Umanistica Digitale – ISSN: 2532-8816 – n. 16, 2023
D. Paci, V. Sarbach-Pulicani - *Una lettura "distante" di* A Muvra*: esplorare la stampa autonomista in lingua còrsa attraverso il topic modeling*
DOI: http://doi.org/10.6092/issn.2532-8816/17642

Una lettura "distante" di *A Muvra*: esplorare la stampa autonomista in lingua còrsa attraverso il topic modeling

Deborah Paci

Università di Modena e Reggio Emilia
deborah.paci@unimore.it

Vincent Sarbach-Pulicani

Université Côte d'Azur
vincent.sarbach-pulicani@univ-cotedazur.fr

Abstract

L'obiettivo di questo contributo è quello di indagare la stampa autonomista attraverso il topic modeling. Verrà confrontata la validità di due metodi di topic modeling, LDA e LSA, su un corpus trilingue - italiano, còrso e francese - non soggetto a lemmatizzazione, non soltanto al fine di interpretare i risultati ottenuti, ma anche di comprendere quale di questi metodi si avvicini maggiormente alle nostre conoscenze pregresse sull'argomento, che sono frutto di una lettura ravvicinata.

Parole chiave: Corsica; Autonomismo; Lingua còrsa; Topic modelling; NLP

The aim of this article is to explore the autonomist press in the Corsican language through topic modelling. In order to do so, the effectiveness of two topic modelling methods, LDA and LSA, will be compared on a trilingual corpus - Italian, Corsican and French - not subject to lemmatization, not only in order to assess the results obtained, but also to understand which of these methods comes closest to our previous knowledge on the subject, which is the result of a close reading.

Keywords: Corsica; Autonomism; Corsican language; Topic modelling; NLP

Introduzione [1]

A Muvra - letteralmente «Il Muflone» - è il titolo del settimanale in lingua còrsa, che fece la sua comparsa nel 1920 attestandosi come l'organo dell'autonomismo còrso. Pubblicato ininterrottamente per quasi vent'anni, nel 1939 fu raggiunto dalla censura governativa poiché reo di veicolare messaggi irredentisti volti ad affermare l'italianità della Corsica e di conseguenza il diritto dell'Italia fascista di annettere l'isola. Sebbene i cosiddetti muvristi si considerassero anzitutto regionalisti che rivendicavano uno status autonomo dell'isola all'interno dello Stato francese, i loro legami con gli irredentisti fascisti erano indiscutibili e ben noti alla polizia francese. *A Muvra* profittò sin dal 1924 dei finanziamenti previsti dal Comitato per la Corsica, un ente istituito dal governo mussoliniano allo scopo di coordinare tutte le iniziative culturali di carattere irredentistico relative all'isola. Benché il Ministero degli Esteri italiano si fosse preoccupato di non far comparire il governo fascista come promotore dell'iniziativa così da evitare di suscitare reazioni indignate da parte del governo francese, i servizi di polizia francese erano perfettamente a conoscenza della rete propagandistica intessuta dal fascismo italiano anche attraverso il finanziamento di *A Muvra*.

Nonostante si presentasse ai suoi lettori come una rivista con finalità culturali, la dimensione politica di *A Muvra* era esplicita. Vi era un'eterogeneità di temi che spaziavano dalla lingua all'istruzione sino alla politica interna ed estera. Anche il formato dei testi pubblicati era variabile: comparivano articoli, poesie, trascrizioni di canti popolari, ma anche lettere aperte e racconti. Questa diversità di forme e contenuti dipendeva in larga misura anche dalla varietà dei collaboratori, che - a dispetto di uno zoccolo duro di autori regolari - risultavano considerevolmente numerosi.

La letteratura specialistica ha prodotto un numero significativo di lavori sull'autonomismo còrso e sulle sue espressioni culturali [30][29][41][42][43][36][22][12][44][39][31], incluse le riviste.[2] Tuttavia, non è ancora stato effettuato uno studio che indaghi la retorica autonomista attraverso l'analisi sistematica di un corpus costituito dai numeri di una rivista.

L'obiettivo di questo contributo è quello di confrontare la validità di due metodi di topic modeling, LDA e LSA, su un corpus trilingue - italiano, còrso e francese - non soggetto a lemmatizzazione, non soltanto al fine di interpretare i risultati ottenuti, ma anche di comprendere quale di questi metodi si avvicini maggiormente alle nostre conoscenze pregresse sull'argomento, che sono frutto di una lettura ravvicinata [32].

L'originalità del nostro studio risiede nella varietà delle scale di analisi che abbiamo previsto, sia che si tratti di autori diversi, di tipologie di articoli o di lingue utilizzate. Inoltre non ci siamo limitati semplicemente a effettuare un'analisi di base di tutti gli articoli per confrontare i due

[1] Questo articolo è il risultato di una collaborazione tra i due autori. Entrambi hanno ideato e discusso collegialmente tutte le parti di cui si compone il testo. In particolare, Deborah Paci ha scritto "Introduzione", "Contesto Storico", "Aspetti linguistici" e "Analisi e risultati". Vincent Sarbach-Pulicani si è occupato della stesura dei seguenti paragrafi: "Metodologia", "Costituzione del set di dati", "Statistiche descrittive", "Selezione del vocabolario", "Note conclusive". Gli autori ringraziano Lorenza Brasile per l'attenta rilettura.

[2] Si veda il numero monografico pubblicato su *Etudes Corse* intitolato "Les Revues corses de l'entre-deux-guerres" (no. 64, dicembre 2007).

metodi al fine di osservare i risultati migliori; dal nostro punto di vista di storici, si tratta piuttosto di confrontare le diverse categorie di analisi per evidenziare le caratteristiche specifiche.

Dapprima verrà ricostruito il contesto storico necessario per inquadrare la genesi e l'evoluzione di *A Muvra*; successivamente verrà tracciato il quadro metodologico; in terzo luogo verrà esposto nel dettaglio il dataset e infine verranno interpretati i risultati.

La stampa dialettale e autonomista è stata indagata attraverso il topic modeling, ossia una tipologia di modellazione statistica tesa a far emergere "argomenti" (topic) astratti che ricorrono in una collezione di documenti. Più precisamente, si tratta di una tecnica di analisi del testo utilizzata nell'ambito dell'intelligenza artificiale e nell'elaborazione automatica del linguaggio naturale. Il suo obiettivo principale è quello di identificare e raggruppare i principali argomenti o temi presenti in un insieme di documenti, basandosi su modelli statistici di apprendimento automatico come LDA (Latent Dirichlet Allocation), che esamineremo in questa ricerca. La sua applicazione metodologica può essere ricondotta alla "distant reading" - "lettura distante" - proposta dallo storico della letteratura Franco Moretti [34]. Questa nuova modalità di lettura e di analisi dei testi implica l'adozione di una posizione distante del lettore rispetto alla collezione di testi da analizzare e al contempo consente di individuare le strutture principali e le parti informative sottostanti l'insieme dei documenti in modo più efficace rispetto alla tradizionale modalità di "lettura ravvicinata" o "lettura squisitamente qualitativa". Il metodo proposto da Moretti si presta anche allo studio di corpora storici [5].

La natura multilingue dei dati che abbiamo analizzato, come vedremo, presenta inoltre una sfida importante, soprattutto nella fase di preparazione e nello specifico quando si tratta di indagare l'idioma còrso, che è considerato una lingua con poche risorse digitali nel campo dell'elaborazione automatica del linguaggio (Under Resourced Language) [45].

A Muvra: lessico e storia in gioco

Contesto storico

Le origini del nazionalismo còrso si collocano nella tarda modernità e precisamente nel XVIII secolo, quando l'isola attraversò, sotto la guida del generale Pasquale Paoli, una fase politica di "quasi-indipendenza" [14]. Il 1729 fu l'anno in cui si consumò la "rivoluzione còrsa" contro la Repubblica di Genova, che reggeva l'isola dal 1284 dopo averla sottratta al controllo di Pisa [2][23]. Ad animare la rivoluzione, che avrebbe avuto un corso lungo oltre quarant'anni, furono gli appartenenti ad un ceto sociale elevato. Fu Paoli, nominato generale della "nazione còrsa", a emanare, nel 1755, una costituzione rappresentativa. Dopo alterne vicende la Repubblica di Genova conferì alla Francia i suoi diritti di sovranità sull'isola. La celebre, quanto mai simbolica, battaglia di Ponte Novu del 1769 decretò la fine il governo di Paoli segnando l'inizio del governo francese sull'isola [11]. Fatta eccezione per il Regno Anglo-Corso (1794-1796), l'isola rimase saldamente nelle mani del governo di Parigi benché permanessero problemi relativi alla gestione dell'ordine pubblico e alla sopravvivenza di alcune pratiche, *in primis* la vendetta [47].

La questione linguistica svolse un ruolo significativo nelle rivendicazioni identitarie della Corsica. Il còrso era praticato da tutti gli strati sociali della popolazione isolana, mentre l'italiano mantenne una co-ufficialità con il francese nella comunicazione scritta almeno fino alla prima metà dell'Ottocento. Ad essere impiegato dall'amministrazione e da coloro che rappresentavano il potere centrale era l'italiano. Il radicamento della lingua e della cultura italiana nell'isola era dovuto non soltanto ai tradizionali contatti con Pisa e alla dominazione secolare della Repubblica

di Genova, ma anche alle difficoltà da parte del governo di Parigi di sviluppare un sistema educativo nell'isola. Questo contribuì a ritardare gli effetti del radicamento della lingua francese [9]. Fu soltanto a partire dall'avvento della Terza Repubblica, dagli anni Settanta dell'Ottocento, che le élite còrse divennero gradualmente consapevoli della loro appartenenza alla corrente delle idee repubblicane: fu allora che si produsse un reale processo di francesizzazione dell'isola [39]. Al contempo emerse, sin dagli anni Ottanta del XIX secolo, un regionalismo còrso che si opponeva allo Stato centrale e che utilizzava la lingua còrsa come uno degli strumenti per veicolare le proprie idee. La difesa della lingua còrsa era tesa a depotenziare l'uso di quella francese e, conseguentemente, a minare l'unità nazionale: "al concetto espresso con la formula 'une Nation, une langue' il nazionalismo insulare oppose quello 'ma Patrie, c'est ma langue', ovverosia la mia identità" [36].

