École nationale des chartes — PSL | Università di Pisa

MASTERS THESIS IN DIGITAL HISTORY

Authors profiling in the Corsican autonomist press during the interwar period

Stylometric analysis and topic modeling on *A Muvra*

Student:

Vincent Sarbach-Pulicani

Graduated with a bachelor's degree in History

Graduated with a master's degree in "History and Civilisations of Europe"

Advisors:

Jean-Baptiste Camps

École nationale des chartes — Université PSL

Alessandro Lenci

Università di Pisa

Humanités numériques et computationnelles
Informatica umanistica

October 4, 2023

Contents

Acknowledgements **6**

Abstracts **7**

Introduction **10**

Rethinking the study of Corsican autonomism **12**

1 A multidisciplinary approach to Corsican studies **13**
1.1 The emergence of a regionalist and identity-based struggle . . 13
1.1.1 The first sign of a cultural affirmation 13
1.1.2 The birth of *A Muvra* 14
1.1.3 Defending the insular particularism 16
1.2 State of the art 19
1.3 Methodological aims 22

2 More than a newspaper, a real cultural movement **24**
2.1 The language diversity 24
2.2 The depth of the typology 30
2.3 Which authors to study? 35
2.3.1 The pseudonyms 35
2.3.2 The candidates 37

The technical stakes of the Muvra study **39**

3 Retrieving the images **40**
3.1 The layout of the *Muvra* 40
3.2 The importance of document provenance 43
3.2.1 Collecting the newspapers from *Gallica* 43

3.2.2 Web scraping from the *Archives départementales de Corse du Sud* . . . 44
3.3 Do we need to clean the images? . . . 46

4 Collecting textual data **49**
4.1 The challenge of newspapers segmentation . . . 49
4.1.1 Technical difficulties . . . 49
4.1.2 Training a model . . . 51
4.2 Making data available . . . 54
4.2.1 Data extraction . . . 54
4.2.2 Data structuration . . . 58
4.2.3 Data standardisation . . . 59

5 Exploring our data **61**
5.1 General informations . . . 61
5.2 Diachronic comparative statistics . . . 65

A Muvra, a symbol of pluralism **69**

6 Identifying author's pseudonyms by their writing style **70**
6.1 Choosing the method . . . 70
6.2 Are anthologists right? The case of *P.diB.* . . . 74
6.2.1 Explanations and first tests . . . 74
6.2.2 Confirming results . . . 77
6.3 Who's hiding behind the mysterious *Altore*? . . . 81
6.3.1 An active contributor . . . 81
6.3.2 A pseudonym of the playwright *U Sampetracciu*? . . . 83

7 The specific roles for the pseudonyms used **93**
7.1 General informations about topic modeling . . . 93
7.2 *Altore*, the perfect cover? . . . 96
7.3 Petru Rocca and its pseudonyms . . . 102
7.4 A diversified heart of authors . . . 107

Conclusion **112**

Bibliography **117**

Acknowledgements

I would first like to thank my two research supervisors, Jean-baptiste Camps in Paris and Alessandro Lenci in Pisa, for their help and availability over the last years.

Je pense bien évidemment à mes parents et toute ma famille pour leur soutien, malgré la distance et les différentes péripéties induites par ce double diplôme réalisé à l'Université de Pise et à l'École nationale des chartes.

Vorrei ringraziare i vari docenti e ricercatori italiani per i loro preziosi consigli e la loro pazienza durante il mio soggiorno in Toscana, tra cui Alessandro Bondielli, Laura Passano, Angelo Mario Del Grosso e Federico Boschetti.

Un autre remerciement à mes différents collègues et amis du master qui sont restés à l'École cette dernière année, en particulier Matenia pour sa positivité et pour toute l'aide qu'elle m'a apporté pour les questions d'HTR.

Ringrazio gli amici italiani di Pisa per l'accoglienza e l'aiuto che mi hanno dato durante tutto l'anno. Penso in particolare a Javier, che m'ha aiutato a correggere l'inglese di alcune parti della tesi.

Infine, cum'è sempre, vuleriu ringrazià i mo amici corsi per a so curiosità è l'aiutu chì m'anu datu in l'imparera di u corsu. Senza i lizzioni di dumenica à u *Little Ajaccio*, sta tesi seria statu più difficiule.

Abstracts

English

With the emergence of nationalism in the 19th century came regionalist movements to assert and claim cultural particularities. Corsica fitted in very well with this dynamic and even presented itself as a favourable location for the development of such ideas. The centralisation of the State around a strong capital and the policies of assimilation of the indigenous populations on the border with France led certain players to defend these particularisms. It was in this context that the Corsican autonomist newspaper *A Muvra* was born in May 1920 in Paris, under the impetus of Petru and Matteu Rocca. For almost 19 years, hundreds of authors participated in the writing of this massive dialectal work. The aim of this dissertation is to carry out author profiling, i.e. to determine the style and subjects covered by an author. To do this, we carry out authorship attribution stylometry on texts using pseudonyms before completing these analyses with topic modelling, indexing of latent topics in a corpus of texts. The aim is to gain a better understanding of the complex sociology behind this rich and varied newspaper, through the use of computational methods.

Français

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. C'est dans ce contexte que naît le journal au-

tonomiste corse *A Muvra* en mai 1920 à paris, sous l'impulsion de Petru et Matteu Rocca. Pendant presque 19 années, des centaines d'auteurs vont se succéder dans la participation à la rédaction de cette oeuvre dialectale massive. L'objectif de ce mémoire de recherche est d'effectuer du profilage d'auteurs, c'est-à-dire déterminer le style et les sujets abordés par un auteur. Pour ce faire, nous effectuons de la stylométrie d'attribution d'autorité sur des textes utilisant des pseudonymes avant de compléter ces analyses avec du *topic modeling*, de la modélisation de sujets latents d'un corpus de textes. Le but est de comprendre un peu mieux la sociologie complexe derrière cette revue riche et variée, grâce à l'emploi de méthodes computationnelles.

Italiano

Con la nascita del nazionalismo nel XIX secolo sono nati i movimenti regionalisti per affermare e rivendicare le particolarità culturali. La Corsica si inserì molto bene in questa dinamica e si presentò addirittura come un luogo favorevole per lo sviluppo di tali idee. La centralizzazione dello Stato attorno a una forte capitale e le politiche di assimilazione delle popolazioni indigene al confine con la Francia portarono alcuni attori a difendere questi particolarismi. Fu in questo contesto che nel maggio 1920 nacque a Parigi, sotto l'impulso di Petru e Matteu Rocca, il giornale autonomista còrso *A Muvra*. Per quasi 19 anni, centinaia di autori parteciparono alla stesura di questa imponente opera dialettale. L'obiettivo di questa tesi è quello di realizzare un profilo di autori, ossia di determinare lo stile e gli argomenti trattati da un autore. A tal fine, eseguiamo una stilometria di attribuzione di autorità su testi che utilizzano pseudonimi, prima di completare queste analisi con la modellazione di argomenti latenti in un corpus di testi, cioè il *topic modeling*. L'obiettivo è quello di comprendere meglio la complessa sociologia che sta dietro a questo giornale ricco e varie, attraverso l'uso di metodi computazionali.

Corsu

Cù la spuntera di u naziunalismu in u XIXsimu seculu sò nati e mosse righjunaliste per affirmà è rivindicà e particularità culturali. A Corsica s'inseriscì bè in sta dinamica è si prisentò cum'è un locu avantaghjosu per u sviluppu di

tal'idee. A centralizzazione di u Statu intornu à una forte capitale è e pulitiche d'assimilazione di e pupulazioni indigene à u cunfinu incù a Francia porteranu alcuni attori à difendà sti particularismi. Fù in stu cuntestu ch'è in u maghju 1920 nascì à Parigi, sottu l'abbriu di Petru Rocca è u so fratellu Matteu, u ghjurnale autonomista corsu *A Muvra*. Per guasi 19 anni, centinaiu d'autori participeranu à a ridazzione di st'opera dialettale maiò. L'ughjettivu di sta tesi è quellu di rializzà un prufilu d'autori, vale à dì ditarminà u stile è l'argumenti trattati da un autore. Cusì, eseguimu una stilumetria d'attribuzione d'autorità nant'à testi ch'è apprudanu pseudonimi, prima di integrà una fasa di *topic modeling*, a mudellizazione d'argumenti latenti in un corpus di testi. L'ughjettivu hè quellu di capisce megliu a cumplessa suciolugia ch'è sta daretu à stu ghjurnale riccu è varie, attraversu l'usu di metodi computazionali.

Introduction

Corsican studies have focused at length on the Corsican autonomist press of the inter-war period, in particular the newspaper *A Muvra*. Founded in 1920 by Petru Rocca and his brother, they were active for nearly 20 years until the outbreak of the Second World War. During these two decades, several hundred authors contributed to the weekly output of the journal. Many of the studies carried out on this movement, which can be described as *muvrism*, are mainly qualitative studies. The historiographical renewal of Corsican automatism began around the 2000s in order to bring a fresh perspective to the subject. This work is part of this desire to revitalise historical studies of contemporary Corsica, and more specifically of the inter-war period. It follows on from previous work, including a first thesis in quantitative history on the newspaper compared with an irredentist periodic, and a first-year thesis in digital humanities on the question of the use of languages in rubrication.

The aim of this thesis is therefore to perform stylometry and topic modelling algorithms in order to profile anonymous authors. This involves attributing authorship to specific pseudonyms and understanding their role. We will therefore carry out a study for certain specific authors that we will see in due course, but also consider them as a group. While the notion of individuality is essential to understand the personal trajectories of the writers of the *Muvra*, the notion of group is just as important when we know the importance of the political activism of the muvrists. Given the large number of authors who have contributed to the journal, we will have to make choices which will be explained in the course of the thesis. The analyses and results are all available on a GitHub repository dedicated to this research[1].

[1]Vincent Sarbach-Pulicani, *Corsican stylometry : ressources and dataset for corsican NLP*, version 2.0.4, June 2022, URL: `https://github.com/vincentsarbachpulicani/Corsican-Stylometry`.

Taking into account all this information, we will try to determine the profile of the main authors of the *Muvra*. The analysis will be done in two complementary phases. First, we will perform a stylometric analysis of authorship attribution for the main authors in order to determine the pseudonyms used by them. Following this task, we will carry out topic modeling in order to highlight the main themes addressed by these authors and the utility of pseudonyms in the case of Corsican autonomists during the interwar period. The idea is thus to understand how pseudonyms have a practical and moral usefulness for the muvrists and how it varies according to the individuality.

This work is divided into three main parts in order to answer these questions. The first takes the form of a preliminary section, the aim of which is to present the issues at stake in this research in greater detail. It covers the state of the art and methodological aspects, as well as the historical matters surrounding the publication of the *Muvra*. The second part is devoted more to the technical challenges associated with the study of such a newspaper and the solutions found to meet our needs. Finally, the third part is devoted to presenting and interpreting the results in a complementary manner.

Rethinking the study of Corsican autonomism

Chapter 1

A multidisciplinary approach to Corsican studies

1.1 The emergence of a regionalist and identity-based struggle

1.1.1 The first sign of a cultural affirmation

With the emergence of nationalisms in the 19th century, regionalist movements of affirmation and claims of cultural particularities were grafted on. Corsica fits in very well with this dynamic and is even a favourable place for the development of such ideas. The centralisation of the state around a strong capital and the policies of assimilation of the indigenous populations on the border with France led some actors to defend these particularisms. With the Jacobin tradition of the "one and indivisible Republic", sometimes widely exaggerated, it is notably from the Second Empire and the reign of Napoleon III that the francization of Corsica took on its full meaning[2]. This resulted in the mandatory learning of French at school instead of Corsican, plans for the economic and industrial revival of the territory or the massive participation of Corsicans in the colonial effort.

From then on, the first fragments of a regionalist fight on the island were born, that is to say by the defence of the Corsican language, central in the preservation of identities. This is the case of Santu Casanova, poet and editor

[2]See: Marco Cini, *Un'integrazione nazionale imperfetta: élite e culture politiche in Corsica nella prima metà dell'Ottocento*, Roma, Viella, 2022.

of the journal *A Tramuntana* of which he is also the founder in 1896. One of the outcomes of the nationalisms of the 19th century is well known: the First World War. The disaster that this event engendered was very much felt on the island, both in demographic and social terms. Even today, the number of islanders who died during what they call the *scumpientu*[3] is difficult to estimate. Moreover, it gave rise to many ideological debates about the sacrifice of the Corsicans for the *Grande Patrie*[4] or in the name of a war that did not concern them.

Nevertheless, the first real mention of the notion of autonomy in Corsica comes with the periodical *A Cispra* published in 1914 in a single issue. Founded by the dialect poets Saveriu Paoli and Ghjacumu Santu Versini, the introduction to this publication mentions autonomy as a salvation for the rebirth of the Corsican nation:

> "So francization must be postponed. So we must ask for the recognition of the Corsican Nation. And when Corsica has become a nation again, when it has become aware of itself, its individualism, after inevitable trials and tribulations, will give an otherwise intense, otherwise admirable performance. With Autonomy, we will have in addition an intellectual life, an economic life, a social life."[5].

. Even though this periodical did not last more than one year, the *Cispra* remained in the memory of their successors as a model to follow for the defence of the Corsican identity in the same way as the *Tramuntana* of Santu Casanova.

1.1.2 The birth of *A Muvra*

The growing distrust towards the French government is embodied by the foundation of the autonomist journal *A Muvra* in 1920 by the war veteran Petru Rocca and his brother Matteu. Written in Corsican language, this weekly

[3]Meaning "disaster" in Corsican language.

[4]Meaning "Great Fatherland" in French language.

[5]*A Cispra*, p. V: « Donc il faut repousser la francisation. Donc il faut demander la reconnaissance de la Nation corse. Et quand la Corse sera redevenue une nation, quand elle aura pris conscience d'elle-même, son individualisme, après d'inévitables tâtonnements donnera un rendement autrement intense, autrement admirable. Nous aurons avec l'Autonomie, nous aurons par surcroît une vie intellectuelle, une vie économique, une vie sociale. ».

magazine represents the hard line of the island regionalism and is influenced by several french nationalists players such as Charles Maurras. Claiming to be a cultural publication, its political dimension thoughout the interwar period made it very controversial inside the island society. This aspect of the "muvrism" is embodied by the political party associated, the *Partitu Corsu d'Azione*[6] founded by the Rocca's brothers in 1922. The authors published in paralell numerous books and almanacs through their own publishing house, the *Stamparia di a Muvra*. Written in French, Corsican or Italian, these publications integrate perfectly the spiritual legacy of Santu Casanova, Saveriu Paoli and Ghjacumu Santu Versini of which the muvrist claim to be clearly the heirs.

In its early years, the *Muvra* did not worry the French authorities very much. However, the radicalisation of the action during the 30's forced the Government to take repressive measures against the activists that led to the censorchip of the journal in 1939. First, their revendications tended to be regionalists but a suite of outside and inside factors to Corsica drove them to pure separatism. The similarity of some arguments and the evident proximity of some muvrists with the Italian authorities increased the mistrust from the *commissaires spéciaux*[7] towards the corsists. As a matter of fact, the intensification of the fascist propaganda toward the rattachment of Corsica to Italy increased all along the period known as the *ventennio fascista*. The island was an important target for the regime for reasons both strategical and ideological. As part of a long doctrinal tradition dating from the 19th century, those known as "irredentists" have largely based their propaganda on the autonomist movements internal to the island society. The italian historian Deborah Paci gives a definition of the irredentism in an article of 2016:

> The notion of irredentism dates back to the 19th century and refers to the cultural and political movement that follows the political doctrine of the annexion of all italian speaking territories ; areas that were not "freed" yet (irredent land).[8].

[6]Meaning "Corsican's Action Party" in Corsican language, on the basis of the sardinian *Partidu Sardu - Partito Sardo d'Azione*. The name changed for *Partitu Corsu Autonomista* in 1929.

[7]Meaning "Specials superintendent". They were agents attached to the Ministry of the Interior. He was in charged to write reports about the political and social activities of a French department including the autonomists.

[8]Déborah Paci, "Le mare nostrum fasciste: l'espace politique et culturel en Corse et à

The theorist of irredentism was the Italian nationalist Scipio Sighele, who believed that the Risorgimento would never be complete until all irredents lands joined the Italian Nation. Moreover, he believed that Italians were not yet "one Italian soul: we have fragmented and weakened the great national pride that would make us strong in the world, into a multitude of futile regional vanities"[9]. Irredentism went hand in hand with the Italian nationalist theories developed in the 19th century to justify the creation of an Italian nation state. With the advent of Fascism, Corsican irredentism materialised with the creation of a *Comitato per la Corsica* with the aim of supporting propaganda for the attachment of the island to Italy. Thanks to the financing of the regime, a whole series of publications were published with the aim of reviving the Italianness of the island's society. In addition to attracting many muvrists to Italy, the committee started to finance the Corsican autonomist newspaper from the mid-1920s. Thus, many topics were raised in their project of autonomy for Corsica, both cultural and political, with the defence of the language first.

1.1.3 Defending the insular particularism

The language quickly becomes a strong argument in the Corsican ideology. Then, using the Corsican language in the written press becomes essential to preserve it, but also to change mentalities in everyday life. This observation is shared by a whole set of local presses, whether they are autonomist newspapers like *A Muvra* or *A Baretta Misgia*, but also the regionalist unionist press like *L'Annu Corsu*[10]. Thus, the appearance of the *Muvra* in the Corsican media landscape is not an exception: the awareness of the necessity to write Corsican to make it live is more general and is shared by several movements with a large variety of political opinions, at least at their beginning. The defence of the island's linguistic heritage involved various demands such as the recognition

Malte à l'époque du fascisme italien," *Mélanges de l'École française de Rome - Italie et Méditerranée modernes et contemporaines*, 128–2 (2016), p. 440: « La notion d'irrédentisme remonte au XIX^{e} siècle et désigne le mouvement culturel et politique ayant pour doctrine politique l'annexion de tous les territoires de langues italiennes ; des espaces qui n'étaient pas encore « libérés » (terre irrédente). ».

[9]Olivier Bosc, "De la foule criminelle à la foule nationaliste: Scipio Sighele, théoricien de l'irrédentisme," *Matériaux pour l'histoire de notre temps*, 43–1 (1996), p. 44.

[10]Periodical founded by Paul Arrighi, it was the main press of the cyrneism, Corsican regionalist movement.

of its existence but also its teaching at school and at the university[11]. But the themes evoked by the muvrists remained vrey broad. Whether it was religious conservatism or the maintenance of the memory of nationalist militants during the Corsican revolutions of the 18th century, the weekly journal enjoyed a wide range of articles and topics, with a very large number of authors.

Studying Corsica means taking into account the cultural and geographical particularities inherent to its condition as an island. However, although this may seem obvious, studying an insular environment requires specific knowledge and tools for reflection so that the interpretation of historical facts is not biased by our "continental" point of view. One of these tools is the concept of "insularism". This term can take on different meanings and should therefore be handled with care. Even if it is quite difficult to find a proper definition of this word in an English dictionanry, we can use a French one. The Larousse dictionary defines insularism as the "tendency of an insular people to lock themselves into their island and to reduce their international relation"[12]. Although relevant, the latter has a very political connotation and seems incomplete from a historiographical angle. Among the works on insularism, we can cite those of François Taglioni[13]. But Deborah Paci also gives an interesting definition:

> The category of analysis constituted by "insularism" allows us to observe that the awareness of belonging to a common territory is the consequence of the interaction between space, culture and the environment.[14].

This notion is important to understand in order to fully understand our subject. If we think about it further, we can admit that the regionalism and identitarianism shown by the muvrists is a direct consequence of this idea of insularism. But the most interesting thing with the study of the Corsican

[11] The University of Corsica was founded for the first time in 1765 by the *Babbu di a Patria* Pasquale Paoli. However, it was dissolved in 1769 when the small island became part of the Kingdom of France and it was not until 1981 that the university was reopened under the name of University of Corsica-Pascal-Paoli

[12] « Tendance d'un peuple insulaire à s'enfermer dans son île et à réduire ses relations internationales. » in: `https://www.larousse.fr/dictionnaires/francais/insularisme/43489`

[13] François Taglioni, "L'insularisme: une rhétorique bien huilée dans les petits espaces insulaires," in *Comme un parfum d'îles*, Paris, Presse Universitaire Paris-Sorbonne (PUPS), 2010.

[14] D. Paci, "Le mare nostrum fasciste: l'espace politique et culturel en Corse et à Malte à l'époque du fascisme italien"…, p. 456.

movement during the interwar period is its collective dimension. Whereas Corsican dialectal production was mainly the result of individual literature and spontaneous actions, the journal *A Muvra* offers us regular publications defined within an editorial setting. But if this collective dimension is assumed by those concerned, the individualism of the muvrists remains essential in the life of the newspaper. This translates into political opinions that can vary, mainly in the attitude to be had towards Fascist Italy, or even in the practice of the language where the idiomatic characteristics of the Corsican language take on its full meaning. And it is this aspect that we will focus on in our study.

1.2 State of the art

Studies on Corsican autonomism are relatively numerous and of high quality. The first academic work on the question is a master's thesis of 1974 written by Daniel Polacci at the University of Aix-Marseille[15]. He made a statistical study centred on the muvrists and the production processes of the periodical. But it is really from the 2000s onwards that more concrete work flourishes. We can cite various quality thesis: the one of Ysée Rogé[16] which was the first work to put corsism and irredentism into perspective, then the one by Deborah Paci[17] who compared Corsican irredentism with Maltese irredentism and finally Ange-Toussaint Pietrera[18] who studied the movement in the context of the emergence of nationalism. But in parallel to these studies, research in late modern history has been particularly flourishing concerning Corsica, especially over the last 20 years. In this regard, the work carried out by Jean-Paul Pellegrinetti is essential. He wrote an excellent synthesis on Corsica and the Third Republic which is still an authority on the subject[19]. But the recent history of Corsica testifies to the sensitivity of the topic in the collective memory of Corsicans. The politicisation of historical studies is a risk that must be taken into account, hence the need to constantly update historiography. Maurice Agulhon, who wrote the preface to Pellegrinetti's book, himself wrote the following:

> Not everyone will like this book, which is republican, and therefore pro-French, and some will no doubt reproach him [...] for having this history of "republican" Corsicans endorsed by a continental preface.[20].

[15] Daniel Polacci, *Les autonomistes corses de l'entre-deux-guerres*, MA thesis, Université d'Aix-Marseille, 1974.

[16] Ysée Rogé, *Le corsisme et l'irrédentisme 1920-1946: histoire du premier mouvement autonomiste corse et de sa compromission par l'Italie fasciste*, PhD thesis, Paris 10, 2008.

[17] D. Paci, *Il mito del Risorgimento mediterraneo: Corsica e Malta tra politica e cultura nel ventennio fascista*, PhD thesis, Université de Nice Sophia-Antipolis, 2013.

[18] Ange-Toussaint Pietrera, *Imaginaires nationaux et mythes fondateurs; la construction des multiples socles identitaires de la Corse française à la geste nationaliste*, PhD thesis, Université Pascal Paoli, 2015.

[19] Jean-Paul Pellegrinetti and 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.

[20] Ibid., p. 9: « 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. »

From a personnal angle, this study is the continuation of a first master's thesis carried out at the University of Strasbourg. The latter dealt with a comparative analysis between the journal *A Muvra* and an Italian irredentist journal, *Corsica antica e moderna*, between 1932 and 1939[21]. Without concentrating on a particular aspect of this corpus, it was more of a close reading of a limited corpus. The idea was to highlight the major ideological differences between the corsists and the irredentists, despite an obvious proximity. If the corsists admitted to being part of a common cultural and linguistic entity with Italy, they did not share the same desire for political unification, even if some autonomists came closer to Fascist ideas just before the beginning of the Second World War.

For several years, there has been a desire in Corsica to structure and study the evolution of the use of the Corsican language. We can notably mention the work of the linguist Marie-José Dalbera-Stefanaggi with her *Nouvel atlas linguistique et ethnographique de la Corse*, the first volume of which was published in 1995[22]. In the republications of this major work in the 2000s, the author incorporated her work on the creation of a *Banque de Données Langue Corse* (BDLC)[23]. This is the first initiative in the will to lemmatise the Corsican language in its diachrony. The contribution of digital humanities to this matter is quite new. The reflections of researchers on the tooling of regional languages using Natural Language Processing have really accelerated since the second half of the 2010s. Our approach is therefore in line with this continuity. Faced with the lack of tools for processing and analysing the Corsican language[24], OCR processing of textual data so that they can be exploited in digital humanities research represents an important issue in the future of the discipline in Corsica. Writing my first research paper required the creation of a comprehensive database of the two journals studied. It was structured according to the following model:

[21] 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*, MA thesis, Université de Strasbourg, 2021.

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

[23] `https://bdlc.univ-corse.fr/bdlc/corse.php`

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

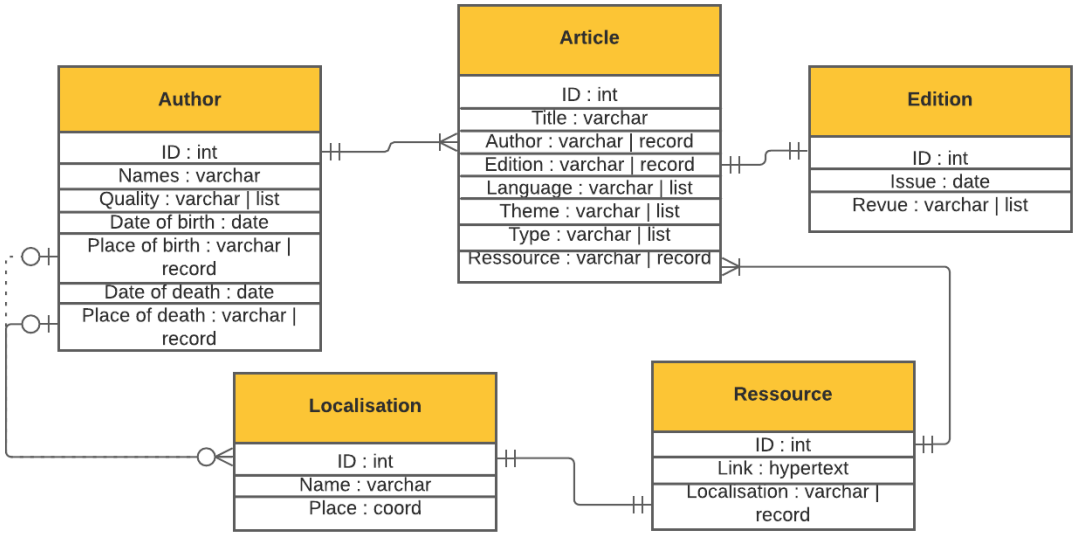


Figure 1.1: Entity-relationship diagram of the A/I database

This database, named *Autonomists/Irredentists Database* (A/I databse), was created with the help of the the archives of the department of South Corsica[25] (*Archives départementales de Corse du Sud*). It consists of 4323 entries grouping together articles, issues, authors or themes. The aim was to make precise statistical studies on this period in order to support my argument. This database was made by hand and may contain errors, representing one of the limits to such a methodology. Our desire to go deeper into this subject therefore requires new digital and computational approaches.

[25] `https://heurist.huma-num.fr/heurist/?db=vsp_presse_corsiste_irredentiste`

1.3 Methodological aims

If our study is part of the historical sciences, the very nature of our subject must include linguistic methodology. As we have already stated, we will carry out a distant reading of our corpus. Our objective here is to extend the observation already made by taking a step back on the corpus and to analyse it as a whole using digital tools. This work is therefore part of a double challenge: that of better understanding the way of thinking of the muvrists in their time but also that of helping to perpetuate the use of digital humanities in Corsican studies. One of the difficulties is the composition of a trilingual corpus. In corpus linguistics, this is not necessarily a problem if the texts are comparable in number, genre and type[26]. In our case, it is difficult to have a totally homogeneous corpus, since the distribution of languages and the thematic and typological aspects are issues of analysis in their own right. Similarly, insofar as this is a corpus of cultural press, the themes are very varied, as are the types of articles. Topic modeling resulting from our corpus will thus have to take into account these metadata for the analysis to be relevant. Nevertheless, if multilingual linguistics lends itself well to large corpora, is it effective on smaller corpora? Indeed, our corpus is composed of 12 issues from the years 1924 and 1925, which corresponds to 169 articles. This question is at the heart of Inga Hennecke's 2018 article:

> In recent decades, the aim of corpus linguistics has been to compile increasingly large corpora in order to be able to carry out quantitative and qualitative research and analysis on as representative a sample as possible. This development went hand in hand with advances in computer science and quantitative linguistics. It is only in the last few years that small corpora have again been taken into consideration by the scientific community.[27].

While our small corpus may appear to be a limitation in our desire to carry out a distant reading of the *Muvra*, the researcher gives us others significant advantages. Even if she focuses on small oral corpora, some of the considerations in this article can be assimilated to our own study, such as the possibility

[26] Isabelle Léglise and 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.

[27] Inga Hennecke, "Petits corpus oraux bilingues et plurilingues–enjeux théoriques et méthodologiques," *Corpus*–18 (2018), p. 2.

of conducting qualitative research in parallel[28]. As we will see later, several NLP methods are used in our study. Although we will have the opportunity to detail it, we will make comparisons between different algorithms in order to try to obtain the best possible results. For topic modeling for example, it is interesting to compare the results between LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis) and we will see that the quality of the textual data is essential in our case. But coupling these results with close reading would validate them. Distant reading is not enough when working on such a subject, it is necessary to provide additional elements in order to fully understand and interpret the outputs obtained. For example, in the case of the modeling of a dominant subject by an author, demonstrating this through articles actually published by the latter is important to illustrate this result.

[28]Ibid., p. 14.

Chapter 2

More than a newspaper, a real cultural movement

2.1 The language diversity

As we have already mentioned, the language occupies an essential place in the muvrist propaganda. The preservation of the Corsican language has always been a major aspect of the fight of the different regionalists, it is the very reason of the existence of these dialectal newspapers. The reciprocal is also admissible if we take into account the republican will to normalise the use of the French language. If the second half of the 19th century is decisive in the teaching of French in schools, this will already goes back to the revolutionary period and the First Republic. As early as the *An II* of the Revolution, the politician Bertrand Barère "had chosen four privileged targets: Lower Brittany, Corsica, the Basque country and Alsace"[29]. For the revolutionaries, this fight against language was above all a fight against Corsican separatism led by Pasquale Paoli. Faced with the threat that the latter represented, knowing that he was supported by the English crown, it seemed necessary to take direct action on the island in order to francize it as much as possible, with the aim of limiting Paoli's influence on the population. It is not necessarily required to return to the question of public education, particularly the upheavals linked to the turn of the 19th century with reformers such as François Guizot or later Jules Ferry. It is nevertheless interesting to look at the place of language teaching

[29]Françoise Mayeur, *Histoire de l'enseignement et de l'éducation III. 1789-1930*, Paris, Editions Perrin, 2004, p. 47.

in secondary education. If, in the inter-war period, the place of humanistic disciplines predominates, the teaching of modern languages is not really in the public eye because its humanistic values are being questioned[30]. What can be said is that the teaching of French was exclusive and the use of the regional language was strictly prohibited, even within the confines of the schools.

