

Profiling Anonymous Authors in the Corsican Autonomist Press of the Interwar Period

Vincent Sarbach-Pulicani

Université Côte d'Azur, Centre de la Méditerranée Moderne et Contemporaine, Campus Carlone, 06100 Nice, France

Abstract

With the emergence of nationalism in the 19th century came regionalist movements to assert and claim cultural particularities. Corsica fitted very well within this dynamic and even presented itself as a favourable location for the development of such ideas. The centralization 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. This paper presents the results of a research that aimed to carry out author profiling, i.e., to determine the style and subjects covered by an author. The goals of this study were to determine the identity behind certain authors and also to highlight the role pseudonyms played in the newspaper's propaganda. We conducted authorship attribution to achieve the first objective before completing these analyses with topic modelling in order to meet the second one.

Keywords

stylometry, topic modelling, corsican studies, under-ressourced languages, computational history

1. Introduction

Corsican studies have focused at length on the Corsican autonomist press of the interwar 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. The historiographical renewal of Corsican automatism began around the 2000s in order to bring a fresh perspective to the subject. We can cite several important thesis: those by Ysée Rogé [23], Deborah Paci [20], and finally Ange-Toussaint Pietrera [22]. 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. His synthesis on Corsica and the Third Republic is still an authority on the subject [21]. This paper is part of this desire to revitalise historical studies of contemporary Corsica. It also occurs in the context of the increasing development of natural language processing (NLP) for the Corsican language, a central technical issue in our study. For several years, there has been

CHR 2023: Computational Humanities Research Conference, December 6 – 8, 2023, Paris, France


✉ vincent.sarbach-pulicani@outlook.com (V. Sarbach-Pulicani)

🌐 <https://github.com/vincentsarbachpulicani> (V. Sarbach-Pulicani)

🆔 0009-0000-2670-7475 (V. Sarbach-Pulicani)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

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*. 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).¹ This is the first initiative to lemmatise the Corsican language in its diachrony and diatopy.² Since the second part of the 2010s, there has been a significant increase in scholars' thoughts on the tooling of regional languages using NLP [16]. Our approach is fully in line with this state of the art. This paper presents is the continuation of a master's thesis written as part of a double degree programme between the École nationale des chartes of Paris and the Università di Pisa [24]. It follows a first thesis that highlighted the major ideological differences between the corsists and the irredentists, despite their obvious proximity. [25] It resulted in the creation a database named *Autonomists/Irredentists Database* (A/I database).³ This work establishes that 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.

Like any political press, *A Muvra* has a large number of anonymous authors writing under pseudonyms. While the possibility of individual authors exists, there is also a good chance that these pseudonyms are the result of recurring authors of the journal publishing under their real names. This raises a number of questions about the identity of these anonymous authors as well as the role that a corsist gives to one or more of his pseudonyms. Several preliminary hypotheses can be proposed at this stage, including the deliberate exaggeration of the number of activists, the need for protection against censorship, or the desire to express varying viewpoints. In order to address these inquiries, we will employ two distinct analytical methods. First, we will utilise stylometry to unveil the identities of anonymous authors, and secondly, we will apply topic modelling to gain insights into the themes associated with these pseudonyms. Subsequently, we will engage in an interpretive phase to discern the purpose and characterization an author assigns to their pseudonym. These dual layers of analysis ultimately encapsulate the concept of author profiling, as previously discussed. The analyses and results are all available on a GitHub repository dedicated to this research [26].

2. Datasets construction: starting from scratch

2.1. The OCR processing

The main issue surrounding the analysis of such a review is the accessibility of the data. In order to carry out the analyses, the data had to be acquired from the digitised images of the newspapers. Segmentation and OCR presented significant challenges, as well as postprocessing and normalisation (see an example of a front page with Figure 12). We were able to locate two online platforms where our documents are available for download. The images come from two sources: the *Bibliothèque nationale de France* (BnF) and the *Archives départementales de Corse*

¹<https://bdlc.univ-corse.fr/bdlc/corse.php>

²This database, which includes a wide range of possibilities, was created on the basis of a vast and particularly impressive field survey.

³https://heurist.huma-num.fr/heurist/?db=vsp_presse_corsiste_irredentiste

du Sud (ADC). So we used Gallica, the digitization platform of the BnF, and THOT, the platform of the ADC.⁴ The fact that these are national institutions means that the digitizations are in the public domain, i.e., open source. After the phase of webscraping, we got a collection of 375 issues of the *Muvra*, i.e., approximately 1500 pages from 1921 to 1931.

One of the problems with having images from two different sources is the quality of the images. This raises the question of whether or not it is appropriate to normalise and clean images in order to facilitate OCR processing. The original idea of our research was to clean the documents using binarization with the Otsu method [19] followed by a despeckling phase. The “speckling” is a type of noise that corresponds to random clusters of black pixels that impair the intrinsic quality of a binarized image [12]. However, the quality of the digitizations, especially 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 digitization. Standardising all the images at the same time requires an initial sorting organised according to identical layouts for a gain in OCR quality that is not necessarily guaranteed. So we decided to prefer quantity over quality, even if the normalisation would occur on the raw data.

One of the major challenges in the world of automatic character recognition today is the segmentation of newspapers. Their complex layout requires the training of complex models that are often specific to a type of newspaper. We decided to train a Kraken segmentation model from the XML files in ALTO format available on Gallica, with the help of the eScriptorium platform [18] and the module ketos. Once the ALTO files were adapted to the good format, we could train the model to segment the images coming from the *Archives départementales de Corse du Sud*. In order to improve the model, it was necessary to use the tool YALTAi [8] developed by Thibault Clérice, which allows the use of YOLOv5 [13], an Ultralytics object detection model, to be adapted for training segmentation models with Kraken. For the text recognition phase, we decided to go for Tesseract-OCR, which includes a Corsican model. We needed to create UZN files readable by this engine in order to follow the coordinates of the image (Figure 1).