Se escludiamo alcuni testi dialettali apparsi sporadicamente, in particolare quelli del poeta Salvatore Viale[3], fu sotto l'impulso di Santu Casanova e del suo *A Tramuntana* che ebbe origine la stampa in lingua còrsa: un'evidenza del fatto che il regionalismo còrso nacque in concomitanza con la fondazione di questo primo periodico. *A Tramuntana* fu un settimanale politico, umoristico, satirico e letterario interamente redatto in lingua còrsa che vide la luce nel 1896 per iniziativa di Pierre-Toussaint Casanova, detto Santu Casanova. L'iniziativa editoriale sorgeva però in un contesto politico e sociale caratterizzato dall'indigenza diffusa e dall'arretratezza sociale e culturale, effetto anche del radicamento del sistema clientelare. Santu Casanova intendeva denunciare la situazione economica, sociale e politica dell'isola: riteneva infatti che la formalizzazione scritta della lingua còrsa avrebbe permesso, da un lato, di rafforzare l'identità del popolo còrso, dall'altro, proprio in virtù di questa presa di coscienza, i còrsi avrebbero potuto esprimersi e denunciare le proprie condizioni [40].

I semi piantati da Casanova con *A Tramuntana* non tardarono a dare i loro frutti. Al termine della Prima guerra mondiale, in Corsica si sviluppò un movimento regionalista che chiedeva una maggiore autonomia, l'istituzione dell'Università nell'isola, la salvaguardia della lingua e della cultura còrsa, nonché il suo insegnamento in ogni grado dell'istruzione scolastica. Il primo conflitto mondiale diede infatti nuova linfa agli impulsi identitari dei nazionalismi periferici. In Corsica, questo si concretizzò con la comparsa nel 1920 di un nuovo giornale in lingua còrsa: *A Muvra* fondato a Parigi per iniziativa dei fratelli Petru e Matteo Rocca. Diretto da Petru Rocca con la collaborazione di alcuni ex combattenti decorati al valor militare, divenne ben presto l'organo di difesa delle rivendicazioni insulari.

La pubblicazione di *A Muvra* fu accompagnata dalla creazione, nell'ottobre 1922, del Partitu corsu d'azione, un partito politico autonomista. Petru Rocca riprese sulle colonne di *A Muvra* i principali temi veicolati da *A Tramuntana*. Le parole d'ordine del pensiero corsista furono: "una Nazione, un Popolo, una Lingua, una Storia e una Religione".

Nel novembre 1926 il Partitu corsu d'azione assunse una nuova denominazione: Partitu corsu autonomista e da quel momento apparvero su *A Muvra* diversi articoli in cui si richiamava l'attenzione ai Quattordici Punti di Woodrow Wilson relativi al diritto di tutti i popoli di disporre

[3] Salvatore Viale riteneva che la "questione corsa" dovesse essere proiettata oltre i confini regionali e che, di conseguenza, fosse necessario collocarla al centro del più generale dibattito sulle politiche economiche e sociali. Questo dibattito trovò eco nelle principali riviste europee, come quella di Vieusseux, *L'Antologia*, molto conosciuta dall'élite di Bastia, di cui Viale era il rappresentante più importante. Si rimanda a [8]. Fu proprio Niccolò Tommaseo, esule in Corsica, a studiare il dialetto còrso, di cui celebrò la ricchezza, grazie all'aiuto di Viale. Si veda [36].

di loro stessi. Questa retorica era tesa a fornire una legittimazione alle aspirazioni autonomiste dei còrsi. Conclusa questa fase autonomista, il Partitu corsu d'azione avrebbe abbracciato la propaganda irredentista condotta dal fascismo italiano, che rivendicava il possesso dell'isola sulla base di argomentazioni di natura storica, culturale e linguistica [36][22][12]. Tuttavia questo movimento filo-fascista era minoritario e pertanto la maggioranza dei còrsi si oppose all'Italia fascista che aveva occupato l'isola nel 1942. Non a caso la Corsica fu il primo dipartimento francese ad essere liberato nell'ottobre 1943 [7][24].

Aspetti linguistici

L'idioma còrso è una variabile centrale all'interno della nostra indagine: da un lato è necessario comprenderne il contesto in cui è stato impiegato, dall'altro guardare alle ragioni del suo utilizzo presso i muvristi. L'esistenza stessa di un giornale in còrso testimonia l'importanza per gli isolani di difendere e promuovere la loro lingua. Oltre alla natura intrinsecamente culturale del regionalismo linguistico, dobbiamo rinvenire implicazioni di carattere identitario, poiché, come ha rilevato Jean-Paul Pellegrinetti, "lingua e popolo sono intimamente legati per la difesa dell'originalità dell'identità" [40]. Un aspetto che implica un vero e proprio ribaltamento dell'ideale universalistico francese, dal momento che la lingua francese diventa una lingua straniera al pari dell'italiano o dell'inglese.

L'uso della lingua còrsa nella stampa era considerato dai muvristi come essenziale per la sua difesa e preservazione. Questo desiderio di unire la comunità nazionale attorno a una lingua comune è un fenomeno ricorrente nei Paesi europei durante il secolo dei nazionalismi. Gli stessi muvristi non contestavano la somiglianza del còrso con l'idioma italiano, come sottolinea il sacerdote Dominique Carlotti, che parlava del dialetto isolano come di un "miscuglio di termini trasversali".[4] E non va taciuto il fatto che la pubblicistica fascista, per avvalorare il diritto dell'Italia a rivendicare l'isola, avesse presentato il còrso come uno dei più antichi e puri dialetti italiani. A questo proposito Gioacchino Volpe scrisse sul periodico *Corsica. Bollettino mensile della Società Gli Amici della Corsica*: "la lingua italiana antica è oggi nel dialetto corso, dialetto che conserva ancora tutta la sonorità e l'armonia dei nostri poeti primitivi, e che soprattutto, come i nostri antichi dialetti, è il depositario dell'anima di un popolo, come mai nessun dialetto italiano lo fu in nessun tempo"[5].

Da un punto di vista strettamente linguistico, il còrso è una lingua italo-romanza molto affine al gruppo linguistico toscano con elementi comuni che provengono dal genovese, dal napoletano e dal sardo [16]. È a sua volta diviso in diverse aree linguistiche, tra cui le due principali sono il Cismuntincu parlato nel nord dell'isola e il Pumuntincu che è predominante nel sud. Sin dal 1922, i muvristi affermarono che il còrso rappresentava "un'entità notevole, una qualità intrinseca, in modo che tutti i Larousse di Francia e Navarra non dicano più che parliamo toscano".[6] Per gli autori autonomisti si rese dunque necessario produrre un numero rilevante di testi scritti in lingua còrsa al fine di legittimare l'uso e la volontà politica di porre tale idioma sullo stesso piano delle lingue maggioritarie, ossia l'italiano e il francese. Questa è stata definita dal semiologo Jean-Marie Klinkenberg la fase difensiva e culturale del regionalismo linguistico [28]. I difensori della lingua còrsa andarono così alla ricerca degli scrittori fondatori assurti a "padri

[4] *A Muvra*, 15/07/1923: "Insiste chì u Corsu è un mischiu di termini straversi, cartaginesi, saracini, aragunesi, greghi, tuscani e francesi".

[5] *Corsica. Bollettino mensile della Società Gli Amici della Corsica*. 1924: 3.

[6] *A Muvra*, 04/06/1922.

della lingua còrsa". Pertanto furono ripubblicati i testi di Pietro Cirneo, storico còrso del XV secolo, o di Salvatore Viale. Questa ricerca filologica fu segnata dalla volontà di standardizzare la lingua di fronte alla grande varietà fonetica e grammaticale del còrso. In questo periodo apparvero le prime grammatiche, come quella di Antoine Bonifacio pubblicata nel 1926 [4]. Tuttavia, i tentativi di standardizzazione non andarono a buon fine: di fatto ogni autore scriveva nel modo in cui si esprimeva oralmente. Questa varietà idiomatica rappresenta anche una sfida nel trattamento della lingua. Per vedere sviluppi concreti nella standardizzazione della lingua còrsa sarebbe stato necessario attendere gli inizi degli anni Settanta, quando il movimento autonomista ritornò in auge, dopo la stagione muvrista, e diede avvio al movimento del *riacquistu*, un risveglio culturale da parte di una nuova generazione di còrsi che si impegnò attivamente per una valorizzazione della lingua e della cultura còrsa [18], [38].

Metodologia

L'evoluzione dell'uso della lingua còrsa è un tema su cui diversi studiosi hanno riflettuto e proposto alcune interpretazioni. A tal proposito risultano imprescindibili i lavori della linguista Marie-José Dalbera-Stefanaggi, autrice del *Nouvel atlas linguistique et ethnographique de la Corse* datato 1995 [15], che hanno condotto alla creazione di una *Banque de Données Langue Corse (BDLC)* [27]. Quest'iniziativa rappresenta anche il primo tentativo di lemmatizzazione di un corpus in lingua còrsa. Soltanto da una decina di anni a questa parte i ricercatori hanno studiato le lingue regionali attraverso i metodi propri dell'elaborazione del linguaggio naturale (NLP, Natural Language Processing). Pertanto il nostro contributo si inserisce in questa corrente di studi che per ulteriori ricerche potrà avvalersi di strumenti per l'elaborazione della lingua còrsa[7]. Ai fini di questo studio abbiamo provveduto ad effettuare analisi computazionali sui dati grezzi. In questo senso, il topic modeling, un metodo probabilistico per far emergere argomenti latenti in un corpus di documenti, sembra essere la modalità più adeguata per indagare un corpus in lingua còrsa.