Indeed, it is easy to understand why the Corsican language is of particular importance in the hearts of Corsican autonomists. In addition to the intrinsically cultural character that is related to linguistic regionalism, there is also a notion of identity inherent in this nationalist ideology. Jean-Paul Pellegrinetti states in an article that "language and people are intimately linked for the defence of identity originality"[31]. Inverting the French universalist ideal, the French language becomes a foreign language in the same way as Italian or English. Its teaching is therefore not rational for young Corsicans who must nevertheless speak it as if it were natural for them. Using the Corsican language in the written press then became essential for its preservation but also with a view to changing the relationship Corsicans have with their own language. And this observation is shared by a whole set of local presses, whether they are autonomist newspapers like *A Muvra* or *A Baretta Misgia*, but also the cyrneist regionalist press like *L'Annu Corsu*. Therefore, it seems to be more of a social fact than a simple Corsican fad. In fact, the lack of language learning becomes a problem for these Corsican newspapers which can only reach a limited public: a public that can read, but above all a public capable of deciphering the Corsican language.

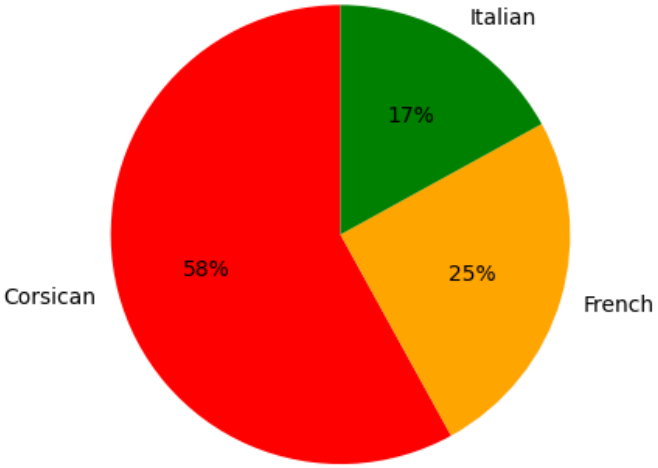
An essential question then arises for our authors: which language for Corsica? If the use of French is in the majority in the first years of publication of the review, Corsican occupies an increasingly preponderant place from the first half of the 1920s. It is difficult to make a precise estimate, since such statistics are not at the heart of this study. However, we can base ourselves on the figures obtained from the year 1932 onwards from my first thesis on this matter[32]. We can also make a comparison from the complete analysis that has

[30]Pierre Albertini and Dominique Borne, *L'école en France XIXe-XXe siècle de la maternelle à l'université*, Paris, Hachette, 1992, p. 102.

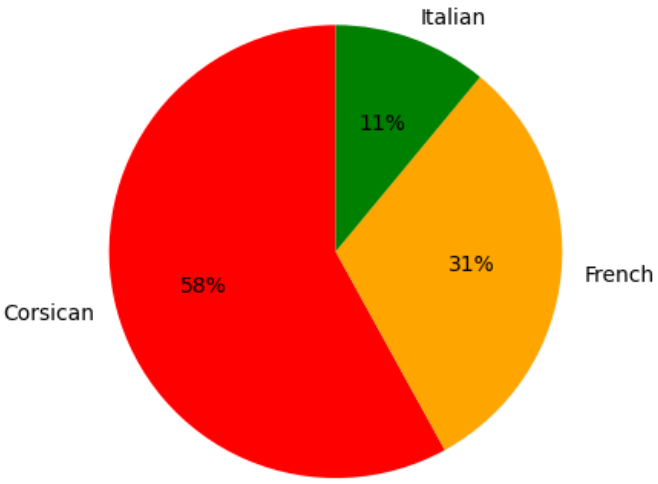
[31]J.P. Pellegrinetti, "Langue et identité: l'exemple du corse durant la troisième république," *Cahiers de la Méditerranée*–66 (2003), p. 267.

[32]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.

be done on the issues coming from the *Bibliothèque nationale de France*[33]:



(a) Issues from 1925 to 1930



(b) Issues from 1932 to 1939

Figure 2.1: Language distributions in the *Muvra* at different time

[33]We will get to it later, but the overall quality of digitalization of the issues coming from the BnF (Gallica) is way better than the Corsican Regional Archive., allowing a good covering of the articles.

It can be observed that from the second half of the 1920s onwards, there is little difference in the distribution of the languages, except for French, which gains slightly on Italian in the following decade. This can be explained by the financing of the newspaper. The *Muvra* was financed by François Coty, a perfumer from Ajaccio who was close to *L'Action Française*, until 1924, which would explain the preponderance of French as the muvrists were still simple regionalists. Then it was Italian funding that supported the magazine via the *Comitato per la Corsica*. The increase in international tensions and the return of censorship in the second half of the 1930s would explain the need for the muvrists to communicate again in French, in order to make themselves understood by their political opponents.

The Corsicans' relationship with the Italian language was somewhat peculiar, leading to increasing confusion about their true political intentions. The standardisation of the Italian language is the result of political and cultural decisions dating back to the 19th century in particular. In Italy, in the 1860s, the debate between ALessandro Manzoni and Graziadio Isaia Ascoli led to the choice of Modern Florentine Tuscan, a central Italo-Romance dialect, for its "prestige"[34]. This was the evolution of the language of Dante Alighieri, a key figure in the Italian *Risorgimento* and associated nationalist movements more generally. It is a phenomenon usually observed in European countries during the century of nationalisms, this desire to unite the national community around a common language. The muvrists themselves do not dispute the resemblance to standard Italian, as the priest Dominique Carlotti points out, referring to the island dialect as a "mixture of transversal terms"[35]. From a strictly linguistic point of view, Corsican is an Italo-Romance language that has had central influences (Pisan Tuscan) but also northern influences (Genoese Ligurian) and southern influences (Neapolitan and Sardinian), but it does belong to the Tuscan group[36]. It is itself divided into different linguistic areas with two main ones, the *Cismontincu* spoken in the North of the island and the *Pumuntincu* spoken in the South. Nevertheless, as early as June 1922, they affirmed their desire to distinguish their language from the standardised

[34]Laura Minervini, *Filologia romanza. Linguistica*, vol. 2, Milano, Le Monnier Università, 2021, p. 90.

[35]*A Muvra*, n°113-1923/07/15: « Insiste chì u Corsu è un mischiu di termini straversi, cartaginesi, saracini, aragunesi, greghi, tuscani e francesi [...]. ».

[36]For more information, see the work of Marie-José Dalbera-Stefanaggi, including: M.J. Dalbera-Stefanaggi, *Essais de linguistique corse*, Ajaccio, Alain Piazzola, 2000.

Italian, Florentine Tuscan:

> To differentiate ourselves, as far as possible, from the Italian with which most of the metatics are too prone to confuse us. Our language must constitute — which is, without question, a real fact — a remarkable entity, an intrinsic quality, so that all the Larousse of France and Navarre no longer say that we speak Tuscan.[37].

The author of this article is Marcu Angeli, a famous Corsican poet and future medical student at the University of Pisa. To see this character speak of the cultural differences between Corsica and Italy is surprising, given that a few years later he became the personal physician of Clara Petacci, Benito Mussolini's mistress, with whom he would form a friendship. But another aspect that the muvrists do not appreciate is the notion of "dialect". If this term is used quite regularly by the various authors and in all its forms (*dialect*, *dialettu* or *dialetto*), it is more the question of the status of Corsica in relation to the majority languages of the Latin sisters. Thus, in 1923, Ghjuvan'Petru Lucciardi wrote this:

> If for that, we were obliged to rely a little often on Latin, you think perhaps that it would lose the right to call itself a language? No, because French, Spanish and other languages that rely heavily on it are not called dialects. It is up to Corsican writers to come to an agreement.[38].

This highlights the importance for regionalists of producing a large quantity of texts of a mainly oral language in the early inter-war period, the process of literarisation of a language. This is part of one of the three phases of "linguistic regionalism" defined by the semiotician Jean-Marie Klinkenberg. Indeed, it could correspond to the defensive and cultural moment:

[37] *A Muvra*, n°65-1922/06/04: « Nous différencier, autant que faire se peut, de l'Italien avec lequel la plupart des métèques sont trop portés à nous confondre. Il faut que notre langue constitue — ce qui est, sans conteste, un fait réel — une entité remarquable, une intrinsèquité, afin que tous les Larousse de France et de Navarre ne radotent plus que nous parlons toscan. ».

[38] *A Muvra*, n°127-1923/10/21: « Si per quessa, fussimu obligati d'appughiacci un pocu spessu sopra u latinu, cridareste forse ch'ella perdissi u dirittu di chiamassi lingua ? Innò, perchè u frencese, u spagnolu e altre lingue chi ci s'appoghianu lergamente, un so micca chiamati dialetti. Tocca a i scrittori corsi a mettesi d'accordu. ».

> The linguistic work on regional languages thus reproduces the work that was then being done on national culture: publishing and making known the authors of the past, illustrating the language in the present. The movement is the work of a category of actors who are at the same time writers and scholars. These writers combine their creative work with their philological quest, carefully distinguishing between the two.[39].

For the corsists, this translates into the valorisation of more or less ancient authors such as the modern historian Pietro Cirneo or the 19th century poet Salvatore Viale. The latter is an interesting case because in addition to criticising the process of francization of the island, his work was considered by corsists and irredentists as one of the "bridges" between France and Italy, in addition to his friendship with many Italian nationalists of the time including Niccolò Tommaseo[40]. This philological quest is also highlighted by the structuring and standardisation of the language, despite the great phonetic and grammatical diversity of Corsican. It was at this time that the first grammars appeared[41]. But this common structure remains marginal, each author writes as he speaks. The will to preserve the Corsican language as well as the different regiolects which make its particularity offers us an important quantity of texts with idiomatic which are useful for our analyses in particular, with regards to stylometry. If the first 20th century represents the beginnings of the standardisation of Corsican, this one obtains an orthographic structure only after the Second World War, during the movement of *riacquistu*[42].

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

[40]M. Cini, *Gli « Studii critici di costumi corsi » di Salvatore Viale. Il processo di modernizzazione della Corsica nel XIX secolo*, Roma, L'Harmattan Italia, 2018 (Il Politico e La Memoria), p. 11.

[41]Antone Bonifacio, *A prima grammatichella corsa*, Bastia, Editions de "l'Annu corsu", 1926.

[42]This term designates the cultural movement in Corsica for the " reappropriation " of the language and of the insular identity: Anne Meistersheim, "Du riacquistu au désenchantement: Une société en quête de repères," *Ethnologie française*, 38–3 (2008).

2.2 The depth of the typology

The typology of the articles in the *Muvra* is particularly varied. In fact, we can find on the same pages classical articles, open letters as well as poems, songs and plays. This great diversity can be explained in part by the phenomenon of literarisation of the Corsican language that we mentioned earlier, with the desire to produce different writings in large quantities. But it is also necessary to take into account the very essence of what *A Muvra* is: a newspaper. In fact, its main role is to transmit information and, in the case of the corsist paper, ideas to its readers. It is therefore normal that it follows the rules and dogmas inherent to the contributors' socio-professional category. Roselyne Ringoot, a PhD lecturer in language sciences and lecturer at Sciences Po Rennes, provides us with some interesting information on this subject:

> In journalism, proximity is established as a law. Sometimes formalised in journalism guides, but more often transmitted and conveyed informally within editorial offices, the law of proximity conditions editorial choices and field practices; "the law of the dead mile" is an overused example.[43].

What we learn from this quote is that while there may be common practices across the profession, these evolve and are adapted according to the needs of newsrooms. Indeed, the categorisation we make as researchers is influenced by our own biases in perceiving what a news story is. Should it be different from a front page story? What is a serial? Does it fall into the category of a column? So many questions that are also influenced by the bias of the subject of our research.

These choices are therefore essential as they will influence the results of our analyses. For example, in the autonomist/irredentist database, the typology was organised according to these categories:

[43]Roselyne Ringoot and Yvon Rochard, "Proximité éditoriale: normes et usages des genres journalistiques," *Mots. Les langages du politique*–77 (2005), p. 73: « La proximité, en journalisme, est érigée en loi. Parfois formalisée dans des guides du journalisme, mais plus souvent transmise et véhiculée de façon informelle au sein des rédactions, la loi de proximité conditionne des choix éditoriaux et des pratiques de terrain ; « la loi du mort-kilomètre » en est un exemple galvaudé. ».

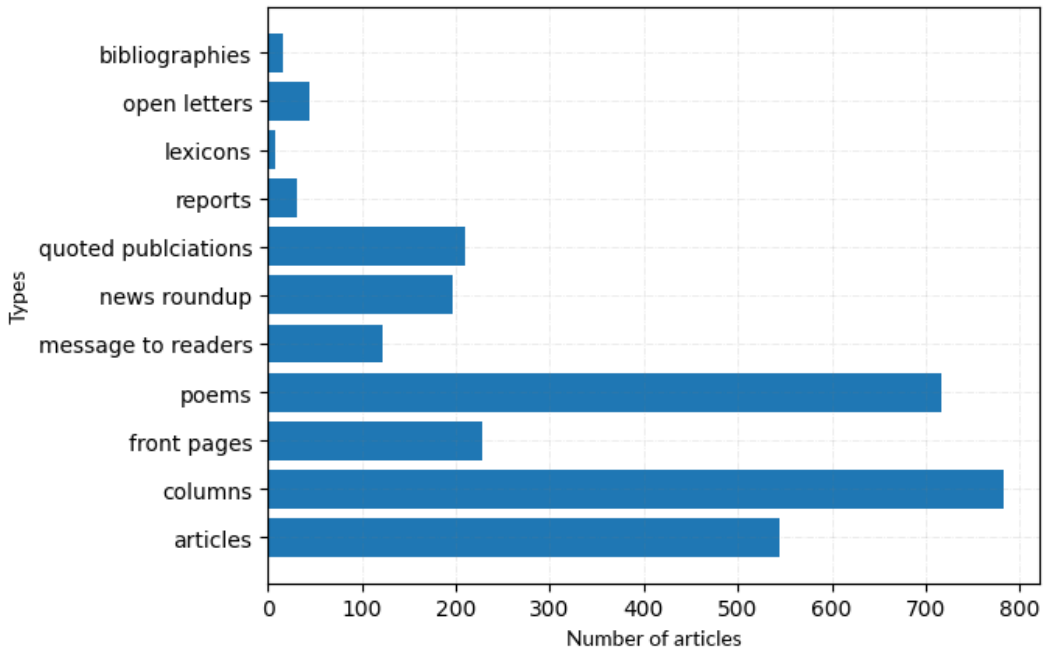


Figure 2.2: Typological categorization in the A/I database

It seems clear that the typology adopted here is not optimal for stylometric analysis or topic modeling. In the first case, the importance is to have textual data resulting from natural language, without necessarily having different poetic licenses that would alter the way of writing. Stylometry using rhymes exists[44] but our study is not concerned with this method. Similarly, bibliographies do not seem particularly relevant for topic modeling. Taking these considerations into account, we can therefore establish our typological categorization chosen for our present study:

- **Article** (art), all "traditional" articles describing an event or a political opinion for example.
- **Chant** (cha), the songs and music present in the newspaper. The scores being sometimes present, cause noise during the OCR.
- **Diverse** (div), any special articles, such as bibliographies or messages to readers.
- **Letter** (let), open letters or articles in the form of a letter.

[44]Voir: Mike Kestemont, Walter Daelemans, and Dominiek Sandra, "Robust rhymes? The stability of authorial style in medieval narratives," *Journal of Quantitative Linguistics*, 19–1 (2012).

- **Play** (the), the plays, having a redundancy of certain names or an alternation between purely literary language and spoken language.
- **Poetry** (poe), poems in verse or prose, are present in very large numbers in the *Muvra*.
- **Serial novel** (feu), in french "*roman-feuilleton*", a journalistic practice that was very present, particularly in the 19th century.
- **Story** (his), narrative forms such as tales and legends.
- **Translation** (trd), ancient or contemporary texts translated from another language.

Let us look at some of these headings. Although most of them are fairly obvious, others have a particular place in the muvrists discursive practices. Poetry and songs occupy a special place in the island's culture through the *paghjelle*, the *chjam'è rispondi* and the *lamenti*. Their transmission in writing is a way of preserving their existence. In the same way, the writing of new poems follows a long tradition of Corsican poets who had already begun the work in the previous century, such as Santu Casanova or Salvatore Viale. The work of the muvrists in this sense was not only limited to the publication of the journal and its almanac, but also by the organisation of annual poetic picnics, the *merendelle d'i pueti corsi*. Studies have already been made by some researchers on Corsican poetry, notably Paul Desanti. The latter has highlighted a series of recurrent themes in the island's poetic production by focusing on three main authors of the inter-war period: Marcu Angeli, Anton'Francescu Filippini and Petru Giovacchini[45]. There is a recurrence of the notion of 'isula persa', the lost island, to evoke the poor economic and social situation of the island on the fringes of the other French metropolitan territories. Desanti relates the production of these *lamenti* to the more global theme of the "literature of abandonment"[46].

Similarly, the writing of plays is part of this same more global context of artistic production. Nevertheless, these are rarely unique creations intended

[45]Paul Desanti, *Trois poètes corses irrédentistes. M. Angeli, P. Giovacchini, A.F. Filippini*, Ajaccio, Albania, 2013.

[46]Idem, *« Gigli di stagnu » di Marco Angeli, un' avvinta literaria*, MA thesis, Università di Corsica Pasquale Paoli, 1997, p. 16: « S'iscriva dunqua senza prublema issa puesia in ciò ch'è Pascal Marchetti chjamarà in *La corsophonie* la « littérature de l'abandon ». E dalli tandu un sensu propiu è solu fascistu ùn saria di sicura criditoghju. ».

for the editorial staff of the Corsican newspaper. Among the many activities of the *Partitu Corsu Autonomista* and the *Muvra*, some contributors have written and directed a number of plays for the *Teatru dialettale di a Muvra*, a theatre company responsible for touring the island to present shows in the Corsican language. The main authors are Simon-Jean Vinciguerra, Simon-Paul Poli and Ghianettu Notini. If the quantity of dramaturgical works does not reach that of the poems, they are perfectly integrated in the will to materialize "cultural ferment aiming at affirming its own cultural and, especially linguistic"[47].

Finally, let us look at a genre that is specific to the press, the serial novel. This genre appeared in the French press in 1836 and consisted of the publication of novels in a newspaper on an episodic basis, with the author sometimes writing his work as he went along. Very popular with the public and with certain authors, such as Eugène Sue, this vision of literature was also criticised by others, such as Victor Hugo[48], for its sensationalist dimension. Generally speaking, Corsica has a particular link with this genre, with the publication in this format of *Le Comte de Monte Cristo* by Alexandre Dumas from 1844 onwards[49], or the short story *Colomba* by Prosper Mérimée in 1840[50]. Muvrists often take the latter as an example to evoke the numerous stereotypes about Corsica that these works have fed, sometimes to the extreme. While the serial novel continued to exist after the Great War, the professor of French literature Lise Queffélec-Dumasy explains that:

> If the serial novel retains a certain importance until the 1940s, we enter, from the beginning of the century, and especially after the First World War, into the modern history of mass culture, of which the serial novel is only the prehistor.[51].

[47] Christelle Hodencq, *Une «certaine» histoire du Théâtre (en) Corse à partir de l'expérience singulière du Teatru paisanu de Dumenicu Tognotti*, MA thesis, Institut d'Études Théâtrales de Paris 3, 2018, p. 16: « Ce rapide panorama montre à quel point la Corse du vingtième siècle est traversée par un ferment culturel visant à affirmer sa propre autonomie culturelle et, spécialement linguistique. La littérature, la poésie, l'édition et, bien qu'en mesure assez faible, la production d'œuvres dramatiques ont matérialisé ce ferment. ».

[48] Paradoxically, *Les Misérables* was adapted into a serial novel in 1888 in *Le Rappel*, 20 years after its first publication.

[49] In *Journal des débats*.

[50] In *La Revue des Deux Mondes*.

[51] Lise Queffélec-Dumasy, *Le roman-feuilleton français au XIXe siècle*, Belphégor: Littérature Populaire et Culture Médiatique, 2008, URL: `http://hdl.handle.net/10222/47746`, « Si le roman-feuilleton conserve une certaine importance jusque dans les années 1940, on

This element makes two things clear. On one hand, the use of such a procedure by the muvrists to put forward texts on Corsica written by foreign authors such as Domenico Guerrazzi[52], may appear as an attempt to reappropriate this genre in order to go against the stereotypes conveyed by the serial novels of the previous century. While on the other hand, it places the publication of the Muvra in a certain temporality. This press and its authors are part of a wider media environment with its own practices and customs, which even they cannot deny. The regional journalistic culture in France is very largely inspired by the large Parisian presses that still exist in the inter-war period, and *A Muvra* is no exception to the rule.

entre, dès le début du siècle, et surtout après la première guerre mondiale, dans l'histoire moderne de la culture de masse, dont le roman-feuilleton n'est que la préhistoire. ».

[52]For example: Domenico Guerrazzi, *Pasquale Paoli ossia la Rotta di Pontenuovo. Racconto Corso del Secolo XVIII*, Milano e Torino, M. Guigoni, 1860.

2.3 Which authors to study?

2.3.1 The pseudonyms

The use of a pseudonym is common in the press of the Third Republic, it is something inherent to the press of opinion. For if Muvra considers itself a cultural journal, there is a very important and assumed political dimension. Maurice Legaa, who wrote a book on the subject in 1986, returns to this essential notion in journalistic and literary production in general.

> The circulation of pseudonyms coincides with the influx of freedom and anarchy; but these energies and impulses do not make pure differences; they are invested in a set of norms, forms and regularities that recompose, on the margins of the official marks of nomination, the hypothesis of a system: a latent institution of pseudonyms, regulating the passage from identities to non-identities and vice versa, exists in our societies.[53].

Faced with the risk of censorship as well as multiple attacks, the corsairs used this process a lot to hide their identity. There is also a poetic dimension in the fact of choosing a pseudonym, paradoxically some authors did not really protect themselves with these, even going so far as to reveal their identity. This is why we can find authors with their associated pseudonyms in many anthologies or bibliographies. Much of the biographical information we have on Corsican authors comes from these extremely valuable documents. The most notable works are the re-edition of Hyacinthe Yvia-Croce's *Anthologie des écrivains corses*[54], Carmine Starace's *Bibliografia della Corsica*[55], and finally, the most recent, the *Antulugia di a Corsica litteraria*[56] published by several researchers in 2020.

[53]Maurice Laugaa, *La pensée du pseudonyme*, Paris, Presses Universitaires de France, 1986 (Écritures), p. 5-6: « La circulation des pseudonymes coïncide avec l'afflux d'une liberté et d'une anarchie; mais ces énergies,ces pulsions n'effectuent pas de pures différences; elles s'investissent dans un jeu de normes, de formes et de régularités qui recomposent, en marge des marques officielles de la nomination, l'hypothèse d'un système : une institution larvée des pseudonymes, réglant le passage des identités aux non-identités, et vice versa, existe dans nos sociétés. ».

[54]Hyacinthe Yvia-Croce, *Anthologie des écrivains corses*, Ajaccio, Ed. Cyrnos et Méditerranée, 1987.

[55]Carmine Starace, *Bibliografia della Corsica*, Milano, Istituto per gli studi di politica internazionale, 1943 (Centro di studi per la Corsica).

[56]Julian Mattei, Petru Santu Menozzi, and A.T. Pietrera, *Antulugia A Corsica Literaria*, Aiacciu, Albiana, 2020.

As we have understood, the choice of pseudonyms to be identified and of potential candidates is likely to depend on very important biases, hence the importance of crossing sources before making this essential choice. But it should be remembered that the authorship attribution is not the only aspect of analysis in our study, it is also necessary to make this choice according to the part on topic modeling. Author profiling will also involve determining the potential usefulness to authors of using a particular pseudonym when writing the article. These biases are also important because the choice of data processed will depend on them.

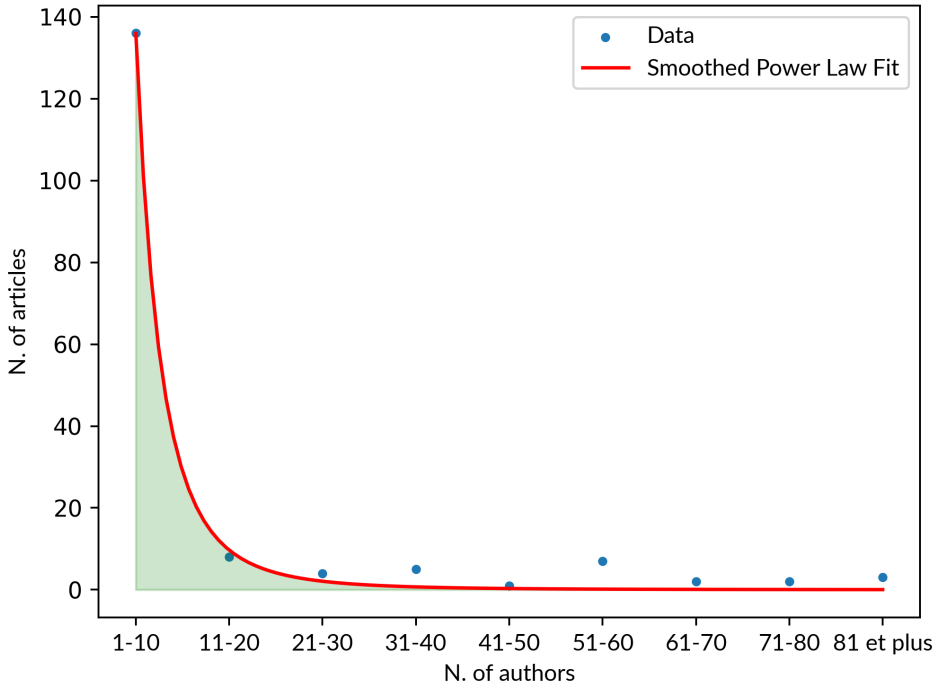


Figure 2.3: Distribution of authors according to the power law

There must be enough textual data for these pseudonyms for the analysis to be viable. As the graph 2.3 shows, if we have a collection of 501 different names, the vast majority of them have authorship of only one or two articles at most, which is far too few. Also, enough has to be written by the potential candidates to determine the authorities, but we will see that later.

2.3.2 The candidates

The choice of authors to be studied must be made from two levels of reflection: first from a technical point of view and then from a historical point of view. An author, in order to be relevant, must have written enough in the *Muvra* to have enough textual data for the stylometric analysis. Then we have to think about the real relevance of the author and his place in the editorial board of the corsist journal. In order to give a first idea of this task, here is a graph representing the distribution of the writing language of the authors for whom we have kept at least 15 articles.

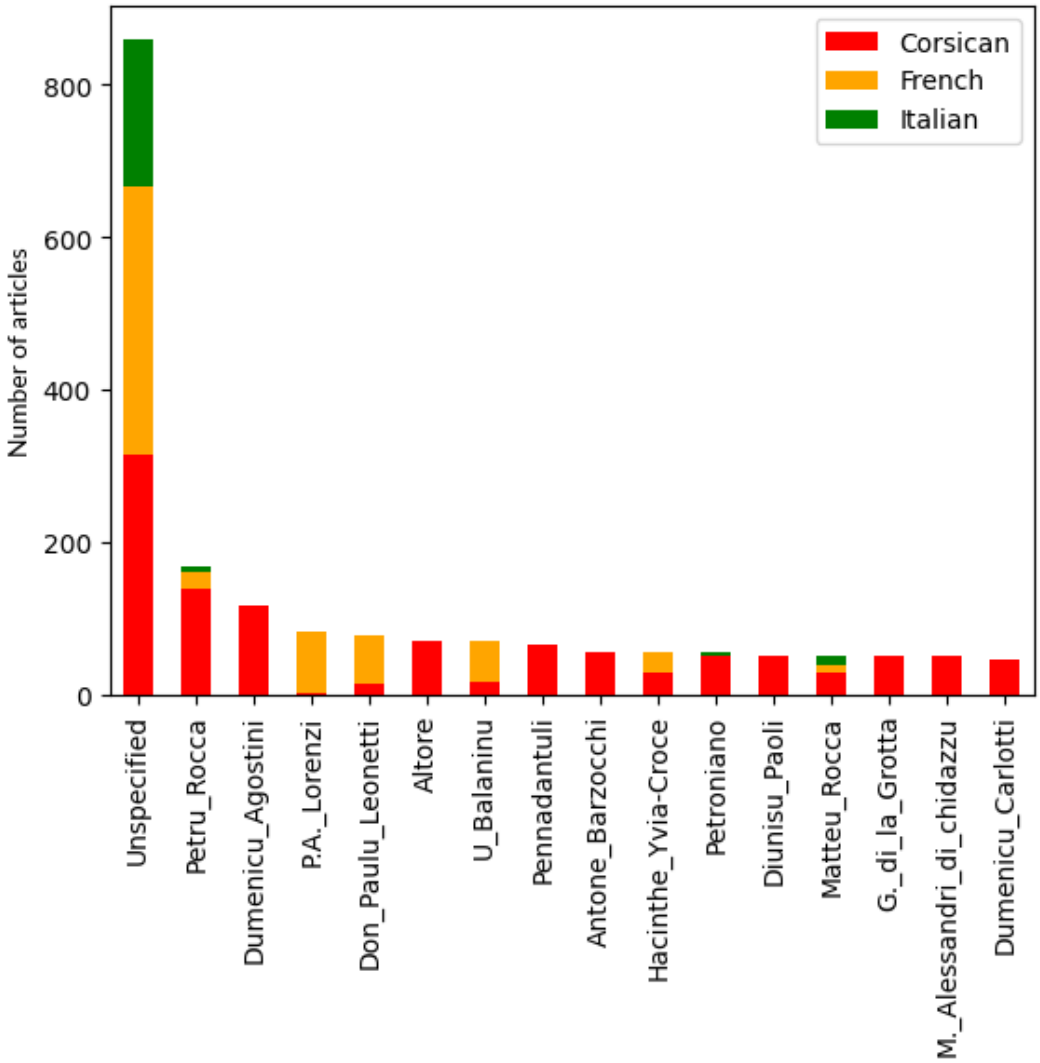


Figure 2.4: Language distribution of articles according to a selection of authors (database)

In fact, there is a very large number of authors and only a few who stand out for their large number of articles written. There is, of course, the author Domenico Guerrazzi, but he lived in the previous century, his texts are reproduced in the form of serials and this makes him a totally anachronistic author and therefore not a potential candidate for the paternity attribution of certain texts.

