Il progetto *ArPeR*. Per un *Archivio dei periodici romaneschi*

Martina Ludovisi
[1] Università di Perugia, Italia, martina.ludovisi@unipg.it

ABSTRACT (ITALIANO)

Scopo del progetto di cui qui si riferisce è realizzare, in un ambiente web integrato e open access, un archivio dei periodici dialettali romani pubblicati tra l'Unità e la Prima Guerra Mondiale (1870-1920), in forma digitale e liberamente consultabile in linea. Non ci si limita però alla sola operazione di digitalizzazione: la possibilità di interrogare il *corpus* secondo diversi parametri (per forme, autori, componimenti, riviste ecc.), obiettivo primario del progetto, consentirà sia di riportare alla luce figure oggi neglette, contribuendo ad ampliare l'attuale conoscenza del panorama dialettale capitolino di fine Ottocento-primo Novecento, sia di porre le basi per ricerche future. Il *corpus* è digitalizzato mediante procedure di OCR e immesso in una piattaforma realizzata per il progetto. I principali risultati attesi sono: costituire una Sala di lettura open access che contenga le riproduzioni digitali dei periodici scelti per la costituzione del *corpus*; realizzare una biblioteca digitale che consenta di interrogare l'intero *database* tramite una ricerca personalizzata; offrire una prima descrizione, specie lessicografica, del romanesco postunitario.
Parole chiave: Periodici romaneschi; Archivio digitale; digitalizzazione; romanesco.

ABSTRACT (ENGLISH)

Paper Title for AIUCD2025. The *ArPeR* Project. For an Archive of Romanesco Periodicals

This paper presents the purposes of the SNSF-Project ArPeR.Archive of Romanesco Periodicals. The project seeks to establish a digital archive of Roman dialect periodicals published between the Unification of Italy and World War I (1870–1920), which will be freely accessible online. By digitizing these dialect texts, the project makes this important material available to the public for the first time. It also provides an opportunity to highlight previously overlooked figures, contributing to a richer understanding of the Roman dialect landscape in the late 19th and early 20th centuries, while laying a foundation for future research. The corpus is being digitized using OCR procedures and entered into a platform created specifically for the project. The main expected results are: the creation of an open-access Reading Room containing digital reproductions of the periodicals selected for the corpus; the creation of a digital library allowing users to query the entire database through personalized searches; and the provision of an initial description, particularly lexicographic, of post-unification Romanesco.

Keywords: Romanesco Periodicals; Digital Archive; Digitization; Romanesco

1. INTRODUZIONE

Fornite le linee generali del progetto, si potrebbe subito osservare che una riproduzione fotografica di alcuni numeri di testate giornalistiche dialettali è già prodotta da piattaforme come Internet culturale o dalle emeroteche digitali delle singole biblioteche; ma i loro archivi, oltre a riprodurre solo pochi fascicoli, non sono agilmente interrogabili, e non consentono alla comunità scientifica di sfruttare appieno il materiale che conservano. L'impossibilità di indagare agilmente tale patrimonio, considerato il cospicuo numero di fascicoli dei diversi giornali pubblicati in quegli anni, è un fattore che scoraggia la ricerca.[1]
Questa lacuna tuttavia può essere colmata: grazie alle attuali risorse informatiche è possibile offrire al fruitore un archivio digitale con un *corpus* testuale praticamente inedito, finora solo parzialmente indagato (cfr. il §2), e facilmente interrogabile dall'utente (cfr. il §3): il progetto ArPeR si muove infatti su un doppio binario: oltre a digitalizzare un materiale di valore storico-linguistico-letterario, si propone di fornire all'utente una lista di testi e relativi autori che, come vedremo (vd. il §4), ha già prodotto l'acquisizione di nuovi dati e l'aggiornamento della bibliografia primaria di alcuni scrittori.

[1] All'interno del progetto è prevista comunque l'acquisizione del materiale già digitalizzato e disponibile in rete, come nel caso della rivista «Fornarina», i cui 21 numeri dell'anno 1883 sono liberamente accessibili (vd. https://books.google.it/books?id=bCklnr0C6M0C&printsec=frontcover&hl=it&source=gbs_ge_summary_r&cad=0#v=onepage&q&f=false). Uno degli obiettivi, infatti, è quello di stabilire connessioni con risorse esterne, già digitalizzate e disponibili online, così da valorizzare e riutilizzare il materiale già esistente.