2.2. Data standardisation

Once we got our raw textual data, we had to classify them according to their language, typology, and author. Then we could perform the cleaning of the textual data, carried out in four main stages:

- The removal of punctuation
- Case reduction
- Normalization of syntax
- Elimination of accents

The most delicate phase in our methodology is the normalisation of the syntax. It is important because, for euphonic reasons, contractions occur in written form in the form of elisions,

⁴<https://gallica.bnf.fr/accueil/en/content/accueil-en?mode=desktop> | http://archives.isula.corsica/Internet_THOT/FrmSommaireFrame.asp

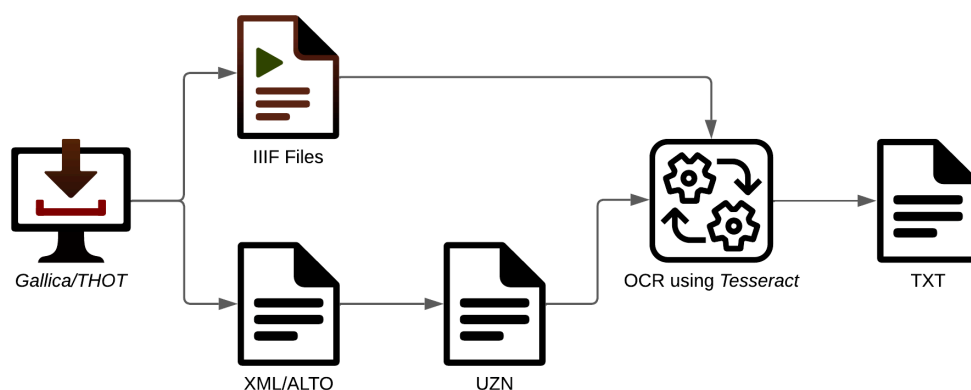


Figure 1: Newspaper page OCR phase pipeline

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 20th 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 [29].

The issue of data normalisation is particularly delicate due to the very nature of our methodology. While topic modelling does not include function words in the analysis because they are meaningless words, stylometry relies mainly on all types of most frequent words. 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, the regiolectal diversity of the Corsican language means that the idiomatic features of the authors are characterised by the great variety of the function words used. A thorough normalisation should not alter our analysis too much, even if it constitutes an improvement perspective for our study.

In the end, we obtained a total of 3 corpora of different sizes with a total of almost 1.5 million words (Table 1), with approximately 56.7% of the words in Corsican, 27% in French, and 16.3% in Italian. The main point of improvement in this method of extracting textual data is the balancing of the corpus. While we were able to obtain almost all the articles in the issues on Gallica, the issues on THOT were selected according to our needs, given the variation in the

Table 1
Length of the Different Corpora.

Corpus	Word count
Corsican	897965
French	380457
Italian	263095
Total	1541517

quality of the images. However, this is still quite sufficient for the type of analysis we are carrying out, focusing on a certain number of authors. Details of the samples selected for this study can be found in the appendix (Table 4).

3. Method proposed: two layers of analysis

The last advances in stylometry have been made with the use of machine learning algorithms. Recent examples include the work of Jean-Baptiste Camps and Florian Cafiero, who used SVM classifier algorithms to identify the authors of the American conspiracy forum *QAnon* [6]. 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 because of the “increase robustness”, they are “known to reduce sparsity and perform well in attribution studies”. 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 [3] and even the rhymes in mediaeval texts like Mike Kestemont did in 2012 [15]. For the previous thesis, we managed to compare the results obtained with the SVM with a metric distance, the Delta score as defined by John Burrow in 2002 [4], in order to confirm them considering the limited length of the corpora. The objective of this double layer of analysis was to confirm the results and determine the best possible approach for our corpus. This paper will focus on the machine learning approach, but the results obtained with Burrow’s Delta that confirmed the SVM methods are available on the GitHub repository [26]. The script being used is the SuperStyl one developed by Jean-Baptiste Camps in 2021 [7]. Whatever the authors and pseudonyms tested, we excluded poetic texts part in prose or verse from the stylometric due to their specificities.

It is very important to vary the 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. After various tests presented in the benchmark (Table 6), we chose those parameters: the statistical units are the most frequent words; we apply the PCA (*Principal Component Analysis*) for dimensional reduction; the cross-validation is carried out with the “Leave-One Out” method; and we balance the dataset with the “upsampling”. This technique consists of isolating a portion of our minority corpus and sampling an equal number of examples from the majority class, as explained by Joseph Barr in 2022 [2]. Once the model has been trained, we apply it to the unseen data. In view of the large number of candidates for the second experiment, we initially subdivided them into two groups in order to obtain more

precise results before carrying out an analysis on the whole corpus.

Concerning topic modelling, 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 LSI (*Latent Semantic Indexing*), 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[9]. In 2020, a group of researchers set out to compare the two methods by training them on a corpus of BBC articles [14]. The results of their research revealed that LSI is more effective when dealing with a large amount of data and fewer iterations than LDA, while the latter is more suitable for smaller corpora. The idea is to present here the most interesting results with an empirical observation of the results obtained as a form of intrinsic evaluation. In the long term, implementing more effective evaluation metrics such as coherence would be very relevant, even if it is not necessary in our case, given that we are modelling general themes rather than assigning a label to each article. To do so, we used the Gensim package for Python, which offers wide possibilities for performing both LSI and LDA techniques. The different experiments presented in the appendix, along with the hyperparameters and methods used, are detailed in the summary table (Table 5). Table 8 serves as a glossary containing pertinent words that were modelled in the course of the experiments.