But this diversity is also the interest of our research. As mentioned earlier, such a press covers a very wide range of authors with very different political opinions, from regionalists to separatists and irredentists. Although the *Muvra* became more radical in its ideas from the 1930s onwards, the editorial board seemed to assume the diversity of opinion of its contributors in the early years of publication:

> *La Muvra!* is a regionalist newspaper: all the opinions expressed in this newspaper are not the expression of the *Partitu*, that's clear! In *Partitu*, as in the Chamber, there is a right, a centre and a left.[57].

The great diversity of opinion and social origin of all its authors makes *A Muvra* a perfect case study for author profiling. Applying computational methods on such a large and varied corpus would allow to enhance these resources. This will allow us to understand a little better the paths of the muvrists and the reasons for the political radicalisation of some towards separatism or irredentism, and on the contrary the relaxation of some others in favour of France.

[57] *A Muvra*, n°110-1923/06/24: « *La Muvra!* est un journal régionaliste: toutes les opinions exprimées dans ce journal, ne sont pas l'expression du *Partitu*, c'est clair ! Il y a dans le *Partitu*, comme à la Chambre une droite, un centre et une gauche. ».

The technical stakes of the *Muvra* study

Chapter 3

Retrieving the images

3.1 The layout of the *Muvra*

The aim of this section, which deals with the more technical aspect of our study, is to explain the process of creating the textual data needed for our analysis. The processing of images is essential because since our data is not available, it is necessary to be able to obtain them through occlusion. In order to do this, high resolution images are required for this process to be effective. This chapter serves to explain the process of obtaining good quality documents. But first, let's take a look at format of the newspaper we are studying.

As we have already seen, the journal *A Muvra* began to be published in Paris in May 1920. Petru Rocca was strongly influenced by the Parisian press, especially as he had spent several years in the French capital before the Great War. He also published his first book in 1913, entitled *Les Corse devant l'anthropologie*[58], already predestining the eugenicist editorial line of the *Muvra* less than a decade later. The fact that the book is dedicated to the memory of grandfather of Rocca and founder of the *Journal de la Corse*, Gabriel Marchi[59], also demonstrates his early interest in the world of the press. It goes without saying that the newspaper's format was influenced by the journalistic context of the time, which was in the midst of a post-war reformation. The end of the 19th century and the beginning of the 20th century saw the domination of the large Parisian presses in the Capital and even in France, giving the

[58]Pierre Rocca, *Les Corses devant l'anthropologie*, Paris, Librairie J. Gamber, 1913.
[59]Ibid., p. 3.

pretty sobriquet of "République du Croissant"[60] to the district where they were installed.

Thus, we can find all the codes of these presses in the format of the *Muvra*. The first issues are in "half-tabloid" format (290 × 210 mm), which corresponds to the year of publication in Paris. It is when they moved to the 38 Cours Grandval in Ajaccio during the year 1921 that the format changes to "berliner", which is an approximate format of 470 × 320 mm. From that year onwards, we find the classic layout of the weekly or daily press of the time, the figure 3.1 on page 42 being an example of a typical front page. The characteristics of a traditional front page are thus clearly visible. We have the title of the newspaper within the *banner headline* (*manchette*), with the subtitle indicated just below. The *ears*, in this case only one, represent a mouflon, the icon of the newspaper. At the top of the page we have the *upper banner* with the issue numbering, the year of publication, the days of publication and finally the date of publication of that particular issue. Below the banner headline is information about prices, subscriptions and the editorial board.

Let's now look at the *body* of the page, the part we will use for our analysis. It is here slightly different from what we can see in the newspapers. We do not find any *headline* (*tribune*), only *columns*. Generally speaking, we can find between two and three *columns* for a typical issue. They can be open letters as well as classical political articles. This less sensationalist aspect of the *Muvra* can be translated into a desire to give the paper a more cultural and reflective angle. If we look at the later issues from the late 1930s, we can see the appearance of *headlines* denouncing the searches and censorship by the French government, like in early 1939[61], which shows the evolution of the movement even in the layout of the paper. Similarly, there are no *footers* on the front page, but they are generally found on the following pages for the serial novels. Finally, there is also a figure which generally corresponds to a caricature of Matteu Rocca, a photograph of a Party meeting or a landscape of Corsica. Overall, this description of the layout shows that the journalistic practices of the muvrists are inspired by the standards of the profession.

[60] Diana Cooper-Richet, *Passeurs culturels dans le monde des médias et de l'édition en Europe (XIXe et XXe siècles)*, vol. 6, Paris, Presses de l'ENSSIB, 2005 (Référence), p. 138.

[61] *A Muvra*, n°686-1929/02/20: "Cume quellu di u 1 ferraghiu, u numeru di A Muvra di u 10 è statu sequestratu per ordine di l'autorità giudiziaria.".

Sest' annata — Numaru 194 Esce a Dumenica Dumenica, 8 Marzu 1925

A MUVRA

GIURNALE DI E PIEVE DI CORSICA

ABBUNAMENTI :
In Corsica. . . . . 12 franchi
In Francia. . . . . 15 »
A l'Esteru. . . . . 25 »
U Numaru : 4 soldi

Direttore : Petru ROCCA

AMMINISTRAZIONE:
38, Corsu Grandval 38. — AIACCIU
Contu currente cu a posta — Marseglia 45,82

SCRIVEMU !

Semu, in litteratura, a u primu periodu : un'alba di rinascita intelletuale, tutta culurita a prumesse, spunta sopr'a i monti superbi di Cirnu. U muvimentu mistralianu corsu, sbucciatu da qualchi lustri appena, oghie diventa più luminosu, e a strada che nò circhemu di traccià in la « macchia folta » di l'Avene, cumencia a parèci più chiara, più sicura.

Simule ad ogni pòpulu in scumbugliu, a Corsica, comu a fece a Rumenia, a Serbia, a Pruvenza e a Catalogna, ha, au paru di a Bretagna autonomista, trè punti a sviluppà :

1°) Bisogna a produce e a produce più chi si pò, a quantità è nicissaria pa u momentu, u tempu, in lu futuru, s'incaricarà di ciarniglianne a qualità e di ritene, cume degni di a pusterità, tale opere chi avaranu in sè tutt'i carattari e e distinzioni d'un capu-lavoru.

Heliade Radulesco, babbu di a litteratura rumena, principiendu u muvimentu di rinascita d'u so paese, allòra sott'a l'influenza d'i principi greghi, pridicava a so cumpaisani : « Scrivite, ò figliòli, purchè ci sia scritti. ».

Stu mottu d'ordine chi dipinghie un'epuca, pudaria esse u nostru. Semu inn'un momentu di frebba e ci vòle a scrive quant'ellu si pò, in corsu.

2°) In sicondu lògu, luttemu contr'a l'influenza straniere, contr'a tuttu stu malsanu frusterisimu di certi e molti chi vurrianu spiantà par sempre, e a nostra lingua, e e nostre custumanze, ultime vistighe d'una razza ammattaghiata.

Travagliemu ad imbellì, a cultivà ghielosamente stu tesoru ch'è a nostra lingua.

3°) Terzamente, ch'in li nostri scritti, sempre ci sia, splinnurente e suvrana, l'anima di Cirnea chi ci deve, sòla, inspirà.

Alluntanemu ogni spirtu *sequanese*. Surgenti di billezza e d'originalità, inde noi, prim'a Diu, un ne manca !

∴

È u sintimentu d'a nostra naziunalità, di ciò che nò semu stati a i tempi gloriosi d'un Sambucucciu ou d'un de' Paoli, è stu sintimentu chi deve risuscità in noi, oghiejornu, e dacci l'indispendenza, almenu spirituale.

Chi la nostr'opara di rissurezione, come a vulia *A Cispra*, sii un'opara naziunale e chi a Corsica ne sorti degna e purificata.

Un ci scurdemu che nò rapprisintemu l'epuca *eroica* di a nostra litteratura in evoluzione ; semu u fiore chi sboccia intantu d'esse u fruttu chi l'Avene — stu gran Magu — farà prestu maturà.

Mauriziu Tarabeni.

1789

Quale strano contrasto ! Quella stessa Francia, che vent'anni prima aveva inviato un esercito per annientare la libertà dei Corsi, elevava ora la loro costituzione sul trono dei suoi re.

Gregorovius — Corsica — d. 239

U TEATRU DI A MUVRA

Caru Direttore,

Sò scalatu in San Petru a tempu a e cutrure, l'aqua-nivata e tutte e zinzalette chi ci poghianu, da foce a tramuntana. U veranu di Campu-d'Unicu m'avia fattu sperà a fine di a cruda stagione, e invece, ecumi furzatu di ricunosce chi u pruverbiu corsu anticu è santu e ghiustu: «l'invernu un lu si rodenu manc u topi»

Ogni male ha u so bene ! Furzatu da u tempurale a stà sempre a u fuculagghiu ne prufittu per occupammi di u nostru **Teatru Corsu di A Muvra**. *L'artisti sò digià mezu che accaparati: e, o furtuna ! ... a crederiste ?... Fra questi, una donna corsa, pront'a saltà in scena per dici qualcosa in corsu sciappatu, senza gniffi a la pinzutesca (cume voi sapete, su puntu era unu di quelli chi mi facia u più paura.)*

Dunque un mancanu più ca i soldi e a bòna vulunté di tutte e penne corse.

Pè i soldi m'indirizzu senza tante fole a tutti l'amici chi a furtuna ha fattu caldi di borsa, e accettu a minima limosina.

In quant'a l'opere, nisun dubbiu pongu sopr'a i fratelli pueti e prosatori perche credu chi tutti devenu capisce a purtata di a miò idea.

Tengu a ripeteli ch'elli facianu u pussibile per chi ogni pezza, sia cumedia, sia tragedia un abbia più di 4 o 5 personaggi, fra i quali unu di sessu feminile.

Si niente di grave un accade, saremu pronti all'istatina.

Postu chi no avemu creatu l'usu d'una merendella annuale, e, chi u lòcu sceltu è San Petru, centru comudu per tutu e per tutti : **U Teatru Corsu di a Muvra**, *darà a so prima raprinsentazione u primu Aostu, in San-Petru di Venacu, davant' ai pueta corsi riuniti in merendella, i cursisti e tutti l'amatori di lingua e letteratura corsa.*

Questa è a miò idea; accetteragghiu sempre però i cunsigli d'i miò più majò, e si c'è megliu obidisceragghiu a l'opinione maestra.

Speru dunque; e a bòna speranza dendumi un pocu d'orgogliu pardunevule, vi dicu a tutti : Viderete !

U Sampetracciu.

Rispundemu a' chiama di nostru fratellu Nottini.

A Muvra *dà una voce a l'amici vicini e luntani.*

A chi vòle vede cresce e fiurisce, U teatru Corsu, U teatru di A Muvra ! *cum'ell'u vòle battizà u Sampetrecciu, mandi cioch'ellu pò ; u più sarà u megliu.*

U nostru prossimu numaru publicarà i prim'arrivi.

ARIA !

Adirizzà e sottuscrizione a u Direttore di A Muvra, 38 Corsu Grandval, Aiacciu.

U NIDU DI L'ACULA

Paoli a Napulione — *Fideghia, per paga di tanta glòria, cioch'elli ne hannu fattu...*

L'ALMANACCU

di

A MUVRA per 1925

A sòla publicazione *corsa* scritta in *corsu* da penne *corse*.

U sòlu libru *corsu* ch'un sia bracanatu di roba pinzuta.

Cumpostu da i più gran scrivani e pueta *corsi* senza a collaborazione di nisun bastardu.

208 pagine — 100 disegni — 8 canzone cu a so musica.

Mandate 4 franchi a u contu currente, Marseglia 45,82, da pudellu riceve ricumandatu pè a posta.

SPICCIATEVI !!!

Raconti di A Muvra

L'ULTIM'OSTIA

A notte fritulosa, pianamente, si pisava da a piaghia chi s'istende in meza luna, tra e cullette di Portu-vecchiu e i monti di Cerbioni; i primi albòri, frigiendu l'alta marina, trimuliavanu e fiurianu di chiarore e punte di a macchia.

Sòlu sòlu, in quellu salvaticume desertu, un prete viaghiava, alla fretta, salutatu, a ogni scornu di u stradone, da i fruli lesti di l'acelli.

— Bisogna ch'iò mi pressi, dicia in se stessu u viaghiadore smaitinatu, a strada è longa; pòi mi sò a diunu... Quand'iò pensu cum'ellu si n'è andatu u miò poveru cunfratellu, cusì bravu e cusì giovanu, mi si spezza u còre ! Chi è a vita di stu mondu ? Un soffiu. Affacca Falgina e u lume è spentu ! Ellu cridia di riturnà subitu e invece... Avale è fatta per ellu. Chi Diu li dia pace e riposu ! Un sò, in chi statu truveraghiu: a chiesa orfana di pastore... orfana, forse, per sempre ? Diceraghiu a messa e cunsumeraghiu e sacre spezie chi si sò fermate. Chi disgrazia ! Eramu dui preti, luntani, belli luntani, eiu quì, ellu quassù in sa cresta ed eccumi sòlu sulettu ! *Hodie mihi, cras tibi!* Povera, povera Corsica !

E u Prete allungava u passu. In stu mentre li s'accosta, quasi silenziosa, un ottomòbile bianca.

— Dunda, o giò curatu, cusì di bon'ora ?

— Vò à dì una Messa in su paese...

— Fatemi u piacè di lasciabbi cunduce a portu; cullate, cullate subitu !

U Prete un si fece prigà, si messe a pusà, pinsendu a u carru di fòcu chi u Signore mandede per purtà u prufeta Elia in Paradisu.

A vittura spianò e so ale d'argentu... Paria un agula marina chi bula a fior di macchia. A un certu puntu, un ghiagaru, rossu rossu, si mette in bia, abbaghiendu cun furore da magnassi i cristiani.

Un ci fù versu di scansallu, passò sottu a vittura; s'intese un culpettu seccu e a bestia si firmò, a crocchie alte, a mezu stradone.

Infratantu u sole dinnurava e culline tonde di a vicinanza, cuparte a villutu; i specchi di i pozzi lucianu cume occhi di bòi, tra i cigli ciuffuti di i iunchi; l'Abatescu e u Travu intunavanu una canzona nòva sottu i listinchi spessi; a Sulenzara turcia u so cullerinu d'acciaghiu rentu e rene fine, a pé di mare. L'ottò, mughiulendu un pocu, si ficcò ind'una strada stretta, sempre a meze macchie, terrorizendu i stòli chiunchiur ladori di i torduli e di e merle. Infine, dòpu centu abbinte a traversu una longa custera, eccuti un pugnu di case e una chiesa grigia in cima d'un pughiale.

U paese s'è appena discitatu. In piazza di a chiesa, sottu un olmu, a una forca pende una campana chi aspetta un campanile. U Prete tira a corda e u batagliòlu, senza sforzu, si mette a stinighirlà e so note linde:

— Vinite, o paisani, c'è una Messa !

Figure 3.1: Example of a front page of the *Muvra* (n°194-8 March 1925)

3.2 The importance of document provenance

We can already locate two online platforms where our documents are available for download. These are *Gallica*, the digitisation platform of the *Bibliothèque nationale de France* (BnF)[62], and *THOT*, the platform of the *Archives départementales de Corse du Sud* (ADC)[63]. The fact that these are national heritage institutions means that the digitisations are in the public domain, open source. In order to exploit good quality images before proceeding with an occlusion phase, we can use IIIF documents, an international interoperability standard for high resolution images[64]. The image recovery stage therefore required two very different methods which we will describe in this section. We will see that these discrepancies had an impact on the logs scanning phase.

3.2.1 Collecting the newspapers from *Gallica*

Let's first look at the documents available on the BnF website. Faced with certain problems of standardisation of the number of IIIF images available for each issue of the journal and in order to be sure to extract all my sources, we use the `IIIF-Crawler` script, developed by Thibault Clérice and Jean-Baptiste Camps[65]. This also allows, in the long run, to import a significant number of good quality images automatically. During an internship carried out in February 2021 with Guillaume Porte, research engineer in digital humanities attached to the UR3400 ARCHE, I had the opportunity to develop a script allowing the automatic retrieval of metadata from Alsatian local history society journals thanks to `ark` identifiers[66]. The realization of this script was part of the *Alsatia Numerica* project aiming at the creation of a portal allowing easy access to all digitized resources concerning medieval history in the Upper Rhine region, such as theses, archives or journal articles[67].

[62] `https://gallica.bnf.fr/accueil/en/content/accueil-en?mode=desktop`

[63] `http://archives.isula.corsica/Internet_THOT/FrmSommaireFrame.asp`

[64] `https://iiif.io/`

[65] Jean-Baptiste Camps and Thibault Clérice, *IIIF-Crawler*, 2019, URL: `https://github.com/Jean-Baptiste-Camps/IIIF-Crawler`.

[66] Permanent identification code used by the BnF for its digitised documents.

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

The idea is to reuse this script and to adapt it in order to retrieve the ark codes of the *Muvra* articles and to import them into a `tsv` file compatible with the `IIIF-Crawler`:

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

In fact, this command above allows you to launch the `IIIF-Crawler` script. The documentation of the script is available on the repository of the project[68]. The `ID` corresponds in our case to the identifier `ark`, the `source` is *Gallica*, the `start` and the `end` are pages 1 and 4 insofar as the publishing format of the Corsican review respects this number of pages throughout its two decades of existence with pages in the "berliner".

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

This command reads a text file which contains 4 columns: the `ID`, the `source`, the `start` and the `end`. Thus, automatically generating a file of this type using a Python script that exploits the *Gallica* API makes it possible to automate the process of retrieving high-resolution images. It was therefore necessary to adapt the script created as part of my internship so that it could retrieve the `ark` code of the journal, and then of each issue. The script then takes care of creating a `tsv` file that can be read by the `IIIF-Crawler` script. The question arose as to whether this should be integrated with the `IIIF-crawler-periodic` script but it was easier to use the initial code directly using a command from the `os` module[69]. With this script, a corpus of 238 issues of the *Muvra* was recovered, i.e. approximately 730 pages, which corresponds to all the copies available on the *Gallica* platform.

3.2.2 Web scraping from the *Archives départementales de Corse du Sud*

The main problem encountered with the digitisation platform of the *Archives départementales de Corse du Sud* is that the API is not open. We therefore had to find a way to find a command capable of downloading the images automatically to avoid doing it by hand. A little digging into the *THOT* website

[68] `https://github.com/Jean-Baptiste-Camps/IIIF-Crawler`

[69] One of the prospects for improving the script is to unify the two scripts in order to avoid dependencies.

traffic data quickly revealed a common template for storing image files. A simple Python script sending web requests to collect this data is then possible. We get this command[70]:

```
requests.get(f"http://archives.isula.corsica/GEDTHOT5/PROD/PRESSE1/148PER/
↪ FRAD02A_148PER/FRAD02A_148PER_00000{PER}/FRAD02A_148PER_00000{PER}_0{ID}/
↪ FRAD02A_148PER_00000{PER}_0{ID}_0{'00'+str(PAGE)}.jpg").content
```

The parameters that interest us are related to the classification system of the Corsican archives. The code `148 PER` corresponds to the reference code of the journal *A Muvra* in the *Archives départementales*, which will then be numbered from 1 to 5 to classify the publication periods of this specific journal. This number corresponds to the `PER` parameter of the command, giving for example `148 PER 1` that corresponds for the issues from 1921 to 1924. The year is selected according to the `ID` parameter; it does not appear on the document reference code because it is the year that is specified in the archive center, so it is specific to the *THOT* platform. Finally, the `PAGE` parameter corresponds to the numbering of the image, the range is predefined in advance.

As we can see, the difficulty does not really lie in the Python code needed to retrieve these images, as was the case for *Gallica*, for example. In reality, it is more a question of understanding the classification system of the *Archives départementales* in order to determine the reference code, and then to find the `url` link allowing the files to be downloaded from the web traffic. Such a command is quite heavy and the ADC servers are probably not used to receiving so many requests at once. This is why we had to space out the image collection times to avoid overloading them.

[70]The command is simplified as it is too long to be fully included.

3.3 Do we need to clean the images?

In fact, this question cannot be resolved before addressing the issue of the OCR processing, which we will see later. Having good quality images is not enough for this step to be completely effective, hence our initial questioning. However, let us focus here on the technical issues related to this question.

To do so, we can use the free software *ScanTailor*[71]. After the first step of retrieving the images, they were sorted into separate folders according to their ark identifier.) To solve this problem, it was necessary to write a script which allows to sort the images obtained by the previous stage according to their numbering, i.e. to join the pages 1, 2 and 3 together, by renaming them according to their ark identifier. This stage is necessary because the tool we want to use can automate the editing of images from a folder. Cleaning up an image by cropping certain edges, for example, is not done in the same way depending on whether it is the front page or the pages containing the advertisements.

Once this first step is done, let's look at the *ScanTailor* software. The idea is to binarise the documents with the Otsu method (set to 30-Thicker). In the output, we have black and white images with a sharper contrast to facilitate the optical detection of the characters but also the column separators of the articles. We are not interested in the *manchette* section of the newspapers, so we can crop all the headlines from the beginning to keep only the headings. In the same way, we are not necessarily interested in advertisements but only in the continuity of the articles on the last page to have the complete content. The figure 3.2 on page 47 is an example of the output of a front page of A Muvra magazine after being processed by the ScanTailor software. Thus, as mentioned earlier, the Otsu method allows the image to be binarised (see the blue box in the figure). It consists of calculating a pixel variance threshold from the image histogram, hence the term thresholding in the software interface. Without dwelling on the mathematical operations behind this algorithm, it should be noted that it deduces whether a given pixel is black or white from the selected threshold. In our case, the threshold chosen is 30 in order to have well readable characters[72].

[71] `https://scantailor.org/`

[72] See: Nobuyuki Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, 9–1 (1979).

The second cleaning phase is the despeckling stage. In a paper published in 2021, Italian researchers define "speckling" as follows:

> 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.[73].

In concrete terms, the "speckling" is a type of noise that corresponds to random clusters of black pixels that impair the intrinsic quality of a binarised image. The despeckling step is therefore necessary to remove these spots. In figure 3.2, this parameter corresponds to the red box that has been set to the maximum according to our own needs.

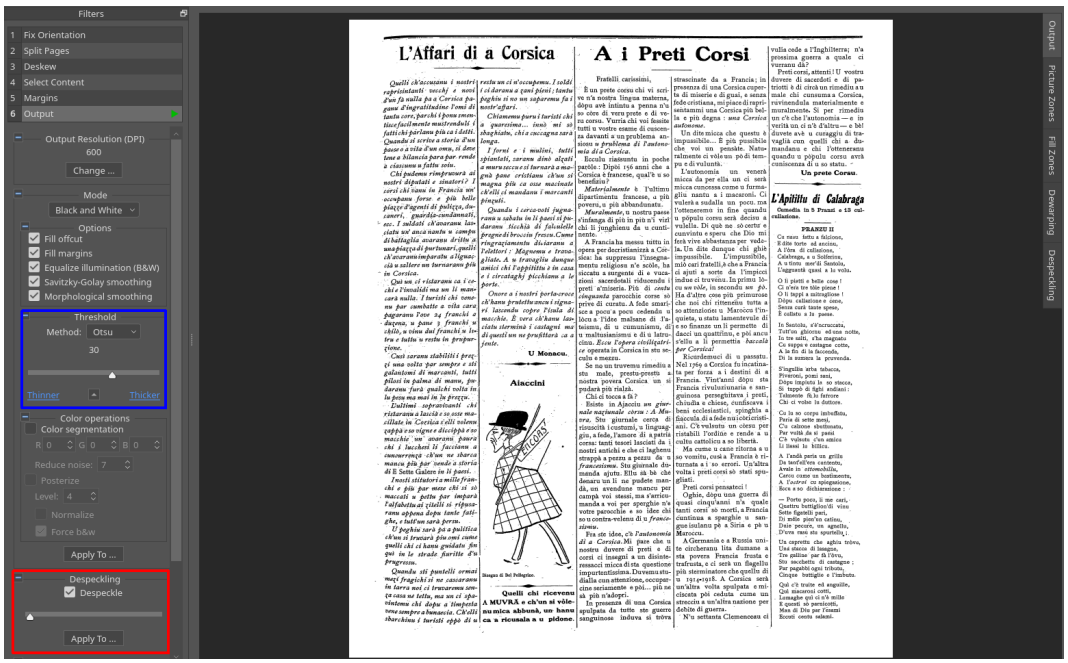


Figure 3.2: Exemple of a cleaned front page using *ScanTailor*

To summarise this section, if we input `jpg` images from the *Gallica* crawling, we get *tif* images as output after cleaning and recalibrating them. The cleaning would then consist of cropping the images to have only the textual

[73]Giulia Fracastoro, Enrico Magli, Giovanni Poggi, Giuseppe Scarpa, Diego Valsesia, and 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), p. 1.

content, binarising the images and adjusting the contrast, and then despeckling off random clusters of pixels. But then what are the problems we might encounter in the image cleaning stage?

Firstly, the images from the *Gallica* and *THOT* digitised archive viewers are very different and the qualities are not the same. If the IIIF images from the BnF have a better intrinsic sharpness of digitisation, there are still some defects. As the scanning work takes a long time on the press, some pages are bent, distorting the image and the scan. The quality of the digitalisations from the *Archives départementales* varies greatly. While sharpness is not the main problem, it is more a question of stains on the paper or pages damaged by time. This is an inherent problem in the conservation of old newspapers, paper is cheap and not made to last over time. The conservation of these documents is therefore difficult and this is reflected in the quality of the digitisations.

The problem also lies in the normalisation and cropping of images. The documents of the ADC have already been binarised, unlike those of the BnF. Moreover, as we have seen previously, the layout of the *Muvra* pages themselves evolves and the format is not fix. Standardising all the images at the same time requires an initial sorting organised according to identical layouts, for a gain in OCR quality which is not necessarily guaranteed, as we shall see later.

Chapter 4

Collecting textual data

4.1 The challenge of newspapers segmentation

4.1.1 Technical difficulties

One of the major challenges in the world of automatic character recognition today is the segmentation of newspapers. Their complex layout, which we described earlier in the case of the *Muvra*, requires the training of complex models that are often specific to a type of newspaper. There are two main steps in the automatic layout analysis, the *Physical Layout Analysis* (PLA) and *Logical Layout Analysis* (LLA). In an article published in 2021, Nicolas Gutehrlé and Iana Atanassova attempt to define these two stages. According to them the PCA consists on:

> *Physical layout analysis* (PLA), which is also sometimes called *document layout analysis*, consists in identifying physical regions of the document, with their text content and boundaries. Such regions can correspond to sections and lines of text, but also to figures, tables, etc. PLA also defines the reading order of the document, which corresponds to the linear order in which the different regions appear. This is particularly important for documents that have multi-column layouts. One commonly used output format of PLA is the `XML ALTO` format.[74].

[74]Nicolas Gutehrlé and Iana Atanassova, "Logical Layout Analysis Applied to Historical Newspapers," in *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, NLP Association of India (NLPAI), 2021, p. 85.

Typically, this phase corresponds to the detection phase of the regions, baselines and masks of a document. The second phase corresponds more to the annotation phase of these previously detected categories.

> *Logical layout analysis* (LLA), sometimes called *logical structure derivation* and *structure understanding*, consists in identifying the document structure elements and their categories i.e. title, header, paragraph, table, etc. Such logical elements can integrate one or more regions in the document that have been identified by PLA.[75]

In practice, the segmentation phase of a document is unique and directly incorporates these two steps in the process. While the question of layout analysis is particularly interesting for newspapers, studies in this area also go hand in hand with the development in recent years of *Handwriting Text Recognition* (HTR), given the complexity of certain old documents such as medieval manuscripts. As Chahan Vidal-Gorène points out in a 2021 article, the manual annotation phase is essential for training models dedicated to under-ressourced languages, a "time-consuming task"[76]. In that way, despite the difference in the periods studied, we find common technical problems with the press, even though the first research into the detection of regions in old newspapers dates back to the end of the 20$^{\text{th}}$ century[77]. The digitisation of newspapers and the collect of textual data is an object of particular interest in current research, as in the ANR *Numapresse* project[78]. However, it is still difficult to find model datasets that are easily available online. The initiative of the *Library of Congress* and the *Newspaper Navigator*[79] can be noted, but nothing really adapted for us at the moment.

The *eScriptorium*[80] platform, which uses the HTR *Kraken* engine, allows

[75] Ibid.

[76] Chahan Vidal-Gorène, Boris Dupin, Aliénor Decours-Perez, and Thomas Riccioli, "A modular and automated annotation platform for handwritings: evaluation on under-resourced languages," in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16*, Springer, 2021, pp. 507–522, p. 507.

[77] N. Gutehrlé and I. Atanassova, "Logical Layout Analysis Applied to Historical Newspapers"…, p? 86.

[78] `http://www.numapresse.org/presentation-du-projet/`

[79] `https://news-navigator.labs.loc.gov/`

[80] Benjamin Kiessling, Robin Tissot, Peter Stokes, and Daniel Stökl Ben Ezra, "eScriptorium: an open source platform for historical document analysis," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, 2019, vol. 2.

recognition models to be trained manually through a user-friendly interface. It gives the possibility to designate its own regions but also to annotate them manually, thus allowing to train the *Physical layout Analysis* and *Logical Layout Analysis* steps. Ideally, the segmentation of the *Muvra* front page would look like the figure if the PCA and LLA were done correctly. The following example (4.1) was produced by manual annotation using the *eScriptorium* platform, without prior automatic segmentation. For this study, we used the Inria instance[81].