2. STATO DELLA QUESTIONE

Le prime ricerche filologiche e linguistiche sul dialetto dell'Urbe risalgono al periodo compreso tra la fine del secolo XIX e l'inizio del XX, quando studiosi quali Ernest Bovet (1898), Tito Morino (1899), Giuseppe Senes (1900) e i più celebri Francesco Sabatini (1890) e Graziadio Isaia Ascoli (1900) si interessarono al romanesco coevo e alla lingua del Belli, attenzione suscitata dalla pubblicazione dell'edizione Morandi (1886-1889) dei *Sonetti*. Posto l'interesse per il romanesco belliano, i successivi lavori scientifici, come quelli di Merlo (1929) e Migliorini (1932),[2] si concentrarono quasi esclusivamente sul romanesco di prima fase, studi che col tempo non hanno fatto altro che moltiplicarsi, complici il dibattito su tempi e modi della toscanizzazione del dialetto capitolino, che ha interessato diversi studiosi (una sintesi si trova in Trifone, 2008) e il rinnovato interesse per le ricerche d'archivio. Dunque, a fronte di un quadro assai mutato per il romanesco di prima fase, lo studio di quello ottocentesco è invece ancora oggi fin troppo incentrato sulla figura del Belli e sul dialetto documentato dai *Sonetti*, benché le vicende postunitarie abbiano avuto importanti ripercussioni sulla varietà capitolina, che proprio in questi decenni comincia a conoscere una serie di trasformazioni strutturali a tutti i livelli di analisi (cfr. *infra* e in particolare Faraoni, 2021). Data la centralità di questo periodo storico anche da un punto di vista linguistico, è evidente che per approfondire la conoscenza del dialetto postunitario la soluzione più efficace è quella di allargare la base documentaria: l'universo delle riviste dialettali dell'Otto e del Novecento è certo un buon punto di partenza. Del resto, già il lavoro di Picchiorri (2019: 478) sui testi romaneschi contenuti nei giornali nati tra il 1848 e il 1849 ha dimostrato come questi «merit[i]no attenzione dal punto di vista linguistico come spaccato realistico [...] del romanesco di metà Ottocento, soprattutto perché presentano un ricco patrimonio lessicale, spesso scarsamente o niente affatto documentato dai repertori, e testimon[i]no l'emersione di alcuni tratti morfosintattici incipienti nel dialetto dell'epoca». Simili risultati sono stati prodotti esaminando solo pochi fogli scritti interamente in dialetto, giacché la stampa periodica romanesca avrà il suo periodo di maggiore splendore solo più tardi, proprio in quei decenni postunitari che il progetto mira a indagare; è certo, quindi, che dai giornali di quest'epoca si possano ricavare dati utili per la ricerca scientifica. D'altra parte, anche Faraoni (2021) ha evidenziato l'esistenza di due direttrici per l'evoluzione del "romanesco postunitario": una che muove in direzione dell'italiano, con il progressivo avvicinamento del dialetto alla lingua standard, un'altra che muove in direzione anti-italiana e che ha consentito la sopravvivenza del romanesco. Di quest'ultima direttrice, che Stefinlongo (1985, 26) identifica con la meridionalizzazione otto-novecentesca, il materiale che si intende raccogliere tramite il presente progetto è testimone d'eccezione, specie dal punto di vista lessicale, con l'intrusione da sud di quel «diuturno flusso di napoletanismi che segna tutta la storia preunitaria e [in particolare] postunitaria del romanesco» (De Mauro, 1989: XXVII). Malgrado ciò, la visione che si ha della varietà capitolina tra Belli (metà Ottocento) e Trilussa (particolarmente attivo nei primi tre decenni del Novecento) è allo stato attuale piuttosto lacunosa: non molti gli studi sistematici sugli autori dialettali dell'ultimo trentennio del sec. XIX, le cui opere risentono spesso del modello belliano che, in parte, àncora il romanesco a una fase pregressa (ne sono un fulgido esempio i debiti belliani contratti da Zanazzo, su cui cfr. Costa, 2011: 35). Quest'ombra però non si allunga sulla pubblicistica dialettale, dove la scrittura in prosa risulta, com'è ovvio, più spontanea e più ricca di nuove voci importate da sud o dalle varietà contermini, dai flussi migratori d'epoca postunitaria e perciò non registrate da Belli. Sul versante storico-letterario, la pubblicistica d'età umbertina interessa per le tematiche affrontate nei diversi periodici e per le relative posizioni politiche: Puglisi (2011), ad esempio, per quanto riguarda il «Rugantino», segnala quelle collegate alla polemica anticlericale, all'impegno politico di Crispi, all'erezione del monumento a Giordano Bruno e, più in generale, ai fatti di cronaca. Un'attenta disamina dei materiali contenuti nei giornali costituirà quindi anche un ausilio per lo studio della situazione politico-culturale e sociale della Roma postunitaria.

