philologique. » [7]

Ce constat est applicable aux muvristes au regard de la floraison de nombreux romans et nouvelles en langue corse à l'entre-deux-guerres et la promotion d'auteurs passés comme Salvatore Viale [8]. À l'inverse, le corse contemporain s'écrit davantage dans

7. Jean-Marie Klinkenberg, « « Grandes langues » et langues minoritaires : deux politiques linguistiques ? », *Lengas. Revue de sociolinguistique*–79 (2016), https ://journals.openedition.org/lengas/1048.

8. Salvatore Viale (1787-1861), ami du nationaliste italien Niccolò Tommaseo, est un poète bastiais réputé pour être le premier auteur à avoir écrit consciemment en langue corse dans son oeuvre *La Dionomachia*.

la vie quotidienne, poussant les auteurs de nos jours à, peut-être, adopter des pratiques orthographiques et grammaticales moins complexes. Dans le schéma de Klinkenberg, le statut du corse peut se situer à cheval entre le moment nationaliste et politique, avec l'arrivée au pouvoir des autonomistes en 2015 faisant du corse leur *leitmotiv* primordial, et le moment prospectif et polycentrique, avec la volonté de promouvoir la langue sous des formes originales [9].



(a) Corpus OLD (b) Corpus NEW

FIGURE 3.1 – Histogramme des MFW-15 de l'analyse exploratoire

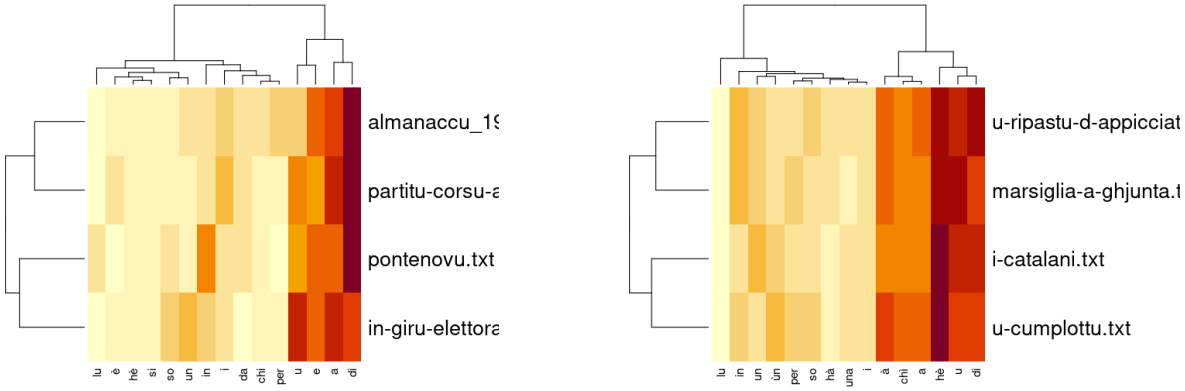
Selon les graphiques 3.1, le mot le plus utilisé dans le corpus OLD est l'article *di* alors que la forme conjuguée *hè* du verbe *esse* (« être ») est le terme le plus employé dans le corpus NEW. Les différences entre les deux corpus ne se font pas uniquement au niveau des mots outils en eux-mêmes. Les graphiques nous montrent également une plus grande homogénéité dans l'utilisation des différents mots outils de la langue corse dans les textes contemporains. À l'inverse, les textes anciens utilisent essentiellement 4 mots : *a*, *di*, *u*, *e*. Encore une fois, faisons attentions aux ambiguïtés induites par cette graphie plus « simpliste » dans le sens où la syntaxe se rapproche de l'italienne, la recherche d'une graphie qui prend en compte les variétés dialectales [10] du corse n'intervient qu'après la Seconde Guerre mondiale [11]. Nous pouvons néanmoins relever les premières tentatives de structuration du langage avec l'ouvrage d'Antoine Boniracia *A prima grammatichella corsa* publié en 1925 aux éditions de *L'Annu Corsu*, dont le nom rappel plaisamment le titre de la première grammaire de *latino volgare* de Leon Battista Alberti, *Grammatichetta della lingua toscana* [12].

9. Sans s'étendre sur ce sujet qui dévie de notre propos, il existe en Corse de plus en plus de voies de promotion de la langue à travers les médias mais aussi sous des formes plus variées : les écoles de chants, les clubs de sports, etc ...

10. Il existe en Corse un certain nombre de régiolectes répartis suivant des aires linguistiques plus ou moins précises. V7oir : M.J. Dalbera-Stefanaggi, *Essais de linguistique corse*, Ajaccio, Alain Piazzola, 2000.

11. Mathée Giacomo-Marcellesi, « Le corse », dans *Histoire sociale des langues de France*, Rennes, Presses universitaires de Rennes, 2013, p. 468.

12. En ce sens, les régionalistes corses se rapprochent de la rhétorique des premiers nationalistes italiens qui dataient les débuts de volonté d'unification italienne avec les auteurs dialectaux du *Quattrocento* comme Dante ou du *Cinquecento* comme Alberti.



(a) Corpus OLD (b) Corpus NEW

FIGURE 3.2 – Carte de chaleur des MFW-15 de l'analyse exploratoire

Les graphiques 3.2 appuient encore plus ce résultat, les cartes de chaleur permettent de mettre en valeur plus clairement deux groupes principaux d'emploi des mots outils dans chaque corpus grâce à leur classification ascendante hiérarchique.

Pour conclure, une telle étude est prometteuse car elle permet de réellement mettre en évidence des évolutions grammaticales dans la diachronie de la langue corse. La structuration d'un jeune langage principalement oral prend du temps et nécessite de la pratique par de nombreux acteurs de plusieurs générations pour qu'elle prenne en consistance. Mais la mise à l'écrit d'une langue est influencée par des facteurs à la fois internes et externes. D'une part, le contexte culturel et social joue un rôle crucial dans ce processus et le *riacquistu*[13] est un terreau favorable après-guerre, permettant l'ouverture du corse littéraire à des catégories sociales moins restreintes[14] qu'auparavant. D'autre part, l'éducation est essentielle et a longtemps favorisé la domination progressive du français sur le corse[15]. Les auteurs contemporains semblent insister sur l'utilisation du verbe « être » et du pronom relatif *chì*, ce-dernier étant peu utilisé par les auteurs autonomistes. La seule concordance se fait au niveau des pronoms définis mais étant la base des groupes nominaux, il y a une forme de logique dans la perpétuation de ces mots outils précis.