La scelta del topic modeling per il nostro studio può essere spiegata in due modi diversi. In primo luogo questo metodo di apprendimento non supervisionato può essere applicato a dati grezzi con le parole-token come unità statistiche. In secondo luogo, da un punto di vista storico, ci permette di spostare l'attenzione sulla relazione tra temi e tipologie negli articoli pubblicati su *A Muvra*.

In questo articolo abbiamo esaminato l'uso di due diversi approcci per determinare quale sia il più rilevante per l'analisi di un corpus trilingue. La Latent Dirichlet Allocation (LDA) è un metodo basato su una matrice termine-documento. Essa si basa sul presupposto che i documenti sono rappresentati come miscele casuali di argomenti latenti, all'interno dei quali ogni argomento è caratterizzato da una distribuzione di parole. La LDA è stata introdotta nel 2003 [3] ma da allora ha conosciuto tutta una serie di miglioramenti, tra cui l'incorporazione di tecniche di inferenza bayesiana [25]. Al contrario, la Latent Semantic Analysis (LSA), prevede la creazione di uno spazio semantico basato su un corpus in cui le somiglianze tra parole o documenti sono calcolate su scala statistica. Introdotta da Susan Dumais nel 1990 utilizza una tecnica matematica definita Singular Value Decomposition (SVD) che consente di ridurre la dimensionalità delle matrici termine-documento e identificare le somiglianze semantiche [17]. Ognuno di questi metodi presenta vantaggi e svantaggi che devono essere presi in considerazione, da qui

[7]BDLC:https://bdlc.univ-corse.fr/tal/index.php?page=home; https://bdlc.univcorse.fr/bdlc/corse.php?page=texte; https://bdlc.univ-corse.fr/concord/; Corpus Canopé: https://bdlc.univ-corse.fr/concord_ccdc/

l'importanza della comparazione insita nel nostro studio [13]. La LSA ha una velocità di calcolo molto più elevata e un'interpretazione più diretta dei risultati. La LDA, invece, è un modello più flessibile. Nel 2020, un gruppo di ricercatori ha inteso confrontare i due metodi addestrandoli su un corpus di articoli della BBC. Gli esiti della loro ricerca hanno rivelato come la LSA sia più efficace se ci troviamo di fronte ad una grande quantità di dati e a un minor numero di iterazioni rispetto alla LDA, mentre quest'ultima è più adatta a corpora di piccole dimensioni [26]. Ai fini di questa ricerca abbiamo deciso di lavorare con il linguaggio Python e specificamente con la libreria Gensim[8]. Questa libreria ha reso possibile, in particolare, di effettuare le nostre analisi utilizzando i due metodi appena menzionati. Pertanto abbiamo tenuto conto di un certo numero di iperparametri offerti da Gensim, come il numero di argomenti, il numero di parole, il numero di iterazioni e il numero di passaggi. Questi iperparametri variano in base al numero di documenti e di parole.

Il corpus

Costituzione del set di dati

La costruzione del set di dati è stata una fase essenziale del nostro studio. Poiché la disponibilità di dati testuali non era garantita, abbiamo dovuto creare un set di dati a partire dai documenti d'archivio digitalizzati. Abbiamo anzitutto proceduto con una conversione automatica di un file contenente l'immagine di un documento in un file di testo, utilizzando un software OCR. Le problematiche legate alla trascrizione automatica dei giornali storici possono avere un impatto sulla qualità dei dati. La fase di standardizzazione è quindi essenziale per elaborare correttamente i dati. Infine, la selezione del vocabolario è un elemento da tenere in considerazione per ridurre i tempi di addestramento dei modelli di topic modeling, conservando al contempo le informazioni rilevanti per la modellazione dei soggetti.

Il nostro set di dati è stato costruito utilizzando la piattaforma di digitalizzazione dell'archivio Gallica gestita dalla Bibliothèque nationale de France (BnF)[9]. Sebbene Gallica detenga quasi tutti i numeri, solo alcuni sono disponibili in formato digitale. Si tratta di 238 numeri pubblicati tra il dicembre 1924 e il 1930, una quantità di documenti piuttosto elevata. Il fatto che l'API di Gallica sia aperta ha reso abbastanza agevole il recupero delle scansioni utilizzando lo standard IIIF per la standardizzazione delle immagini ad altissima risoluzione[10]. Di fronte ad alcuni problemi di standardizzazione del numero di immagini IIIF disponibili per ogni numero della rivista e per essere sicuri di estrarre tutte le fonti, abbiamo utilizzato lo script IIIF-Crawler, sviluppato da Thibault Clérice e Jean-Baptiste Camps [6]. Questo procedimento ci ha permesso di importare automaticamente un grande numero di immagini di buona qualità. L'idea è stata quella di interrogare l'API di Gallica per recuperare gli identificatori ARK perenni per ogni numero e importare tutte queste informazioni in un file TSV che potesse essere letto dallo strumento IIIF-Crawler. In questo modo, tenendo conto delle immagini mancanti, abbiamo ottenuto un corpus completo di 230 numeri di *A Muvra*.

[8] https://pypi.org/project/gensim/

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

[10] https://iiif.io/.

Per la raccolta dei dati testuali, abbiamo deciso di sperimentare una modalità non consueta. Gallica fornisce un OCR delle fonti stampate insieme alla loro digitalizzazione. A questo proposito, non bisogna fare affidamento sul dato di accuratezza indicato dalla piattaforma, che spesso si avvicina al 99%. Una semplice valutazione empirica rivelerà i limiti dell'OCR proposto. Possiamo però affidarci ai file XML/ALTO relativi all'OCR messi a disposizione anche per ottenere informazioni sulla segmentazione, elemento particolarmente complesso da gestire per i giornali, e che siano di ottima qualità come nel caso specifico. Abbiamo quindi estratto da questi file le coordinate delle regioni e le abbiamo incorporate in file di tipo UZN che possono essere letti dal motore di riconoscimento automatico dei caratteri Tesseract-OCR. Questo procedimento ha il vantaggio di tenere conto di più lingue pur avendo un modello dedicato al còrso. Questa fase può essere riassunta dal seguente diagramma della pipeline:

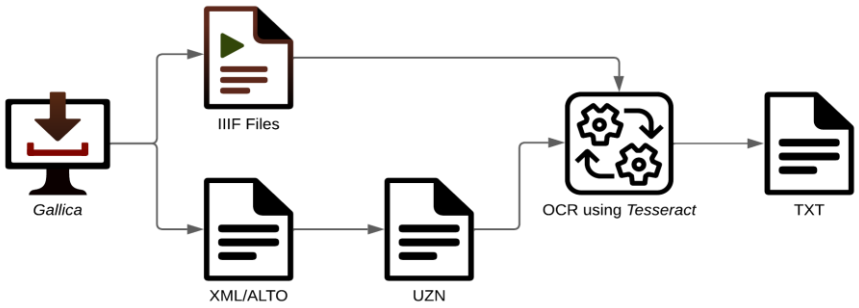

Fig. 1 - Pipeline del processo di raccolta dei dati testuali di *A Muvra*.

Sulla base di un documento di prova, abbiamo dunque ottenuto un'accuratezza OCR di circa il 97% misurando la sua efficienza con la distanza di Levenshtein. Questi dati, tuttavia, devono essere considerati con cautela, perché a differenza dei manoscritti, dove la qualità delle immagini è spesso costante, le scansioni dei giornali possono essere variabili. Pertanto, spesso troviamo angoli danneggiati o leggermente piegati, che creano una quantità significativa di "rumore". La fase di normalizzazione è quindi essenziale per rimuovere la punteggiatura e l'accentazione. Questa normalizzazione altamente granulare spiega la scelta della distanza di Levenshtein per valutare la trascrizione in modo estrinseco. Questa misura è ideale per confrontare due stringhe di caratteri e ha una base teorica particolarmente affidabile [48]. Tuttavia è necessaria anche una valutazione intrinseca per sapere se il rumore generato ha inquinato troppo i nostri dati. A tal fine, abbiamo determinato l'affidabilità del vocabolario rappresentandolo con la legge di Zipf. Questa legge presuppone che la frequenza di una parola sia inversamente proporzionale al suo rango. Anche se datata, questa legge è ancora valida, in particolare nella linguistica quantitativa, le cui implicazioni sono descritte in un articolo di Marcelo A. Montemuro [33]. Scalando la legge in senso logaritmico, abbiamo ottenuto un grafico che rappresenta approssimativamente una linea retta se il vocabolario è affidabile e se non si ha una sovrarappresentazione di apici e parole a bassa frequenza dovuta a un OCR scadente.