The vocabulary plays an essential role in topic modelling. The words chosen to be taken into account in topic modelling must not be too numerous, as training 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. Unlike stylometry, function words are of no interest because they are considered to be empty words, i.e., words without a significant meaning but serve to add details to the sentence [1]. We had to create a specific list of stopwords for our Corsican corpus (Table 7) due to the absence of a basic language toolkit [17]. The list creation process occurred in two phases: initially, it involved comparing it with an Italian list that contained overlapping stopwords with Corsican. Following that, it consisted of the examination of various corpora, including the *Muvra* dataset. This examination led to the identification of the most frequent words, followed by a selection between stopwords. 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 [10] with the following formula:

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

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.

4. Results

4.1. The two pseudonyms chosen

The aim is to test our methodology on two different pseudonyms. The first, *P. di B.*, allows us to check the reliability of our tools on a relatively small corpus in Corsican by confirming the identity of the author. The second, *Altore*, gives us the opportunity to test these tools on a completely unknown author, leaving us free to interpret and choose the candidates.

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* [28]. This pseudonym is believed to be the initials of his mother’s surname, Maria Saveria Rocca-Pozzo di Borgo. The latter had remained very close to her sons Petru and Matteu, even publishing drawings in the *Muvra*. 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.

The other pseudonym seen in this paper is *Altore*. It 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 the 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.

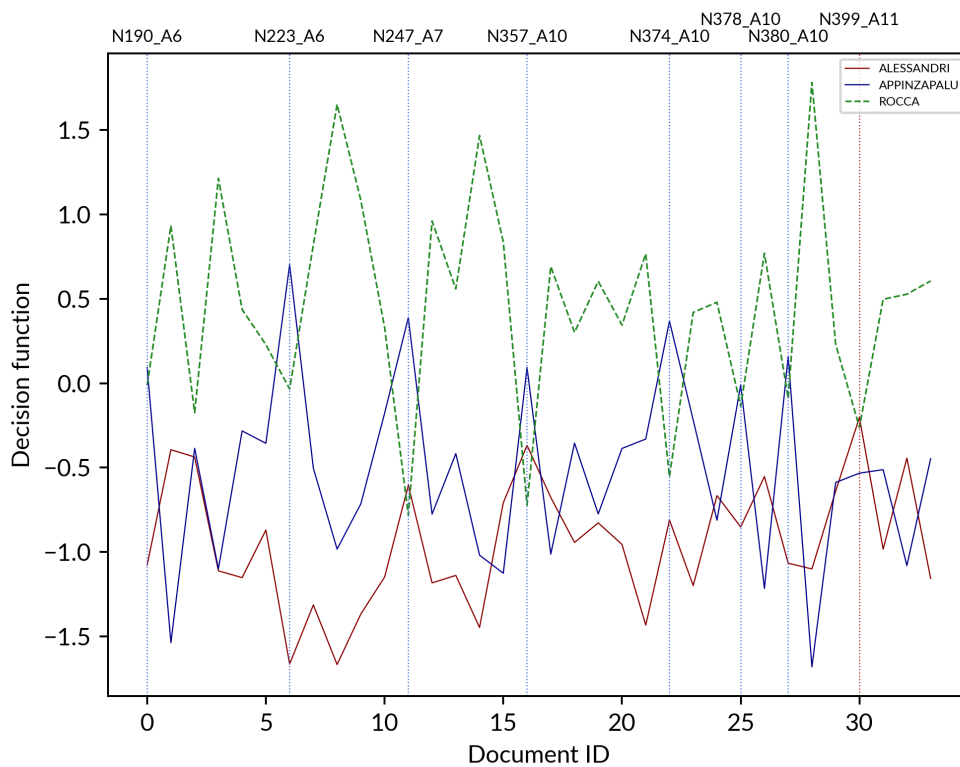
Concerning the candidates, apart from Petru Rocca, who seems obvious to include in the analysis given the information we provided earlier, we decided to choose two other potential authors. The first is Martinu Appinzapalu, a pseudonym of the Corsican priest Dumenicu Carlotti and symbol of the religious aspect of the insular’s autonomist struggle at the time, who published numerous articles throughout the paper’s existence and was part of the *Partitu Corsu d’Azione*, the political party attached to the *Muvra*. The second is Marcellu Alessandri di Chidazzu, one of the authors most involved in the writing and a fervent defender of the irredentist cause.

4.2. First experiment: *P. di B.*

The evaluation of the trained model is presented in the table 2, we got an accuracy of **0.95**. We then obtain a file with the predictions of the author of the articles and the results of the decision function that “tells us how close each sample is to the hyperplane separating each class” [5]. A negative value means that the sample is outside; a positive value means it is inside. The higher the score, the greater the probability that this sample has been written by the candidate. By applying this function to our study, we get the figure 2. 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. 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

Table 2Detailed Class Scores using SVM for *P. di B.* – Leave-One-Out – PCA – Word-Tokens – Upsampling.

Candidate	Precision	Recall	F1-score	Support
ALESSANDRI	0.96	0.90	0.93	50
CARLOTTI	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

**Figure 2:** Value of the Decision Function for *P. di B.*

more interesting to study is the behaviour of the curves on the decision function graph. On average, the decision function scores are much higher for Petru Rocca’s texts.

Petru Rocca is an expert in this field, as nearly five 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. In order to optimise the performance of our stylometry models, several parameters

need to be taken into account, such as the number of k topics, iterations, words, and passes. 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. If we can reference 139 articles written by Rocca in total, we performed the LDA on sub-corpora according to language and pseudonym (Figures 4, 5, 6, 7, 8). 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 pseudonyms seem to 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 attributed to propaganda objectives. 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*.⁵ The list includes Dumenicu Carlotti, Eugeniu Grimaldi, Petru Rocca, and a certain Pasquale Manfredi.