Mais comment appliquer ces considérations au sujet principal de notre recherche ? Déjà, cela nous a permis de mettre en évidence les particularités syntaxiques d'un corpus

13. Mouvement culturel d'après-guerre dont le but était la « réappropriation » des traditions corses par sa population.

14. D'un point de vue sociologique, les muvristes étaient principalement des lettrés ayant fait des études ou travaillant parfois dans la fonction publique. Pour en savoir plus, voir : V. Sarbach-Pulicani, *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*..., p. 32 à 39.

15. Après la Révolution française, il y a eu une volonté en France de généraliser l'usage de la langue française au sein de la population. Voir : Eugène François Xavier Gherardi, *Précis d'histoire de l'éducation en Corse. Les origines : de Petru Cirneu à Napoléon Bonaparte*, Ajaccio, CRDP de la Corse ; A Meridiana, 2011.

en langue corse du XX^{e} siècle. Celles-ci sont importantes dans le cas d'un étude stylométrique où, comme déjà évoqué précédemment, les mots outils sont essentiels. Mais cela reste utile, au point de vue sémantique et pas syntaxique, pour notre modélisation de sujets car il est nécessaire de supprimer ces mots. Relever les MFW d'un corpus conséquent a permis de compléter une liste de mots outils dédiée au corse [16].

16. Dans le dossier ressources. V. Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP...*

Chapitre 4

Topic modeling : faire émerger des sujets latents

Après avoir fait une première phase de *text mining* afin de comprendre les particularités syntaxiques d'un corpus en corse, considérée comme une langue « peu dotée » [1] en outils numériques, nous pouvons nous atteler à la modélisation sémantique de sujets. Notre approche est double afin de voir quel outil est le plus pertinent. pour cela, nous avons utilisé le langage R [2] et le langage Python [3] dans une logique de complémentarité des résultats.

4.1 Visualisation des données

La volonté d'employer différentes méthodes de *topic modeling* [4] est né d'une nécessité de complémentarité dans notre approche. L'objectif de base était de travailler à partir de fonctions R mais face à la redondance de certains sujets modélisés, croiser ces résultats avec d'autres obtenus avec la librairie *Top2Vec* de Python paraissait pertinent. Ainsi, dans un premier temps, le choix a été fait de travailler à partir de la méthode *Spectral* [5] et d'isoler notre corpus de rubriques selon leur langue d'écriture. Trier notre corpus de cette manière correspond à l'objet de notre problématique, afin de mettre en évidence d'éventuels champs sémantiques propres au langage utilisé dans la rubrication. La pertinence des sujets sortis dépendaient également des mots outils supprimés (surtout en langue corse) et du nombre de sujets que nous voulions (paramètre désigné par la variable

1. L. Kevers, F. Gueniot, A. G. Tognotti, *et al.*, « Outiller une langue peu dotée grâce au TALN : l'exemple du corse et BDLC »...
2. https://www.r-project.org/
3. https://www.python.org/
4. Méthode probabiliste permettant de déterminer des sujets et thèmes abstraits de documents.
5. Algorithmes d'inférence « spectrale » basés sur l'hypothèse de séparabilité. Voir : Moontae Lee, David Bindel et David Mimno, « Robust Spectral Inference for Joint Stochastic Matrix Factorization », dans *Proceedings of NIPS 2015*, 2015.

k dans les graphiques). Pour ce mémoire, nous avons choisi des sorties à 10 sujets [6]. Ainsi, nous obtenons les résultats suivants :