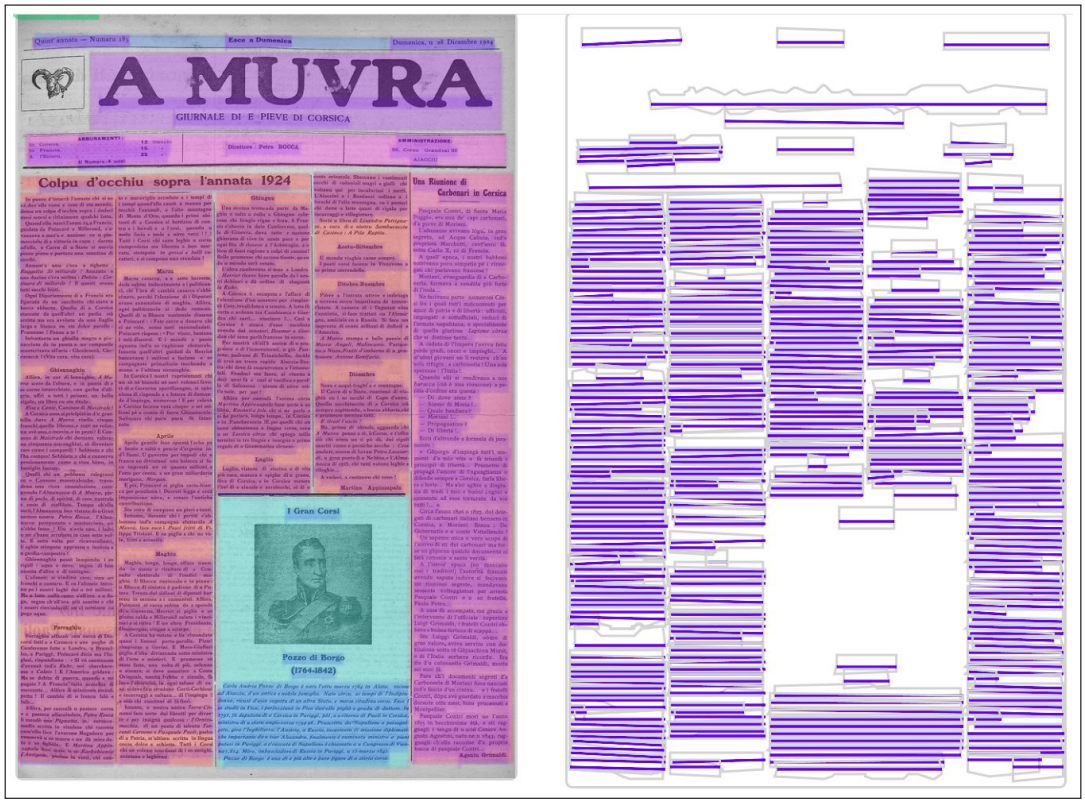


Figure 4.1: Exemple of segmentation for a front page of the *Muvra*

4.1.2 Training a model

As we have seen, the *eScriptorium* platform includes the possibility of training segmentation and character recognition models. This stage involves the manual annotation of pages, as was done in figure 4.1. This step, which is particularly time-consuming, can nevertheless be done in a different approach. The platform uses the HTR engine *Kraken*, so we can train a model directly

[81] `https://escriptorium.inria.fr/`

with the package, as long as we get the format `mlmodel` supported by *eScriptorium*. It is possible to train a model using `XML` files in `ALTO` (*Analyzed Layout and Text Object*)[82] format which are used to report on the layout of a segmented document. But which files should be used? The Gallica platform also offers an OCR service for digitised printed sources, accessible through its API, which we have used before. While the intrinsic accuracy of OCR is not necessarily good, which we will see later, the segmentation is particularly accurate, in this case the PCA phase.

The training of the model was done in collaboration with Angelo Mario Del Grosso, researcher at the Computational Philology Laboratory (*CoPhiLab)* of the CNR-ILC (*Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli"*) of Pisa and Federico Boschetti, researcher at the Università Ca' Foscari of Venice. After having downloaded all the `ALTO` files from Gallica (738 files), the idea was to use the `ketos` module dedicated to segmentation in *Kraken* on these files using the following command:

```
ketos segtrain -f ALTO -q early -N 100 --lag 10 -o vincentmodel *.xml
```

An `XSLT` transformation sheet had to be applied in order to adapt the BnF `ALTO` files into `ALTO` files readable by the module. This was mainly a question of version compatibility but also of preserving the elements that really interest us. Indeed, the importance is the detection of the regions, the detection of the baselines and the masks would be done directly during the segmentation using the basic engine of *Kraken* The training of the model required the use of a server in order to use its GPU. To do this, Jean-Baptiste Camps, researcher at the École nationale des chartes, took care of training the model using the Inria (*Institut national de recherche en sciences et technologies du numérique*) servers used by the École nationale des chartes.

Although a good start, the model was not yet powerful enough. Another way was to repeat this methodology but using another more powerful segmentation engine and then adapt it so that it was readable by *Kraken*. So it was necessary to use the tool *YALTAi*[83] developed by Thibault Clérice which allows

[82] We can find interesting information on the *Library of Congress* website: `https://www.loc.gov/standards/alto/`

[83] T. Clérice and Ronan Chauhan, *YALTAi, You Actually Look Twice At it*, version v0.0.1rc4, 2022, URL: `https://github.com/PonteIneptique/YALTAi`.

the use of *YOLOv5*[84], an Ultralytics object detection model, to be adapted for training segmentation models *Kraken*. Once the model was trained, we just had to match the names of the image files with those of the `ALTO` files and then run the following command:

```
yaltai kraken --device cpu -I patḧ--suffix ¨xmlsegment --yolo ./best.pt
```

The model, which is more efficient, allows a faithful segmentation as we can see on the figure 4.2. However, there is still the problem of annotation during the *logical layout analysis*, with the order of the blocks of text sometimes being wrong, even though this is the detection scale of the regions. We will see later that this has an impact not on the intrinsic quality of the OCR but rather on the normalisation of the textual data.

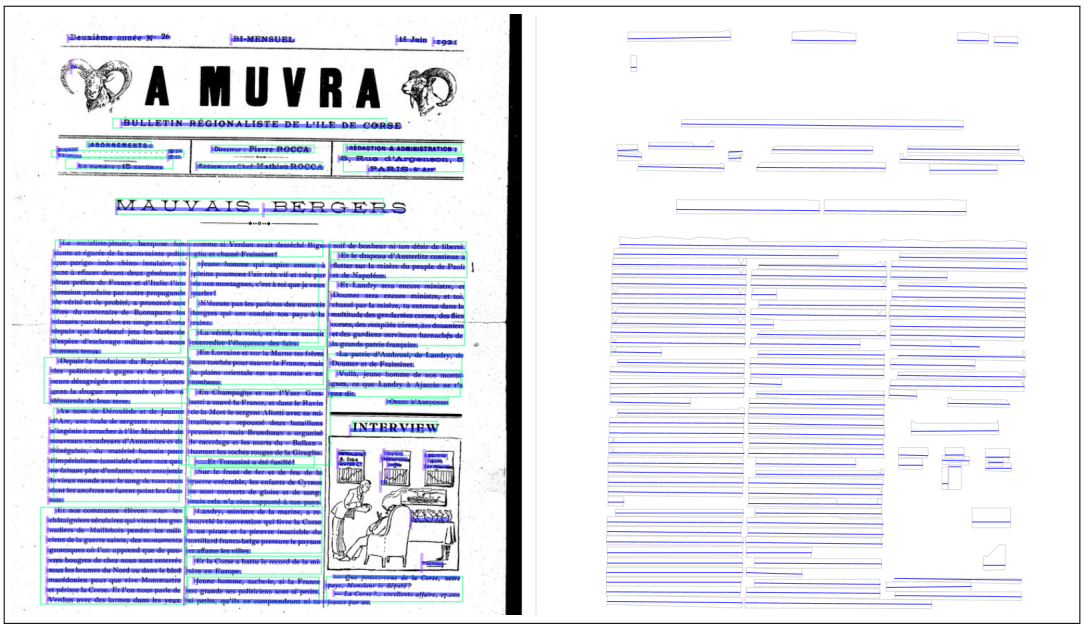


Figure 4.2: Result of the segmentation after the model training

[84]Glenn Jocher, *YOLOv5 by Ultralytics*, version 7.0, 2020, DOI: 10.5281/zenodo.3908559.

4.2 Making data available

4.2.1 Data extraction

The question of the choice of OCR quickly arose once our sources were segmented, both in terms of the engine and the model. While an OCR of the printed sources is made available by *Gallica* alongside their digitisation, it is clearly not of sufficiently high quality. From this perspective, one should not rely on the accuracy figure indicated by the platform, which is often around 99%. A simple empirical evaluation allows us to realise the limits of the proposed OCR. The choice of the OCR engine was first made with *Kraken*, via the *eScriptorium* platform using the *NFC DAHN manuscript*[85] model developed by Floriane Chiffoleau and found on the HTR-United[86] GitHub depository. It also allowed us to stay on the same platform in order to facilitate and centralize the process. Nevertheless, as it was not very efficient for presses written in Corsican, I turned to the *19th century prints - HTRcatalogs Artlas* model. Much more efficient than the previous one, this one seemed to be a good source of work to train a new model specific to Corsican presses in order to have an optimal efficiency rate. Unfortunately this required some fine tuning of the model so that it could be specifically adapted to my type of sources.

The choice was finally made to start with the automatic transcription of newspapers with *Tesseract-OCR*. This engine was not chosen for the segmentation phase because the layout analysis of this engine is very basic despite the existence of the PSM (*Page Segmentation Mode*). But we can rely on the `ALTO` files relating to the OCR also made available to obtain information on segmentation, a particularly complex element to manage for newspapers and of very good quality in this particular case. We therefore extracted the coordinates of the regions from these files and incorporated them into `uzn` type files which can be read by the automatic character recognition engine *Tesseract-OCR*, which has the advantage of taking into account several languages while having a model dedicated to Corsican. This step can be summarised by the following pipeline diagram.

[85] Floriane Chiffoleau, *dahncorpus*, version 1.0.0, Mar. 2021, DOI: `10.5281/zenodo.5911868`.

[86] Alix Chagué and T. Clérice, *HTR-United: Ground Truth Resources for the HTR and OCR of patrimonial documents*, version 0.1.56, URL: `https://github.com/HTR-United/htr-united`.

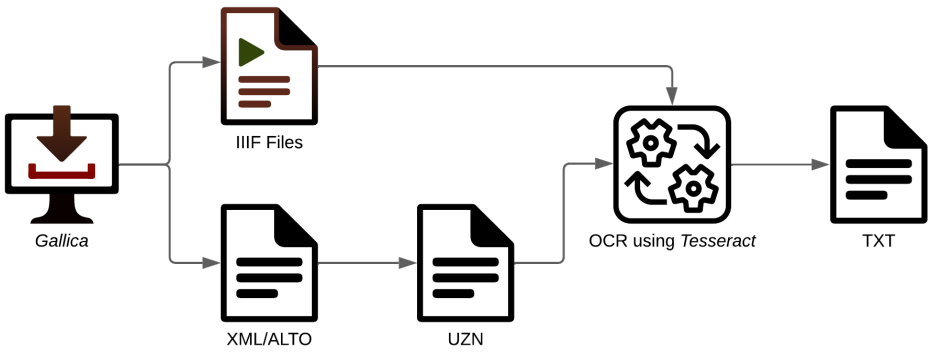


Figure 4.3: Pipeline of the *Muvra* textual data collection process

The method is more or less the same with the sources from the *Archives départementales de Corse du Sud* which we have segmented using our trained model. The importance is to recover the coordinates of the regions and to give an identifier to this precise region. Then we launch the terminal command for transcription with *Tesseract* by directly including the image to be analysed and the corresponding `uzn` file with the coordinates of the regions. As previously mentioned, the baselines are detected directly by the OCR engine. Here is an example of the notation of a `uzn` file:

```
363 1090 787 57 block_67
421 1065 723 44 block_64
595 1186 357 32 block_70
469 1112 721 52 block_66
```

And here is the *Tesseract* command to start the OCR process:

```
tesseract jpeg_files.jpeg uzn_files.uzn --psm 4 -l fra+cos+ita
```

We do not use the *Tesseract* Python library as the terminal command is more than sufficient for our purposes. Thus, based on a test document, we obtain an OCR accuracy of about 97% by measuring its efficiency with the Levenshtein distance. However, these figures should be treated with caution because, unlike manuscripts where the image quality is often consistent, newspaper scans can be variable. We therefore often find damaged or slightly bent corners of folded pages, creating a significant amount of noise. The normalisation phase is therefore essential to remove punctuation or accentuation. This

normalisation at a high granularity explains the choice of the Levenshtein distance to extrinsically evaluate the transcription. This measure is ideal for comparing two strings of characters and has a particularly reliable theoretical basis[87].

However, an intrinsic evaluation is also necessary to know whether the noise generated has not polluted our data too much. To do this, we will determine the reliability of the vocabulary by representing it with Zipf's law. As a reminder, this law assumes that the frequency of a word is inversely proportional to the rank of the word. Although old, this law is still important, especially in quantitative linguistics, whose implications are described in an article by Marcelo A. Montemuro[88]. By scaling up the law to the logarithmic scale, we should then have a graph representing approximately a straight line if the vocabulary is reliable and we do not have an overrepresentation of hapax and low frequency words resulting from poor OCR.

The graphs 4.4 present on the page 57 make it possible to represent the various corpora at our disposal according to the language: Corsican, French and Italian. As we can see, the Zipf's law representation seems to confirm the good scores obtained during the automatic press transcription. We even notice common phenomena such as the fact that the most frequent words seem to stand out, representing the tool words whose overall frequency is much higher than that of the other words.

[87]Li Yujian and Liu Bo, "A normalized Levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, 29–6 (2007).

[88]Marcelo A Montemurro, "Beyond the Zipf–Mandelbrot law in quantitative linguistics," *Physica A: Statistical Mechanics and its Applications*, 300–3-4 (2001).

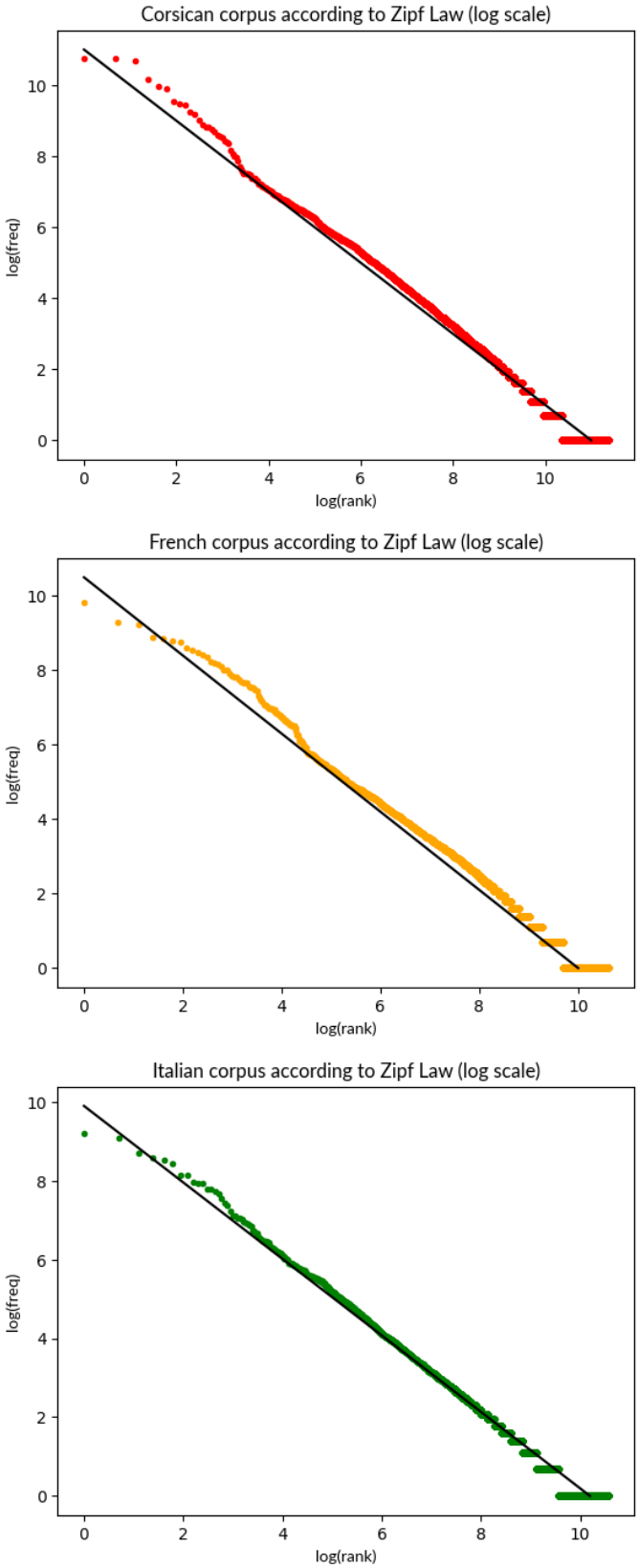


Figure 4.4: Reliability of the corpora according to Zipf's law (log scale)

4.2.2 Data structuration

Now that we have the raw text, we need to make it usable for analysis. As we saw earlier, an OCR is never 100% reliable. The trick is to determine, via an initially manual approach, recurring errors in order to correct them automatically. However, the standardisation of the data set is also a methodological issue, particularly with regard to the contracted forms of stopwords. It is necessary to encode the texts in `XML` before cleaning the data and publishing them. The question of the tagging format then arose. The choice of `TEI`[89] seems obvious in the case of making the data available, but a customised format proved more interesting from the point of view of exploiting the data. In our analysis methodology, only certain information seemed relevant to be kept, namely:

- Language
- Author
- Typology
- Content

With these considerations in mind, here is an example of how to encode a *Muvra* issue:

```
<rubrique lan="co" regio="inc" typ="art">
    <index> 4 </index>
    <texte> u teatru corsu... </texte>
    <langue> co </langue>
    <auteur> Maistrale </auteur>
    <type> art </type>
    <regiolecte> Inconnu </regiolecte>
</rubrique>
```

This allows us to parse the `XML` files more easily when we are in the analysis phase. But this does not prevent us from also having to structure this data in `TEI`.

[89] `https://tei-c.org/`

4.2.3 Data standardisation

Once we had separated the different texts and characterised them according to their language, we could apply a text cleaning function. The idea is to use the Python library `lxml` and to parse the various `XML` files created during the previous step. We must extract the content that is located between the text tags and apply the appropriate cleaning function for the language used. Once this has been done, a text file with the cleaned content must be returned as output. To identify the language of the content, we must rely on the value of the attribute `@lan` of the article in question, hence the importance of having well-normalised `XML` files. The cleaning is then carried out in four main stages:

1. The removal of punctuation
2. Case reduction
3. Normalization of syntax
4. Elimination of accents

For punctuation, it was necessary to make a custom list rather than using a list already provided by Python libraries such as `string`. Indeed, the term "punctuation" also includes some special characters. The removal of characters is essential to have a clean dataset and a tokenisation that considers only words. Similarly, the steps of case reduction and accent removal are fairly obvious when counting word occurrences.

The most delicate phase in our methodology is that of the normalization of the syntax. It is important because for euphonic reasons, contractions occur in written form in the form of elisions which reflect the discourse practices of speakers of Corsican. For example, the expression *s'è ellu hè* ("if he is") becomes *s'ell'è* in writing. Inversely, restoring the original form of the elision requires taking into account the context of gender and number: *ell'* can give *ellu*, *ella*, *elli* or *elle*. There is also the question of the normalisation rule, should we base ourselves on the syntax of the 20þ century or on the current one? Moreover, a certain number of ambiguities can creep into such a correction, such as the word *e*, which, depending on the context, can mean either "the" or "and". We should not forget to take into account that Corsican is a "*langue*

par élaboration" or *Ausbau* language and that, consequently, the syntax has a complexity due to the distinct instantiations according to the authors. In sociolinguistics, this type of language is a variant of a structured language (such as Italian) and set up as a distinct elaborated language Viaut (Alain), "Marge linguistique territoriale et langues minoritaires," *Lengas. Revue de sociolinguistique*–71 (2012).

The issue of data normalisation is particularly delicate due to the very nature of our methodology. While topic modeling does not include stopwords in the analysis because they are meaningless words, stylometry relies mainly on these words to attribute the authorship of a text. Indeed, to what extent should we normalise the data? Do we lose information if we normalise the syntax of certain terms or do we gain information? The choices that have been made are recorded in the Python file dedicated to data cleaning. This is nevertheless an important bias for our analysis. Fortunately, as we have seen before in the case of Corsican, the regiolectal diversity means that the idiomatic features of the authors are characterised by the great variety of the stopwords used. A thorough normalisation should not alter our analysis too much, even if it constitutes an improvement perspective to our study.

Chapter 5

Exploring our data

5.1 General informations

The length of the corpora and the differences according to the language in which they are written are among the essential information to be known. Here is a table summarising the length of the corpora according to their language, followed by two graphs to better represent this information.

	Corsican	**French**	**Italian**	**Total**
Word count	897965	380457	263095	1541517

Table 5.1: Length of the different corpora

This table shows the large amount of data available for further research. It should be remembered that stylometry requires a relatively large amount of data to function. Our corpus is thus composed of a total of about 1.5 million words distributed among three different corpora: one in Corsican, one in French and one in Italian. This corpus is an aggregation of two datasets obtained in different ways: one from the *Gallica* platfform and one from the *THOT* one. The methodology for obtaining them differs as does the selection of the sections. The figures 5.1 shows the predominance of Corsican over the other languages, which is quite logical considering that the *Muvra* is above all a dialectal press. French and Italian also play an important role and these two languages represent almost 40% of our total corpus.

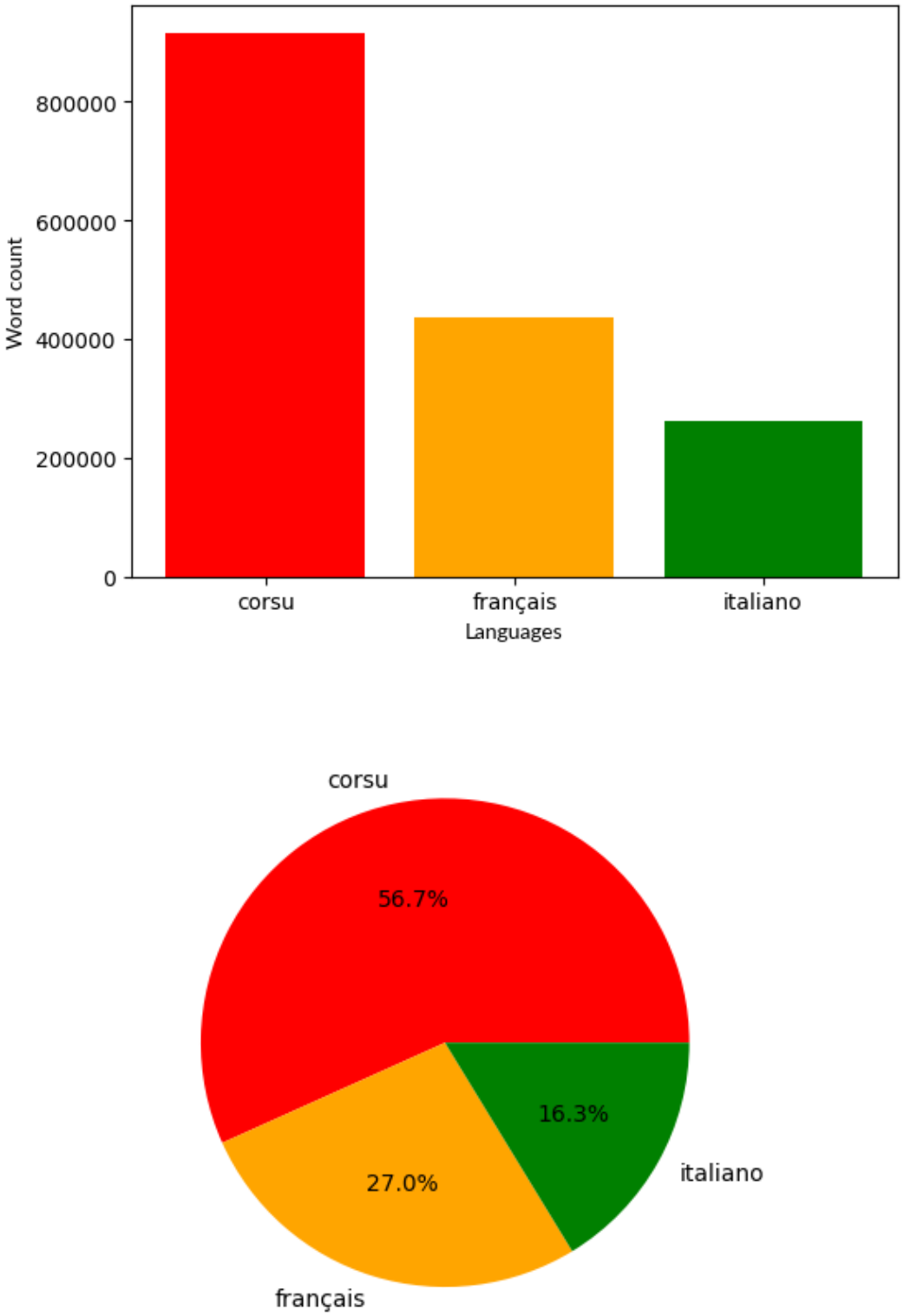


Figure 5.1: Language distributions in the dataset

Let us now look at the distribution of articles according to their typology. It can be seen on the figure 5.2 that classical articles represent the majority of the typology alongside poetic productions. Poetry, which is very numerous, nevertheless accounts for almost as many words as the so-called "miscellaneous" (*divers*) articles or the serials.

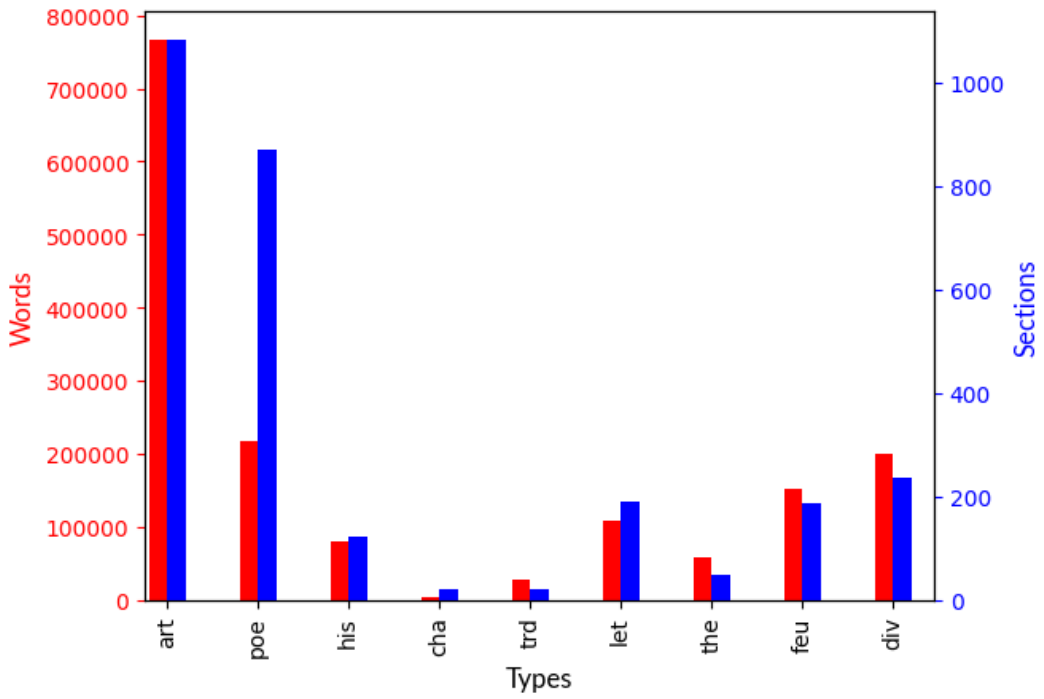


Figure 5.2: Distribution of article typology and their share in the total of words

The relationship is even more interesting to study thanks to the ratio between words and articles in the table 5.2 on page 64. The poetic works and songs are on average shorter than the average, which seems quite logical given the nature of these productions. Similarly, plays will be longer as will translations, the latter often being translations of old books or long articles from Italian periodics. Classic articles seem to be about average, with an average of around 708 words per article, which is just the right amount to convey an idea or some information to the reader.

	art	poe	his	cha	trd	let	the	feu	div
Words	766965	217445	79782	4983	27861	107953	57881	152794	200611
Sections	1084	870	124	22	21	191	48	186	238
Ratio w/s	708	250	643	227	1327	565	1206	821	843

Table 5.2: Distribution of words and section

5.2 Diachronic comparative statistics

The aim of this chapter is to describe the data available to us for future analysis. The idea is also to make a slight comparison between the figures in this dataset and those presented in the chapter 2, on page 24, dedicated to the description of the *Muvra* during the 1930s with the help of the Autonomists/Irredentists' database. In fact, the comparison is only relevant between the years 1925-1930 and 1932-1939 because the methodology of recovering textual data on the first method made it possible to produce statistics on the entirety of the articles present in the *Muvra*, as has already been done for the figure 2.1 on page 26.

So let's start with the statistics for the type of articles in our dataset. It's worth remembering that the dataset is just a sample of articles chosen according to biases defined by our own needs. It will give an overview of our selection, but not of all the issues of the *Muvra* that have been published.

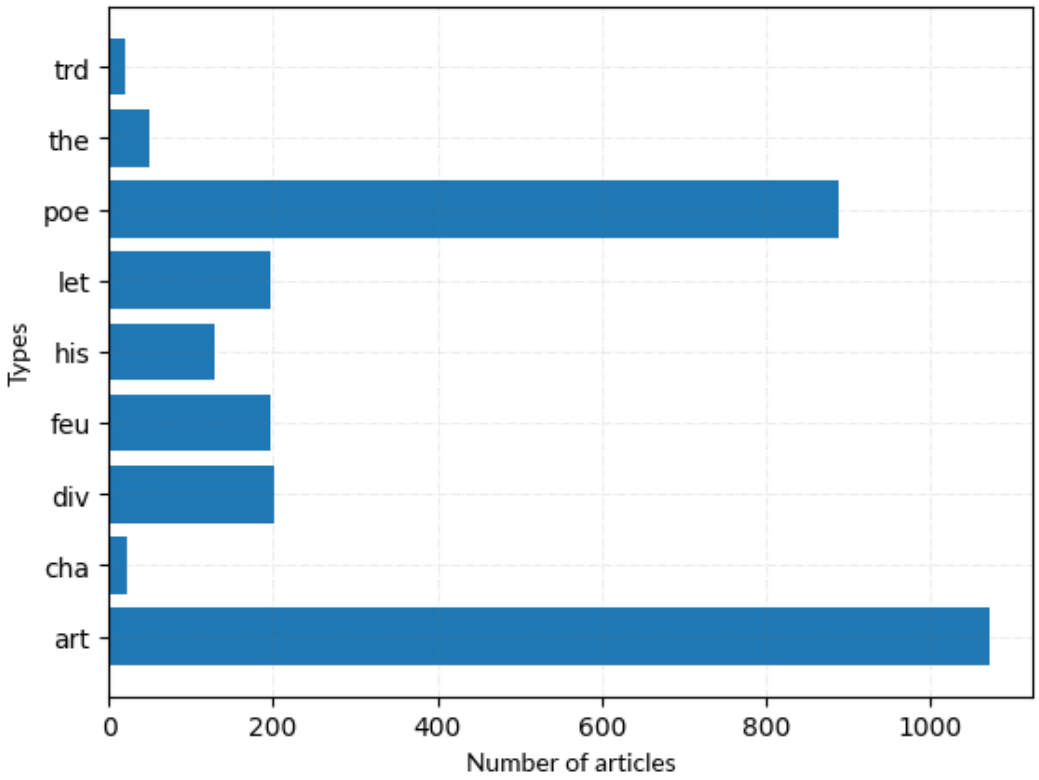


Figure 5.3: Typology of articles of our dataset

	art	poe	his	cha	trd	let	the	feu	div
Value	1071	22	202	198	128	198	888	50	21
%	38.55	0.79	7.27	7.13	4.61	7.13	31.97	1.8	0.76

Table 5.3: Distribution of sections types

Our classification is still based on that established in the introductory chapter. This graph refers to the graph 2.2 on page 31. We can see that we have a majority of articles and poems compared with the data collected for the database. Our dataset is in fact much less homogeneous than a classic *A Muvra* issue might be. As explained above, we have made a selection of texts specific to our needs. In other words, we need texts that are syntactically similar for stylometric analysis or topic modelling. If our dataset is relatively unbalanced, it's because we don't have a corpus representing all the issues of the newspaper, our reference horizon being limited to certain authors only.