La mancanza di uno strumento di consultazione che consenta di estrapolare dai testi giornalistici romaneschi di fine Ottocento informazioni mirate e allo stesso tempo permetta la visione del suddetto materiale è stata alla base dell'ideazione di *ArPeR*.

3. METODOLOGIE E ARCHITETTURA DELL'OPERA

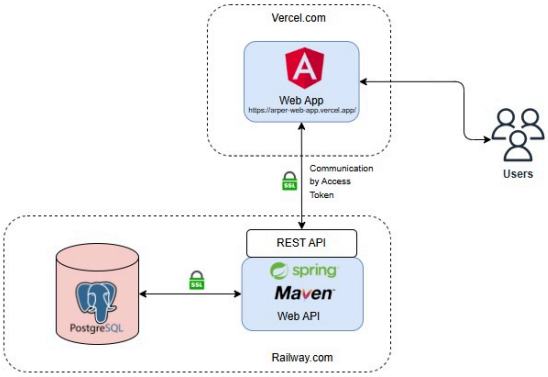
Per iniziare, è utile esaminare la struttura dell'opera. Il *corpus* è costituito dai fascicoli di alcuni dei più importanti giornali dialettali d'epoca postunitaria, elencati qui di séguito in ordine cronologico: 50 fascicoli di «Don Pirloncino» (30 lug. 1871 - 2 gen. 1887), 279 di «Capitan Fracassa» (25 mag. 1880 – 10 ott. 1905), 35 di «Fornarina» (7 gen. 1882 - 3 giu. 1883), 654 fascicoli del «Rugantino» (18 set. 1887 - 28

[2] È più tardo il fondamentale lavoro di Ernst, 1970.

dic. 1915), 20 di «Casandrino in dialetto romanesco» (8 agosto – 14 ottobre 1897) e 299 di «La frusta» (gen. 1873 - 31 dic. 1873).[3] Considerata la consistenza documentaria, al momento della selezione si sono privilegiati i periodici conteneti un numero maggiore di testi in dialetto; nonostante ciò, il *database* è destinato a un continuo ampliamento e, in una fase successiva, sarà possibile aggiungere testi e riproduzioni di altri giornali. Una volta definite le specifiche tecniche per i formati dei file digitali – selezionate sulla base di standard internazionali – il progetto si compone di tre fasi operative principali:

1) Strutturazione del lavoro: inserimento dei dati in un apposito file Excel per facilitare il processo di organizzazione dei numeri; identificazione dei problemi di indicizzazione (file mancanti); individuazione dei numeri assenti nelle biblioteche.
2) Acquisizione del materiale: è la fase in cui le varie testate vengono scannerizzate. Successivamente, si valuta la qualità delle immagini digitali e, se non conformi alle specifiche approvate, si procede con una nuova acquisizione.
3) Trascrizione (e correzione) dei testi del *corpus*: il materiale digitalizzato viene trascritto tramite metodologie e tecnologie di OCR open source e commerciali (come Tesseract,[4] ABBYY FineReader[5] ed eScriptorium[6]) scegliendo il formato plain text, che permette di preparare i testi a una successiva fase di revisione. Tra le soluzioni impiegate, ABBYY FineReader, il più impiegato, è stato oggetto di un addestramento mirato, basato sulla correzione manuale di un campione selezionato di pagine particolarmente rappresentative del *corpus*. Questo processo ha consentito al software di apprendere le specificità grafiche e tipografiche dei periodici romaneschi ottocenteschi, migliorando progressivamente la precisione del riconoscimento. L'addestramento ha riguardato in particolare l'identificazione di caratteri irregolari, l'interpretazione di impaginazioni non standard e la gestione delle forme dialettali che, in assenza di norme ortografiche condivise, si presentano spesso con elevata variabilità.
4) Inserimento: i documenti in formato testo vengono salvati nella piattaforma creata per il progetto, un database relazionale PostgreSQL, insieme alla relativa riproduzione fotografica.[7]