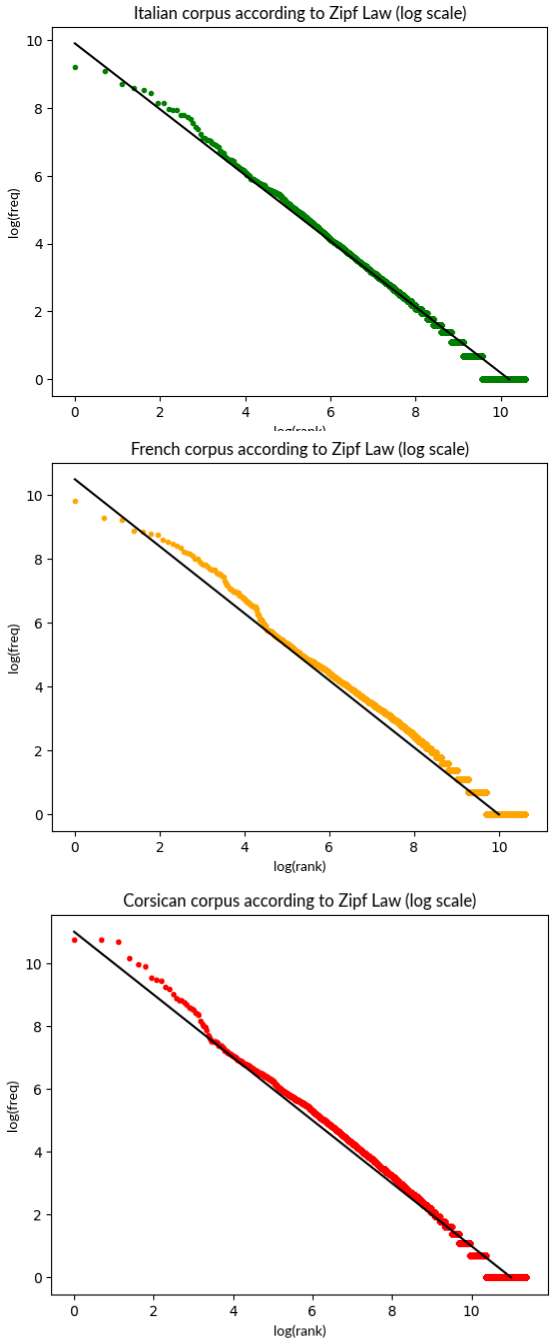

Fig. 2 - I diversi corpora secondo la legge di Zipf in scala logaritmica

Come si può osservare, la rappresentazione della legge di Zipf sembra confermare i buoni punteggi ottenuti durante la trascrizione automatica. Possiamo altresì notare fenomeni comuni: ad esempio il fatto che le parole più frequenti sembrano spiccare, rappresentando parole funzionali la cui frequenza complessiva è molto più alta di quella delle altre parole.

Statistiche descrittive

Dopo aver ottenuto i dati testuali, siamo stati in grado di effettuare statistiche esplorative per riassumere e dare una visione complessiva dei dati a nostra disposizione. Il topic modeling non richiede necessariamente una quantità massiccia di token per funzionare, per cui sono stati sufficienti i numeri di *A Muvra* pubblicati tra il 1925 e il 1930. Tuttavia, occorre tenere presente che questa circostanza può influenzare i risultati ottenuti, ma su questo torneremo più avanti nel nostro articolo. Di seguito riportiamo una tabella che riassume la lunghezza dei corpora in base alla lingua, seguita da due grafici che rappresentano in maniera più immediata queste informazioni:

	Còrso	**Francese**	**Italiano**	**Totale**
Lunghezza del corpus	897965	380457	263095	1541517

Tab. 1 - Lunghezza del corpus a seconda della lingua

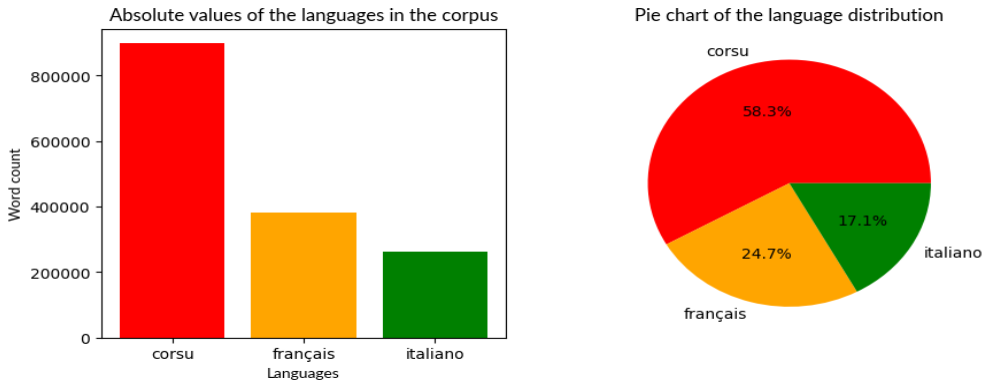

Fig. 3 - Rappresentazione statistica del corpus

Il còrso predomina sulle altre lingue, il che è abbastanza logico visto che *A Muvra* è anzitutto un giornale dialettale. Anche il francese e l'italiano svolgono un ruolo importante, rappresentando insieme quasi il 40% del nostro corpus totale. Vediamo ora la ripartizione degli articoli per tipologia.

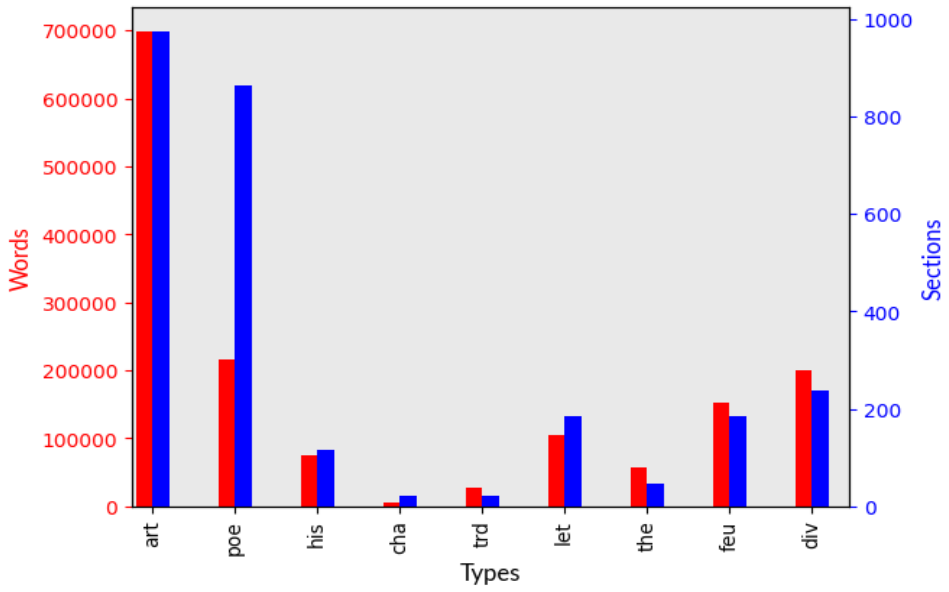

Fig. 4 - Ripartizione delle tipologie di articoli e la loro distribuzione sul totale delle parole

La tipologia degli articoli è stata scelta sulla base dell'osservazione empirica. L'elevato numero di tipologie di articoli riflette la molteplicità degli interessi del giornale, che vuole essere da un lato una pubblicazione di carattere culturale, dall'altro un organo politico. Di seguito sono elencate le tipologie osservate:

- Articolo (*art*): tutti gli articoli "tradizionali" che descrivono un evento o, ad esempio, un'opinione politica.
- Canti (*cha*): il testo dei brani musicali presenti nel giornale. Sono esclusi gli spartiti musicali poiché queste generano "rumore" durante il processo di ocerizzazione.
- Miscellanea (*div*): sezioni speciali, come bibliografie o glossari.
- Lettere (*let*): lettere aperte o articoli in forma di lettera.
- Opere teatrali (*the*): opere in cui alcuni nomi sono ridondanti o alternano il linguaggio puramente letterario a quello parlato.
- Feuilleton (*feu*): pratica giornalistica molto diffusa, soprattutto nel XIX secolo.
- Storia (*his*): forme narrative come racconti e leggende.
- Traduzioni (trd): testi antichi o contemporanei tradotti da un'altra lingua.

Gli articoli classici costituiscono la maggior parte dei contributi. La poesia, molto numerosa, rappresenta tuttavia un numero di parole quasi pari a quello degli articoli miscellanei o feuilleton. In questo articolo non ci occuperemo tuttavia di una selezione di autori perché ciò significherebbe entrare troppo nel dettaglio, con il rischio di trascurare il focus della trattazione.

Selezione del vocabolario

Il vocabolario utilizzato per il topic modeling è particolarmente importante, in quanto determina le parole da prendere in considerazione nella fase di apprendimento. Le parole scelte per il topic modeling non devono essere troppo numerose, poiché il modello può richiedere un tempo di apprendimento piuttosto lungo. Il numero di documenti e il vocabolario scelto rivestono

pertanto un ruolo centrale nei vari bias che verranno applicati in questa fase della nostra analisi. Queste parole funzionali costituiscono un "rumore" che non è necessariamente auspicabile e che può alterare i nostri risultati. Il nostro proposito è stato quello di eliminarle per ridurre le dimensioni del vocabolario.

La questione delle parole funzionali nel caso del còrso risulta di particolare interesse. Mentre per l'italiano e il francese gli elenchi sono facilmente disponibili, lo stesso non vale per il còrso. L'identificazione di parole-vuote in lingue con poche risorse digitali si presenta come una sfida importante. Alcuni ricercatori si sono proposti di identificarle, utilizzando, ad esempio, tecniche di clustering [20]. Sebbene il còrso abbia il vantaggio di essere una lingua italoromanza e quindi condivida con l'italiano un certo numero di parole funzionali, questo non è sufficiente. Per tale ragione occorre creare un elenco personalizzato che comprenda le variazioni diacroniche e dialettali. Tra queste, parole come "aghju" (io ho), che negli anni Venti poteva essere scritto "aghiu", o "per" (per), che si può trovare nella forma "pè" o "par". Le regole ortografiche del còrso presentano ancora una volta sfide interessanti. Si pone dunque il seguente interrogativo: è opportuno standardizzare la lingua per favorirne la diffusione e la comprensione oppure è preferibile preservarne la varietà? Oltre alla questione delle parole funzionali, dobbiamo considerare il caso degli hapax o poco frequenti, delle parole frequenti che non sono stop words, come "corsica", un termine impiegato di frequente data la natura stessa del nostro corpus. Una soluzione è quella di includere la nozione di entropia statistica nella scelta del vocabolario, come è stata presentata da Susan Dumais [19] attraverso la seguente formula:

$$E = 1 - \sum_j \frac{p_{ij} log(p_{ij})}{log(ndocs)} \text{ dato } p_{ij} = \frac{tf_{ij}}{gf_i}$$

In questa equazione, la variabile *ndocs* rappresenta il numero di documenti, *tf* è la frequenza del termine *i* nel documento *j* e *gf* è la frequenza globale del termine. L'idea è quella di calcolare l'entropia di ogni parola del corpus e di selezionare il vocabolario entro un intervallo definito. In teoria, con l'intervallo giusto, non sarebbe stato necessariamente indispensabile rimuovere le parole funzionali dal corpus, perché la loro entropia sarebbe stata molto più alta della media. Tuttavia, è stato preferibile eliminarle prima, in quanto ciò non ha influito sul punteggio degli altri termini. Un altro criterio per distinguere le parole importanti da quelle che non lo sono è il loro peso, che consiste nel moltiplicare l'entropia puntuale *E* di un termine *i* per la sua frequenza complessiva *gf*. Ciò significa che due termini possono avere la stessa entropia ma pesi radicalmente diversi.