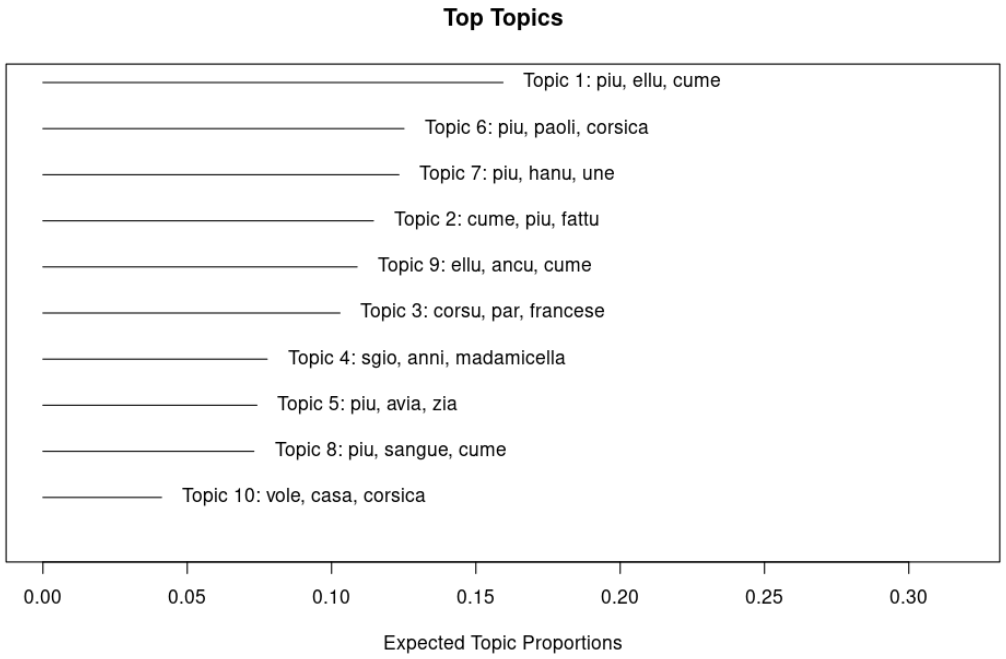


FIGURE 4.1 – *Top topics* des rubriques corses, $k = 10$, avec les mots outils

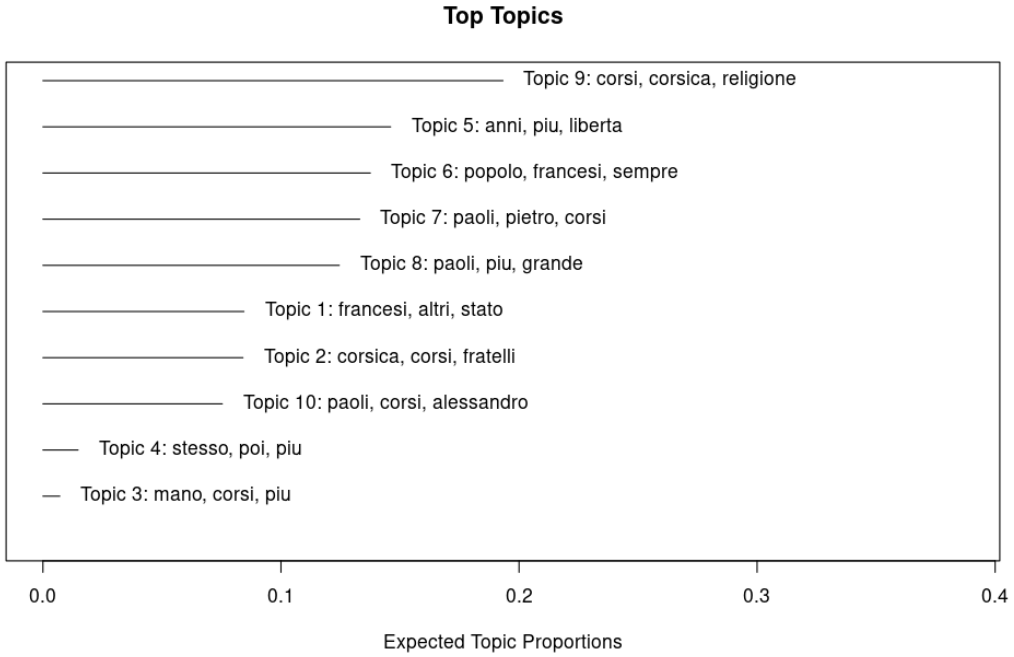


FIGURE 4.2 – *Top topics* des rubriques italiennes, $k = 10$, avec les mots outils

6. Vis-à-vis de la pertinence des sujets sortis, mais se rendre dans le dossier graphs situé dans topic_modeling pour voir les autres essais. V. Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP...*

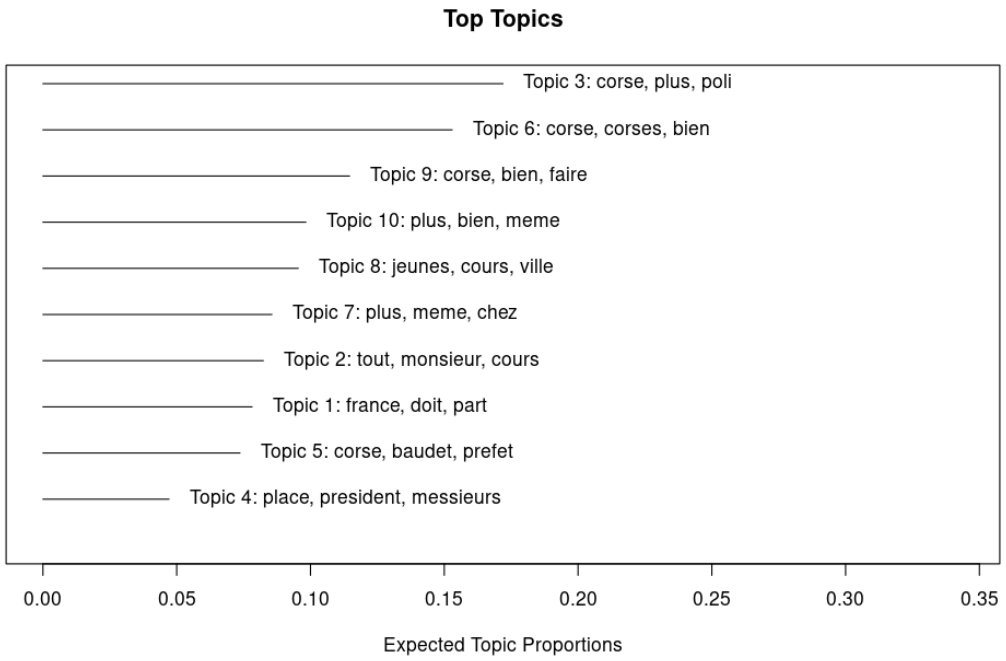


FIGURE 4.3 – *Top topics* des rubriques françaises, $k = 10$, avec les mots outils

En étudiant ces graphiques, on remarque assez rapidement deux problèmes principaux. Déjà, il est clair que la présence massive des mots outils dans les sujets modélisés constitue un problème dans l'interprétation des données, ce qui corrobore le choix de leur suppression. Ensuite, certains termes qui n'entrent pas dans la catégorie des mots outils sont clairement redondant comme « paoli » ou homonymiques comme les déclinaisons du mot « corse », ce qui était relativement prévisible. Comme nous l'avons vu précédemment, en se basant sur les mots outils mis en évidence par l'étape du lexicométrie, nous avons pu établir une liste de mots permettant d'affiner nos résultats obtenus dans la phase de *topic modeling*. Nous obtenons donc ces graphiques :