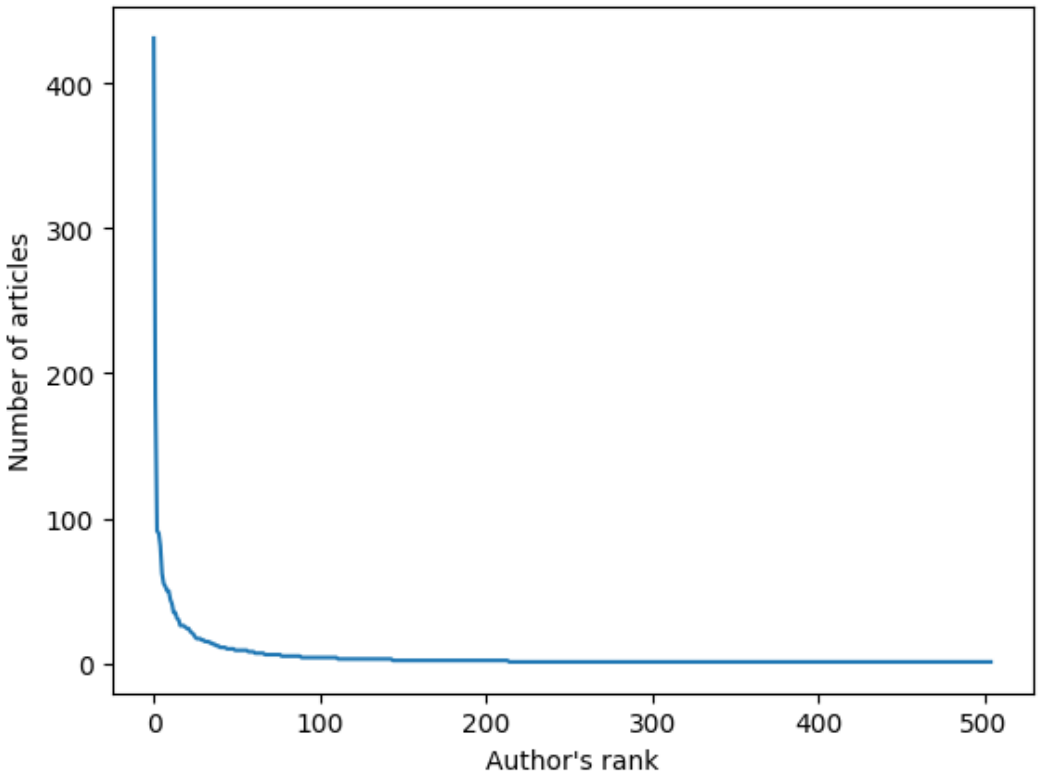


Figure 5.4: Number of articles written by an author or a pseudonym

This plot ties in with what has already been said above about the need to choose only certain authors and not all of them. The vast majority of authors have contributed very little to the writing of the *Muvra*, with one or two articles maximum. They are often one-off poets sending in a contribution to be published on the third page of the journal with the other poems.

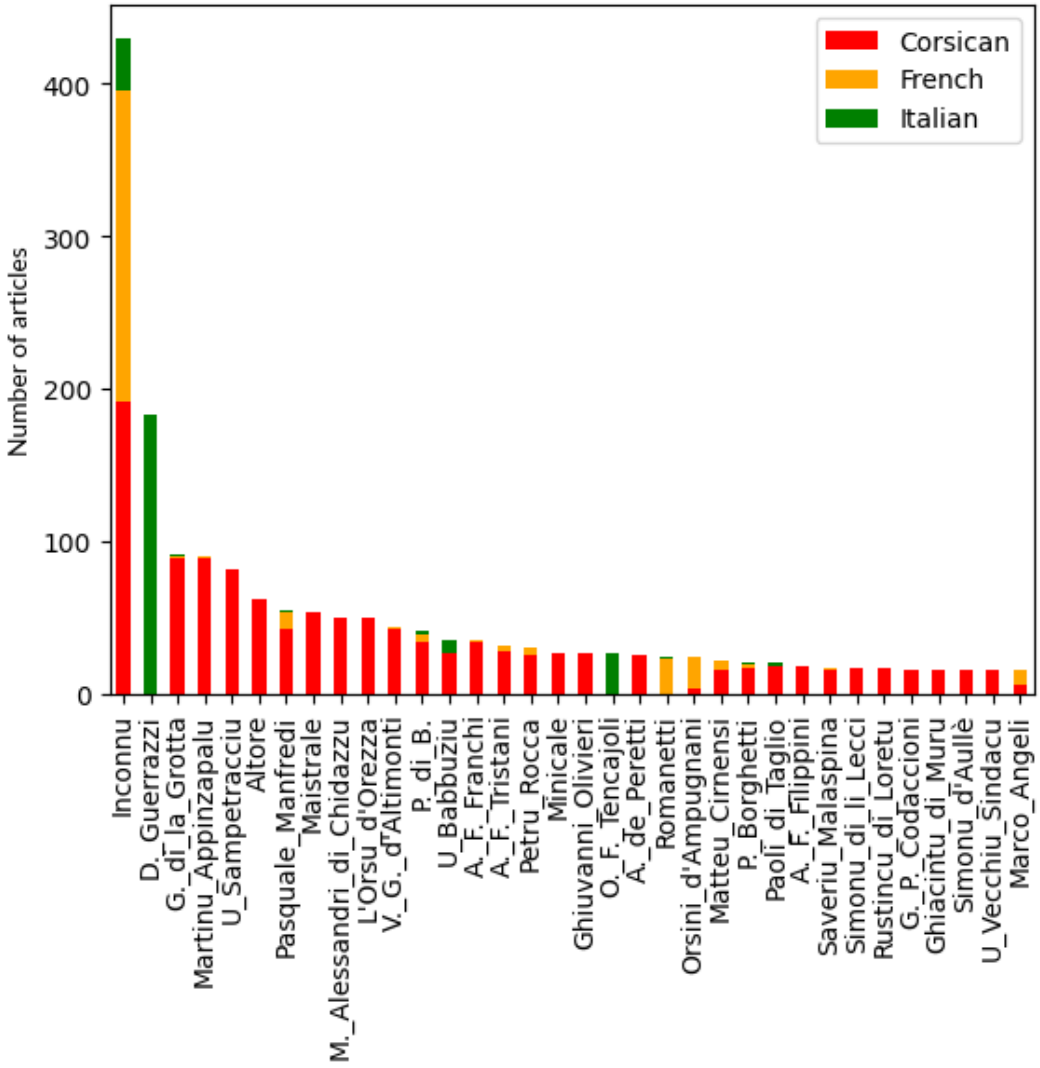


Figure 5.5: Language distribution of articles according to a selection of authors (dataset)

This graph completes figure 5.4 on information concerning authors. The selection was made with the authors having written at least 15 articles in our dataset, which corresponds to 33 different authors or pseudonyms, as well as anonymous texts or texts whose author is not explicitly specified. The Corsican language remains in the majority, even if there are certain specific

authors whose majority language is French or Italian. As Domenico Guerrazzi is a writer who predates the *Muvra*, he will be of interest exclusively in the section on topic modeling, as he is not a potential candidate for stylometry.

A Muvra, a symbol of pluralism

Chapter 6

Identifying author's pseudonyms by their writing style

6.1 Choosing the method

In literary studies, stylometry is defined as the measurement of the style of an author or genre. It can have several applications, including attributing authorship to texts, as is the case in our own study. The desire to identify authors by measuring the style of a text dates back to the 19th century, with numerous studies being carried out across Europe. In 1899, Paul Tennery published an article outlining the history of this discipline, which was still pioneering at the time. He cited the work of the German Wilhelm Dittenberger and gave an initial definition of the methodology to be applied:

> Seventeen years ago, Dittenberger proposed, in order to determine the chronology of the *Dialogues of Plato*, to follow a method consisting in making statistics of certain stylistic features, having no philosophical significance, but relating, for example, to the more or less frequent use of equivalent expressions or words.[90].

After the Second World War, the development of computational methods and computer science led to a renewed approach to stylometry in the humani-

[90]Paul Tannery, "La stylométrie ses origines et son présent," *Revue Philosophique de la France et de l'Étranger*, 47 (1899), p. 159: « Il y a dix-sept ans que Dittenberger a proposé, pour déterminer la chronologie des *Dialogues de Platon*, de suivre une méthode consistant à faire la statistique de certaines particularités de style, n'ayant aucune signification philosophique, mais relatives, par exemple, à l'emploi plus ou moins fréquent d'expressions ou de mots équivalents. ».

ties and social sciences. In 1963, the study of the authorship of the *Federalists Papers* represented a major turning point in the use of stylometry to attribute authorship[91].

Nowadays, stylometry has made many advances in methodology, particularly with the use of high-performance machine learning algorithms. Recent examples include the work of Jean-Baptiste Camps and Florian Cafiero, who used SVM classifiers algorithms to identify the authors of the American conspiracy forum *QAnon*[92]. This means that we can now tackle the question of the statistical units to be analysed with our algorithms, whether using machine learning techniques or distance metrics. The two French researchers chose to work on character 3-grams in order to "increase robustness, ad they are known to reduce sparsity and perform well in attribution studies"[93]. In reality, the features to be analysed vary according to the nature of the corpus and the quality of the data. One example is the measurement of verses in poetic works to measure an author's style[94] and even the rhymes in medieval texts like Mike Kestemont did in 2012[95]. This also raises the question of how to identify the authorship of a document. For some years now, the scientific community has been distinguishing between *authorship attribution* and *authorship verification*. The difference can be summed up as follows:

> Author identification (or authorship attribution) aims to reveal information about the individual(s) who wrote a text. [...] In authorship verification, texts of known authorship by one author are presented to a system, which is then tasked to verify whether another text has also been written by that same author.[96].

[91] Frederick Mosteller and David L Wallace, "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers," *Journal of the American Statistical Association*, 58–302 (1963).

[92] J.B. Camps and Florian Cafiero, "Who could be behind QAnon? Authorship attribution with supervised machine-learning," *arXiv*, arXiv:2303.02078 (2023).

[93] Ibid., p. 10.

[94] Valérie Beaudouin and François Yvon, "Contribution de la métrique à la stylométrie," in *Actes des 7èmes Journées Internationales d'Analyse Statistique des données textuelles (JADT)*, 2004, vol. 1.

[95] M. Kestemont, W. Daelemans, and D. Sandra, "Robust rhymes? The stability of authorial style in medieval narratives"…

[96] Efstathios Stamatatos, M. Kestemont, Krzysztof Kredens, Piotr Pezik, Annina Heini, Janek Bevendorff, Benno Stein, and Martin Potthast, "Overview of the authorship verification task at PAN 2022," in *CEUR Workshop Proceedings*, CEUR-WS, 2022, vol. 3180, p. 2301.

The advantage of verification for our own study is that it can perform on short texts of completely different genres. This is largely compatible with the fact that we are working on newspaper articles of a very varied typology, as we saw earlier. However, with the progress made in particular with the use of SVMs, these problems are less of an issue in the case of identification, allowing us to use algorithms of this type.

There have been several different methods for measuring the distance between texts. The idea is to supplement the results obtained with SVM classifiers with specific metrics. The first stylometric analyses were carried out by calculating the number of occurrences of a selection of tool words and features. This count was then projected onto a two-dimensional plane in order to study the curves of composition. This method, theorised by T. C. Mendenhall in 1887[97], has the merit of being a pioneering work in the field. More recently, the linguist Adam Kilgarriff published an article in 2001 in which he proposed using chi-square to calculate the distance between vocabularies in two corpora. However, the author also points out the limitations of this method:

> Two limitations on the validity of the method are, first, there are different ways in which corpora can be different. [...] Second, if the corpora are small and the difference in proportions between the corpora is also small, it is not clear that all the "gold standard" assertions are in fact true.[98].

Finally, the last method we will discuss in this section consists of calculating the Delta score as defined by John Burrow in 2002[99]. The advantage of this method over chi-square lies in the weight given to the most frequent words. The measurement rules of the different corpora thus make it possible to limit the bias of particularly frequent tool words. This is particularly useful in our case with regard to the ambiguities that exist in the normalisation of our data as seen previously on page 4.2.3. This method was used for the Arabic

[97]Thomas Corwin Mendenhall, "The characteristic curves of composition," *Science*–214s (1887).

[98]Adam Kilgarriff, "Comparing corpora," *International journal of corpus linguistics*, 6–1 (2001), p. 121.

[99]John Burrows, "'Delta': a measure of stylistic difference and a guide to likely authorship," *Literary and linguistic computing*, 17–3 (2002).

language in 2014, demonstrating the flexibility of the statistical model[100]. The idea is therefore to analyse our texts using the Burrows-Delta method and SVM classifiers in order to cross-reference the results.

[100]Ammar Adil Abdul Razzaq and Tareef Kamil Mustafa, "Burrows-Delta method fitness for Arabic text authorship Stylometric detection," *International Journal of Computer Science and Mobile Computing*, 3–6 (2014).

6.2 Are anthologists right? The case of *P.diB.*

6.2.1 Explanations and first tests

The pseudonym *P. di B.* is a name that appears fairly regularly in the writings of the *Muvra*. A number of articles were published under this pseudonym and it is generally accepted that it is actually Petru Rocca, as mentioned by Carmine Starace in the pages of his *Bibliografia della Corsica*[101]. This pseudonym is believed to be the initials of his mother's surname, Maria Saveria Rocca-Pozzo di Borgo[102]. The latter had remained very close to her sons Petru and Matteu, even publishing drawings in the *Muvra*, such as this drawing of an old shepherd from 1934[103].

VECCHIU PASTORE

Disegnu di Maria Saveria Rocca-Pozzo di Borgo.

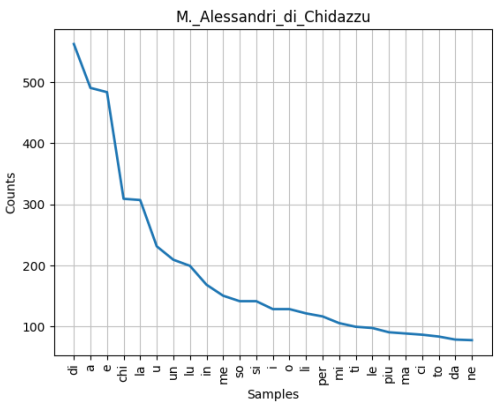
Figure 6.1: *Vecchiu pastore*, Maria Saveria Rocca-Pozzo di Borgo

[101] C. Starace, *Bibliografia della Corsica...*, p. 711.

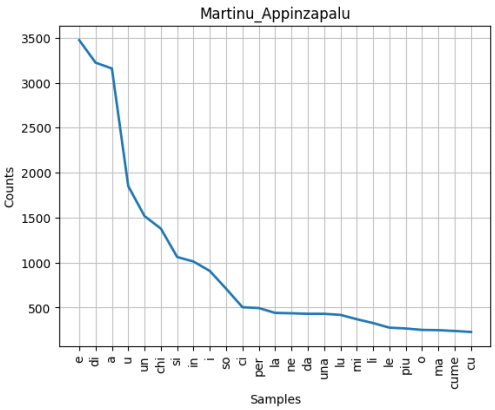
[102] While it is known that Maria Saveria was a close relative of Petru, there is still some doubt as to whether she was his mother, although it is certain that his mother bore the name Pozzo di Borgo

[103] *A Muvra*, n°539-12/30/1934: "Vecchiu pastore".

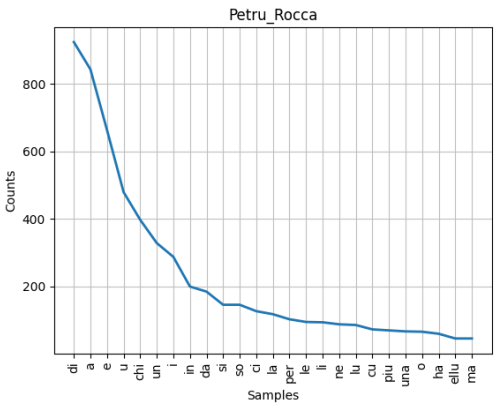
Confirming the writings of contemporary actors from this period also makes it possible to verify the rigour of their anthological work. It is also an excellent way of testing our methodology in a more or less reliable setting. In this first case study, we will try out the three vocabulary distance measurement methods seen in the previous section. We will cross-reference these results with those obtained when training an SVM model with the *SuperStyl* module developed by Jean-Baptiste Camps[104]. Finally, the algorithms for measuring distances are adapted from François Dominic Laramée's publication on the *Programming Historian* website, validated by Folgert Karsdorp, Jan Rybicki and Antonio Rojas Castro[105].



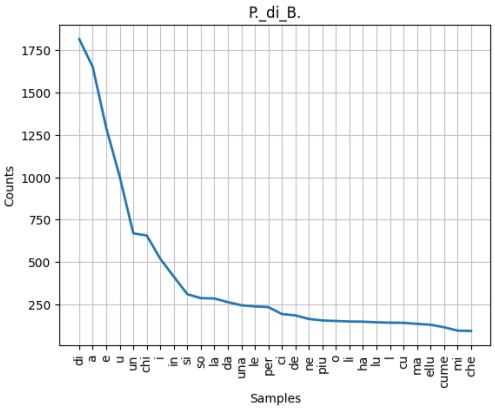
(a) Marcellu Alessandri di Chidazzu



(b) Martinu Appinzapalu



(c) Petru Rocca



(d) **P. di B.**

Figure 6.2: Medenhall's Characteristic Curves of Composition for *P. di B.*

[104]J.B. Camps, *SUPERvised STYLometry (SuperStyl)*, version 0.9.0, Oct. 2021.

[105]François Dominic Laramée, "Introduction to stylometry with Python," in *Programming Historian*, copyrights=CC-BY 4.0, 2018, URL: `https://doi.org/10.46430/phen0078`.

As we can see, distinguishing between the different composition curves can be quite complex. Taking into account the shape of these curves, we can see that, overall, the characteristics of Petru Rocca's corpus are the most similar among the other candidates. However, if we look a little more closely, we can see that certain tool words are better ranked according to their position on Martinu Appinzapalu's graph. Overall, this gives an initial indication of the complexity of the exercise, while also showing the limitations of this method. Although it is a good start, this method is too subject to variations in the size of the corpus, for example. Let's now look at the method proposed by Adam Kilgarriff:

	Petru Rocca	**Martinu App.**	**M. Alessa. di C.**
Chi-square	1548.15	3764.45	4122.63

Table 6.1: Kilgarriff's Chi-Square Score for *P. di B.*

Again, the results seem to confirm the same findings. Of all our candidates, Petru Rocca's style is the closest to that of *P. di B.*. There are several reasons why we chose to compare these specific candidates. Petru Rocca, as explained above, is the author to whom those of *P. di B* are generally attributed. For the others, it is a choice linked to the rate of publication. Under this pseudonym, the author published simultaneously and in greater quantity than a number of authors including Marcellu Alessandri di Chidazzu. Martinu Appinzapalu, whose real name was Dumenicu Carlotti, was a close collaborator of the *Muvra*, who was responsible for the *merendelle* of Corsican poets, a poetry competition in Corsican language. A priest by profession, he published mainly under his pseudonym, although he did write a few articles under his real name[106]. He only left the ranks of the Corsican poets in the mid-1930s because of his radicalisation with the irredentist *Corsica antica e moderna*.

[106]J. Mattei, P. S. Menozzi, and A.T. Pietrera, *Antulugia A Corsica Literaria...*, p. 533.

6.2.2 Confirming results

6.2.2.1 Burrow's Delta

Now that we have carried out our initial analyses with the first two methods of measuring distance between texts, we can set about confirming these results with the more reliable methods. We will first look at Burrow's Delta, which we will describe in a little more detail in this section. The idea is to calculate the proportion of n stopwords in a corpus, along with their mean and standard deviation. Using these elements, we first calculate the Z-score, which measures the difference between the proportion of the word x in a text and its overall proportion in the entire corpus C, known as its norm. Here is the formula for the Z-score, with μ representing the mean of x and σ representing the standard deviation:

$$Z_x = \frac{C_x - \mu_x}{\sigma_x}$$

Once this Z-score is calculated, we can use the result to calculate the delta score in order to compare the different texts with the target corpus of texts, in our case that of *P. di B.* The formula can be written as follows, where t represents the document analysed:

$$\Delta_C = \sum_x \frac{|Z_{C(x)} - Z_{t(i)}|}{n}$$

Thus, the closer the score is to 0, the closer the candidate's style is to the style of the anonymous author. In the case of *P. di B.*, we obtain the following results:

	Petru Rocca	**Martinu App.**	**M. Alessa. di C.**
Delta score	0.983	1.118	1.418

Table 6.2: Burrow's Delta Score for *P. di B.*

Once again, the results tend to show that Petru Rocca was indeed hiding behind the pseudonym of *P. di B.*, confirming previous empirical observations as well as the claims of contemporary authors of the time. To obtain the results

of the 6.2 table, we have chosen to take into account the 50 most frequent words, therefore representing mainly stopwords. Varying this parameter can produce more precise results, but without drastically changing the results. Finally, let's take a look at stylometric analysis using an SVM classifier, the *SyperStyl* module by Jean-Baptiste Camps mentioned earlier.

6.2.2.2 SVM Classifier *SuperStyl*

The Support Vector Machines (SVMs) are a class of supervised machine learning models used for classification and regression. They are particularly effective for solving binary classification problems, but can also be extended to handle multi-class classification. They seek to find an optimal hyperplane in a high-dimensional feature space that best separates data points of different classes. The hyperplane is chosen to maximise the margin, defined as the distance between the hyperplane and the nearest data points of each class, but flexible enough to allow misclassification of some observations. In the case of our stylometric analysis, the dimensions vary according to the number of statistical units to be taken into account, whether characters n-grams or word-tokens.

It is very important to vary the various hyperparameters available to us in order to optimise machine learning. To do this, the *SuperStyl* algorithms allow us great flexibility in the options to be taken into account. For example, we can choose the type of statistical unit, the sample size for rolling stylometry, the dimensional reduction method or the type of dataset balancing. In the case of *P. di B.* , if several tests have been carried out (all the tests can be found in the corresponding file on the *GitHub* repository), we have started with these hyperparameters: the statistical units are the stopwords, we apply the PCA (*Principal Component Analysis*) for dimensional reduction, we balance the dataset with the "upsampling" and the cross-validation is carried out with the "Leave-One Out" method. Once the model has been trained, the evaluation of which is presented in the table 6.3, we apply it to the unseen data.

	Precision	Recall	F1-score	Support
ALESSANDRI	0.96	0.90	0.94	50
APPINZAPALU	0.95	0.97	0.96	89
ROCCA	0.92	0.96	0.94	50
macro avg	0.94	0.94	0.94	164
weighted avg	0.95	0.95	0.94	164

Table 6.3: Detailed class scores using SVM for *P. di B.* — Leave-one-Out — PCA — Word-tokens — Upsampling

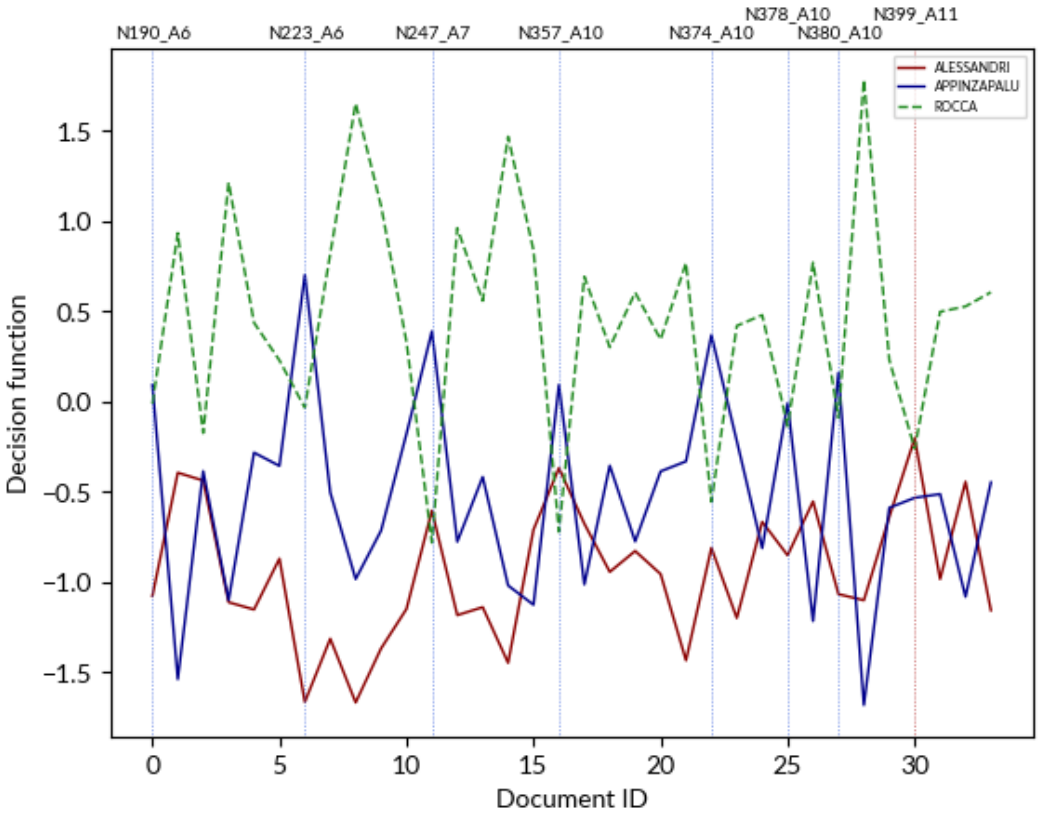


Figure 6.3: Value of the decision function for *P. di B.*

We then obtain a file with the predictions of the author of the articles with the results of the decision function. Florian Cafiero and Jean-Baptiste Camps describe this function by these words:

> The decision function tells us how close each sample is to the hyperplane separating each class. A negative value means that the sample is outside, a positive, inside. The higher the score, the deeper inside the class is located a dot, which can be interpreted as a strength of the authorial markers or an increase in the confidence of the classifier.[107].

By applying this function to our study, we get the graph 6.3. We have also added the identifiers of articles written by *P. di B.* whose authorship has not been attributed to Petru Rocca. On the whole, however, almost all the articles were attributed to Petru Rocca, once again confirming the results obtained previously. Of the 34 articles in the test corpus, 26 are attributed to the director of the *Muvra*, i.e. 76% of them. But what is even more interesting to study is the behaviour of the curves on the decision function graph. This graph shows that there is little doubt about the authorship of this pseudonym. On average, the decision function scores are much higher for Petru Rocca's texts.

[107]F. Cafiero and J.B. Camps, "'Psyché'as a Rosetta Stone? Assessing Collaborative Authorship in the French 17th Century Theatre," in *Proceedings of the Conference on Computational Humanities Research 2021*, CEUR-WS, 2021, vol. 2989, p. 382.

6.3 Who's hiding behind the mysterious *Altore*?

6.3.1 An active contributor

Let's now turn our attention to the author behind the pseudonym *Altore*. He is directly inspired by the lake of the same name in the Asco valley, in the old Caccia *pieve* within the region of the same name. *Altore* is the author of *Lettere aiaccine*, the letters from Ajaccio, which often appeared on the front page of the newspaper. In this format, he covers all the subjects of society and politics in general, in an open, family-friendly letter format. Our corpus contains 62 of these letters, all written in Corsican language. The difficulty with this part of our study is that we have no information or clues about the real author behind this pseudonym. Nevertheless, its presence on the front pages of many issues at least testifies to the importance attached to this particular section and therefore to its author.