4.3. Second experiment: *Altore*

Table 3

Detailed Class Scores for *Altore* — Leave-One Out — PCA — Word-Tokens — Upsampling

Candidate	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

In the same way as we confirmed Petru Rocca's authorship of the texts of *P. di B.*, we carried out the stylometric analysis of those of *Altore* using the SVM classifier. For the candidates, we chose a wide range of possible authors among the most important ones in the *Muvra*. For

⁵A *Muvra*, n°527-1934/09/01-10.

this experiment, the first sub-group mentioned above was made up of Ghjanettu Notini, Victor Gianviti, Dumenicu Antone Versini and Marcellu Alessandri. The second was made up of Simon Ghjuvanni Vinciguerra, Orsu Francescu Piazzoli, Petru Rocca and Dumenicu Carlotti.

We thus obtain an accuracy of about **0.86** and a model quite good, as seen on the table 3. This test bears witness to another important aspect of stylometry that has not yet really been addressed in this paper: 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 [11] where he stated that “the effectiveness of attribution depends on corpus size and particularly on the number of authors tested”.

The results of the decision function (Figure 3) show us that Ghjanettu Notini is the most likely candidate among the panel of candidates. But 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 December 4, 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, he was the founder of the *Teatru corsu di A Muvra* in the early years of the newspaper and a loyal contributor.

We can notice certain terms that come up frequently on the wordcloud that visualises the results of topic modelling on *Altore* (Figure 9), 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. For *U Sampetracciu* (Figures 10, 11), 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*” (*mother*).

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 the *lamentu*, a poetic style cherished by the *muvrists*.

5. Further research

While this research shows promise, it is important to acknowledge its limitations, which are closely intertwined with its strengths. In the long run, it would be pertinent to develop a dedicated OCR model for recognising printed Corsican text. Additionally, exploring the possibility

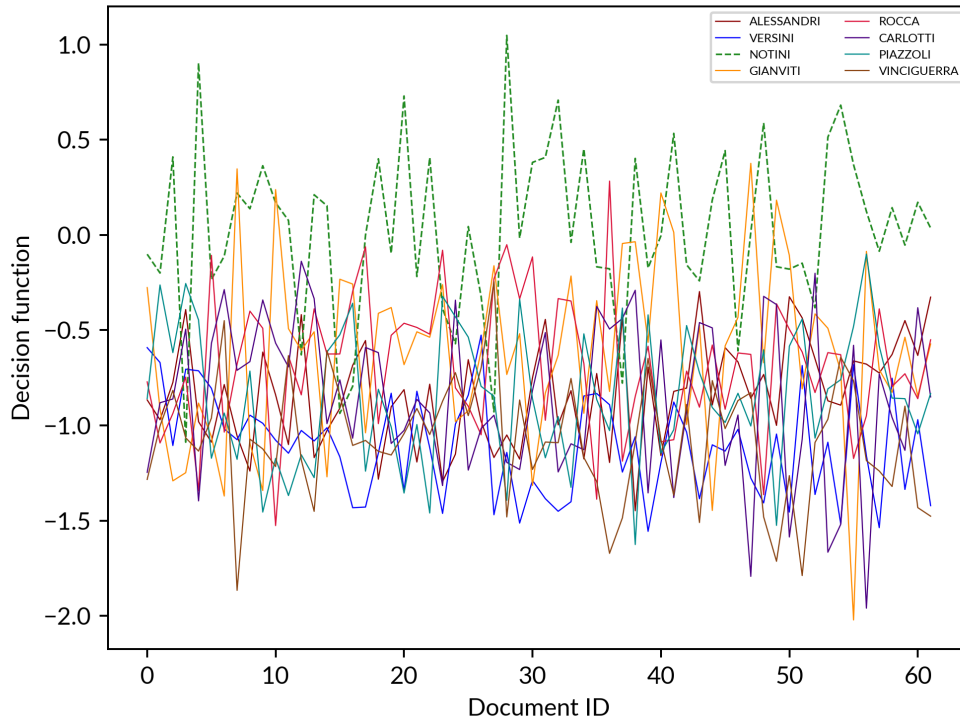


Figure 3: Value of the Decision Function for *Altore*

of fine-tuning the segmentation model to enhance its effectiveness holds significant potential. This article has highlighted the constraints of using topic modelling techniques, which may not be the most suitable approach for detecting word characteristics. Considering this, alternative methods like frequency-based analysis could be more appropriate, given our knowledge of the specific vocabulary found in the *Muvra* dataset. Moreover, the time invested in removing stopwords might have been unnecessary, as demonstrated by the experiments conducted by Alexandra Schoffield and her colleagues [27]. Lastly, in terms of stylometric analysis, it is essential to conduct it on the entire newspaper corpus to validate the obtained results, and this should coincide with a more careful selection of candidates for analysis.

6. Conclusion

The dual nature of pseudonym usage can also be clarified by considering how we employ it. 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 *USampetracciu* 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. On the contrary, the use of a pseudonym may not have a purely ideological role but a more propagandist one, as in the case of *P. di B* for Petru Rocca.

Studying a weekly newspaper spanning almost 20 years represents a real technical challenge that forces us to make choices. Confronted with the intricate nature presented by the numerous metadata within our dataset, we 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. It would be possible to perform a stylometric analysis on all anonymous authors or topic modelling on every combination of articles, but it would be time-consuming and represent a possible improvement to this research. In addition to determining the authorship of certain pseudonyms and the role of others, the question was also to work on an under-resourced 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 include more resources to allocate to the research, this subject has a lot of potential.

Acknowledgments

I would like to thank Jean-Baptiste Camps and Alessandro Lenci for their supervision of this research. Although this paper is the conclusion of a two-year dissertation, it is also the fruit of cooperation with several researchers, including Angelo Mario Del Grosso and Federico Boschetti, members of the CNR Pisa.