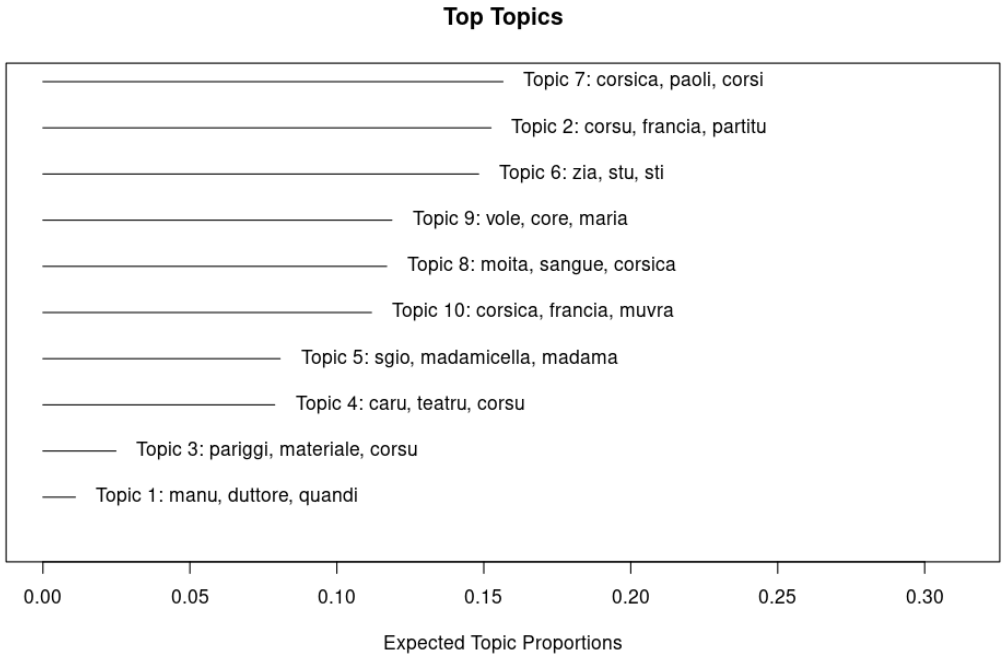


FIGURE 4.4 – *Top topics* des rubriques corses, $k = 10$, sans les mots outils

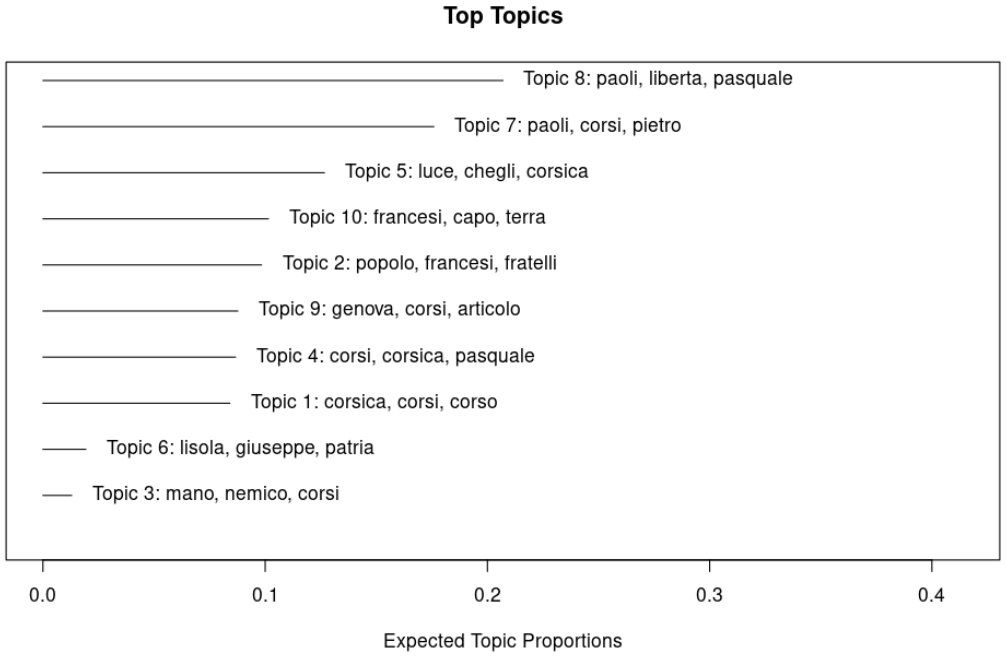


FIGURE 4.5 – *Top topics* des rubriques italiennes, $k = 10$, sans les mots outils

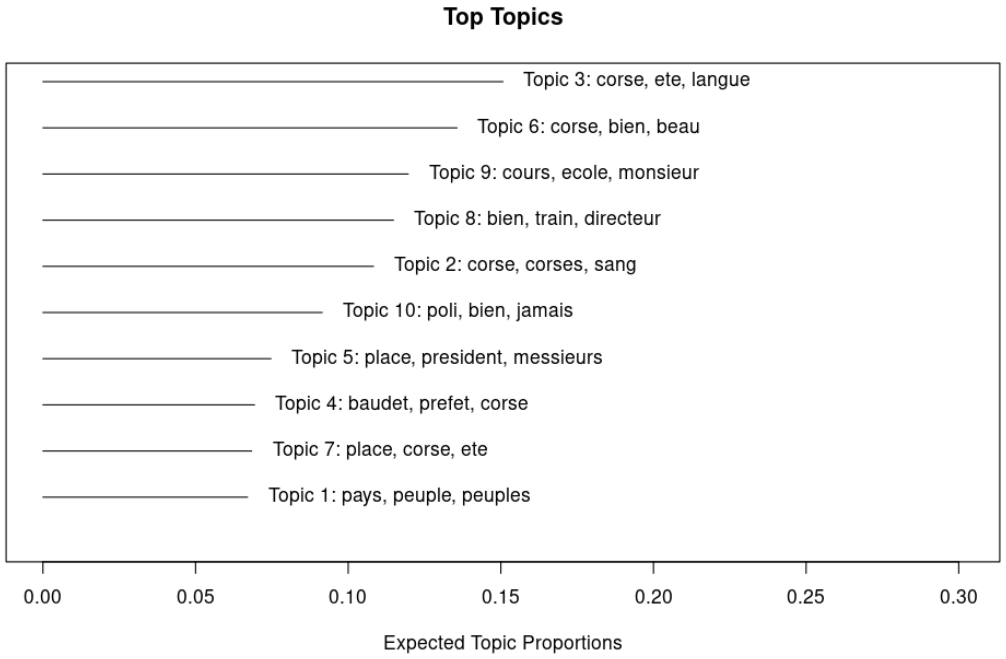


FIGURE 4.6 – *Top topics* des rubriques françaises, $k = 10$, sans les mots outils