I testi delle riviste, correttamente trascritti e accompagnati dal documento riprodotto nella sua veste originaria, sono resi accessibili e interrogabili dagli utenti mediante un'interfaccia web.
L'applicazione web è composta da due fondamentali componenti: un'interfaccia utente implementata utilizzando Angular (front-end) e un'API web (back-end) costruita tramite il framework Spring Boot. L'interazione fra i due sistemi avviene in modalità sicura, attraverso l'uso di un meccanismo di autenticazione basato su token di accesso (login dell'utente, limitato al pannello di controllo), che assicura l'integrità e la privacy dei dati scambiati. Dal punto di vista della distribuzione, l'applicazione front-end è ospitata sulla piattaforma Vercel, mentre l'API back-end è distribuita tramite Railway: si tratta, per entrambe, di soluzioni cloud che offrono infrastrutture scalabili e ad alte prestazioni; una scelta che permette da un lato di ottimizzare la gestione delle risorse, dall'altro di garantire una risposta efficiente alle richieste dei fruitori. Di séguito, un grafico che illustra i rapporti descritti:



[3] Si tratta di numeri approssimativi, giacché si annoverano all'interno del *corpus* esclusivamente i fascicoli che contengono testi dialettali. Per il «Rugantino» e il «Casandrino» vuol dire considerarli tutti, essendo periodici redatti esclusivamente in dialetto; diversamente, per «Fornarina» ci sono state esclusioni.
[4] https://github.com/tesseract-ocr/tesseract
[5] https://pdf.abbyy.com/it/
[6] https://gitlab.com/scripta/escriptorium
[7] Attraverso Apache PDFBox (https://pdfbox.apache.org/), libreria java open source, è inoltre possibile inserire nella piattaforma ArPeR l'eventuale pdf a nostra disposizione ed estrarre, visualizzare e correggere il testo.

Figura 1. Grafico dei rapporti

4. PRIMI RISULTATI E PROSPETTIVE ATTESE

La piattaforma per l'interrogazione della banca dati testuale è tuttora in fase di controllo e di implementazione; di séguito un esempio del prototipo (figura 2):

Figura 2. Esempio di una ricerca per forme

Il motore di ricerca consente l'impiego dei caratteri jolly e la possibilità di effettuare una ricerca espansa, ovvero senza tener conto degli accenti: la ricerca di una forma dà inoltre accesso ai dati a essa relativi (ad es. autori che la impiegano, riviste e fascicoli dove compare) con la possibilità di visualizzare sia contesti immediati, di lunghezza prestabilita, sia contesti allargati.[8] Sarà possibile, poi, effettuare indagini incrociando più parametri: per esempio, basterà selezionare un autore per trovare le occorrenze di quella forma nei testi a lui attribuiti o attribuibili, raffinando così la propria ricerca.[9] Peraltro, il collegamento testo-autore, effettuato dall'operatore nel pannello di controllo di ArPeR, è un'operazione con risvolti non trascurabili: infatti, sebbene la funzione primaria della presente banca dati sia quella di fornire un *corpus* di testi dialettali che serva di base per ricerche linguistiche e lessicografiche, la piattaforma ArPeR restituisce per ogni autore testi che a volte non erano stati considerati parte della loro bibliografia primaria. Del resto, l'indagine sugli autori e sui relativi pseudonimi, imprescindibile per la creazione delle schede "Autori",[10] ha già integrato la bibliografia sinora nota per Zanazzo (vd. figura 3): i componimenti *Bbone Feste* e *Er parlà de l'Africani*, benché pubblicati nel «Rugantino», non sono stati inclusi nelle antologie complete postume più importanti, quella del figlio Alfredo (Zanazzo, 1921) e quella di Orioli (1968).