References

- [1] R. Arun, V. Suresh, and C. V. Madhavan. “Stopword graphs and authorship attribution in text corpora”. In: *2009 IEEE international conference on semantic computing*. Ieee. 2009, pp. 192–196. DOI: 10.1109/icsc.2009.101.
- [2] J. R. Barr, M. Sobel, and T. Thatcher. “Upsampling, a comparative study with new ideas”. In: *2022 IEEE 16th International Conference on Semantic Computing (ICSC)*. 2022, pp. 318–321. DOI: 10.1109/icsc52841.2022.00059.
- [3] V. Beaudouin and F. Yvon. “Contribution de la métrique à la stylométrie”. In: *Actes des 7èmes Journées Internationales d'Analyse Statistique des données textuelles (JADT)*. Vol. 1. 2004, pp. 107–118. URL: <https://imt.hal.science/file/index/docid/741596/filename/JADT%5C%5F133%5C%5FBeaudouinYvonDef20030116.pdf>.
- [4] J. Burrows. “‘Delta’: a measure of stylistic difference and a guide to likely authorship”. In: *Literary and linguistic computing* 17-3 (2002). DOI: 10.1093/lc/17.3.267. URL: <https://academic.oup.com/dsh/article-abstract/17/3/267/929277>.

- [5] 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*. Vol. 2989. Ceur-ws. 2021, pp. 377–381. URL: <http://star.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-2989/long%5C%5Fpaper51.pdf>.
- [6] F. Cafiero and J.-B. Camps. “‘Who could be behind QAnon? Authorship attribution with supervised machine-learning’”. In: *arXiv Cornell University abs/2303.02078* (2023). DOI: 10.48550/arXiv.2303.02078.
- [7] J.-B. Camps. *SUPERvised STYLometry (SuperStyl)*. Version 0.9.0. 2021. URL: <https://github.com/SupervisedStylometry/SuperStyl/>.
- [8] T. Clérice and R. Chauhan. *YALTAi, You Actually Look Twice At it*. Version v0.0.1rc4. 2022. URL: <https://github.com/PonteIneptique/YALTAi>.
- [9] T. Cvitanic, B. Lee, H. I. Song, K. Fu, and D. 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. URL: <https://par.nsf.gov/biblio/10055536>.
- [10] S. Dumais. “Enhancing performance in latent semantic indexing (LSI) retrieval”. 1992. URL: <http://www2.denizyuret.com/ref/dumais/Enhancing%5C%5FLSI%5C%5F%5C%5F%5C%5FDumais%5C%5F1991.pdf>.
- [11] M. Eder. “Does size matter? Authorship attribution, small samples, big problem”. In: *Digital Scholarship in the Humanities* 30.2 (2015), pp. 167–182. DOI: 10.1093/llc/fqt066. URL: <https://academic.oup.com/dsh/article-abstract/30/2/167/390738>.
- [12] G. Fracastoro, E. Magli, G. Poggi, G. Scarpa, D. Valsesia, and L. Verdoliva. “Deep learning methods for synthetic aperture radar image despeckling: An overview of trends and perspectives”. In: *IEEE Geoscience and Remote Sensing Magazine* 9.2 (2021), pp. 29–51. DOI: 10.1109/mgrs.2021.3070956. URL: <https://ieeexplore.ieee.org/document/9416740>.
- [13] G. Jocher. *YOLOv5 by Ultralytics*. Version 7.0. 2020. DOI: 10.5281/zenodo.3908559. URL: <https://github.com/ultralytics/yolov5>.
- [14] Y. Kalepalli, S. Tasneem, P. D. P. Teja, and S. Manne. “Effective comparison of LDA with LSA for topic modelling”. In: *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. Ieee. 2020, pp. 1245–1250. DOI: 10.1109/iciccs48265.2020.9120888. URL: <https://ieeexplore.ieee.org/abstract/document/9120888>.
- [15] M. Kestemont, W. Daelemans, and D. Sandra. “Robust rhymes? The stability of authorial style in medieval narratives”. In: *Journal of Quantitative Linguistics* 19-1 (2012), pp. 54–76. DOI: 10.1080/09296174.2012.638796. URL: <https://www.tandfonline.com/doi/full/10.1080/09296174.2012.638796>.
- [16] L. Kevers, F. Gueniot, A. G. Tognotti, and S. R. 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, pp. 371–380. URL: <https://hal.science/hal-02452276/>.

- [17] L. Kevers and S. R. Medori. “Towards a Corsican Basic Language Resource Kit”. In: *12th Language Resources and Evaluation Conference (LREC 2020)*. 2020. URL: <https://hal.science/hal-02865699/>.
- [18] B. Kiessling, R. Tissot, P. Stokes, and D. S. B. Ezra. “eScriptorium: an open source platform for historical document analysis”. In: *International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2. Ieee. 2019, pp. 19–24. DOI: 10.1109/icdarw.2019.10032. URL: <https://ieeexplore.ieee.org/abstract/document/8893029>.
- [19] N. Otsu. “A threshold selection method from gray-level histograms”. In: *IEEE transactions on systems, man, and cybernetics* 9-1 (1979), pp. 62–66. URL: <https://cw.fel.cvut.cz/b201/%5C%5Fmedia/courses/a6m33bio/otsu.pdf>.
- [20] 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. URL: <https://www.theses.fr/2013NICE2012>.
- [21] J.-P. Pellegrinetti and A. Rovere. *La Corse et la République. La vie politique, de la fin du second Empire au début du XXIe siècle*. Paris, Média Diffusion, 2013, 688 p.
- [22] A.-T. Pietrera. “*Imaginaires nationaux et mythes fondateurs; la construction des multiples socles identitaires de la Corse française à la geste nationaliste*”. PhD thesis. Université de Corse Pascal Paoli, 2015. URL: <https://www.theses.fr/2015CORT0008>.
- [23] Y. 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, 1 vol. (882 p.) URL: <http://www.theses.fr/2008PA100048>.
- [24] V. Sarbach-Pulicani. “*Authors profiling in Corsican autonomist press during the interwar period. Stylometric analysis and topic modeling on “A Muvra”*”. MA thesis. École nationales des chartes (PSL) and Università di Pisa, 2023. DOI: 10.5281/zenodo.8381161.
- [25] 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.
- [26] V. Sarbach-Pulicani. *Stylometry and topic modelling in Corsican language*. Version 2.0.4. 2022. URL: <https://github.com/vincent-sarbach-pulicani/Corsican-Stylometry>.
- [27] A. Schofield, M. Magnusson, and D. Mimno. “Pulling Out the Stops: Rethinking Stopword Removal for Topic Models”. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 432–436. URL: <https://aclanthology.org/E17-2069>.
- [28] C. Starace. *Bibliografia della Corsica*. Centro di studi per la Corsica. Milano, Istituto per gli studi di politica internazionale: Istituto per gli studi di politica internazionale, 1943.
- [29] A. Viaut. “Marge linguistique territoriale et langues minoritaires”. In: *Lengas. Revue de sociolinguistique*. 71. Presses universitaires de la Méditerranée, 2012, pp. 9–28. URL: <https://journals.openedition.org/lengas/301>.