Toujours dans cette logique de complémentarité des résultats, l'utilisation d'algorithmes sur Python a permis de mettre en évidence des nuages de mots particulièrement pertinents pour les sujets modélisés. L'avantage de la librairie est qu'elle fonctionne via des mots-clés qui, une fois déterminés, font émerger les sujets qui s'y rapprochent le plus. Néanmoins, du fait de l'entraînement d'un modèle propre au corpus, la librairie *Top2Vec* se heurte assez rapidement au problème de la taille des corpus. Si le modèle pour la langue corse est convaincant, celui pour la langue italienne l'est un peu moins [7]. Pour déterminer les mots-clés que nous utilisons, nous pouvons nous reposer sur les thèmes mis en évidence lors de mon mémoire de master en histoire contemporaine. Par exemple, observons les résultats avec le thème « histoire » [8] :

7. Sur les 169 rubriques, 110 sont en corse, 44 en français et seulement 15 en italien.

8. La liste des thèmes est disponible sur la base de données Heurist ou ici : Id., *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*..., p. 54.

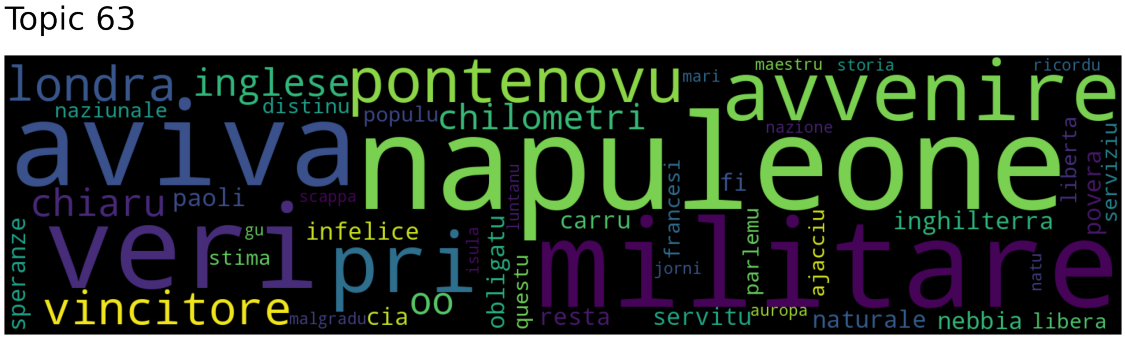


FIGURE 4.7 – Nuage de mots du sujet le plus pertinent avec le modèle corse, mot-clé = *storia*

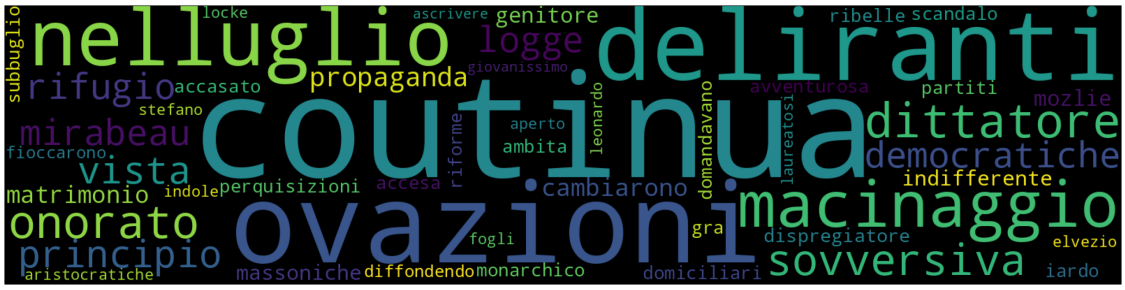


FIGURE 4.8 – Nuage de mots du sujet le plus pertinent avec le modèle italien, mot-clé = *storia*

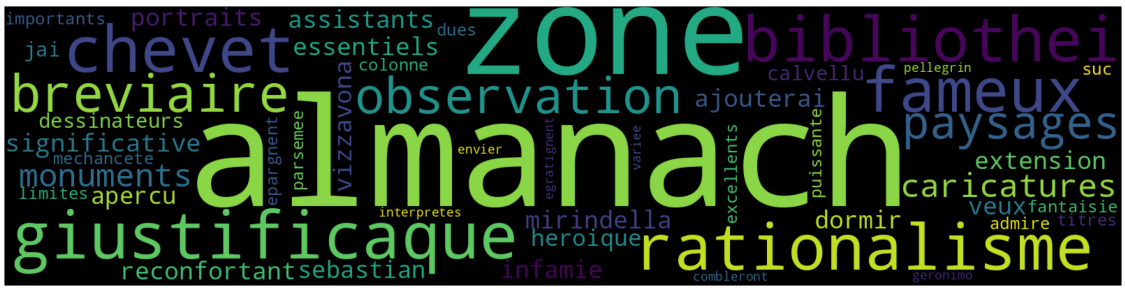


FIGURE 4.9 – Nuage de mots du sujet le plus pertinent avec le modèle français, mot-clé = *histoire*

Les combinaisons de paramètres à prendre en compte et de mots-clés sont nombreuses mais elles ne sont pas toutes pertinentes. Dans ce cas précis, on voit bien la limite de l'emploi de Python avec le nuage de mot sur les rubriques italiennes qui ne fait pas forcément beaucoup sens par rapport aux autres graphiques. Mais maintenant que nous avons des résultats, il est temps de les interpréter et de les expliquer.