Analisi e risultati

I risultati della nostra indagine sono frutto dell'impiego di una distant reading realizzata attraverso due metodologie che fanno capo al topic modeling: LDA e LSA. Abbiamo deciso di estrapolare sei topic per ciò che riguarda il corpus còrso e francese, mentre per quello italiano abbiamo isolato quattro topic. Infine abbiamo generato un corpus in lingua francese costituito unicamente dagli articoli relativi agli anni 1926-1927, nel corso dei quali il Partitu corsu d'azione si trasformò in Partitu corsu autonomista e avviò un dialogo con gli altri movimenti etno-regionalisti francesi.

Il numero di topic scelto dipende dal set di dati a nostra disposizione: il nostro corpus è abbastanza ampio ma tende a evocare temi più o meno simili. Dal momento che il corpus in lingua italiana è meno esteso, i risultati ottenuti applicando gli stessi parametri utilizzati per il corpus in lingua francese e còrsa forniscono risultati meno interessanti. Si è pertanto reso necessario modificare i parametri relativi al corpus in lingua italiana e in lingua francese (1926-1927). Questo ci induce a tenere conto della cosiddetta nozione di "bias di conferma" nel topic modeling menzionata da Stefano Sbalchiero e Eder Maciej [46]. Dal momento che il bias di conferma è quell'attitudine secondo la quale gli individui tendono a spiegare ogni cosa all'interno delle loro conoscenze pregresse e acquisite [35], occorre considerare altresì la variabile relativa al giudizio umano che si riflette nell'affidabilità dei risultati.

Se analizziamo il corpus còrso, attraverso la LDA, osserviamo una prevalenza di argomenti connessi alla sfera intima e familiare. Nel topic 1 emergono le espressioni "paese", "casa", "core" [cuore] ma anche "paoli" (topic 1, 5) in riferimento a Pasquale Paoli che è considerato il "padre" della nazione còrsa, noto anche come "Babbu di a Patria". Nel topic 2 rinveniamo vocaboli quali "mamma", "amore", "dolce", "focu" [fuoco], "tempu" [tempo], "notte", "fiore" ma anche "ziu" [zio] e "pipparellu". Quest'ultimo è il protagonista di una storia intitolata *Ziu Pipparellu* il cui autore si firma "U Patriotta"[11]. Non soltanto il topic 2 ma anche i topic 3, 4 e 5 sono riferiti alla letteratura dialettale. Nel topic 3 ritroviamo l'espressione "prusarpina" in riferimento a "Prusarpina-Pulitica, ossia una fiera di poesie còrse che si teneva nel 1928, nel topic 3 "merru" [sindaco], il protagonista della commedia in 4 atti del poeta còrso Maistrale dal titolo *A cumuna di Parapiglia*, o ancora nel topic 4 "corduella", località conosciuta anche con il nome Cordovella e luogo in cui è ambientata una leggenda pubblicata su *A Muvra* il 27 giugno 1926. Nei topic 2 e 5 troviamo riferimenti agli autori di *A Muvra*: nel primo caso "casinca" che indica l'antica pieve della Corsica situata nella parte nordorientale dell'isola ma che si riferisce anche ad un autore di *A Muvra*: Sambucucciu di Casinca; nel secondo "codaccioni", anch'egli autore di diversi contributi firmati G. P. Codaccioni. Si discosta un po' dagli altri topic, il topic 6 in cui emergono temi più politici come dimostrato dalla presenza di espressioni quali "lingua", "dialettu" [dialetto], "statu" [stato], "pulitica" [politica].

Se utilizziamo il metodo dell'LSA osserviamo risultati un po' differenti o quantomeno che presentano sfumature diverse. Benché la sfera intima emerga, come dimostrato dal topic 1 in cui ricorrono espressioni quali "casa", "paese", "core" [cuore], tuttavia non si colloca in una posizione centrale. Al contrario, rinveniamo un'attenzione particolare rivolta a Pasquale Paoli. Questo riferimento al generale della 'nazione còrsa" è associato ad una dimensione religiosa come è dimostrato dalle espressioni presenti nel topic 2 - "babbu" [babbo], "merusaglia" (la città natale di Paoli), "fede", "patria", gloria" - ma anche nel topic 6 dove troviamo termini quali "santa", "san", "generale". Nel topic 3 i vocaboli rimandano all'esperienza del governo guidato da Paoli, come dimostrato dalla presenza di "pasquale", "paoli", "babbu" [babbo], "autonomia", "guvernu" [governo]. Il topic 4 traccia una continuità nella rivendicazione culturale e politica còrsa, dal governo di Paoli al particolarismo linguistico rappresentato dalla rivista *A Tramuntana* di Santu Casanova e che trova in *A Muvra* il luogo in cui poter promuovere la lingua còrsa: difatti troviamo espressioni come "cursichella" (canzone antifrancese scritta da Petru Rocca), "sampetracciu", che era il titolo di una commedia in quattro atti *U Sampetracciu*, "casanova", ovverosia Santu Casanova, "lingua", "dialettu" [dialetto], "cirnu" che è l'antico nome della Corsica, "pontenovu" in riferimento alla celebre battaglia che avrebbe decretato la fine del governo paolino. La battaglia di Ponte Novu è al centro del topic 5 dove rinveniamo espressioni

[11] *A Muvra*, 08/06/1930.

quali "pontenovu", "ponte", "cumitatu" [comitato], "croce". Tra le iniziative promosse dal Partitu corsu d'azione vi fu infatti l'erezione di una croce commemorativa nei luoghi in cui l'armata di Paoli fu battuta dalle truppe francesi. Il 1° aprile 1923 sulle prime tre colonne di *A Muvra* venne pubblicato un appello intitolato *Ponte Novu! À tous les Corses!* per raccogliere il denaro necessario a erigere una croce a Ponte Novu. Il Leitmotiv di *A Muvra* era «Ponte Novu! Mancu una Croce!». Così si leggeva: "il 9 maggio 1769 le milizie di Pasquale de Paoli furono sconfitte a Pontenovu dall'esercito del generale de Vaux. Dopo una campagna micidiale segnata dall'indomito eroismo della nostra razza; dopo le vittorie còrse di Borgu! Barbaggiu, Furiani e Moncale, 35.000 mercenari con un centinaio di cannoni schiacciarono i nostri 8.000 miliziani. 2.000 montanari morirono tra il Nebbio e il Custere, più di 800, pressati nel ponte, furono massacrati lì dal fuoco penetrante di una formidabile fanteria e artiglieria"[12].

Erigendo una croce in memoria dei còrsi caduti durante la battaglia di Ponte Novu, i muvristi volevano mettere in evidenza il valore e la rilevanza della resistenza mostrata dalle truppe di Paoli e porre una linea di continuità con l'azione politica e culturale svolta dal Partitu corsu d'azione.

Fig. 5 - Topic LDA lingua còrsa

[12] *A Muvra*, 01/04/1923.

Fig. 6 - Topic LSA lingua còrsa

Il corpus in lingua italiana indagato attraverso i due metodi restituisce risultati piuttosto simili che si riferiscono al periodo della dominazione genovese e alla stagione paolina, come possiamo appurare dalla presenza delle espressioni "paoli" (LDA, topic 1, 2, 3, 4; LSA, topic 1, 2, 3), "pasquale" (LSA, topic 4), "genova" (LDA, topic 2, 4; LSA, topic 1, 2, 3), "genovesi" (LDA, topic 3), "liberta" (LDA, topic 2, 3, 4; LSA, topic 1, 4), "patria" (LDA, topic 1, 2, 3, 4; LSA, topic 1). Infine è presente nel nostro corpus il riferimento alla letteratura còrsa e in particolare alla *Storia Populare di Corsica illustrata*, come testimonia il vocabolo "altobello" che è il nome del protagonista di questa storia (LDA, topic 1, 2, 3; LSA, topic 1, 2, 4).

Fig. 7 - Topic LDA lingua italiana

Fig. 8 - Topic LSA lingua italiana

Il corpus francese processato attraverso la LDA e la LSA offre risultati abbastanza simili per quanto vi siano alcune sfumature che determinano esiti leggermente differenti. A prevalere in tutti i topic è la questione alsaziana [21]. Per comprendere l'attenzione riservata da *A Muvra* alla questione dell'Alsazia e della Lorena, occorre dunque ricostruire a grandi linee il quadro storico. Dopo l'epilogo della guerra franco-prussiana nel 1871 l'Alsazia e la parte nordorientale della Lorena furono inglobate nell'Impero tedesco. Tra la fine degli anni Settanta e gli inizi degli anni Ottanta dell'Ottocento nacque in Alsazia un partito autonomista che, pur mostrandosi leale nei confronti della Germania, chiedeva alcune forme di autonomia per l'Alsazia-Lorena. La Dieta dell'Impero convocata il 4 luglio 1879 stabilì per venire incontro alle richieste di questo partito l'istituzione di un governo regionale autonomo con un luogotenente, un ministero e una giunta regionale (Landesausschuss). Nel 1911 la nuova Costituzione estese i poteri del Landesausschuss e istituì una Camera alta e una Camera bassa che deteneva alcune prerogative legislative in materia locale. Quando, nel primo dopoguerra, l'Alsazia e la Lorena ritornarono alla Francia per effetto del Trattato di Parigi del 1919, la République mise in atto una politica di assimilazione con l'invio nella regione di funzionari provenienti da altre regioni francese e ponendosi l'obiettivo di imporre alla popolazione rurale la pratica di una lingua, quella francese, che non avevano mai parlato.