Appendix

Table 4
Details of the Samples Used for the Study

CORPUS	Number of texts	Average length of texts	Number of words
ALESSANDRI	50	265	13231
ALTORE	62	690	42807
CARLOTTI	89	684	60890
GIANVITI	43	834	35870
NOTINI	81	608	48900
P. DI. B	34	844	28707
PIAZZOLI	50	284	14181
ROCCA	25	559	13969
VERSINI	53	486	25783
VINCIGUERRA	89	307	26757

Table 5
Hyperparameters of the Topics Presented in this Article

Topics	Method	<i>k</i> topics	Language	Iterations	Passes	Words	Target
Figure 4	LDA	3	Corsican	600	20	30	Petru Rocca
Figure 5	LDA	3	French	600	20	30	Petru Rocca
Figure 6	LDA	3	Corsican	600	20	30	Pasquale Manfredi
Figure 7	LDA	3	French	600	20	30	Pasquale Manfredi
Figure 8	LDA	3	Corsican	600	20	30	P. di B.
Figure 9	LDA	4	Corsican	250	20	30	Altore
Figure 10	LSI	3	Corsican	600	20	20	U Sampetracciu
Figure 11	LDA	3	Corsican	500	20	30	U Sampetracciu

Table 6
Benchmark of the Stylometric Analysis with Most Relevant Tests

Experiment	Candidates	Statistical Unit	Cross validation	Balancing	Accuracy
1	All	Character 3-grams	K-Fold 5	Downsampling	0,66
1	All	Character 3-grams	K-Fold 10	Downsampling	0,72
1	All	Character 3-grams	K-Fold 15	Downsampling	0,76
1	All	Character 3-grams	K-Fold 15	Upsampling	0,65
1	All	Character 3-grams	K-Fold 20	Downsampling	0,77
1	All	Character 3-grams	Leave-One Out	Downsampling	0,81
1	All	Character 3-grams	Leave-One Out	Upsampling	0,8
1	All	Most Frequent Words	Leave-One Out	Downsampling	0,66
1	All	Most Frequent Words	Leave-One Out	Upsampling	0,87
Training 1	All	Most Frequent Words	Leave-One Out	Upsampling	0,95
2	Sub-group 1	Most Frequent Words	K-Fold 35	Upsampling	0,89
2	Sub-group 1	Most Frequent Words	Leave-One Out	Upsampling	0,9
2	Sub-group 2	Most Frequent Words	K-Fold 20	Upsampling	0,88
2	Sub-group 2	Most Frequent Words	Leave-One Out	Downsampling	0,63
2	Sub-group 2	Most Frequent Words	Leave-One Out	Upsampling	0,9
2	All	Most Frequent Words	K-Fold 125	Upsampling	0,84
Training 2	All	Most Frequent Words	Leave-One Out	Upsampling	0,86