4.2 Interprétation des résultats

En observant les sujets modélisés dans le graphique 4.4, on remarque que les thèmes les plus abordés dans leur ensemble (voir les graphiques 4.1, 4.2 et 4.3) sont les thèmes historiques. Cela peut déjà s'expliquer par la taille du corpus. Ce dernier contient un numéro dédié au bicentenaire de la naissance de Pasquale Paoli [9]. Cela influe bien évidemment sur les résultats étant donné la taille réduite du corpus mais cela s'explique d'un point de vue qualitatif. En effet, Pasquale Paoli est considéré comme le plus grand héros corse par les muvristes, l'un des trois avec Napoléon Bonaparte et Sampiero Corso. Cela témoigne de l'importance du *Babbu* dans l'imaginaire autonomiste et de son rôle dans les mythes fondateurs de la nation corse. Cette pratique de surreprésentation d'une figure historique constitue ainsi une forme de « statuomanie » qu'Ange-Toussaint Pietrera définit comme tel :

> « Choisir d'entreprendre la statuomanie d'un héros implique a fortiori de faire fructifier sa légende. C'est aussi et surtout l'occasion de sensibiliser le public à ses exploits, augmentant par là même toute l'émulation à son égard, et faire ainsi fructifier l'attente. Cette étape discursive sera désignée sous le terme de « puissance d'évocation ». Les promoteurs chantent ses exploits, mentionnent sa « galaxie humaine », y désignent lieutenants et traîtres, mais n'en oublient pas l'aspect revendicatif de leur démarche. » [10]

Mais la modélisation ne met pas en évidence que les grands axes de l'histoire de l'île, on remarque que les écrivains en langue corse s'intéresse aussi à l'histoire locale comme par exemple avec le *Topic 8 - moita, sangue, corsica* qui peut clairement faire référence au siècle des révolutions corses [11]. On a donc l'évocation d'une grande histoire violente et glorieuse personnifiée par des personnages, comme Paoli, mais aussi par des lieux. Le nuage de mots 4.7 semble confirmer cette question avec des termes comme *militare* ou encore *pontenovu* en rapport avec la fameuse bataille décisive contre l'armée française en 1769. Mais les rubriques corses semblent relativement variées dans le contenu publié dans les colonnes de la revue avec des allusions au théâtre dialectal d'*A Muvra* [12]. Il y a également des éléments plus classiques d'un journal, à l'instar du *Topic 5 - sgio, madamicella, madama* (« monsieurs », « mademoiselle » et « madame ») qui fait référence

9. Exemplaire du 5 avril 1925 (Sest'annata | Numaru 198).

10. A.T. Pietrera, « La construction des héros corses durant la Troisième République. Le cas de Sampiero et Paoli », dans *Héros, mythes et espaces. Quelle place du héros dans la construction des territoires*, 2016, p. 24.

11. Moïta est un village situé en Castagniccia, dans l'Est de la l'île. Cette région faisait partie des plus agitées pendant les troubles du XVIII^e siècle et Moïta ferait partie des premiers villages à rejoindre Paoli dans sa lutte.

12. L'une des créations de Petru Rocca avec son journal est le *Teatru dialettale di A Muvra*, qui avait pour but la mise en scène de pièces de théâtre en langue corse. Pour en savoir plus, voir : Christelle Hodencq, *Une « certaine » histoire du Théâtre (en) Corse à partir de l'expérience singulière du Teatru paisanu de Dumenicu Tognotti*, mém. de mast., Université de Paris 3, 2018.

aux notices d'actualités (*Nutizie di Corsica*) dans lesquelles on annonce par exemple des décès ou des mariages.

De même, il semblerait que la langue italienne ne soit utilisée que pour parler de sujets historiques. La plupart des articles italiens sont issus de publication de l'historien italien Oreste Ferdinando Tencajoli [13]. Cela fait néanmoins sens, l'action des irrédentistes a longtemps été principalement culturelle pour éviter d'éventuels incidents diplomatiques avec la France, en particulier dans la décennie 1920.

Les thèmes abordés par la langue française sont, en revanche, relativement intéressants. Le nuage de mots 4.9 montre que le sujet en lien avec le mot-clé « histoire » évoque principalement les almanachs. On peut en déduire que la langue française sert notamment à la publicité démontrant une volonté de toucher un plus grand nombre de lecteurs. Mais des thèmes davantage politiques sont évoqués comme avec le *Topic 4 - baudet, prefet, corse* évoquant de manière sarcastique les relations tumultueuses entretenues avec la haute magistrature française et les autonomistes corses [14].

13. On sait peut de choses de ce personnage. Historien spécialisé en histoire médiévale et religieuse, il a beaucoup travaillé sur la Corse et sa relation avec Gênes. Il a effectué quelques voyages dans l'île et a même fait l'objet d'une expulsion au cours de l'année 1932. Il devient rapidement un auteur régulier pour la revue irrédentiste *Corsica antica e moderna* dans laquelle il s'attelle à raconter l'histoire des cardinaux originaires de Corse.

14. Le préfet représente aux yeux des corsistes la plus haute autorité de l'État en Corse et fait l'objet de satyres régulières. L'allusion à l'âne est un comparaison récurrente dans le vocabulaire muvriste pour tourner en ridicule les adversaire du comité éditorial, en témoigne d'ailleurs le titre de notre mémoire.

Conclusion

Les premiers résultats que nous avons pu exprimer sur ce corpus réduit démontre des perspectives intéressantes pour la suite. Il semblerait que la thématique que les auteurs souhaite aborder influence le langage utilisé pour le faire. Si le corpus est principalement constitué de rubriques en langue corse, cela s'explique déjà par la volonté des auteurs de la diffuser plus largement. Mais elle remplit également une fonction plus pratique qu'idéologique : outre le fait d'être la langue d'expression des oeuvres poétiques, le corse s'emploie dans les rubriques de tous les jours et tout le journal se structure autour de celle-ci.

En revanche, la langue italienne a un emploi très particulier. Déjà, les muvristes ne l'utilise que très peu, le corpus italien étant constitué principalement de reprises de publications d'auteurs italiens dans d'autres journaux comme *L'Archivio storico di Corsica*, revue livournaise irrédentiste, ou *L'Idea nazionale*, périodique romain nationaliste. Ainsi, les *topics* qui ressortent le plus sont très orientés vers l'histoire et la culture ce qui coïncide avec l'orientation des publications irrédentistes des années 1920. De plus, les rares fois où les muvristes écrivent directement en italien, c'est pour s'adresser à ce public à fortiori très limité d'académiciens et journalistes originaires de la Péninsule.