[8] Si è scelto di non inserire una ricerca di tipo fuzzy, in quanto tale modalità, se utile per ricerche di tipo lessicografico, rischia di restituire risultati troppo ampi e non controllabili, inadatti a indagini linguistiche mirate che richiedono corrispondenze esatte, come nel caso dell'analisi di tratti fonologici o morfologici specifici. Si è pertanto optato per una ricerca esatta seguendo i criteri adottati nella costruzione del *corpus OVI*, ormai consolidati nella prassi degli studi linguistici e filologici, dove è consuetudine operare su corrispondenze formali precise per garantire l'affidabilità dell'analisi.

[9] Riguardo agli scritti finora anonimi, l'attribuzione avverrà su base stilistica e si ricorrerà alla formula [*nome dell'autore* (attrib.)] per registrare tale attribuzione nel *database*. Sarà fondamentale, in quest'ultima fase del progetto, l'uso del software Stylo impiegato presso l'Università Jagellonica di Cracovia per le ricerche stilometriche (cfr. Eder, Kestemont & Rybicki, 2016).

[10] Ciò è possibile specie grazie alle schede di Majolo Molinari (1963) sulla stampa periodica romana dell'Ottocento e sui suoi protagonisti.

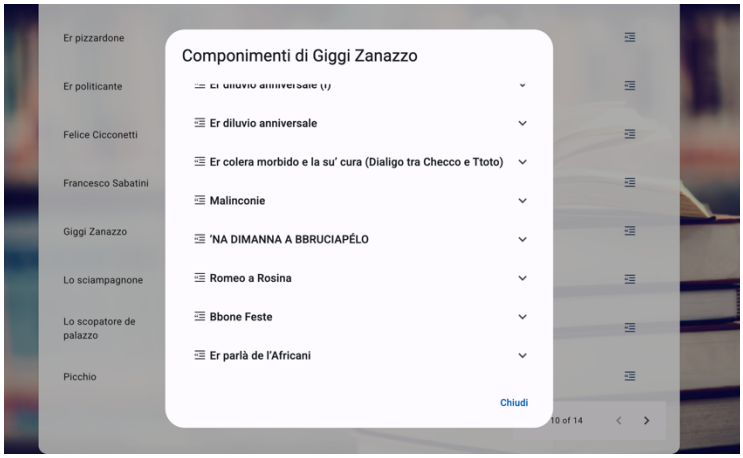


Figura 3. Lista testi di un autore

Se è possibile aggiornare la bibliografia primaria uno dei maggiori autori dialettali di quest'epoca, si può ben immaginare il margine di miglioramento della situazione dei poeti "minori" (per es. Giaquinto, Ilari e Francati; sul «Rugantino», inoltre, scrive il giovane Trilussa).
Non va dimenticato, in conclusione, che la prospettiva *open source* permette interventi di aggiornamento e di perfezionamento di fatto infiniti: si pensi alla possibilità di inserire una ricerca per lemmi[11] o all'apertura dell'orizzonte cronologico verso i decenni precedenti e successivi, così da creare, col tempo, un *Archivio unico dei periodici romaneschi*.

RINGRAZIAMENTI

Il progetto (n. P500PH_222332) è finanziato dal Fondo Nazionale Svizzero per la Ricerca Scientifica per il periodo 2024-2026, nell'ambito della borsa Postdoc Mobility. Si ringrazia Michele Loporcaro per il sostegno dato nel corso dell'avviamento del progetto e in particolare Giulio Vaccaro per l'insostituibile guida nell'ideazione della struttura della piattaforma. L'intera realizzazione dell'output digitale si deve a Fabrizio Nicolò.

BIBLIOGRAFIA

Ascoli, G. I. (1900). Varia (Del romanesco ancora), *Archivio Glottologico Italiano*, 15, 323–325.