Figure 4: LDA Analysis on *Petru Rocca*



Figure 5: LDA Analysis on *Petru Rocca*



Figure 6: LDA Analysis on *Pasquale Manfredi*



Figure 7: LDA Analysis on *Pasquale Manfredi*

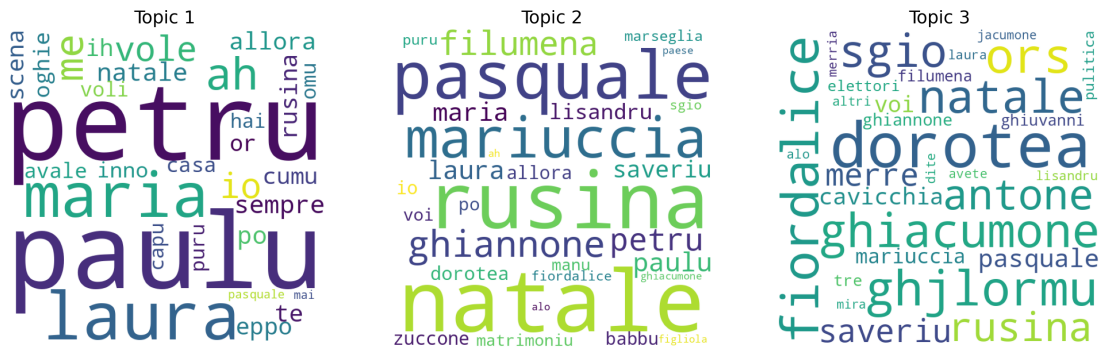


Figure 10: LSI Analysis on *U Sampetracciu*

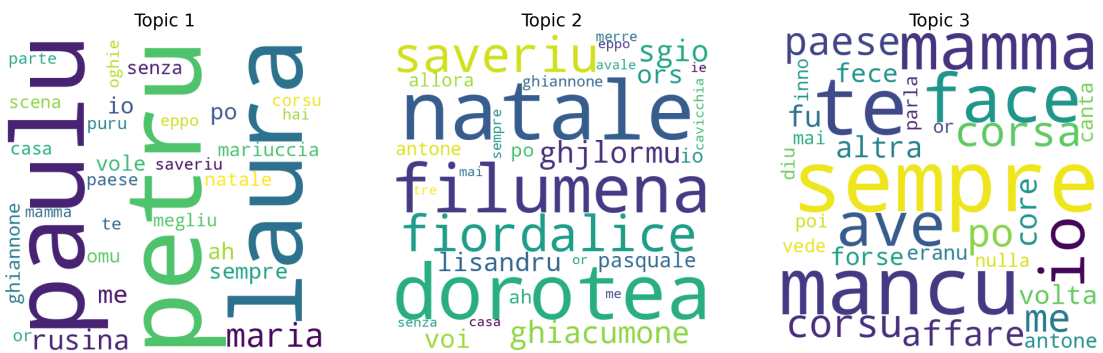


Figure 11: LDA Analysis on *U Sampetracciu*

Table 7

List of Regiolectal and Diachronic Corsican Stopwords

li	com	le	e	u	a
la	ancu	unn	ellu	elli	iddu
cu	ju	ja	cume	ella	elle
quellu	quelle	quelli	quella	so	aghju
aghju	he	simu	avemu	site	hannu
hanu	pe	par	seraghju	sera	sara
oramai	tuttu	chi	quessu	quessa	quesse
quessi	bellu	bonu	bona	bon	boni
bone	boni	be	benche	mo	lu
idda	incu	nostru	vostru	cusi	cun
st	bi	micca	altru	to	avia
stu	quandu	dopu	ca	ava	sottu
pocu	tutt	ind	inde	unu	tantu
ssu	idde	dui	ghje	fattu	vo
eiu	gran	bella	eccu	cum	nun
quantu	nantu	caru	cara	cari	mi
esse	sti	sta	ste	ti	vi
ssa	ssu	ssi	sse	che	perche
dunque	coi	noi	di	da	si
ci	un	una	ma	nostri	nostra
nostre	si	qui	ogni	cio	piu
per	ha	qualchi	ne	in	fa
tu	tutte	tutta	tutti	era	no
dinu	dino	sopra	sotto	mio	mo
so	quand	duve	me	seraghju	po
voi	non	parchi	ad	de	pa

Table 8
Glossary of Relevant Words Present in the Topics

Word	Language	Translation	Word	Language	Translation
affaire	french	business	manu	corsican	hand
affaire	corsican	business	marseglia	corsican	Marseille
ami	french	friend	matrimoniù	corsican	mariage
amore	corsican	love	megliu	corsican	better
article	french	article	merre	corsican	mayor
babbu	corsican	father, dad	ministru	corsican	minister
barbare	french	barbarian	minuranze	corsican	minority
bien	french	good	moda	corsican	mode
canta	corsican	to sing	mondu	corsican	world
centrale	corsican	central	monsieur	french	sir
chemin	french	path	nasitortu	corsican	hook nose
concours	french	competition	naziunale	corsican	national
confrere	french	fellow	oghie	corsican	today
contre	french	against	omu	corsican	man
core	corsican	heart	paese	corsican	country, village
corse	french	Corsica/n	parigi	corsican	Paris
corsu/a/e/i	corsican	Corsican	parti	french	party (political)
croce	corsican	cross	passager	french	passenger
cumitatu	corsican	board	patrie	french	homeland
cummissione	corsican	commission	pays	french	country
cumpagnu	corsican	comrade	poetes	french	poets
cumpare	corsican	mate	politique	french	politic
directeur	french	director	populu	corsican	people
droit	french	right	postal/aux	french	postal, mailing
elettori	corsican	electors	presse	french	press
esprit	french	spirit	prete	corsican	priest
fedede	corsican	faith	prima	corsican	first
federazione	corsican	federation	primavera	corsican	spring
fonctionnaire	french	civil servant	prisidente	corsican	president
français	french	French	prix	french	price
francese	corsican	French	projet	french	project
franchi	corsican	franc (currency)	prova	corsican	try
francia	corsican	France	pueti	corsican	poets
fuir	french	to run away	pulitica	corsican	politic
gauche	french	left	raghione	corsican	reason
giurnale	corsican	journal	razza	corsican	race
gouvernement	french	government	sangue	corsican	blood
guerra	corsican	war	santu/a	corsican	saint
governo	corsican	government	scena	corsican	scene
histoire	french	history	separatisti	corsican	separatists
honneur	french	honor	sgio, scio	corsican	sir
ile	french	island	sicondu	corsican	second
isula	corsican	island	stampa	corsican	press
italie	french	Italy	statu	corsican	state
italien/ne	french	Italian	surete	french	security
jente	corsican	people	teatru	corsican	theatre
jeune	french	young	temps	french	time
jornu	corsican	day	varghiolu	corsican	smallpox
jour	french	day	vergogna	corsican	shame
legge	corsican	law	vita	corsican	life
liberta	corsican	liberty	vitesse	french	speed
lingua	corsican	langage	vole	corsican	to want



A MUVRA

GHIURNALE DI E PIEVE DI CORSICA

ABBONAMENTI :
In Corsica, 8 franchi
In Altro, 9 franchi
U Numeru : 3 soldi

Direttore : Petru ROCCA
Capo Redattore : Mattieu ROCCA

AMMINISTRAZIONE :
99, Corsu Grandval, 99
AJACCIU

Pour nos Détracteurs

** Quale semu noi, e quale capisciarà cio chi noi semu ?
Noi semu quelli chi s'abedemu
ch'ogni populu ricerca l'Indi-
pendenza; ma travagliemu à a
nostra schiavitù.
Noi semu quelli chi ogni volta
ch'un nostru paisanu alxau capu
per fà da bè, li minemu in le
Juntanelle.
Noi semu quelli chi laudemu
Varia e l'acquadi u nostru paese,
a fertilita diaso tarra, a dulcezza
di i so frutti; ma un ci demu
a pena di tira prufittu di ste
ricchezze.
Noi semu u solu populu chi
prifrisce a schiavitù à a libertà
e chi si rallegra d'esse calcicu-
tu.... **

Est-ce assez dur?... Peut-on reprocher plus vivement à des compatriotes leur apathie et leur je m'enfichisme? Peut-on montrer plus de courage, mettre davantage les points sur les i? Les lignes en langue corse qu'on vient de lire, si elles ne brillent pas par la richesse du vocabulaire ou de la syntaxe, constituent cependant un suprême avertissement au peuple... syrien.