Enfin, la langue française possède un rôle communicatif, destinée à atteindre le plus grand nombre. Si comme le corse elle sert à aborder des sujets politiques, elle est davantage utilisée pour s'adresser directement à des membres de l'administration ou à des adversaires politiques. Le français est utilisé de plusieurs manières mais sert avant tout à diffuser des idées plus largement qu'avec le corse, allant finalement un peu à contresens de l'idéologie corsiste. L'intérêt de faire du *topic modelling* sur *A Muvra* n'était pas de montrer les principales thématiques abordées, mais d'aller un peu plus dans le détail de ces thématiques pour mieux faire ressortir les différences. Et le cas des rubriques françaises montrent bien cette nuance.

La frontière entre les langues est poreuse et n'est pas totalement imperméable aux exceptions. L'avantage d'effectuer une lecture distante sur un corpus de presse est d'aller outre ces spécificités pour avoir une approche plus globale de la situation. Mais il s'agit aussi de la principale limite de notre étude. Pour faire une lecture distante efficace de la revue *A Muvra*, le corpus devrait être beaucoup plus conséquent que ces quelques numéros, même si cela nous a déjà donné une vue d'ensemble assez pertinente.

Il reste de nombreux points d'améliorations pour cette recherche. Dans un premier temps, il faut améliorer la qualité de nos données afin de pouvoir les structurer plus facilement. Puis dans un second temps, il faut mettre à notre disposition des données en plus grand nombre en particulier pour effectuer une analyse stylométrique. Ces premiers résultats démontrent néanmoins le potentiel du traitement automatique du langage sur la

langue corse, même si le développement d'outils numériques est un enjeu crucial pour faciliter le travail de traitement des données. Mais nous avons pu voir l'intérêt des humanités numériques dans l'étude de la presse dialectale autonomiste de l'entre-deux-guerres.

Bibliographie

Bernhard (Delphine) et Ligozat (Anne-Laure), « Es esch fàscht wie Ditsch, oder net ? Étiquetage morphosyntaxique de l'alsacien en passant par l'allemand », dans *TALARE 2013*, 2013.

Bernhard (Delphine) et Soria (Claudia), « Traitement automatique des langues peu dotées », *Traitement Automatique des Langues*, 59–3 (2018).

Camps (Jean-Baptiste), Clérice (Thibault) et Pinche (Ariane), « Noisy medieval data, from digitized manuscript to stylometric analysis : Evaluating Paul Meyer's hagiographic hypothesis », *Digital Scholarship in the Humanities*, 36–Supplement_2 (2021).

Chagué (Alix) et Clérice (Thibault), *HTR-United : Ground Truth Resources for the HTR and OCR of patrimonial documents*.

Chiffoleau (Floriane), *dahncorpus*, version 1.0.0, mars 2021, doi : 10.5281/zenodo.5911868.

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

— *Essais de linguistique corse*, Ajaccio, Alain Piazzola, 2000.

Dalbera-Stefanaggi (Marie-José) et Medori (Stella Retali), « Trente ans de dialectologie corse : autour du programme Nouvel Atlas Linguistique et Ethnographique de la Corse et Banque de Données Langue Corse », dans *Tribune des chercheurs en linguistique*, Société des Sciences Historiques et Naturelles de la Corse, 2013.

Delporte (Christian), Blandin (Claire) et Robinet (François), *Histoire de la presse en France : XXe-XXIe siècles*, Paris, Armand Colin, 2016.

Eder (Maciej), « Rolling stylometry », *Digital Scholarship in the Humanities*, 31–3 (2016).

Fracastoro (Giulia), Magli (Enrico), Poggi (Giovanni), Scarpa (Giuseppe), Valsesia (Diego) et Verdoliva (Luisa), « Deep learning methods for synthetic aperture radar image despeckling : An overview of trends and perspectives », *IEEE Geoscience and Remote Sensing Magazine*, 9–2 (2021).

Getz (Jasmine), « Histoire, poésie, transmission », *Les Temps Modernes*–4 (2011).

Gherardi (Eugène François Xavier), *Précis d'histoire de l'éducation en Corse. Les origines : de Petru Cirneu à Napoléon Bonaparte*, Ajaccio, CRDP de la Corse ; A Meridiana, 2011.

Giacomo-Marcellesi (Mathée), « Le corse », dans *Histoire sociale des langues de France*, Rennes, Presses universitaires de Rennes, 2013.

Hennecke (Inga), « Petits corpus oraux bilingues et plurilingues–enjeux théoriques et méthodologiques », *Corpus*–18 (2018).

Hodencq (Christelle), *Une « certaine » histoire du Théâtre (en) Corse à partir de l'expérience singulière du Teatru paisanu de Dumenicu Tognotti*, mém. de mast., Université de Paris 3, 2018.

Kevers (Laurent) et Medori (Stella Retali), « Copyright in the context of tooling up Corsican and other less-resourced languages », dans *International Conference on Language Technologies for All (LT4All), Enabling Linguistic Diversity and Multilingualism Worldwide*, 2019.

— « Towards a Corsican Basic Language Resource Kit », dans *12th Language Resources and Evaluation Conference (LREC 2020)*, 2020.

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

Klinkenberg (Jean-Marie), « « Grandes langues » et langues minoritaires : deux politiques linguistiques ? », *Lengas. Revue de sociolinguistique*–79 (2016), https ://journals.openedition.org/lengas/1048.

Kremnitz (Georg), Broudic (Fañch) et Garabato (Carmen Alén), *Histoire sociale des langues de France*, Rennes, Presses universitaires de Rennes, 2013.

Lee (Moontae), Bindel (David) et Mimno (David), « Robust Spectral Inference for Joint Stochastic Matrix Factorization », dans *Proceedings of NIPS 2015*, 2015.

Léglise (Isabelle) et Alby (Sophie), « Les corpus plurilingues, entre linguistique de corpus et linguistique de contact : réflexions et méthodes issues du projet CLAPOTY », *Faits de langues*, 41–1 (2013).

Minervini (Laura), *Filologia romanza. Linguistica*, t. 2, Milano, Le Monnier Università, 2021.

Otsu (Nobuyuki), « A threshold selection method from gray-level histograms », *IEEE transactions on systems, man, and cybernetics*, 9–1 (1979).