Bovet, E. (1898). *Le peuple de Rome vers 1840 d'après les sonnets en dialecte trastévérin de G. G. Belli*. Attinger.

Costa, C. (2011). «Nun se venne, è sincera, forte, onesta». La poesia romanesca di Giggi Zanazzo. In Onorati F. e Scalessa G. (Eds.), *Le voci di Roma. Omaggio a Giggi Zanazzo, Atti del convegno di studi* (pp. 33–45*)*, Roma: Il cubo. ISBN: 9788897431022.

De Mauro, T. (1989). Per una storia linguistica della città di Roma. in De Mauro, T. (Ed.). *Il romanesco ieri e oggi* (pp. 13-37). Roma: Bulzoni. ISBN: 978-88-7119-021-1.

Eder, M., Kestemont, M. e Rybicki, J. (2016). Stylometry with R: A package for Computational Text Analysis, in *The R Journal* 8(1), 107-121. DOI: 10.32614/RJ-2016-007.

Ernst, G. (1970). *Die Toskanisierung des romischen Dialekts im 15. und 16. Jahrhundert*. Niemeyer. https://doi.org/10.1515/9783111328492

Faraoni, V. (20-21 maggio 2021). *Il romanesco prima e dopo Porta Pia*. [Conference presentation] Dalla Roma pontificia alla Roma italiana. Le istituzioni culturali e la città, Roma, Università La Sapienza.

Majolo Molinari, O. (1963). *La stampa periodica romana dell'Ottocento*. Istituto Nazionale di Studi Romani. ISBN: 9788873114727.

Merlo, C. (1929). Vicende storiche della lingua di Roma I. Dalle origini al sec. XV. *L'Italia dialettale*, 5, 172-201. (Reprinted from *Saggi linguistici*, pp. 33–62, by C. Merlo, Ed., 1959, Pisa: Pacini-Mariotti).

[11] Allo stato attuale non è prevista una ricerca per lemmi data l'elevata richiesta di tempo e risorse, sia in modalità automatica (che richiederebbe addestramento e revisione) sia manuale. L'implementazione potrà eventualmente essere considerata, come detto, in una fase futura del progetto.

Migliorini, B. (1932). Dialetto e lingua nazionale a Roma. *Capitolium*, 10, 350–356. (Reprinted from *Lingua e cultura*, pp. 109–123, by B. Migliorini, Ed., 1948, Roma, Tumminelli).

Morandi L. (Ed.). 1886-1889. I sonetti, Città di Castello: Lapi.

Morino, T. (1899). *La grammatica del dialetto romanesco secondo i sonetti di G. G. Belli. I. Suoni - Forme*. Colitti.

Orioli. G. (ed.). 1968. *Poesie romanesche di G. Zanazzo*. Roma: Avanzini e Torraca.

Picchiorri, E. (2019) *«Nun vorrebbia che fusse na cianchetta der nemico». Il romanesco nei giornali della repubblica romana».* in Vaccaro, G. (Ed.). *Marcello 7.0. Studi in onore di Marcello Teodonio* (pp. 477–487), Roma: Il cubo. ISBN: 9788897431091.

Puglisi, P. (2011). «Ggente mia, garbata e bbella. Zanazzo giornalista». In Onorati F. e Scalessa G. (Eds.), *Le voci di Roma. Omaggio a Giggi Zanazzo, Atti del convegno di studi* (pp. 33–45*)*, Roma: Il cubo. ISBN: 9788897431022.

Sabatini, F. (1890). *Il volgo di Roma*. Loescher.

Senes, G. (1900). *Anatomia filologica del dialetto romanesco fatta sulla Scoperta dell'America di Cesare Pascarella*. Florenz.

Stefinlongo, A. (1985). Note sulla situazione sociolinguistica romana. Preliminari per una ricerca. *Rivista italiana di dialettologia*, 9, 43–67.

Trifone, P. (2008). *Storia linguistica di Roma*. Carocci. ISBN: 9788843044597.

Zanazzo, A. (ed.) 1921-1922. *Versi romaneschi editi ed inediti*, di G. Zanazzo. M. Carra e C. di Luigi Bellini.