Eh ! oui, tous les peuples opprimés, Corses comme Syriens, Egyptiens comme Irlandais, parlent un même langage, souffrent d'un mal identique. Le document ci-dessus qui n'est autre qu'un passage d'article extrait d'un journal de Beyrouit et directement traduit de l'arabe en corse, nous fournit la preuve évidente de la communauté d'âme, de cœur et aussi, avouons-le, de déchéance des peuples opprimés.

Pourquoi alors, certains de nos compatriotes tels que Jean Wallis et P. O. Poli, nous refuseraient-ils le droit de parler au titre de sans-patrie et de déshérités? Quelle plume corse en dépit de toutes les réticences de ces messieurs, n'eût pas signé le dit document, n'en eût pas paraphé tous les termes, légitimé toutes les audaces? Son auteur cependant n'est pas corse... il est sy-

rien, c'est à dire soumis à la très douce, très paternelle et très bienfaisante autorité du général Gouraud....

Certes, nous, n'entendons pas établir entre Corses, Syriens, Irlandais, Marocains, Bretons, etc... une identité d'esclavage et d'asservissement. La formule de l'oppression dont souffre le peuple insulaire à été définie par les régionalistes, elle ne compromet la bonne foi de personne nime met directement en cause cette France dont la mission civilisatrice est pourtant de jour en jour plus contestée.

Ni séparatistes ni irrédentistes! se sont empressés de clamer certains de nos compatriotes chatouilleux. Nous sommes parfaitement d'accord. Mais que ces messieurs susceptibles, veuillent reconnaître à leur tour que pour la grosse majorité des Corses, les mots d'Indépendance et de Liberté, d'Esclavage et de Domination ont toujours un sens très réel, très profond, très visible.....

Mathieu ROCCA

L'EXTRAORDINAIRE APPAKITION

Art floou, or art floou not, fatal vision ?
SHAKESPEARE

Le dieu-huitième jour de janvier 1922, vers 10 heures du soir, à Ajaccio, alors qu'une troupe de comédiens jouait une revue locale, un fait des plus extraordinaires se produisit :

Un homme d'assez belle taille, coiffé d'une perruque à canons, cravaté de blanc, revêtu d'un pourpoint bleu et de hauts de chausse noirs, la dentelle aux poignets et les escarpins aux pieds, sortit subrepticement des coulisses et se plaça stupéfié comme la Mort, au beau milieu d'un simili peron qui - du coup - résonna comme un cerceuil....

Quale sarà ?

Il avait le teint rosé, les yeux drôlement ouverts, les lèvres lippues, le menton abondant, la poitrine fière, les jambes plutôt courtes.

A la minute même où il était entré en scène, deux allégores représentant Ajaccio et Bastia caquetaient comme de petites folles. Leur querelle avait pour motif une question d'amour-propre; Bastia jalousait Ajaccio, sa Préfecture, sa Marine, son siège épiscopal, et Ajaccio, fraîche et rose, lui répondait du tac au tac.... Quand au sublime Inconnu d'outre-tombe, il se contentait de dominer.

Quale sarà ?
C'était un contemporain de Washington et de Lafayette; la puissance métaphysique que son être dégageait vous libérait temporairement du Temps et de l'Espace, pour vous précipiter dans le champ magnétique de l'Histoire....

Puis soudain, ce holoide effrayant d'orgueil et de placidité, se mit avec une grande noblesse....

Il parla.
Gigantesquement campé entre les allégores d'Ajaccio et de Bastia, il disait son courroux, semonçait ses filles bien aimées....

Quale sarà ?

Sa bouche d'Orphée nouvellement revenu des champs élyséens vomissait le reproche avec mesure, tact, amour, paternité.

Bon appétit mes filles! O vous sœurs rivales!

Vertueuses cités! Voilà votre façon de servir, Servantes qui ruinez la maison!

Rougissez toutes deux devant la Corse entière,

D'un glorieux passé pitieuses héritières!

Regardez! regardez! quelle pitié mes filles!

Les frondaisons du maquis semblaient s'étendre sur sa tête auguste, entourer son être astral, pour laisser retentir dans nos seuls coeurs, au seul usage de notre entendement corse, l'écho de sa mâle parole.

Quale sarà ?

Et, lentement à la faveur de l'ombre et de la conspiration, d'autres allégores de cités cyméennes se serrèrent autour du *Padre della Patria*, mort élyséen en visite chez la grande morte élyséenne.

Michal CORANO

En Corse... toi ? Revue locale en deux actes de Jean Maki, actuellement représentée au théâtre Napoléon à Ajaccio à notre confrère Michel Corano la matière de son émouvant article. C'est dire assez si cette manifestation d'art théâtral, où le nationalisme corse eut sa large part puisque le talentueux artiste Battesti y interpréta Pascal Paoli, a été goûtée par tous les patriotes corses sans distinction de clan ni de parti. Nous donnerons d'ailleurs un long compte rendu d'*En Corse... toi ?* dans notre prochain numéro.

U Yeru Corsu



Sare stu guardia-prighe....

...C stu lavuratore ?

Figure 12: Example of a Muvra Front Page