Paci (Déborah), « Le dialogue des élites méditerranéennes à travers les médias au XIXe siècle : le cas de Malte et de la Corse », *Cahiers de la Méditerranée*–85 (2012).

— *Il mito del Risorgimento mediterraneo : Corsica e Malta tra politica e cultura nel ventennio fascista*, thèse de doct., Université de Nice Sophia-Antipolis, 2013.

— « Le mare nostrum fasciste : l'espace politique et culturel en Corse et à Malte à l'époque du fascisme italien », *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines*, 128–2 (2016).

Pellegrinetti (Jean-Paul), « Sociabilité républicaine en Corse de 1870 à 1914 : Mutation d'une société », *Cahiers de la Méditerranée*, 56–1 (1998).

Pellegrinetti (Jean-Paul) et Rovere (Ange), *La Corse et la République. La vie politique, de la fin du second Empire au début du XXIe siècle*, Paris, Média Diffusion, 2013.

Pietrera (Ange-Toussaint), *Imaginaires nationaux et mythes fondateurs ; la construction des multiples socles identitaires de la Corse française à la geste nationaliste*, thèse de doct., Université Pascal Paoli, 2015.

— « La construction des héros corses durant la Troisième République. Le cas de Sampiero et Paoli », dans *Héros, mythes et espaces. Quelle place du héros dans la construction des territoires*, 2016.

— « La Corse contemporaine au prisme du XVIIIe siècle français : de l'enracinement républicain à l'affirmation nationaliste », *Astérion. Philosophie, histoire des idées, pensée politique*–24 (2021).

Polacci (Daniel), *Les autonomistes corses de l'entre-deux-guerres*, mém. de mast., Université d'Aix-Marseille, 1974.

Poli (Jean-Pierre), *Autonomistes corses et irrédentisme fasciste (1920-1939)*, Ajaccio, Éd. DCL, 2007.

Porte (Guillaume), « Alsatia Numerica », *Source(s). Cahiers de l'équipe de recherche Arts, Civilisation et Histoire de l'Europe*–2 (2013).

Retali-Medori (Stella), « Le programme Nouvel Atlas Linguistique et ethnographique de la Corse-Banque de Données Langue Corse (NALC-BDLC) », dans *«Identité linguistique de la Corse et de la Sardaigne : aires, strates, systèmes dans l'espace insulaire et roman/Linguistic identity of Corsica and Sardinia : areas, strata, systems in the insular and Romance space»*, 2021.

Retali-Medori (Stella) et Kevers (Laurent), « La morphologie dans la Banque de Données Langue Corse : bilan et perspectives », *Corpus*–23 (2022).

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 doct., Paris 10, 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*, mém. de mast., Université de Strasbourg, 2021.

— *Corsican stylometry : ressources and dataset for corsican NLP*, version 2.0.4, juin 2022, URL : https://github.com/vincentsarbachpulicani/Corsican-Stylometry.

Vergez-Couret (Marianne), « Tagging occitan using french and castillan tree tagger », dans *Less Resourced Languages, new technologies, new challenges and opportunities*, 2013.

Viaut (Alain), « Marge linguistique territoriale et langues minoritaires », *Lengas. Revue de sociolinguistique*–71 (2012).

Table des figures

1 Schéma de la base de données relationnelle 6
2 *Pipeline* de notre projet de recherche 8

1.1 Exemple d'utilisation du logiciel ScanTailor 13

2.1 Exemple de segmentation d'une page de une d'un numéro d'*A Muvra* . . . 16

3.1 Histogramme des MFW-15 de l'analyse exploratoire 27
3.2 Carte de chaleur des MFW-15 de l'analyse exploratoire 28

4.1 *Top topics* des rubriques corses, $k = 10$, avec les mots outils 31
4.2 *Top topics* des rubriques italiennes, $k = 10$, avec les mots outils 31
4.3 *Top topics* des rubriques françaises, $k = 10$, avec les mots outils 32
4.4 *Top topics* des rubriques corses, $k = 10$, sans les mots outils 33
4.5 *Top topics* des rubriques italiennes, $k = 10$, sans les mots outils 33
4.6 *Top topics* des rubriques françaises, $k = 10$, sans les mots outils 34
4.7 Nuage de mots du sujet le plus pertinent avec le modèle corse, mot-clé = *storia* 35
4.8 Nuage de mots du sujet le plus pertinent avec le modèle italien, mot-clé = *storia* 35
4.9 Nuage de mots du sujet le plus pertinent avec le modèle français, mot-clé = *histoire* 35

Table des matières

Résumé i

Abstract i

Remerciements ii

Introduction 2

0.1 L'émergence d'une lutte régionaliste et identitaire . . . 2
0.1.1 Les prémices d'une affirmation culturelle . . . 2
0.1.2 La naissance du muvrisme . . . 2
0.1.3 Le rôle central de la langue . . . 3
0.2 État de l'art . . . 3
0.2.1 Un sujet d'étude particulier ? . . . 3
0.2.2 La Corse et les humanités numériques . . . 4
0.2.3 Travaux individuels . . . 5
0.3 Objectifs . . . 6
0.3.1 Approches méthodologiques . . . 6
0.3.2 Problématisation . . . 8

I Préparation des données 10

1 Traitement des images 11
1.1 Récupération des images . . . 11
1.2 Phase de nettoyage . . . 13

2 Mise à disposition des données textuelles 15
2.1 Le défi de l'océrisation . . . 15
2.1.1 Un cas d'analyse de segmentation complexe . . . 15
2.1.2 Le choix du moteur d'OCR . . . 17
2.2 Exploitation des données brutes . . . 18
2.2.1 Structuration en XML . . . 19

2.2.2 Nettoyage automatique du texte 21

II Valorisation des données 23

3 Analyse lexicométrique d'un corpus en langue corse 24

3.1 Méthodologie et limites 24

3.2 Des premiers résultats à saisir 26

4 *Topic modeling* : faire émerger des sujets latents 30

4.1 Visualisation des données 30

4.2 Interprétation des résultats 36

Conclusion 39

Bibliographie 41

Table des figures 45