Gli amministratori francesi erano inconsapevoli di trovarsi di fronte a territori con una storia e una cultura segnata da secoli in cui vi erano stati contatti con il mondo renano e dall'appartenenza alla religione cattolica. Pertanto la lingua maggiormente praticata era un dialetto locale e il tedesco era assai diffuso presso la popolazione.

Sin da subito gli autonomisti richiesero il rispetto del particolare regime religioso, che era stato concordato da Napoleone I e dalla Chiesa, secondo il quale lo Stato avrebbe dovuto garantire il sostentamento del clero e il fatto che gli studenti delle scuole religiose ricevessero un'educazione

secondo i principi della propria confessione. A questo si aggiungevano richieste relative al diritto di praticare e ricevere l'insegnamento nella lingua tedesca [1].

Quando a salire al governo in Francia fu il Cartel des Gauches emersero nuove tensioni che conobbero il loro apice nel 1924. Fu in quel frangente che l'allora presidente del consiglio Edouard Herriot intese smantellare lo status confessionale delle scuole della regione. Le proteste furono tali che il governo fu costretto a ritirare il provvedimento. Proliferarono quotidiani autonomisti come *Die Zukunft* e nel novembre 1927 fu fondato il partito autonomista Landespartei per iniziativa di Karl Roos. Dall'analisi del nostro corpus francese emerge un'attenzione specifica a questo periodo storico. Il movimento autonomista alsaziano e il suo organo di stampa *Die Zukunft* furono allora sospettati di agire di concerto con gli irredentisti d'oltre Reno per elaborare un piano teso alla separazione dalla Francia. Dal 1° al 24 maggio 1928 un numero considerevole di autonomisti furono accusati di complotto contro la sicurezza dello Stato e furono processati a Colmar. Tuttavia, in occasione delle elezioni generali, che si tennero dal 22 al 29 maggio 1928, gli autonomisti in Alsazia ottennero un ottimo risultato elettorale. Ad essere eletti furono proprio due di coloro che erano stati tratti in arresto per il complotto, Eugène Ricklin e Charles Philippe Roos. Benché fossero stati assolti, entrambi si videro privati del loro mandato parlamentare. Nel nostro corpus spiccano le espressioni "ricklin" (LDA, topic 1; LSA, topic 1, 2, 3, 5), "zukunft" (LSA, topic 2,5, 6), "schall" (LDA, topic 1; LSA, topic 2, 3, 5, 6), "complot" [complotto] (LDA, topic 1; LSA, topic 1, 2, 6). Ciò dimostra un'attenzione da parte di *A Muvra* agli sviluppi che stava assumendo la questione alsaziana. Consigliere generale, deputato al Landesausschuss e al Reichstag, presidente del Landtag Alsazia-Lorena (1912-1918) e deputato autonomista di Altkirch (1928), Eugène Ricklin rivestì un ruolo importante nell'ambito dell'autonomismo alsaziano anche grazie alla sua veste di direttore del giornale *Die Zukunft*. Allo stesso modo Paul Schall, caporedattore de *Die Zukunft* , nonché uno degli imputati al processo di Colmar, fu una figura di rilievo del movimento alsaziano. A Schall si deve la creazione del Comité des minorités nationales de France, un raggruppamento che riuniva rappresentanti catalani, bretoni, fiamminghi e còrsi. Il 12 settembre 1927 a Quimper si svolse il Congrès national des minorités nationales de France, a cui partecipò oltre a Schall anche Petru Rocca in rappresentanza del Partitu corsu autonomista.

Un altro tema che emerge in maniera preponderante nel corpus francese è quello connesso all'Affaire Berthon. *A Muvra* dedicò ampia eco alla vicenda di André Berthon, avvocato e deputato alsaziano che aveva difeso gli autonomisti al processo di Colmar [13], come dimostra la ricorrenza dell'espressione "berthon" (LDA, topic 1, 4; LSA, topic 1, 2, 3, 5). Malgrado l'ideale vicinanza con le idee espresse da Berthon, *A Muvra* ne stigmatizzò il comportamento perché questi in un'udienza che riguardava un divorzio in cui era coinvolto un còrso, il dottor Alessandri, aveva basato la difesa della sua assistita mettendo in rilievo "il carattere còrso, il suo temperamento di Corso e i costumi corsi"[14] come a denigrare la cultura còrsa.

Il metodo LSA sembra garantire una maggiore coerenza all'interno di ciascun topic. A tal proposito è indicativo il topic 4 che fa riferimento alla questione del trasporto marittimo tra

[13] BERTHON Pierre Marie André. Accessed July 8, 2023. https://www.alsace-histoire.org/netdba/berthon-pierre-marie-andre/

[14] *A Muvra*, 27/10/1929. L'articolo in questione si intitolava "L'Avucatu Berthon, difensore di a Minuranza Alsaziana, parte in guerra contru a Minuranza Corsa. Toute la Lumière sur l'Affaire Berthon" firmato da Pierre Alessandri, professore di liceo a Nizza.

Corsica, Italia e Francia[15], come è testimoniato dalle espressioni "services" [servizi], "tarifs" [tariffe], "lignes" [linee], "maritimes" [marittime], "convention" [convenzione], "voyageurs" [viaggiatori]. Lo stesso argomento è rinvenibile anche nel topic 2 indagato attraverso la LDA ma è, per così dire, reso meno riconoscibile perché associato ad altre espressioni: nello stesso topic troviamo infatti "services", "tarifs", "lignes", "maritimes" ma anche e, in misura maggiore, vocaboli quali "ecrivains" [scrittori], "siecle" [secolo], "ile" [isola].

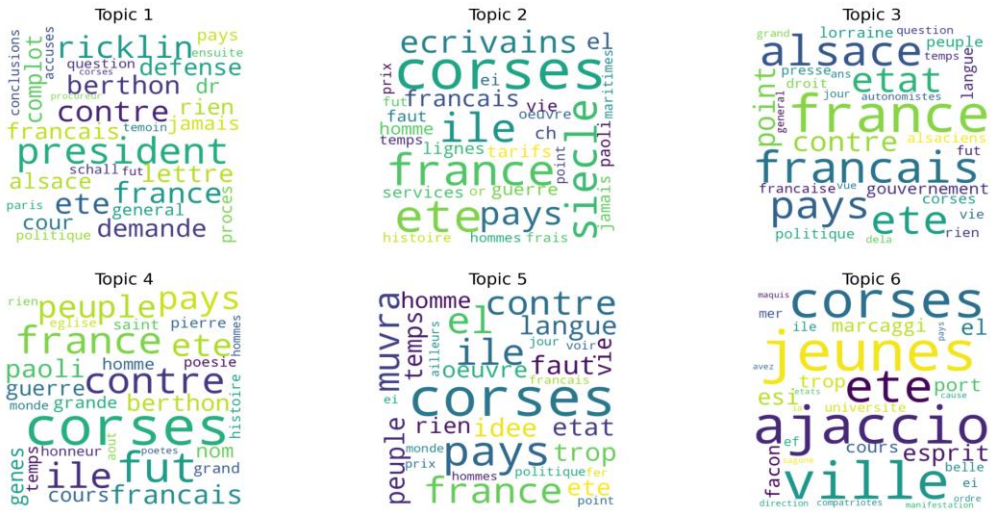

Fig. 9 - Topic LDA lingua francese

[15] *A Muvra*, 14/03/1926.

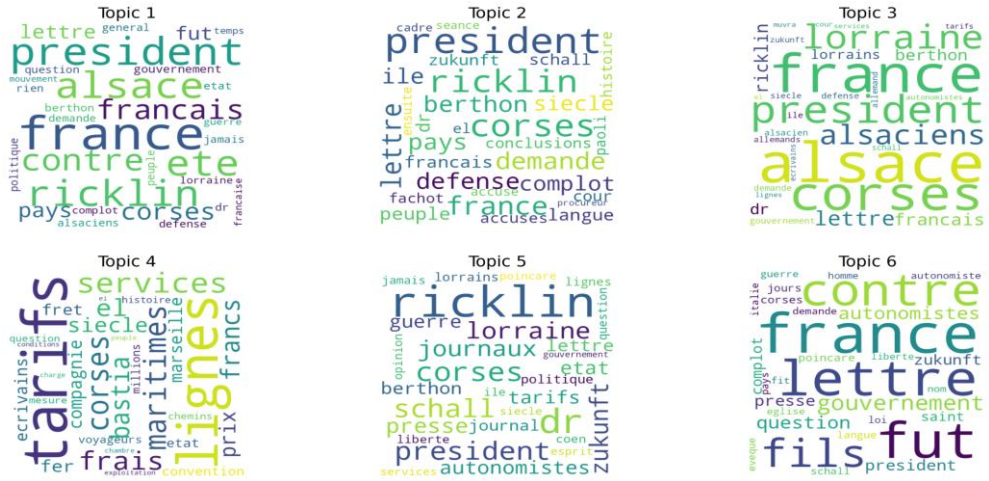

Fig. 10 - Topic LSA lingua francese

Se confrontiamo i due metodi, osserveremo che la LSA coglie maggiormente la dimensione politica relativa alla questione alsaziana. Questo è ancor più evidente se isoliamo all'interno del nostro corpus francese gli articoli pubblicati nel biennio 1926-1927.