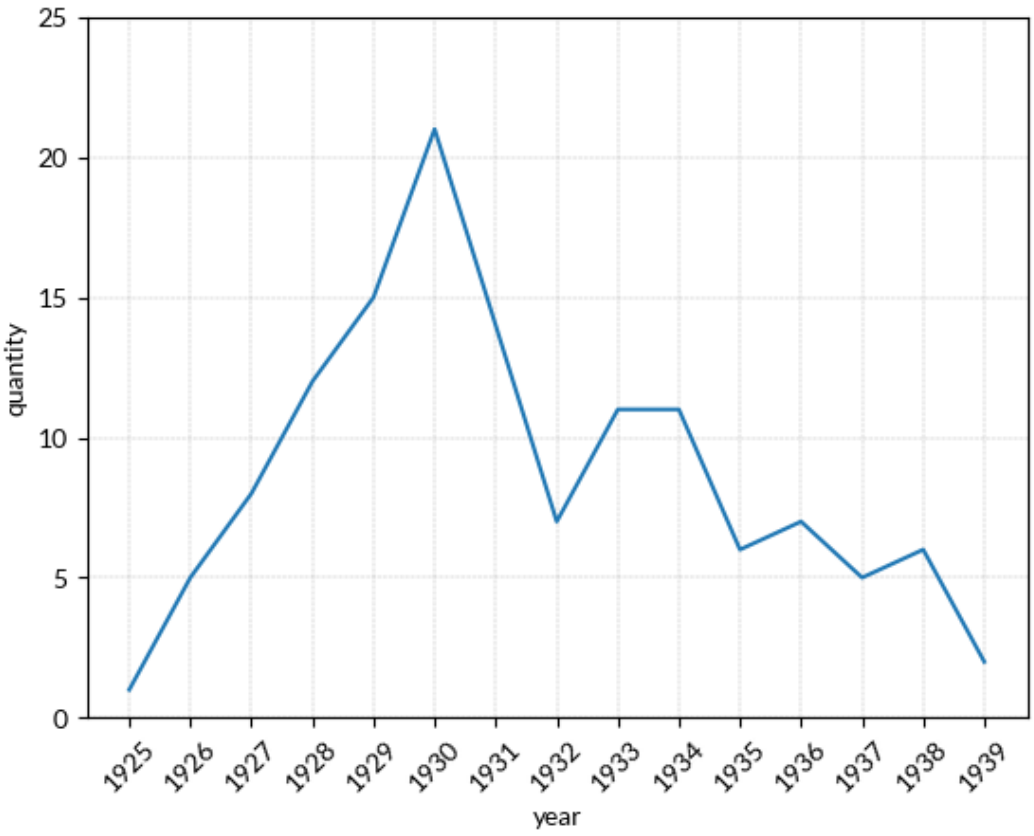


Figure 6.4: Quantity of *Lettere Aiaccine* published by *Altore*

The graph 6.4 shows the importance of *Lettere Aiaccine* over the years. Although the author published less and less from 1936 onwards, he was particularly active between 1925 and 1939. This graph is based on our corpus and the A/I database. Although these figures may be incomplete, as manual recording is very error-prone, the figures should not vary particularly much. From this curve, we can deduce that the author is probably close to the editorial board because of his high level of activity. Then, we have to choose candidates who have played an important role in the redaction of *Muvra* or the life of *Partitu Corsu Autonomista*. So we are hoing with the following authors:

- **Petru Rocca**
- **Dumenicu Carlotti** (Martinu Appinzapalu)
- **Ghjanettu Notini** (U Sampetracciu)
- **Orsu Francescu Piazzoli** (L'Orsu d'Orezza)
- **Marcellu Alessandri**
- **Dumenicu Antone Versini** (Maistrale)
- **Simon'Ghjuvanni Vinciguerra** (Ghiuvanni di a Grotta)
- **Victor Gianviti** (V. G. d'Altimonti)

Between them, these 8 authors represent approximately 480 of the 8247 articles in our corpus in Corsican, i.e. about 6% of them, which is still a more than acceptable number. Let us bear in mind the difficulty of the exercise: it is not a question here of confirming the authorship of an author, but rather of identifying an anonymous author without any clues. With a collection of almost 500 possible authors, the biases to be applied are highly discriminating. The potential results we could extract would not represent an absolute truth, but that is the nature of stylometry. This method does not act as proof but as a confirmation bias, which we supplement with elements that would tend to confirm these results.

6.3.2 A pseudonym of the playwright *U Sampetracciu*?

In the same way as we confirmed Petru Rocca's authorship of the texts of *P. di B.*, we will carry out the stylometric analysis of those of *Altore* using Burrow's Delta method as well as the SVM classifier. We will detail the different hyperparameters used to conduct the analysis and how we have tried to optimise them as much as possible in order to obtain the most accurate results possible.

We'll start by looking at the results of measuring distances between texts using the Burrow Delta technique. The parameter to be taken into account here is the number of most frequent words to be used. We can see quite quickly that a number of tokens are particularly present without being stopwords, such as "*corsica*" or "*corsu*". Nevertheless, these terms are particularly relevant given the very nature of the newspaper, being an autonomist newspaper about Corsica. This is what we can see in the figure 6.6 where we started with the 50 most frequent words. With those results, this setting of the hyperparameter should be enough to measure effectively the different distances.

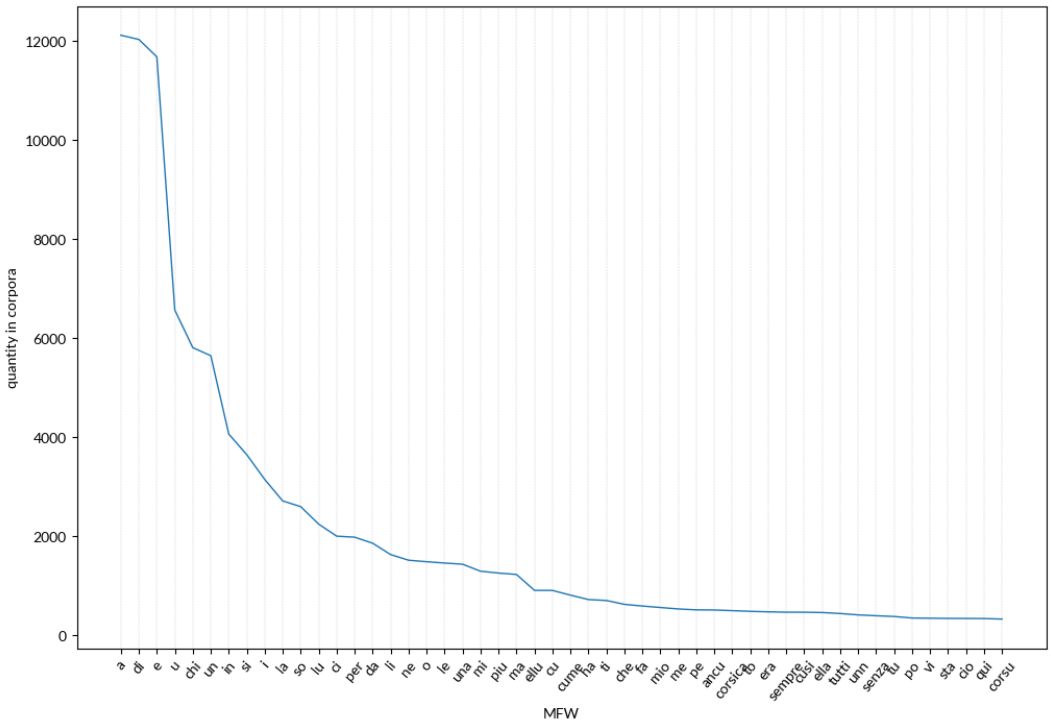


Figure 6.5: Top 50 Most Frequent Words in the whole corpus

With these parameters, we obtain the following results with the calculation of the Burrow's Delta score:

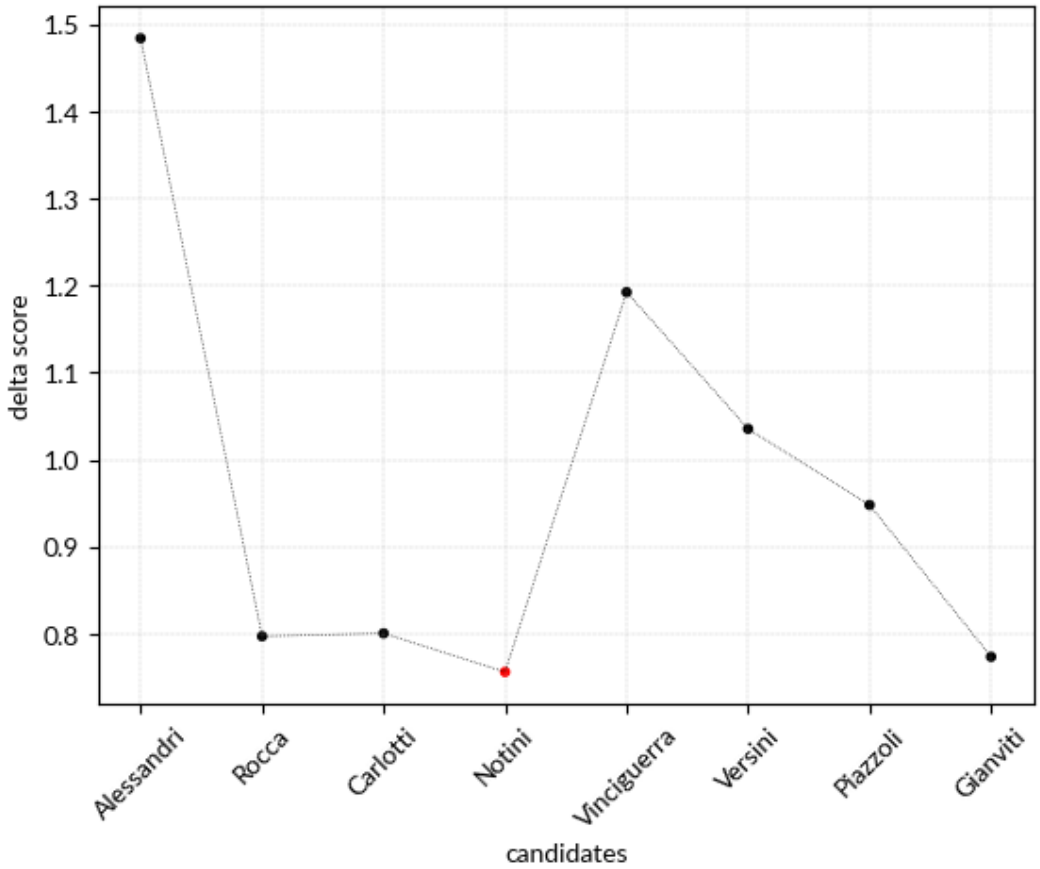


Figure 6.6: Visualisation of Burrow's Delta Score results for *Altore* — MFW 50

	Roc	**Car**	**Not**	**Pia**	**Ale**	**Ver**	**Vin**	**Gia**
Delta	0.797	0.801	0.756	0.948	1.484	1.035	1.192	0.774

Table 6.4: Burrow's Delta Score for *Altore*

As we can see, it would appear that the most likely candidates are Petru Rocca, Dumenicu Carlotti, Victor Gianviti and Ghjanettu Notini, the latter having the score closest to 0 at around 0.756. These numbers show the limitations of Burrow's Delta and the methods for measuring distance between texts in general. This can be seen by modifying the quantity of most frequent words

to be taken into account in the analysis. If we set the algorithm to 100, Notini and Gianviti remain dominant, but the latter would be the closest to Altore's texts (figure 6.7). However, with a limit of 100, the words taken into account have a deeper meaning as with the words "*sgiò*" (*sir*), "*core*" (*heart*) or "*casa*" (*home*). As our study is not a semantic analysis, these particular terms may distort the results. A pseudonym is used in particular to differentiate the author's words from his own signature, so it is normal that words with meaning are not representative of the author's style.

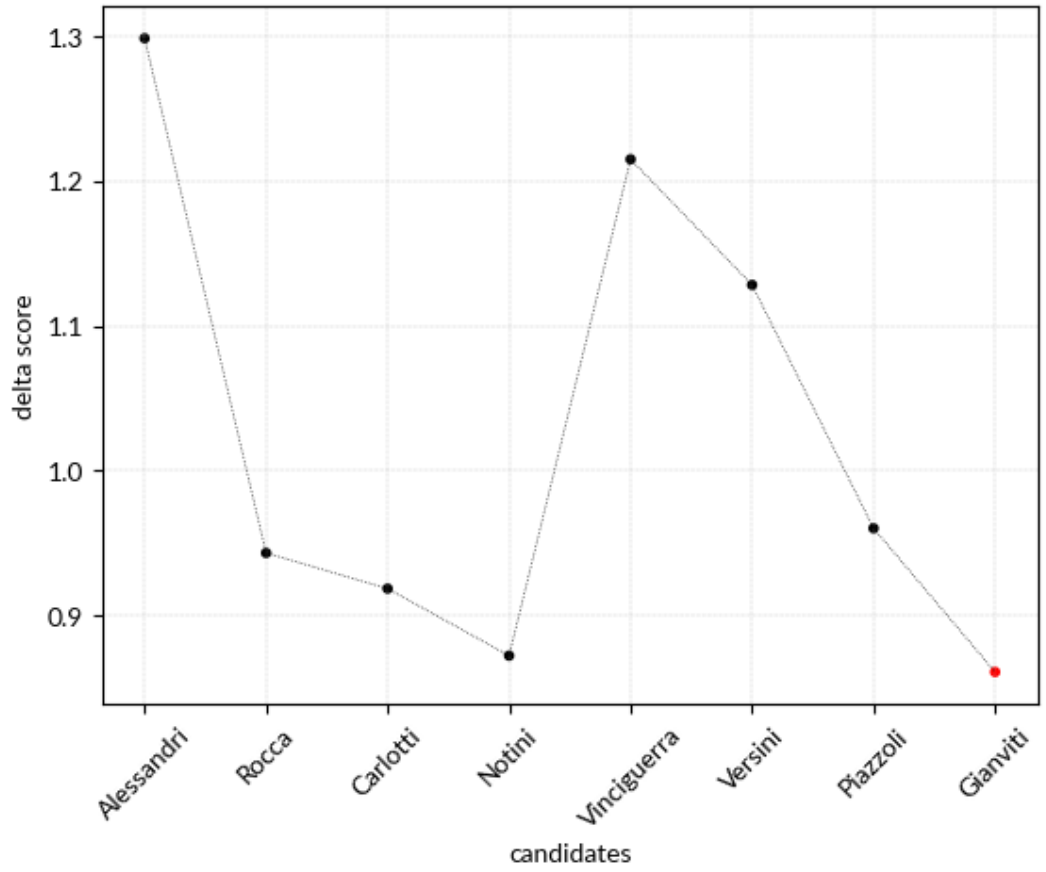


Figure 6.7: Visualisation of Burrow's Delta Score results for *Altore* — MFW 100

It is therefore necessary to supplement this analysis with the results obtained with the SVM classifier, given that Burrow's Delta is too sensitive to the semantic and typological varieties of the articles in our corpus. This analysis was carried out in two main stages: firstly, by splitting the training corpus in two and then by calculating the results using the entire corpus. This choice was made in an attempt to optimise the accuracy of the results. After trying

to vary certain parameters, we kept the same ones as for the *P. di B.* analysis, only varying those for the cross-validation.

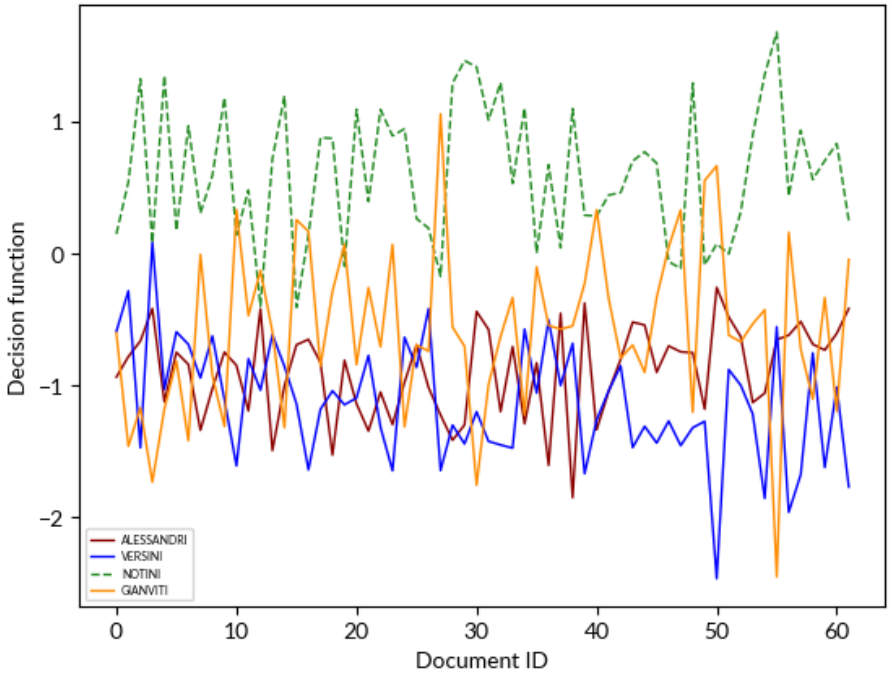
	Precision	**Recall**	**F1-score**	**Support**
ALESSANDRI	0.93	0.76	0.84	50
VERSINI	0.96	0.94	0.95	53
NOTINI	0.91	0.93	0.92	81
GIANVITI	0.81	0.98	0.88	43
macro avg	0.90	0.90	0.90	227
weighted avg	0.91	0.90	0.90	227

Table 6.5: Detailed class scores ALE-VER-NOT-GIA — Leave-one-Out — PCA — Word-tokens — Upsampling

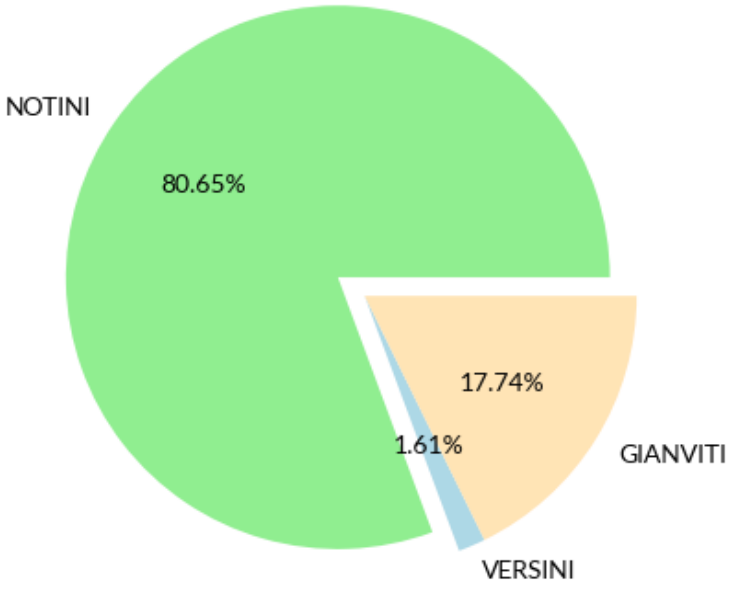
	Precision	**Recall**	**F1-score**	**Support**
ALESSANDRI	0.93	0.74	0.82	50
VERSINI	0.94	0.94	0.94	53
NOTINI	0.90	0.90	0.90	81
GIANVITI	0.77	0.95	0.85	43
macro avg	0.89	0.88	0.88	227
weighted avg	0.89	0.89	0.88	227

Table 6.6: Detailed class scores ALE-VER-NOT-GIA — K-Fold 35 — PCA — Word-tokens — Upsampling

With the scores of the tables 6.5 and 6.6, we obtain an accuracy of **0.90** and **0.89** respectively. In the same way as for the analysis on *P. di B.*, the scores are globally much more interesting with the cross-validation using the Leave-one-Out method.



(a) Value of the decision function



(b) pie-chart of the distribution

Figure 6.8: Authorship attribution of *Altores*'s texts for ALE-VER-NOT-GIA

As we can see, training on the reduced corpus but including Gianviti and Notini shows that the latter's style is much closer to that of *Altore*. In addition to almost 80% of the texts attributed to *Sampetracciu*, we can clearly see that the decision function curve behaves better than that of *d'Altimonti*, with almost exclusively positive values. The ambiguity between the two authors persists somewhat, but it must be admitted that the results are much clearer and more descriminating than with Burrow's Delta method. The relevance of the model is also confirmed by the much less equivocal results with the second part of the corpus, as we can see in figure 6.9.

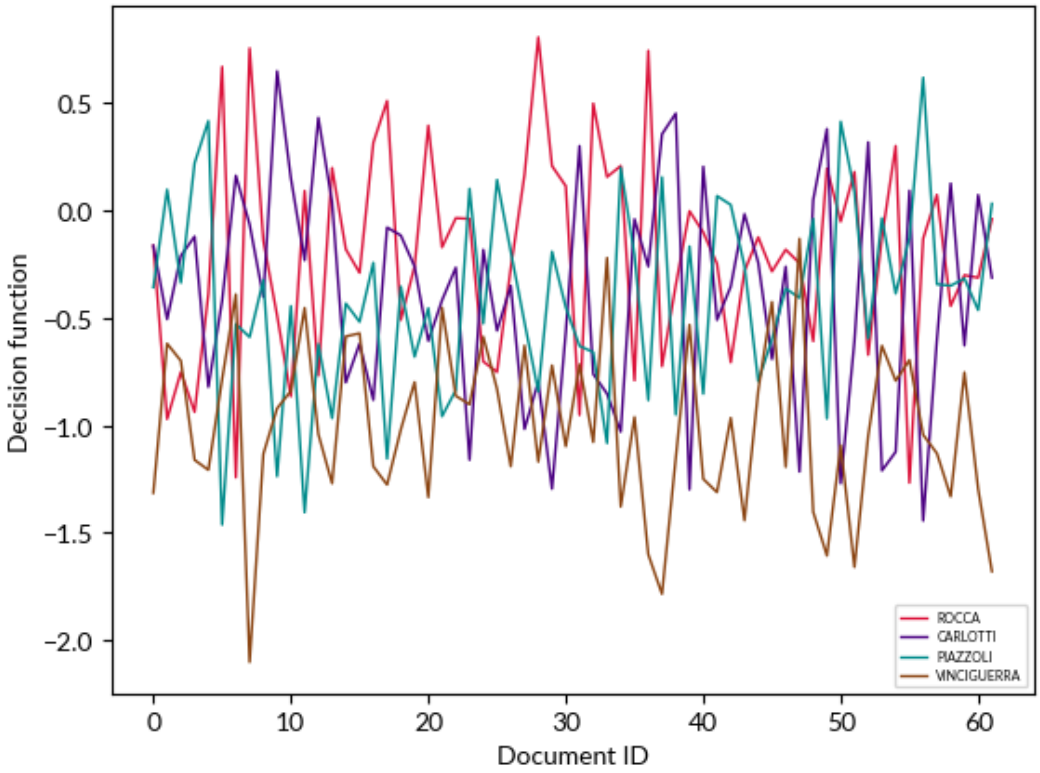


Figure 6.9: Value of the decision function for ROC-CAR-PIA-VIN

These initial results therefore show a trend towards the two authors who already use the pseudonyms *U Sampetracciu* and *V. G. d'Altimonti*. However, the initial results of the analysis show that the figures obtained with Burrow's Delta with the correct parameters are relatively reliable, since the same trends were found. It is now time to look at the results obtained with the SVM classifier on the entire training corpus.

	Precision	**Recall**	**F1-score**	**Support**
VINCIGUERRA	0.92	0.91	0.22	89
PIAZZOLI	0.83	0.68	0.75	50
ALESSANDRI	0.95	0.82	0.88	50
VERSINI	0.96	0.89	0.92	53
CARLOTTI	0.86	0.96	0.90	89
ROCCA	0.69	0.88	0.77	25
NOTINI	0.81	0.74	0.77	81
GIANVITI	0.76	0.95	0.85	43
macro avg	0.85	0.85	0.85	480
weighted avg	0.86	0.86	0.86	480

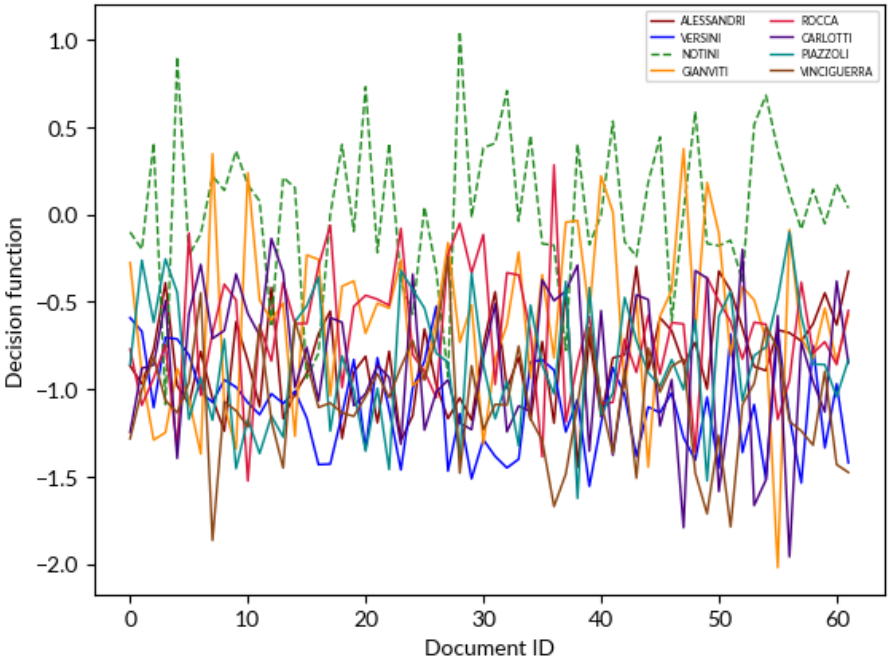
Table 6.7: Detailed class scores ALL CORPUS — Leave-one Out — PCA — Word-tokens — Upsampling

Although the results are not bad, they are nonetheless lower than those obtained with a prior subdivision of the training corpus and therefore of the number of candidates. We thus obtain an accuracy of about **0.86**. This test bears witness to another important aspect of stylometry which has not yet really been addressed in this dissertation: the notion of corpus size as a function of the number of candidates. This echoes the article by Eder Maciej published in 2015 at Oxford University:

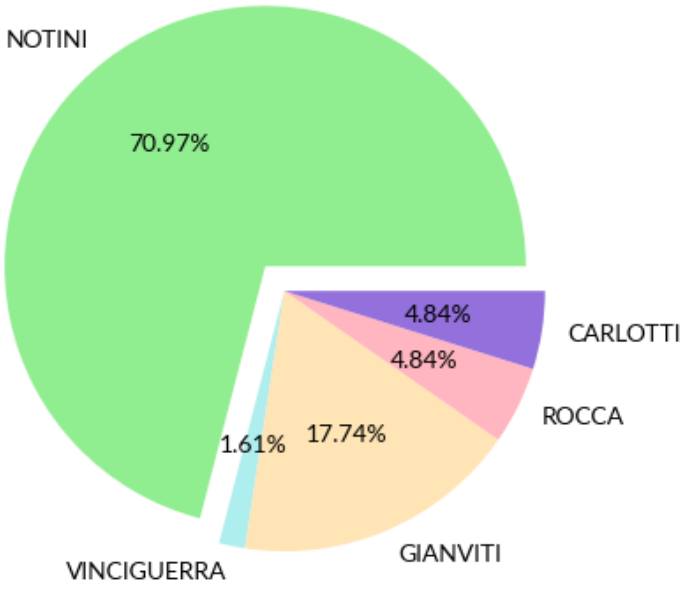
> It has been revealed that the effectiveness of attribution depends on corpus size and particularly on the number of authors tested. In short, a 2-class authorship attribution case needs less textual data than a 100-class case.[108].

We obtain the following results, bearing in mind that more data would be preferable to obtain better results with 8 classes (candidates).

[108]Maciej Eder, "Does size matter? Authorship attribution, small samples, big problem," *Digital Scholarship in the Humanities*, 30–2 (2015), p. 177.



(a) Value of the decision function



(b) pie-chart of the distribution

Figure 6.10: Authorship attribution of *Altores*'s texts for ALL CORPUS

Once again, the results show us that Ghjanettu Notini is the most likely candidate among the panel of candidates. We can then ask ourselves the following question: is he really the author behind *Altore*? As mentioned above, it is difficult to answer this question with absolute certainty. Stylometry, like any computational method used in the field of digital humanities, also requires more in-depth research with "close reading". Numbers are not proof. Ghjanettu Notini was born on 4 December 1890 in San Petru di Venacu, in the old pieve of Venacu in Corsica's *Curtinese* region. Interestingly enough, this region of central Corsica is relatively close to Lake Altore. He was a Corsican poet and writer who contributed for many years to the *Muvra* under the pseudonym *U Sampetracciu*. Nicknamed the "Corsican Molière", according to Ghjacumu Thiers[109], he was the founder of the *Teatru corsu di A Muvra* in the early years of the newspaper and a loyal contributor.

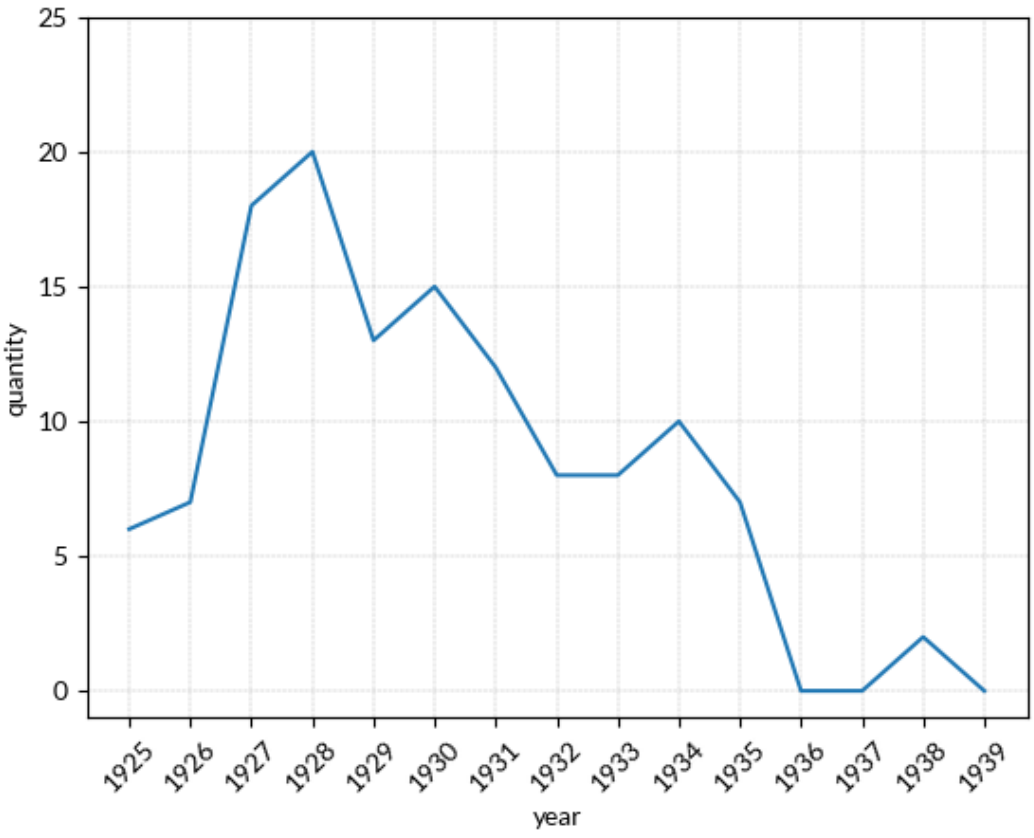


Figure 6.11: Number of articles published bu Ghjanettu Notini between 1925 and 1939

[109]See: https://www.interromania.com/corsu-cismuntincu/literatura/teatru/identite-culturelle-et-theatre-en-corse-628.html

With figure 6.11, we can see that the trends are somewhat similar, with a peak in publication between 1925 and the early 1930s, before a regression and then a fall in participation in 1936. This was the pivotal period when the *Muvra* became more radical, moving ever closer to Fascist Italy and the irredentists. It was also at this time that many contributors turned their backs on the autonomist paper, either because they no longer agreed with the editorial line or simply by fear of government reprisals. The fact that Notini is not known for being a fervent supporter of the irredentist cause also tends to confirm this hypothesis. Now that we have been able to shed some light on the identity of certain authors, let's look at the role of pseudonyms for authors.