Il nostro corpus indagato attraverso la LDA esprime un'attenzione specifica all'autonomia come esito della volontà del popolo còrso: ciò è dimostrato dalla presenza di espressioni quali "autonomie" [autonomia] (topic 1, 3), "interets" [interessi] (topic 1), "esprit" [spirito] (topic 2). Al contrario lo stesso corpus analizzato attraverso la LSA restituisce un'immagine diversa tutta consacrata alla questione alsaziana e all'affaire Berthon: ritroviamo dunque espressioni quali "richlin" (topic 1, 2, 3), "alsace" [Alsazia] (topic 1, 2, 3), "berthon" (topic 1, 2), "zukunft" (topic 1, 3), "schall" (topic 3).

Fig. 11 - Topic LDA lingua francese (1926-1927)

Fig. 11 - Topic LSA lingua francese (1926-1927)

Note conclusive

L'esame della stampa autonomista còrsa attraverso il ricorso al topic modeling ci ha consentito di individuare i temi generali affrontati dai muvristi all'interno dell'organo di stampa del Partitu corsu d'azione: *A Muvra*. La natura stessa di questa pubblicazione ha orientato i risultati ottenuti, come è dimostrato dalla frequenza del termine "corsica". Gli argomenti che emergono sono inoltre influenzati dalla tipologia degli articoli. Questo spiega perché gli argomenti in italiano tendono a concentrarsi maggiormente su temi culturali o storici. Lo studio di *A Muvra* attraverso un approccio multilingue ha altresì dimostrato come gli autori degli articoli avessero optato per l'impiego dell'idioma còrso quando si trattava di affrontare argomenti afferenti alla sfera dialettale e familiare. L'utilizzo delle lingue francese e italiana era dovuto, al contrario, alla volontà di trattare questioni di natura più prettamente politica. I temi modellati esprimono infatti fortemente la dimensione politica del periodico, benché i muvristi ne evidenziassero anche l'aspetto culturale. Sono ricorrenti i riferimenti alla questione alsaziana e, più in particolare, al processo di Colmar, tema che stava molto a cuore a Petru Rocca.

Da un punto di vista metodologico la nostra indagine ha inteso proporsi come un primo lavoro pionieristico su un corpus in lingua còrsa. Tuttavia, vi sono ancora margini di miglioramento, in particolare per quanto riguarda la qualità dell'OCR, essenziale per ottenere risultati soddisfacenti. Sarebbe infatti utile la creazione di un modello specifico per l'OCR relativo a corpora storici scritti in lingua còrsa, che fornisca un metodo efficace per valutare i risultati ottenuti. Pertanto questo studio evidenzia le prospettive di miglioramento nonché i limiti della metodologia impiegata. Il metodo di valutazione dei modelli addestrati ci consente infatti di perfezionarli, al fine di ottenere soggetti modellati migliori. In linea generale, effettuando una "extrinsic evaluation" [valutazione estrinseca] dei risultati utilizzando i bias di conferma, le parole chiave generate rimangono particolarmente rilevanti. Tuttavia, osserviamo come abbiano fatto la loro ricomparsa alcuni vocaboli come "qu" in francese, che è una forma contratta di "que" [che]. Pertanto un migliore pre-trattamento dei dati prima dell'analisi ci permetterebbe in prospettiva di perfezionare i nostri risultati. Allo stesso modo, sarebbe particolarmente rilevante modificare le scale di analisi, ad esempio prendendo in considerazione soltanto alcuni anni. Si è tentato di farlo con questo corpus focalizzando l'attenzione sugli anni 1926-1927, nel corso dei quali, come abbiamo visto, il Partitu corsu d'azione avviò una trasformazione interna al partito sposando in toto la causa autonomista. Tuttavia abbiamo osservato come i risultati non siano stati pienamente

soddisfacenti. A lungo termine, l'ampliamento del corpus consentirebbe di entrare più nel dettaglio per comprendere i cambiamenti temporali dei temi trattati. Fatte queste premesse metodologiche possiamo osservare come l'esplorazione della stampa autonomista còrsa tra le due guerre attraverso il topic modeling si sia rivelata promettente dal momento che i risultati ottenuti hanno presentato un avanzamento nella conoscenza e nell'interpretazione della retorica corsista.

Bibliografia

[1] Bastianelli, Rodolfo. 2015. "La questione dell'Alsazia e della Lorena," *Rivista di Studi Politici Internazionali* 82, no. 4.

[2] Beri, Emiliano. 2011. Genova e il suo regno. Ordinamenti militari, poteri locali e controllo del territorio in Corsica fra insurrezioni e guerre civili (1729–1768). Novi Ligure: Città del silenzio edizioni.

[3] Blei, David, Ng, Andrew, and Jordan, Michael. 2003. "Latent dirichlet allocation," *Journal of machine Learning research* 3: 993-1022.

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

[5] Brauer, René, and Mats Fridlund. 2013. "Historicizing topic models, a distant reading of topic modeling texts within historical studies." In *Cultural Research in the Context of "Digital Humanities": Proceedings of International Conference 3-5 October 2013, St Petersburg*, edited by Nikiforova, Larisa V. and Natasha V., 152-163. St. Petersburg: Russian State Herzen University.

[6] Camps, Jean-Baptiste, and Thibault Clérice. 2019. *IIIF-Crawler*. Accessed July 8, 2023. https://github.com/Jean-Baptiste-Camps/IIIF-Crawler.

[7] Chaubin, Hélène. 2005. *Corse des années de guerre 1939 -1945*. Paris: Éditions Tirésias.

[8] Cini, Marco. 2003. Une île entre Paris et Florence: culture et politique de l'élite corse dans la première moitié du 19e siècle. Ajaccio: Albiana.

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

[10] Cini. Marco, and Ange Rovere. 2001. "Pascal Paoli de l'Histoire aux mythes," *Panoramique(s)* 53: 90-101.

[11] Cini, Marco, and Francis Beretti. 1998. *La nascita di un mito: Pasquale Paoli tra '700 e '800*. Pisa: BFS.

[12] Cuzzi, Marco. 2007. "La rivendicazione fascista della Corsica (1938-1943)," *Recherches Régionales – Alpes – Maritimes et Contrées limitrophes* 48, no. 3: 57-71.

[13] Cvitanic, Toni, Bumsoo Lee, Hyeon Ik Song, Katherine Fu, and David Rosen. 2016. "LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents." In *International Conference on Case-Based Reasoning*. Accessed July 8, 2023. https://par.nsf.gov/servlets/purl/10055536.

[14] Dal Passo, Fabrizio. 2007. Il Mediterraneo dei lumi. Corsica e democrazia nella stagione delle rivoluzioni. Napoli: Bibliopolis.

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

[16] Dalbera-Stefanaggi, Marie José. 2000. *Essais de linguistique corse*. Ajaccio: Alain Piazzola.

[17] Deerwester, Scott, Dumais, Susan, Furnas, Georges, Landauer, Thomas, and Harshman, Richard. 1990. "Indexing by latent semantic analysis," *Journal of the American society for information science* 41, no. 6: 391-407.

[18] Dottelonde, Pierre. 1984. *Histoire de la revendication corse 1959-1974: du département français à la nation corse*. Tesi di dottorato. Paris: Institut d'Études Politiques de Paris.

[19] Dumais, Susan. 1992. "Enhancing performance in latent semantic indexing (LSI) retrieval," *Technical Report TM-ARH-017527*, Bellcore: Morristown.

[20] Fayaza, Faathima, and Fathima F. Farhath. 2021. "Towards stop words identification in Tamil text clustering," *International Journal of Advanced Computer Science and Applications* 12, no. 12: 524–259.

[21] Fischer, Christopher J. 2014. Alsace to the Alsatians?: Visions and Divisions of Alsatian Regionalism, 1870-1939. New York: Berghahn Books.

[22] Giglioli, Alessandra. 2001. Italia e Francia 1936-1939. Irredentismo e ultranazionalismo nella politica estera di Mussolini. Roma: Jouvence.

[23] Graziani, Antoine Marie. 1997. La Corse génoise: économie, société, culture; période moderne 1453–1768. Ajaccio: Piazzola.

[24] Gregori, Sylvain. 2014. "(Ré)écrire l'histoire de la Résistance corse: de l'enjeu mémoriel à l'essai historiographique." In *Chercheurs en Résistance: Pistes et outils à l'usage des historiens*, edited by Blanc, Julien, and Cécile Vast. Rennes: Presses universitaires de Rennes. Accessed July 7, 2023. https://books.openedition.org/pur/49018

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

[26] Kalepalli, Yaswanth, Shaik Tasneem, Pasupuleti Durga Phani Teja, and Suneetha Manne. 2020. "Effective Comparison of LDA with LSA for Topic Modelling." In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 1245-1250, Madurai.

[27] Kevers, Laurent, Florian Guéniot, Aurelia Ghjacumina Tognotti, and Stella Retali-Medori. 2019. "Outiller une langue peu dotée grâce au TALN: l'exemple du corse et BDLC." In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN) PFIA 2019*, vol. 2, 371–380, Toulouse: ATALA.

[28] Klinkenberg, Jean-Marie. 2016. "Grandes langues » et langues minoritaires : deux politiques linguistiques ?," *Lengas. Revue de sociolinguistique* 79. Accessed July 8, 2023. https://journals.openedition.org/lengas/1048.

[29] Leca, Antoine. 1992."A Muvra ou le procès de la France par les autonomistes corses (1920-1939)." In *Actes du colloque de Toulouse 11-12-13 avril 1991. État et pouvoir. L'idée européenne*, 327-350. Aix-Marseille: Presses Universitaires d'Aix-Marseille.

[30] Leca, Antoine. 1993. "A Muvra ou l'autonomisme corse de la réhabilitation de l'Italie à la tentation fasciste (1920-1939)." In *Actes du colloque de Nice. 17-18-19 septembre1992, État et pouvoir (II)*, 405-430. Aix-Marseille: Presses Universitaires d'Aix-Marseille.

[31] Lepeltier, Marie-Claude. 2014. "«A Muvra», 1920-1939: la caricatura in Corsica," *Diacronie. Studi di storia contemporanea* 17, no. 1: 1-21.

[32] Minello, Giorgia, and Deborah Paci. 2022. "A benchmark corpus for topic modeling on the origins of modern antisemitism," *Umanistica Digitale* 13: 117-151.

[33] Montemurro, Marcelo A. 2001. "Beyond the Zipf–Mandelbrot law in quantitative linguistics," *Physica A: Statistical Mechanics and its Applications* 300, no. 3-4: 567-578.

[34] Moretti, Franco. 2000. "Conjectures on world literature," *New left review* 1.

[35] Nickerson, Raymond S. 1998. "Confirmation Bias: A Ubiquitous Phenomenon in Many Guises," *Review of General Psychology* 2, no. 2: 175–220.

[36] Paci, Deborah. 2012."Le dialogue des élites méditerranéennes au XIXe siècle. Le cas de Malte et de la Corse," *Cahiers de la Méditerranée* 42: 20-26.

[37] Paci, Deborah. 2015. Corsica fatal, Malta baluardo di romanità. L'irredentismo fascista nel mare nostrum (1922-1942). Firenze: Le Monnier-Mondadori Education.

[38] Paci, Deborah. 2023. "'Je suis corse, un homme de village': Towards a study of contemporary Corsican nationalism (1959-1998)," *History: The Journal of the Historical Association* 108, no. 383: 556-580.

[39] Pellegrinetti, Jean-Paul, and Ange Rovere. 2004. La Corse et la République. La vie politique de la fin du second Empire au début du XXIe siècle. Paris: Éditions du Seuil.

[40] Pellegrinetti, Jean-Paul. 2003. "Langue et identité: l'exemple du corse durant la troisième république," *Cahiers de la Méditerranée* 66: 265-277.

[41] Poli, Jean-Pierre. 2007. Autonomistes corses et irrédentisme fasciste, 1920-1939. Ajaccio: Éditions DCL.

[42] Pomponi, Francis (ed.). 1979. *Le Mémorial des Corses, L'Île éprouvée: 1914-1945*, vol. IV. Ajaccio: Le Mémorial des Corses.

[43] Rogé, Ysée. 2008. Le corsisme et l'irrédentisme 1920-1946: histoire du premier mouvement autonomiste corse et de sa compromission par l'Italie fasciste, Tesi di dottorato. Paris: Université de Paris Nanterre.

[44] Roux, Christophe. 2014. Corse française et Sardaigne italienne. Fragments périphériques de construction nationale. Paris: L'Harmattan.

[45] Sarbach-Pulicani, Vincent. 2023. "Profiling Anonymous Authors in the Corsican Autonomist Press of the Interwar Period." In *Proceedings of the 4th Computational Humanities Research Conference 2023*, edited by Šeļa, A., Jannidis, F. and Romanowska, I., 78-99. Paris.

[46] Sbalchiero, Stefano, and Maciej Eder. 2020. "Topic modeling, long texts and the best number of topics. Some Problems and solutions," *Quality & Quantity* 54: 1095–1108.

[47] Wilson, Stephen. 1995. Vendetta et banditisme en Corse au XIXe siècle. Ajaccio: Albiana.

[48] Yujian, Li, and Liu Bo. 2007. "A normalized Levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence* 29, no. 6: 1091-1095.

Annessi

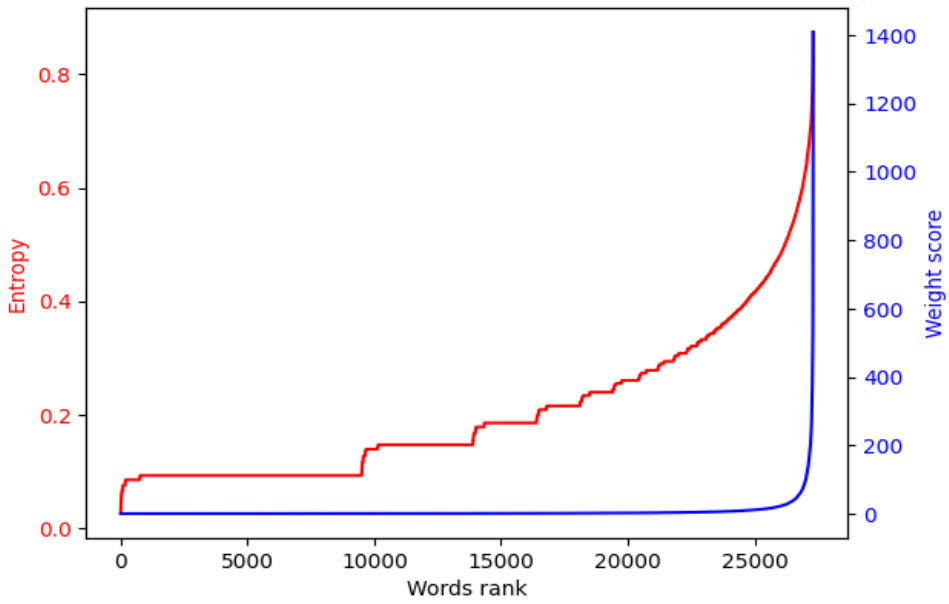

Fig. 12 - Entropia e weight score del vocabolario nel corpus in lingua còrsa

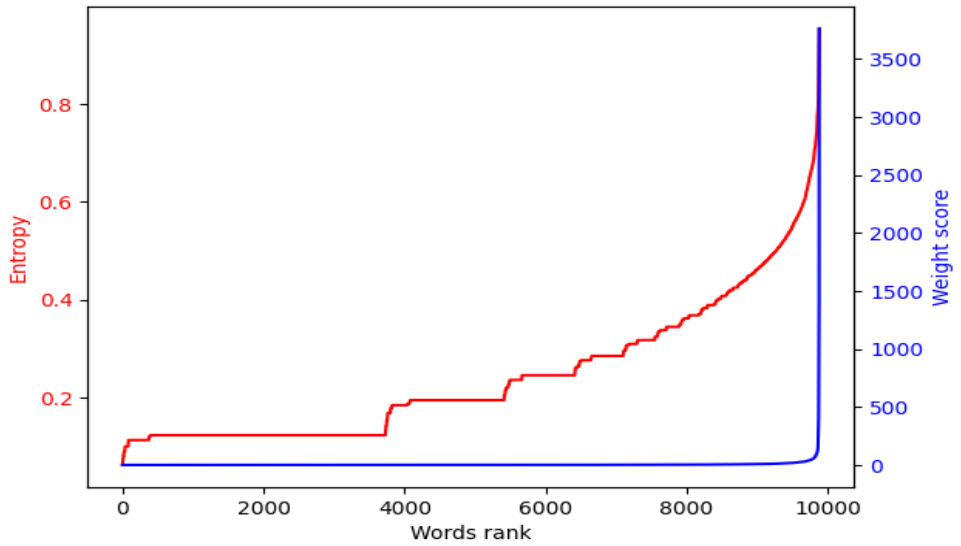

Fig. 13 - Entropia e weight score del vocabolario nel corpus in lingua francese

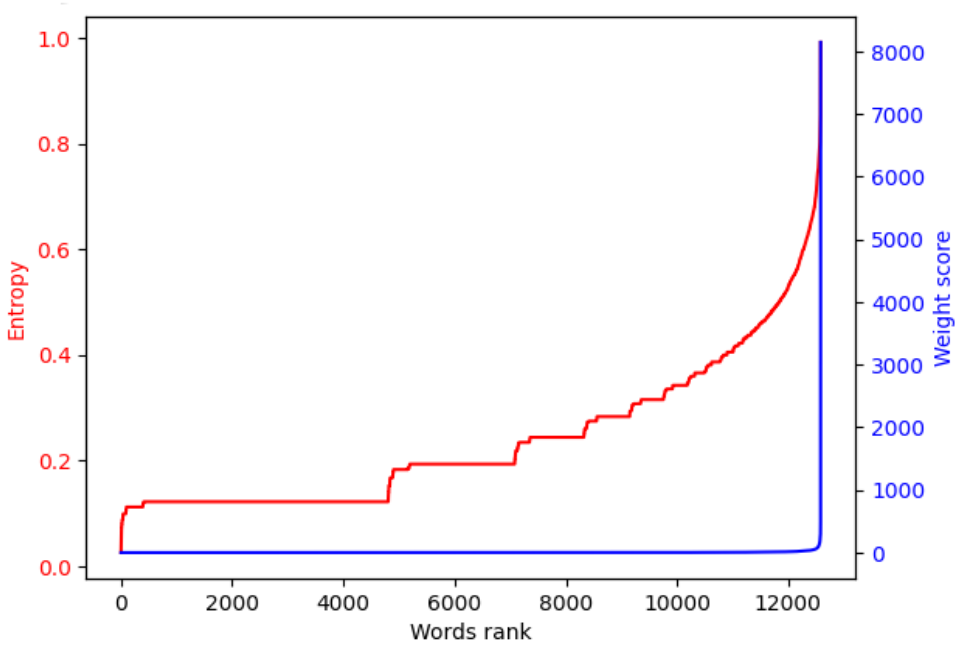

Fig. 14 - Entropia e weight score del vocabolario nel corpus in lingua italiana

li	com	le	e	u	A
la	ancu	unn	ellu	elli	iddu
cu	ju	ja	cume	ella	elle
quellu	quelle	quelli	quella	so	aghju
aghiu	he	simu	avemu	site	hannu
hanu	pe	par	seraghiu	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

Tab. 2 - Lista delle parole funzionali