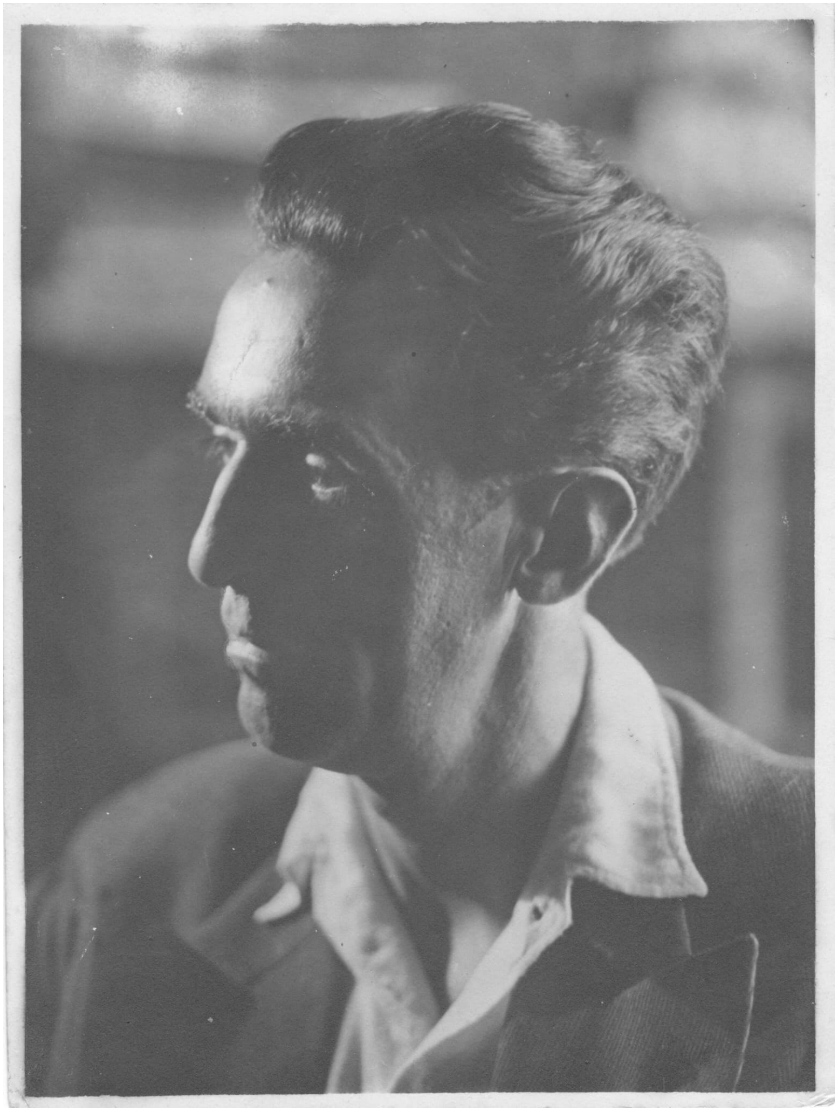


Figure 6.12: Picture of Ghjanettu Notini in the 30's

Chapter 7

The specific roles for the pseudonyms used

7.1 General informations about topic modeling

Topic modeling is a text analysis technique used in artificial intelligence and automatic natural language processing. Its aim is to identify and group together the main topics or themes present in a set of documents. Topic modeling is generally based on statistical models, such as the Latent Dirichlet Allocation (LDA) model, which is one of the most commonly used algorithms. This model considers that each document is a combination of several 'topics' and that each topic is a distribution of words. The topic modeling process consists of extracting significant information from a corpus of unlabelled texts. It identifies the most frequent keywords in each topic and then assigns a probability to each document for each topic. This enables documents to be organised according to their predominant topics, making it easier to explore and analyse large sets of texts. The use of this method is particularly relevant in our case because it should enable us to model the most relevant latent topics from our different corpora by applying different levels of distinction to them. There are two main methods of topic modelling, which we will describe below.

The Latent Semantic Analysis (LSA) approach was introduced by Susan Dumais in 1990[110], when the technique began to be developed in the NLP

[110]Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and

field. It uses a mathematical technique called Singular Value Decomposition (SVD) to reduce the dimensionality of term-document matrices and identify semantic similarities. Latent Semantic Analysis (LDA) was introduced in 2003 by David Blei, who considers each document as a distribution of subjects and each subject as a distribution of words[111]. It is used to infer distributions of subjects in documents and distributions of words in subjects. This method was improved a year later by Thomas Griffiths and Mark Steyvers by incorporating Bayesian inference techniques[112].

As we saw earlier, the LDA (*Latent Dirichlet Allocation*) is a method based on a term-document matrix. This method is based on the assumption that "documents are represented as random mixtures of latent topics, where each topic is characterised by a distribution of words". The LSA (*Latent Semantic Analysis*), on the other hand, consists of creating a semantic space based on a corpus in which similarities between words or documents are calculated on a statistical scale. Each of these methods has its own advantages and disadvantages that need to be taken into account, hence the importance of the notion of comparability inherent in our study[113]. Because of the intrinsic differences between algorithms, these advantages vary enormously. For example, LSA has a much faster calculation speed or a more direct interpretation of results. LDA, on the other hand, is a more robust and flexible model. In 2020, a team of researchers attempted to compare the two methods by training them on a corpus of BBC articles. According to them, LSA is more effective with a lot of data and fewer iterations than LDA, the latter being better suited to smaller corpora[114].

As we have seen, vocabulary plays an essential role in these analyses. The words chosen to be taken into account in topic modelling must not be too

Richard Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, 41–6 (1990).

[111]David M Blei, Andrew Y Ng, and Michael I Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, 3–Jan (2003).

[112]Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum, "Integrating topics and syntax," *Advances in neural information processing systems*, 17 (2004).

[113]Toni Cvitanic, Bumsoo Lee, Hyeon Ik Song, Katherine Fu, and David Rosen, "LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents," in *International Conference on Case-Based Reasoning*, 2016.

[114]Yaswanth Kalepalli, Shaik Tasneem, Pasupuleti Durga Phani Teja, and Suneetha Manne, "Effective comparison of LDA with LSA for topic modelling," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2020.

numerous, as learning the model can be extremely time-consuming. The number of documents and the vocabulary chosen will therefore play a central role among the various biases to be applied in this phase of the thesis. Unlike stylometry, stopwords are of no interest because they are considered to be empty words. On the contrary, these words constitute noise which is not necessarily desirable and which can distort our results. The idea is therefore to remove them in order to reduce the vocabulary. But there is also the case of hapax or infrequent words, as well as frequent words that are not stopwords, such as "*corsica*" in this case. One solution is to include the notion of statistical entropy in the choice of vocabulary as presented by Susan Dumais in a 1992 article[115] with the following formula:

$$E = 1 - \sum_{j} \frac{p_{i,j} log(p_{i,j})}{log(ndocs)} \; and \; p_{i,j} = \frac{tf_{i,j}}{gf_i}$$

In this equation, *ndocs* represents the number of documents, *tf* is the frequency of the term i in the document j and *gf* is the overall frequency of the term i. The idea is to calculate the entropy of each word in the corpus and to select vocabulary within a defined interval.

[115]Susan Dumais, *Enhancing performance in latent semantic indexing (LSI) retrieval*, 1992.

7.2 *Altore*, the perfect cover?

To begin the topic modeling analyses, we will use the latest results obtained with the pseudonym *Altore*, whose author could be Ghjanettu Notini. This is a good transition to explain our methodology but it is also a good example of author profiling, the initial objective of this work. The first step was therefore to remove the stopwords from the corpus of articles. The difficulty presented here is that we are working on a language with few digital resources. This is reflected not only in the lack of PoS tagging or lemmatisation tools, but also more basic tools such as a list of Corsican stopwords. We therefore had to create a custom list which will also be available for download from the GitHub repository in order to benefit research on the Corsican language. Identifying stopwords in under-resourced languages is a major challenge for these specific languages, with some researchers going to great lengths to identify them, using clustering techniques for example[116]. Fortunately, Corsican is an Italo-Romance language and therefore shares a number of stopwords with Italian. This list tries to include as many graphic varieties of these words as possible in space and time: space, through the different regiolects used in the *Muvra*, and time, through the evolution of grammatical rules. This list was therefore created on the basis of observations of the corpus with certain particular forms to be taken into account, for example with "*aghju*" and "*aghiu*" (*I have*) or "*inde*", "*ind*" or "*indu*" (*inside/in*). Running the topic modelling algorithms also enabled us to identify certain relatively frequent words that we hadn't necessarily suspected. This variety of stopwords prevented us from using a generic list in Italian.

Once the stopwords had been removed from our corpora, we could move on to the vocabulary selection phase by calculating the point entropy of each word in a corpus of texts. Theoretically, with the right interval, we could not have deleted the stopwords from the corpus because their entropy would have been much higher than the average. But in the interests of rigour, it was preferable to remove them beforehand, as this would not affect the score of the other terms. Another criterion for distinguishing between important and unimportant words in a text is their weight, which amounts to multiplying the

[116]Faathima Fayaza and F Fathima Farhath, "Towards stop words identification in Tamil text clustering," *International Journal of Advanced Computer Science and Applications*, 12–12 (2021).

point entropy of a term by its overall frequency. So two terms can have the same entropy but radically different weights.

Now that the basics have been introduced, let's turn our attention to the author *Autore*. We saw earlier that it would appear that the person behind this pseudonym could be Ghjanettu Notini, as known as *U Sampetracciu*. Let's try to understand why such an author would need to hide behind a false identity in order to publish. The graph 7.1 represents the calculation of the point entropy of each term as well as their weight.

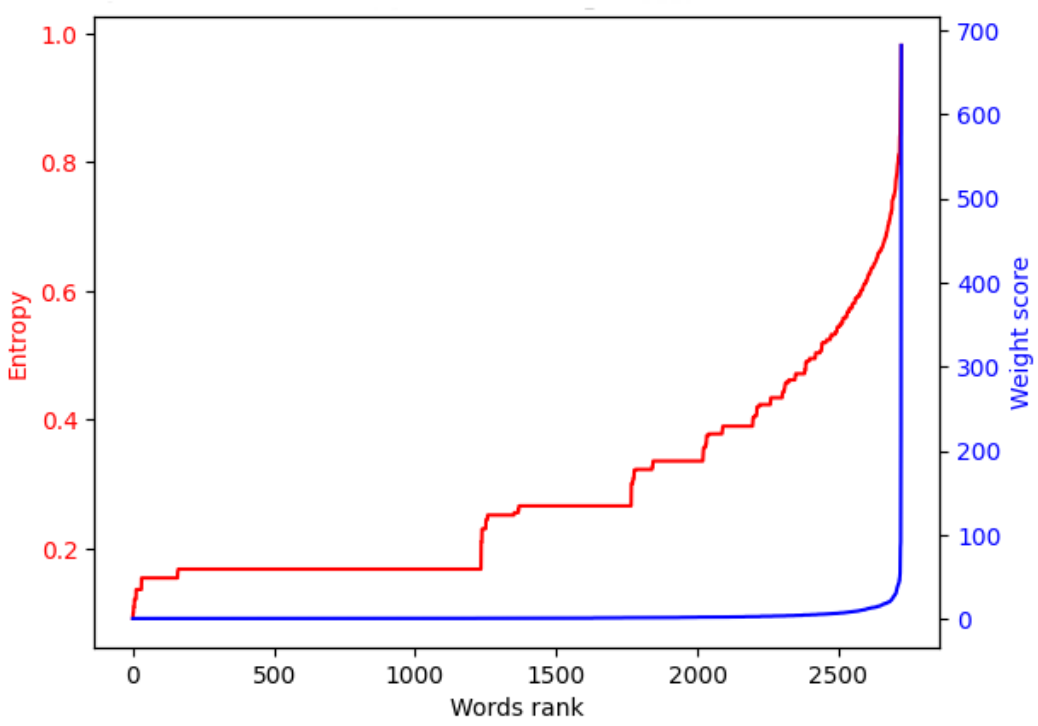


Figure 7.1: Entropy and weight score of *Altore*'s vocabulary

As we can see, the limit of the entropy score is around 0.85 because we have removed the stopwords beforehand. Similarly, the dynamics of the word weight curve are quite interesting. The vast majority of words have a very low weight, demonstrating the over-representation of a certain number of very specific terms. The aim is therefore to find an interval straddling the two curves, taking into account words that are frequent enough to be relevant but not over-represented to avoid too much noise in the results. Thus, in view of the results obtained with this calculation, the choice was made to select an

interval between **0.3** and **0.85** on the entropy scale. We therefore obtain a list of words to be deducted from our total vocabulary of **9455** words, reducing our total vocabulary to **7655** words.

Once we have selected our vocabulary, we can start training the LDA and LSA models. To do this, we'll use Python's *Gensim* library[117]. The idea is to play with the parameters of the different functions to obtain the best possible results. The intrinsic evaluation of a topic modeling model is not straightforward, since everything depends on the use we make of it. In our specific case, we are not looking to annotate a corpus of texts with their most likely topic, but rather to identify trends in the type of subject tackled, as a whole, by a certain author. The "coherence rate" exists but it's not enough to be sure that our model is good, that's why it is necessary to validate it also in an extrinsic way with an empirical aspect: does this topic generated makes sense? Visualising topics is therefore essential. That's why we've opted for wordcloud-style graphs and tables showing the 10 most frequent words in the topics.

Figure 7.2: LDA anaylisis on *Altore* — k=4 — iter400 — pass20

[117]See: `https://pypi.org/project/gensim/`

Figure 7.3: LDA anaylisis on *Altore* — k=4 — iter250 — pass20

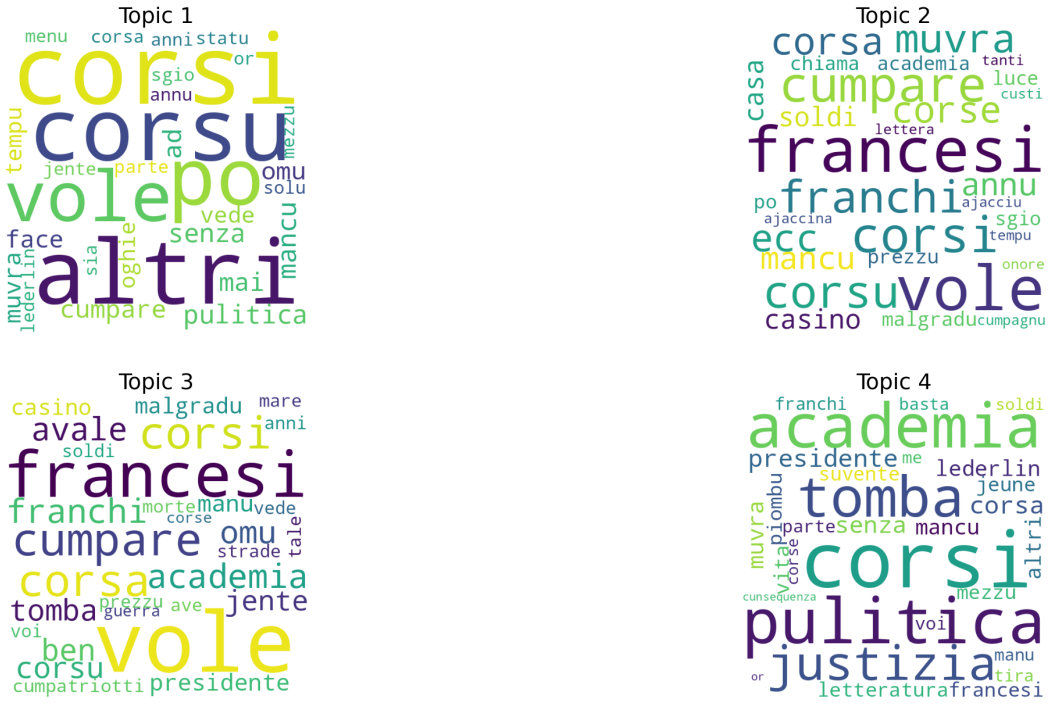
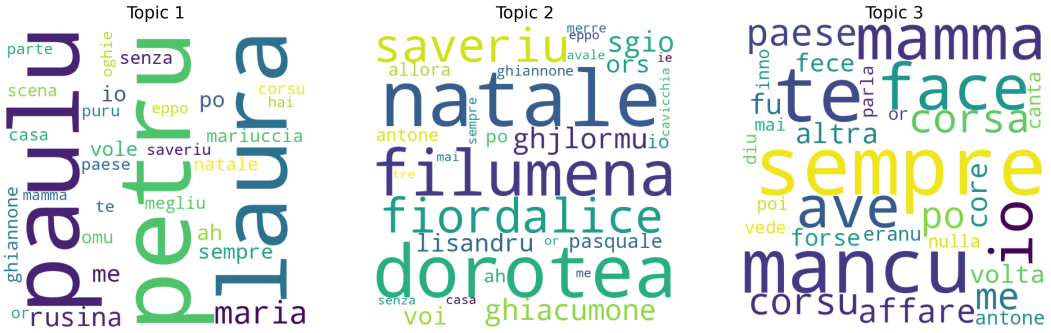
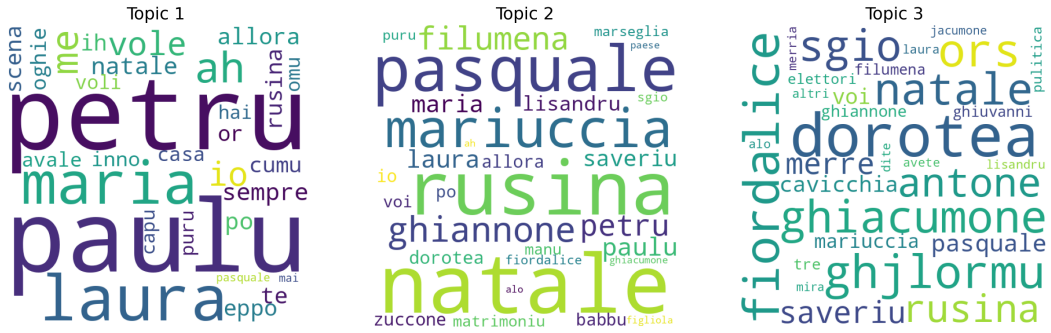


Figure 7.4: LSA anaylisis on *Altore* — k=4 — words20

What we can observe is that the LDA seems to give more interesting results when 250 iterations are performed on the dataset instead of 400, representing the base value of the *Gensim* function parameter. Whether for the LDA or the LSA, it is easy to notice certain terms that come up frequently, such as "*corsu*" or "*corsica*". This brings us face to face with our vocabulary selection methodology. These words are very frequent but remain essential in the context of a Corsican autonomist newspaper. Nevertheless, certain trends stand out, with political issues omnipresent in these *lettere aiaccine*. In particular, there is the notion of the French politician and industrialist Paul Lederlin, who was elected Senator for Corsica in 1930[118]. The francization of Corsican politics was a very important theme for the muvrists, who regretted that the island was becoming nothing more than an electoral breeding ground for continental politicians. This was part of the general theme of the "*isula persa*", with Corsica representing nothing more than an *El Dorado* for certain opportunists.



(a) LDA — k=3 — iter500



(b) LSA — k=3

Figure 7.5: Top topics for *U Sampetracciu*

[118]Patrick Cabanel and André Encrevé, *Dictionnaire biographique des protestants français de 1787 à nos jours*, vol. 3 H-L, Paris, Les éditions de Paris / Max Chaleil, 2022, p. 697.

So with the these figures, we see that the plays written by Ghjanettu Notini are particularly dominant in the detection of topics. This can be seen thanks to the large number of first names, typical of the theatrical style which incorporates a lot of dialogue. Other elements highlight this, such as the presence of the onomatopoeia "*Ah*" or the term "*scena*" (*scene*). We can also observe the poetic dimension of Notini's work with Topic 3 of the LDA: we find there the lexical field typical of Corsican poems with the importance of the "*mamma*", the *mother* or the mention of local flora with the "*fiordalice*" (*Pancratium illyricum*). This is a fairly rare Mediterranean flower that is only found in a few places around the world. In France, this plant is found only in Corsica, rapidly becoming a symbol of the island's particularism in the eyes of the muvrists.

Can we talk about a cover, then, if Ghjanettu Notini does indeed use the pseudonym *Altore*? It seems fairly obvious that the Corsican author seems more inclined to evoke political and topical themes with the pseudonym. He does this in a very particular literary style, that of the open letter, which corresponds quite well to Notini's great talent for writing. However, Notini did not hesitate to raise these intrinsically political issues in his plays. Likewise, his poetry does not appear to be a simple ode to the beauty of Corsica, but a complete reworking of the island's poetic traditions through the prism of *lamentu*.

7.3 Petru Rocca and its pseudonyms

In order to complete the analyses carried out on the pseudonym *P. di B.*, it is time to apply the same methodology as for Notini to the various identities of Petru Rocca. Petru Rocca is an expert in this field, as nearly 5 different identities are attributed to him in the various anthologies and studies carried out on him. We find his signature Petru Rocca or Pierre Rocca and the pseudonyms *Pasquale Manfredi*, *P. di B.* and *P. di C.* In view of the stylometric results, we can assume that these various identities attributed to him are indeed his own. From there, we can prepare our corpus accordingly.

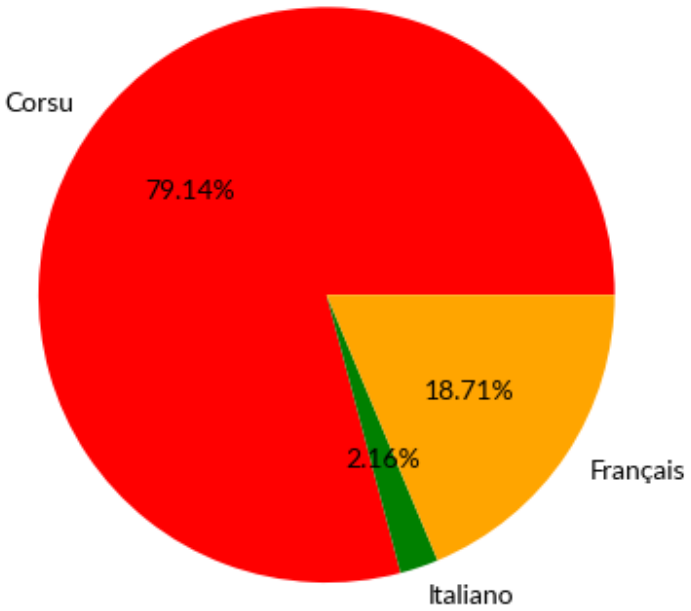


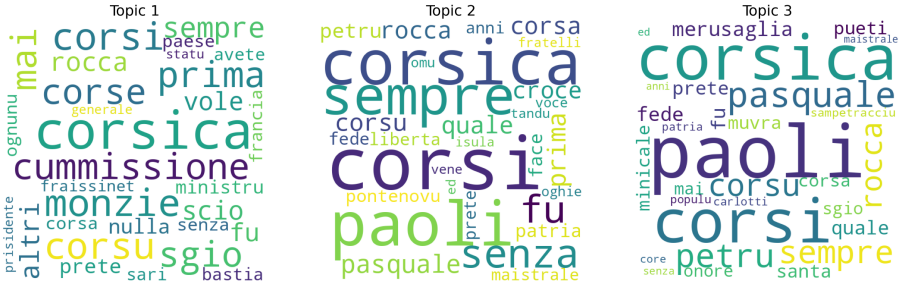
Figure 7.6: Distribution of languages used by Rocca in all his texts

The graph 7.6 highlights the fact that Petru Rocca writes mainly in Corsican, although he does leave an important place for French. He also writes a little in Italian, but there are too few texts to be relevant. From a methodological point of view, we won't do it the same way this time, as the basic vocabulary isn't large enough to be a handicap in training the model. Similarly, we will only be using LDA to process a smaller corpus than previously. If we can reference **139** articles written by Rocca in total, we will perform the algorithm on sub-corpuses according to language and signature.

(a) COS — k=3 — iter600

(b) FRA — k=3 — iter600

Figure 7.7: LDA analysis on *Petru Rocca*



(a) COS — k=3 — iter600

(b) FRA — k=3 — iter600

Figure 7.8: LDA analysis on *Pasquale Manfredi*

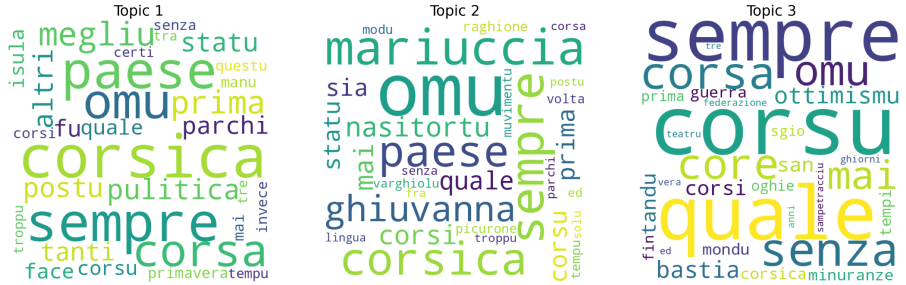


Figure 7.9: LDA analysis on *P. di B.* — COS — k=3 — iter600

It is important to note that for reasons of data quantity, we have grouped together in the same sub-corpus the texts signed by Petru Rocca and Pierre Rocca as well as the texts signed by *P. di B.* and *P. di C.* We assume that these have the same utility, but this is obviously a point to be improved in further analyses of the question. The first observation to make is that Pierre Rocca doesn't hesitate to tackle topical subjects and doesn't hesitate to denounce what he considers to be injustices under his real identity. He regularly talks about France and the State, but also about the Corsican national movement. It would seem, however, that he uses the Corsican language to also address cultural themes with the notion of "*lingua*" (*language*) and "*diu*" (*god*) in topic 2 of the figure 7.7a. For French, Petru Rocca mentions politicians in topic 3 of the figure 7.7b. These include Louis Marlier, prefect of Corsica from 1924 onwards, who cut his teeth in the general security service[119], and Léon Daudet, a Maurassian-inspired politician and supporter of the *Action française*, with whom the *Muvra* maintains good relations[120]. Petru Rocca also seems rather severe on the question of Corsican emigration to the continent, given the language used in topic 2 of the figure 7.7b with the notion of "*fuir*" (*flee*), "*terre*" (*land*) and "*ailleurs*" (*elsewhere*). The Corsicans' departure for mainland France is a very important subject in the muvrist discourse, which they describe as exile, as shown by this text written in 1922:

> Flee or shiver, arms folded across the ground. Flee; or wither, yellow, age, drained like a grape without juice; give in to the count-

[119]The fact that he was director of the "Sûreté Nationale" and then prefect of Corsica is quite interesting because it shows how much the government cared about Corsica's separatism and irredentism. See: René Bargeton, *Dictionnaire biographique des préfets, septembre 1870-mai 1982*, Paris, Archives nationales, 1994.

[120]See: François Broche, *Léon Daudet: le dernier imprécateur*, Paris, Éditions Robert Laffont, 1992.

> less vibrios of the race's strong and precious blood. Flee. Flee! or bleach the humus from your vertebrae. Flee...[121].

The topics modelled for *Pasquale Manfredi* do not vary greatly in the types of subjects tackled, even if the content varies. Petru Rocca seems to use this pseudonym to be very politically committed and to denounce political or economic practices deemed to go against the interests of Corsicans. This is the case, for example, with the ongoing attacks on the Fraissinet shipping company, which runs services between Marseille, Toulon and the port of Bastia. *Manfredi* also seems to evoke this theme in Corsican and French, with topic 1 in figure 7.8a and topic 2 in figure 7.8b respectively. The Fraissinet family controlled many shipping lines, with the Corsican line being created in 1862 and remaining active until 1948. It was Alfred Fraissinet, followed by his son Jean from 1927, who decided to concentrate on the Mediterranean basin. Fraissinet's prices and unsavoury business practices had made it a major target for the muvrists. Jean Fraissinet wpointed out a few decades later that "it was jokingly said that there were three scourges in Corsica: malaria, *libeccio*... and Fraissinet!"[122].

Finally, Petru Rocca wrote almost exclusively in Corsican under the name *P. di B.*, apart from a few writings in Italian. Once again, the subjects covered seem to be political and cultural, the two themes often being linked in general. The topic 2 in Figure 7.9 expresses this well with the term *lingua* (*language*) associated with the term *statu* (*state*). We can also see the term *nasitortu* (*crooked nose*) which is surely a reference to the anti-Semitic imaginary which can easily be associated with the *Muvra*. This is not particularly surprising when you consider the close links corsists had with the French far right, as mentioned above. This anti-Semitism was expressed in articles, but also in the many caricatures that appeared from the 1930s onwards, drawn by Matteu Rocca[123].

[121] *A Muvra*, n°59-1922/04/09: « Fuir, ou grelotter, bras en croix sur la terre. Fuir ; ou dépérir, jaunir, se flétrir, drainé comme un raisin sans suc ; céder aux vibrions innombrables de sang fort et précieux de la race. Fuir. Fuir ! ou blanchir l'humus de ses vertèbres. Fuir... ».

[122] Roland Caty and Eliane Richard, "Les Fraissinet, une famille d'armateurs protestants marseillais," *Bulletin de la Société de l'Histoire du Protestantisme Français (1903-)*, 135 (1989), p. 233: « On disait plaisamment qu'il existait, en Corse, trois fléaux: la malaria, le *libeccio*... et Fraissinet ! ».

[123] See: Marie-Claude Lepeltier, "La Caricature insulaire à travers le journal *A Muvra*, 1920-1939," *Études corses. Les revues corses de l'entre-deux-guerres*–64 (2007).

Thus, as we have seen, the pseudonyms do not serve to categorise themes and types but allow Petru Rocca to evoke a wider spectrum of specific subjects that remain around political and cultural current affairs. Similarly, the use of language doesn't seem to be part of any attempt to separate themes, with French and Corsican acting more as a complement to each other, even if the local dialect seems to be used more to address cultural notions. How then to explain the use of several pseudonyms to express himself in his own newspaper? Let's not forget that he is in fact the director of the *Muvra*. This can be explained by purely propaganda and publicity reasons for the weekly. Indeed, even though there are a large number of contributors, there are very few who are really involved in the corsist struggle over the long term. For Rocca, it would be a question of inflating the numbers of contributors a little in order to get a more substantial core of regular authors to appear. It's not all ideology, and there are sometimes simpler justifications to understand the muvrists' approach. This reason can also be seen in the public demonstrations organised by the autonomists. Thus, in 1934, a number of participants are mentioned in the sixth edition of the *merendelle d'i pueti còrsi*[124]. The list includes Dumenicu Carlotti, Eugeniu Grimaldi, Petru Rocca and a certain Pasquale Manfredi...

[124] *A Muvra*, n°527-1934/09/01-10.

7.4 A diversified heart of authors

Performing topic modelling on all the authors of the *Muvra* would appear to be a rather difficult task, given their diversity. It also poses a problem of data quantity and vocabulary thickness, as we saw in the previous section. To try and understand the authors and their role a little more, we can proceed according to their participation rate. As we mentioned at the beginning of this work, the *Muvra* is made up of a core group of authors who publish regularly. The vast majority of authors publish very little in comparison. This part will be devoted to this core group of writers by analysing the subjects they deal with most, taking into account typology and language. To do this, we will be looking at the 20 most regular editors of the autonomist newspaper, including Petru Rocca, Dumenicu Carlotti or Dumenicu Andreotti, also known as *Minicale*. The case of the latter is particularly interesting because, unlike the other authors from higher social backgrounds, Andreotti is a shepherd[125]. He is recognised by his peers for his talent as an improviser when he sings, and was one of the initiators of the movement to reclaim the island's polyphonic culture after World War two.

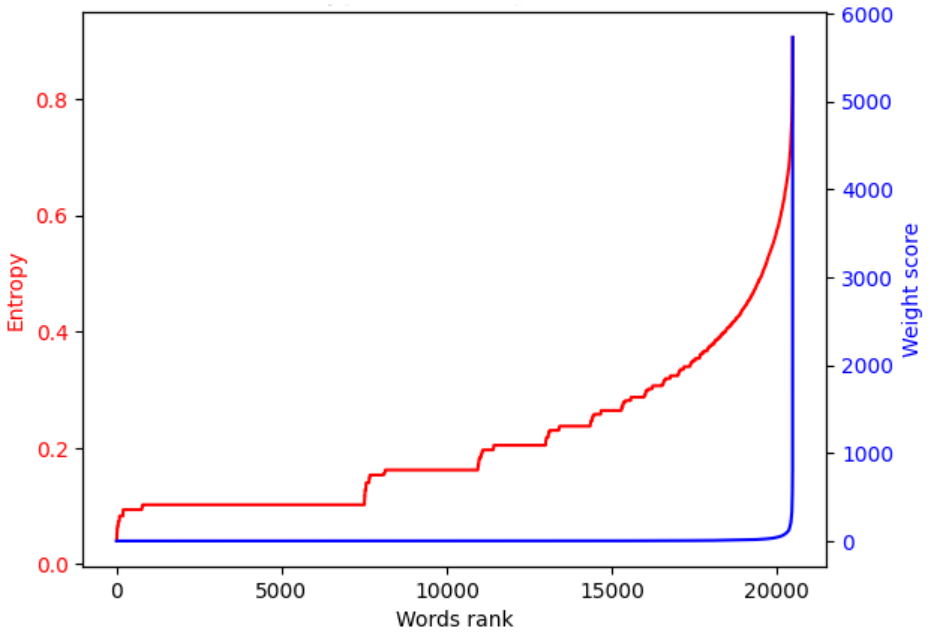


Figure 7.10: Entropy and weight score of TOP20 author's vocabulary (corsican)

[125]J. Mattei, P. S. Menozzi, and A.T. Pietrera, *Antulugia A Corsica Literaria...*, p. 524.

The first interesting point to note is that despite a much larger basic vocabulary, given the 876 articles instead of 62 for *Altore*, the entropy and weight score curves follow the same dynamics. The fact that we have a vocabulary of **51929** is problematic because it seems far too large to conduct anamyses, particularly for LDA. Insofar as this technique works with a document-term matrix, the computation time is exponential with respect to the size of the vocabulary. The idea is to reduce this vocabulary to approximately **30000** tokens. With a first interval defined between **0.3** and **0.8**, as for the previous analyses, we obtain a list of **13774** words to delete which is not sufficient. But if we decrease the interval with an entropy between **0.35** and **0.85**, we obtain a list of **14988** words to delete which brings us closer to the desired count. With the progressive adjustment of stopwords over the course of the analyses, we obtain a total vocabulary of **36812** words, which seems to be perfect for our topic modelling. Several parameter combinations were made to obtain the best possible results and we started with the 4 classes to be determined with the training, 500 iterations for the LDA and 25 words to be presented for the LSA.

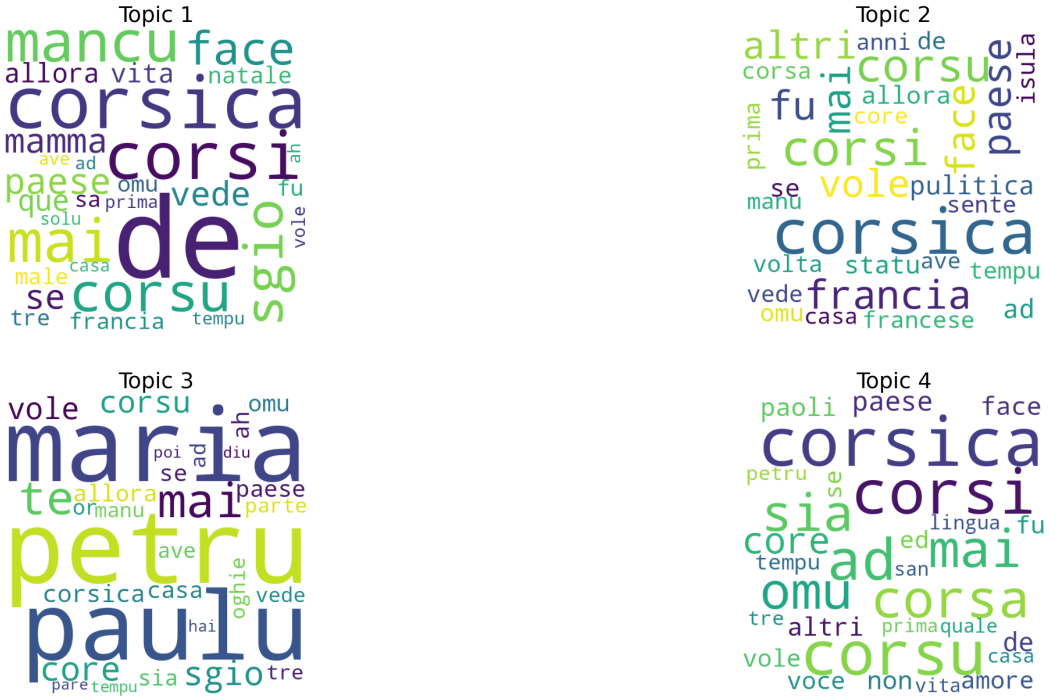


Figure 7.11: LDA anaylisis on TOP20Authors — k=4 — iter500 — pass20

Figure 7.12: LSA anaylisis on TOP20Authors — k=4 — words25

As we might have expected, LDA seems to be a little more effective in this particular context, given that we have much more data than previously. The results do not seem to vary enormously, and this can be explained by the fact that the newspaper has a single political line, so the themes are always more or less the same. However, we can observe certain peculiarities that bear witness to the typological diversity of the newspaper. The topic 1 of the LDA on the wordclouds 7.11 demonstrates the poetical and cultural dimension of the newspaper. In particular, we find the notion of "*mamma*", as we saw earlier, but also terms relating to belonging to national community: following the example of Italian, the term "*paese*" (*country*) has a double meaning, as it can mean the nation or simply the village. These two notions intersect and the muvrists play on them, adding depth to the traditionalism claimed by corsist doctrine. This can be seen, for example, in this poem by *Ghjuvanni di la Grotta* entitled *Fegatelli*, published in the columns of *Muvra* in 1929. *Ghjuvanni di la Grotta*, otherwise known as Simon-Jean Vinciguerra, had a trajectory that differed from the other muvrists. He was a Communist who quickly joined the Resistance when the war broke out[126]:

[126]Ibid., p 651.

Mama, by the gallows, bent over,
Chewed the meat of our pig ;
In the village, it wasn't the worst
We'll have sausages all year round![127]

In these words we find everything we mentioned earlier. The relationship with the mother is central, like a point of reference, in the same way as the village, which could be seen as the country, the nation to which the author belongs. The most important thing here is to have enough food for the year, a kind of return to the rustic way of life in the old island villages. This text is an ode against modernity and capitalism, which the muvrists complain about repeatedly. Given the figure 7.11, it seems that this was a very important theme in the eyes of Corsican autonomists. This text is interesting because it bears witness to the values shared by the various writers despite their radically different ideological opinions, as shown by the overt anti-communism of certain muvrists such as Lucien Orsini. It also bears witness to the diversity of the *Muvra* and the sociological depth of its authors. The abundance of names that emerge in the results of the analyses is certainly linked to plays, but these remain in the minority. On the other hand, poets tended to correspond with each other directly through the newspaper's columns, in the pure tradition of *chjam'è respondi*. This particularly emotive poetry was denounced by the newspaper's editorial board in 1935, which recommended that poets "deal with virile subjects and leave out serenades, amorous disputes and the lamentations of cats and donkeys"[128].

The other central element is the political and historical aspects of the journal. Whereas topic 2 of the figure 7.11 mentions the state ("*statu*") or more generally France ("*francia*"), topic 3 of the figure 7.12 mentions faith ("*fede*") and Pasquale Paoli. We are entering completely into the classic political discourse of the editors of the Muvra and members of the *Partitu Corsu Autonomista*. We regularly come across the notion of "companion of faith" ("*cumpagni di fede*") to designate the corsists who were politically involved. It is really something specific to this heart of authors that we were talking about. Indeed, when events are organised, the same participants are regularly found,

[127] *A Muvra*, n°342-1929/01/27: « Mamma, vicinu a lu troppu, ingrunchiata, tazza la carne d'u nostru maghiale ; Di lu paese, unn'era lu più male, salàme n'averemu per l'annata! ».
[128] *A Muvra*, n°581-1935/12/29: « Trattà sugetti virili e lascià corre i serinati, i cuntrasti d'amore, i lamenti di jatti e di sumeri. ».

as at the *terza meredenella d'i pueti corsi* in 1926 in honour of the *Babbu di a Patria* Pasquale Paoli. This phenomenon, which combines religious and political discursive practices, is the "sacralisation of politics", originally theorised by the Italian historian Emilio Gentile[129]. These topics clearly demonstrate the forum that the *Muvra* represents for disseminating not only political ideas but also the muvrist political culture. They are much more than a simple opinion press, they are a complete political and ideological press. These pretensions explain in particular the trajectories of the various authors over time. Although the *Muvra* aims to be a newspaper that embraces different opinions, as we mentioned earlier, it eventually adopted a hard line fairly quickly, leaving a number of authors out in the cold. And there were many of them: Lucien Orsini was dismissed for taking too strict a stance against communism or Maistrale joined in 1928 the cyrneists of *L'Annu Corsu* when the autonomists moved closer to Fascist Italy. The diversity of the *Muvra* is a fact, but this diversity was severely tested in the face of the radicalisation of discourse and ideas common to Europe during the interwar period.

[129]Emilio Gentile, *Politics as religion*, Princeton, Princeton University Press, 2006.

Conclusion

A Muvra is as fascinating as it is complex to study. This political newspaper is characterised by great diversity on every scale. Already on a typological scale, with a huge variety of different types of text. We can find classic articles as well as plays, poems and novels. That's the ambiguity of the magazine: it's a political press that claims to be cultural. But the diversity is also reflected in the languages used. The majority of texts are written in Corsican, French and Italian, with a few in English and Spanish. Behind these choices lie real political intentions. The defence of language is essential for the Muvrists, and the use of a specific language is a sign of bias. But this typological and linguistic diversity is also the result of the sociological diversity of the authors. Each contributor to the *Muvra* bears witness to his own relationship with Corsicanism, whether through his political ideology or his artistic style. After all, *A Muvra* is first and foremost a platform from which people who share the same vision of Corsica can express themselves, whatever their political opinion. Despite these unique characteristics, this autonomous newspaper remains part of a timeline with specific journalistic practices, such as the use of pseudonyms. Generally used to hide one's identity so as to be able to publish without fear of reprisals from governments, it also seems to be a form of poetic licence for our editors. Indeed, for some of them, their pseudonyms were known to everyone, particularly contemporary anthologists and Carmine Starace is a perfect example. The latter published his *Bibliografia della Corsica* in 1939, in a very turbulent complex on the eve of the Second World War with specific directives from the Fascist government. However, the fact that the Italian author was right in attributing the pseudonym *P. di B.* to the director of the *Muvra* Petru Rocca also bears witness to the rigour of his titanic work. The duality in the use of pseudonyms can also be explained by the use we make of them. An identity can be used to evoke more sensitive subjects that we wouldn't discuss without it. Ghjanettu Notini makes no secret of the fact that he is *U*

Sampetracciu when he writes his plays and poetry. Even if he tackles specific political themes, he never goes too far and effectively protects himself from criticism behind his dramatic work. But it's thanks to his hypothetical identity as *Altore* that Notini can really express his intentions, with more assertive political discourse and fewer filters. These are subjects that are also evoked by a core of authors who are recurrent in the *Muvra*. There are politicians like Petru Rocca and Eugeniu Grimaldi, the priest Dumenicu Carlotti, playwrights like Simon'Ghjuvanni Vinciguerra and poets like Marcu Angeli and Dumenicu Andreotti. These characters were not only the main players in the life of the newspaper, but also in the cultural life of the island. We therefore find themes linked to history and politics that are inherent to their predominant role. They also share common ground with the rarer contributors through the themes of their poetry. While a huge number of subjects are evoked in several languages and forms throughout these two decades, it is this diversity that also represents the main limitation of this work.

Studying a weekly newspaper spanning almost 20 years represents a real technical challenge that forces us to make choices. It is impossible to carry out a stylometric analysis on all the anonymous authors, just as it is complex to carry out topic modelling on all the combinations of articles according to their type, author, language or even date of publication. We therefore had to make choices and apply biases in order to obtain an overview of what computational methods can offer in the study of such a corpus. In addition to determining the authorship of certain pseudonyms and the role of others, the question was also to work on an under-ressourced language. The aim is to encourage this type of study in areas other than pure linguistics, as can be done at the Università di Corsica. While the complexity of the subject is a fact, it does not prevent us from obtaining coherent and promising results for the future. With better preparation of the data, as part of a broader project that would not just be part of a Master's thesis, this subject has a lot of potential. The study of la *Muvra* is rich, and it is an excellent gateway for the development of digital humanities within Corsican studies.

List of Figures

1.1 Entity-relationship diagram of the A/I database 21

2.1 Language distributions in the *Muvra* at different time 26
2.2 Typological categorization in the A/I database 31
2.3 Distribution of authors according to the power law 36
2.4 Language distribution of articles according to a selection of authors (database) 37

3.1 Example of a front page of the *Muvra* (n°194-8 March 1925) . 42
3.2 Exemple of a cleaned front page using *ScanTailor* 47

4.1 Exemple of segmentation for a front page of the *Muvra* 51
4.2 Result of the segmentation after the model training 53
4.3 Pipeline of the *Muvra* textual data collection process 55
4.4 Reliability of the corpora according to Zipf's law (log scale) . . 57

5.1 Language distributions in the dataset 62
5.2 Distribution of article typology and their share in the total of words 63
5.3 Typology of articles of our dataset 65
5.4 Number of articles written by an author or a pseudonym . . . 66
5.5 Language distribution of articles according to a selection of authors (dataset) 67

6.1 *Vecchiu pastore*, Maria Saveria Rocca-Pozzo di Borgo 74
6.2 Medenhall's Characteristic Curves of Composition for *P. di B.* 75
6.3 Value of the decision function for *P. di B.* 79
6.4 Quantity of *Lettere Aiaccine* published by *Altore* 81
6.5 Top 50 Most Frequent Words in the whole corpus 83
6.6 Visualisation of Burrow's Delta Score results for *Altore* — MFW 50 84

6.7 Visualisation of Burrow's Delta Score results for *Altore* — MFW 100 85
6.8 Authorship attribution of *Altores*'s texts for ALE-VER-NOT-GIA 87
6.9 Value of the decision function for ROC-CAR-PIA-VIN 88
6.10 Authorship attribution of *Altores*'s texts for ALL CORPUS . . 90
6.11 Number of articles published bu Ghjanettu Notini between 1925 and 1939 91
6.12 Picture of Ghjanettu Notini in the 30's 92

7.1 Entropy and weight score of *Altore*'s vocabulary 97
7.2 LDA anaylisis on *Altore* — k=4 — iter400 — pass20 98
7.3 LDA anaylisis on *Altore* — k=4 — iter250 — pass20 99
7.4 LSA anaylisis on *Altore* — k=4 — words20 99
7.5 Top topics for *U Sampetracciu* 100
7.6 Distribution of languages used by Rocca in all his texts 102
7.7 LDA analysis on *Petru Rocca* 103
7.8 LDA analysis on *Pasquale Manfredi* 103
7.9 LDA analysis on *P. di B.* — COS — k=3 — iter600 104
7.10 Entropy and weight score of TOP20 author's vocabulary (corsican) 107
7.11 LDA anaylisis on TOP20Authors — k=4 — iter500 — pass20 108
7.12 LSA anaylisis on TOP20Authors — k=4 — words25 109

List of Tables

5.1 Length of the different corpora 61
5.2 Distribution of words and section 64
5.3 Distribution of sections types 66

6.1 Kilgarriff's Chi-Square Score for *P. di B.* 76
6.2 Burrow's Delta Score for *P. di B.* 77
6.3 Detailed class scores using SVM for *P. di B.* — Leave-one-Out — PCA — Word-tokens — Upsampling 79
6.4 Burrow's Delta Score for *Altore* 84
6.5 Detailed class scores ALE-VER-NOT-GIA — Leave-one-Out — PCA — Word-tokens — Upsampling 86
6.6 Detailed class scores ALE-VER-NOT-GIA — K-Fold 35 — PCA — Word-tokens — Upsampling 86
6.7 Detailed class scores ALL CORPUS — Leave-one Out — PCA — Word-tokens — Upsampling 89

Bibliography

Abdul Razzaq (Ammar Adil) and Mustafa (Tareef Kamil), "Burrows-Delta method fitness for Arabic text authorship Stylometric detection," *International Journal of Computer Science and Mobile Computing*, 3–6 (2014).

Albertini (Pierre) and Borne (Dominique), *L'école en France XIXe-XXe siècle de la maternelle à l'université*, Paris, Hachette, 1992.

Bargeton (René), *Dictionnaire biographique des préfets, septembre 1870-mai 1982*, Paris, Archives nationales, 1994.

Beaudouin (Valérie) and Yvon (François), "Contribution de la métrique à la stylométrie," in *Actes des 7èmes Journées Internationales d'Analyse Statistique des données textuelles (JADT)*, 2004, vol. 1.

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

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

Blei (David M), Ng (Andrew Y), and Jordan (Michael I), "Latent dirichlet allocation," *Journal of machine Learning research*, 3–Jan (2003).

Bosc (Olivier), "De la foule criminelle à la foule nationaliste: Scipio Sighele, théoricien de l'irrédentisme," *Matériaux pour l'histoire de notre temps*, 43–1 (1996).

Broche (François), *Léon Daudet: le dernier imprécateur*, Paris, Éditions Robert Laffont, 1992.

Burrows (John), "'Delta': a measure of stylistic difference and a guide to likely authorship," *Literary and linguistic computing*, 17–3 (2002).

Cabanel (Patrick) and Encrevé (André), *Dictionnaire biographique des protestants français de 1787 à nos jours*, vol. 3 H-L, Paris, Les éditions de Paris / Max Chaleil, 2022.

Cafiero (Florian) and Camps (Jean-Baptiste), "'Psyché'as a Rosetta Stone? Assessing Collaborative Authorship in the French 17th Century Theatre," in *Proceedings of the Conference on Computational Humanities Research 2021*, CEUR-WS, 2021, vol. 2989.

Camps (Jean-Baptiste), *SUPERvised STYLometry (SuperStyl)*, version 0.9.0, Oct. 2021.

Camps (Jean-Baptiste) and Cafiero (Florian), "Who could be behind QAnon? Authorship attribution with supervised machine-learning," *arXiv*, arXiv:2303.02078 (2023).

Camps (Jean-Baptiste) and Clérice (Thibault), *IIIF-Crawler*, 2019, url: `https://github.com/Jean-Baptiste-Camps/IIIF-Crawler`.

Camps (Jean-Baptiste), Clérice (Thibault), and 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).

Caty (Roland) and Richard (Eliane), "Les Fraissinet, une famille d'armateurs protestants marseillais," *Bulletin de la Société de l'Histoire du Protestantisme Français (1903-)*, 135 (1989).

Chagué (Alix) and Clérice (Thibault), *HTR-United: Ground Truth Resources for the HTR and OCR of patrimonial documents*, version 0.1.56, url: `https://github.com/HTR-United/htr-united`.

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

Cini (Marco), *Gli « Studii critici di costumi corsi » di Salvatore Viale. Il processo di modernizzazione della Corsica nel XIX secolo*, Roma, L'Harmattan Italia, 2018 (Il Politico e La Memoria).

— *Un'integrazione nazionale imperfetta: élite e culture politiche in Corsica nella prima metà dell'Ottocento*, Roma, Viella, 2022.

Clérice (Thibault) and Chauhan (Ronan), *YALTAi, You Actually Look Twice At it*, version v0.0.1rc4, 2022, url: `https://github.com/PonteIneptique/YALTAi`.

Cooper-Richet (Diana), *Passeurs culturels dans le monde des médias et de l'édition en Europe (XIXe et XXe siècles)*, vol. 6, Paris, Presses de l'ENSSIB, 2005 (Référence).

Cvitanic (Toni), Lee (Bumsoo), Song (Hyeon Ik), Fu (Katherine), and Rosen (David), "LDA v. LSA: A comparison of two computational text

analysis tools for the functional categorization of patents," in *International Conference on Case-Based Reasoning*, 2016.

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) and 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," in *Tribune des chercheurs en linguistique*, Société des Sciences Historiques et Naturelles de la Corse, 2013.

Deerwester (Scott), Dumais (Susan T), Furnas (George W), Landauer (Thomas K), and Harshman (Richard), "Indexing by latent semantic analysis," *Journal of the American society for information science*, 41–6 (1990).

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

Desanti (Paul), *« Gigli di stagnu » di Marco Angeli, un' avvinta literaria*, MA thesis, Università di Corsica Pasquale Paoli, 1997.

— *Trois poètes corses irrédentistes. M. Angeli, P. Giovacchini, A.F. Filippini*, Ajaccio, Albania, 2013.

Dumais (Susan), *Enhancing performance in latent semantic indexing (LSI) retrieval*, 1992.

Eder (Maciej), "Does size matter? Authorship attribution, small samples, big problem," *Digital Scholarship in the Humanities*, 30–2 (2015).

— "Rolling stylometry," *Digital Scholarship in the Humanities*, 31–3 (2016).

Fayaza (Faathima) and Fathima Farhath (F), "Towards stop words identification in Tamil text clustering," *International Journal of Advanced Computer Science and Applications*, 12–12 (2021).

Fracastoro (Giulia), Magli (Enrico), Poggi (Giovanni), Scarpa (Giuseppe), Valsesia (Diego), and 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).

Gentile (Emilio), *Politics as religion*, Princeton, Princeton University Press, 2006.

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," in *Histoire sociale des langues de France*, Rennes, Presses universitaires de Rennes, 2013.

Griffiths (Thomas), Steyvers (Mark), Blei (David), and Tenenbaum (Joshua), "Integrating topics and syntax," *Advances in neural information processing systems*, 17 (2004).

Gutehrlé (Nicolas) and Atanassova (Iana), "Logical Layout Analysis Applied to Historical Newspapers," in *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*, NLP Association of India (NLPAI), 2021.

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*, MA thesis, Institut d'Études Théâtrales de Paris 3, 2018.

Izbassarov (Tleusher) and Turan (Cemil), "Understanding Authorship Attributions of Kazakh Texts via Distance Measures," in *2022 International Conference on Smart Information Systems and Technologies (SIST)*, 2022.

Jocher (Glenn), *YOLOv5 by Ultralytics*, version 7.0, 2020, doi: 10.5281/zenodo.3908559.

Kalepalli (Yaswanth), Tasneem (Shaik), Teja (Pasupuleti Durga Phani), and Manne (Suneetha), "Effective comparison of LDA with LSA for topic modelling," in *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2020.

Kestemont (Mike), Daelemans (Walter), and Sandra (Dominiek), "Robust rhymes? The stability of authorial style in medieval narratives," *Journal of Quantitative Linguistics*, 19–1 (2012).

Kestemont (Mike), Stover (Justin), Koppel (Moshe), Karsdorp (FB), and Daelemans (Walter), "Authorship Verification with the Minmax Metric," in *Digital Humanities 2016: Conference Abstracts*, Kraków, Jagiellonian University & Pedagogical University, 2016.

Kevers (Laurent) and Medori (Stella Retali), "Copyright in the context of tooling up Corsican and other less-resourced languages," in *International*

Conference on Language Technologies for All (LT4All), Enabling Linguistic Diversity and Multilingualism Worldwide, 2019.

Kevers (Laurent) and Medori (Stella Retali), "Towards a Corsican Basic Language Resource Kit," in *12th Language Resources and Evaluation Conference (LREC 2020)*, 2020.

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

Kiessling (Benjamin), Tissot (Robin), Stokes (Peter), and Ezra (Daniel Stökl Ben), "eScriptorium: an open source platform for historical document analysis," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, IEEE, 2019, vol. 2.

Kilgarriff (Adam), "Comparing corpora," *International journal of corpus linguistics*, 6–1 (2001).

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), and Garabato (Carmen Alén), *Histoire sociale des langues de France*, Rennes, Presses universitaires de Rennes, 2013.

Lam (Ho Ngoc), Nhu (Vo Diep), Dien (Dinh), and Nhung (Nguyen Tuyet), "Identifying Authors Based on Stylometric Measures of Vietnamese Texts," in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 2020.

Laramée (François Dominic), "Introduction to stylometry with Python," in *Programming Historian*, copyrights=CC-BY 4.0, 2018, url: `https://doi.org/10.46430/phen0078`.

Laugaa (Maurice), *La pensée du pseudonyme*, Paris, Presses Universitaires de France, 1986 (Écritures).

Lee (Moontae), Bindel (David), and Mimno (David), "Robust Spectral Inference for Joint Stochastic Matrix Factorization," in *Proceedings of NIPS 2015*, 2015.

Léglise (Isabelle) and 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).

Lepeltier (Marie-Claude), "La Caricature insulaire à travers le journal *A Muvra*, 1920-1939," *Études corses. Les revues corses de l'entre-deux-guerres*–64 (2007).

Mattei (Julian), Menozzi (Petru Santu), and Pietrera (Ange-Toussaint), *Antulugia A Corsica Literaria*, Aiacciu, Albiana, 2020.

Mayeur (Françoise), *Histoire de l'enseignement et de l'éducation III. 1789-1930*, Paris, Editions Perrin, 2004.

Meistersheim (Anne), "Du riacquistu au désenchantement: Une société en quête de repères," *Ethnologie française*, 38–3 (2008).

Mendenhall (Thomas Corwin), "The characteristic curves of composition," *Science*–214s (1887).

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

Montemurro (Marcelo A), "Beyond the Zipf–Mandelbrot law in quantitative linguistics," *Physica A: Statistical Mechanics and its Applications*, 300–3-4 (2001).

Mosteller (Frederick) and Wallace (David L), "Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers," *Journal of the American Statistical Association*, 58–302 (1963).

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*, PhD thesis, 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).

— "Langue et identité: l'exemple du corse durant la troisième république," *Cahiers de la Méditerranée*–66 (2003).

Pellegrinetti (Jean-Paul) and 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*, PhD thesis, Université Pascal Paoli, 2015.

— "La construction des héros corses durant la Troisième République. Le cas de Sampiero et Paoli," in *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*, MA thesis, 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).

Queffélec-Dumasy (Lise), *Le roman-feuilleton français au XIXe siècle*, Belphégor: Littérature Populaire et Culture Médiatique, 2008, URL: `http://hdl.handle.net/10222/47746`.

Retali-Medori (Stella), "Le programme Nouvel Atlas Linguistique et ethnographique de la Corse-Banque de Données Langue Corse (NALC-BDLC)," in *«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) and Kevers (Laurent), "La morphologie dans la Banque de Données Langue Corse: bilan et perspectives," *Corpus*–23 (2022).

Ringoot (Roselyne) and Rochard (Yvon), "Proximité éditoriale: normes et usages des genres journalistiques," *Mots. Les langages du politique*–77 (2005).

Rocca (Pierre), *Les Corses devant l'anthropologie*, Paris, Librairie J. Gamber, 1913.

Rogé (Ysée), *Le corsisme et l'irrédentisme 1920-1946: histoire du premier mouvement autonomiste corse et de sa compromission par l'Italie fasciste*, PhD thesis, 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*, MA thesis, Université de Strasbourg, 2021.

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

Singh (Rishi R), Koundal (Deepika), and Tiwari (Rajeev), "Linguistic Approach for Authentic Authorship," in *International Conference on Emerging Technologies: AI, IoT and CPS for Science & Technology Applications*, CEUR-WS, 2021, vol. 3058.

Stamatatos (Efstathios), Kestemont (Mike), Kredens (Krzysztof), Pezik (Piotr), Heini (Annina), Bevendorff (Janek), Stein (Benno), and Potthast (Martin), "Overview of the authorship verification task at PAN 2022," in *CEUR Workshop Proceedings*, CEUR-WS, 2022, vol. 3180.

Starace (Carmine), *Bibliografia della Corsica*, Milano, Istituto per gli studi di politica internazionale, 1943 (Centro di studi per la Corsica).

Taglioni (François), "L'insularisme: une rhétorique bien huilée dans les petits espaces insulaires," in *Comme un parfum d'îles*, Paris, Presse Universitaire Paris-Sorbonne (PUPS), 2010.

Tannery (Paul), "La stylométrie ses origines et son présent," *Revue Philosophique de la France et de l'Étranger*, 47 (1899).

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

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

Vidal-Gorène (Chahan), Dupin (Boris), Decours-Perez (Aliénor), and Riccioli (Thomas), "A modular and automated annotation platform for handwritings: evaluation on under-resourced languages," in *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part III 16*, Springer, 2021, pp. 507–522.

Yujian (Li) and Bo (Liu), "A normalized Levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, 29–6 (2007).

Yvia-Croce (Hyacinthe), *Anthologie des écrivains corses*, Ajaccio, Ed. Cyrnos et Méditerranée, 1987.