Umanistica Digitale – ISSN: 2532-8816 – n. 10, 2021
Alberto Baldi – Text analysis *ed editoria digitale: proposte di integrazione con valutazioni critiche. A proposito del carteggio tra Ignazio Silone e la Arnoldo Mondadori Editore*
DOI: http://doi.org/10.6092/issn.2532-8816/12762

Text analysis ed editoria digitale: proposte di integrazione con valutazioni critiche. A proposito del carteggio tra Ignazio Silone e la Arnoldo Mondadori Editore

Alberto Baldi

Università di Firenze
alberto.baldi@unifi.it

Abstract

Le tecniche di *text analysis* trovano ampia applicazione nell'umanistica digitale, in particolare a supporto della critica letteraria su vasti *corpora* di fonti primarie. In questo articolo si ripercorre, in prospettiva editoriale, la sperimentazione di tecniche di analisi computazionale dei testi – nella fattispecie *topic modeling* e *sentiment analysis* – su un *corpus* di carte d'archivio, il carteggio tra Ignazio Silone e la Arnoldo Mondadori Editore (396 pezzi epistolari). I risultati emersi, pur con i limiti documentati nell'articolo, permettono di ipotizzare come questi strumenti, al di fuori della critica letteraria *strictu sensu*, possano ricoprire un ruolo anche nello studio e, soprattutto, nell'edizione di epistolari d'autore, e come i loro *output*, integrati a metodologie e tecniche del *digital scholarly editing*, possano rendere più produttiva l'esperienza di fruizione per i lettori.

Text analysis techniques are widely applied in digital humanities, in particular to support literary criticism on large *corpora* of primary sources. In this paper we retrace, from an editorial perspective, the experimentation of computational text analysis techniques – in this case topic modeling and sentiment analysis – on a *corpus* of archival documents, the correspondence between Ignazio Silone and Arnoldo Mondadori Editore (396 pieces of correspondence). The results that emerged, even with the limitations documented in the article, allow us to hypothesize how these tools, outside of literary criticism *strictu sensu*, can also play a role in the study and, above all, in the edition of author's letters, and how their output, integrated with methodologies and techniques of digital scholarly editing, can make the experience more productive for readers.

1. L'edizione digitale del carteggio Silone-Mondadori

Il presente articolo discute gli esiti di una sperimentazione di tecniche per il *topic modeling* e per la *sentiment analysis* sul carteggio tra Ignazio Silone e la Arnoldo Mondadori Editore, suo

principale editore in Italia, proponendo possibili integrazioni tra *text analysis* ed edizioni digitali.[1] Lo scambio tra Silone e la Mondadori conta 396 pezzi epistolari, per lo più dattiloscritti,[2] distribuiti in un arco cronologico dal 1946 al 1977, e coinvolge, oltre a Silone e ad Arnoldo e Alberto Mondadori, importanti personalità della letteratura, della cultura e dell'editoria del Novecento italiano, tra cui Vittorio Sereni, Niccolò Gallo, Sergio Polillo, Marco Forti.

Il dialogo di Silone con la Mondadori è ovviamente incentrato sulle edizioni italiane dei suoi libri e l'aspetto tecnico-editoriale è preponderante: questioni di natura contrattuale (come la negoziazione delle percentuali, la gestione dei diritti, la tiratura, le copie omaggio per l'autore), promozionale (la distribuzione dei volumi, l'impatto delle campagne pubblicitarie, il coinvolgimento di recensori), economica (avvisi di retribuzione, richieste di anticipi, sollecitazioni) ricorrono per tutto il carteggio, dimostrando la particolare cura dello scrittore nella gestione dei suoi interessi professionali. A ciò si affianca spesso l'elemento eminentemente letterario, soprattutto nella valutazione delle opere proposte – come nel caso del rifiuto da parte di Alberto Mondadori di accogliere la prima redazione della *Volpe e le camelie* per la collana del Saggiatore Biblioteca delle Silerchie o viceversa degli entusiasmi per opere quali *La scuola dei dittatori*, accolta con grande interesse dopo il lavoro di revisione del testo italiano inedito – o nella scelta delle collane di destinazione. Particolare rilievo hanno anche le dinamiche connesse con la diffusione dei libri siloniani all'estero, di cui resta traccia nei riferimenti alle trattative con editori stranieri (Harper & Row, Oprecht, Grasset…) per la cessione dei diritti e per la scelta dei traduttori.

2. *Text analysis* del carteggio Silone-Mondadori

In margine a una lettura diacronica del rapporto tra Silone e la Mondadori, condotta ripercorrendo i principali snodi che si incontrano nel loro trentennale dialogo epistolare e concretizzatasi in un capitolo introduttivo al carteggio e nelle note di commento alle lettere, grazie all'analisi automatica dei dati testuali si è sperimentata una lettura "computazionale" della corrispondenza intercorsa tra lo scrittore e il suo editore, nel tentativo di verificare eventuali punti

[1] Tale sperimentazione è parte della mia tesi di dottorato che ha avuto come oggetto finale l'edizione digitale del carteggio tra Silone e la Arnoldo Mondadori Editore reinterpretando il modello di *knowledge site* proposto da Peter Shillingsburg ([21]; [22]). La tesi ha come titolo "'Una potenza lontana e misteriosa'. Il carteggio tra Ignazio Silone e la Arnoldo Mondadori Editore come *knowledge site*: tra *digital scholarly editing* e *text analysis*" ed è collegata al sito epistolariosilone.it (ospitato sui server del Laboratorio Disit del Dipartimento di Scienze dell'Informazione dell'Università di Firenze), attualmente ad accesso ridotto per ragioni di diritti. L'edizione è stata realizzata utilizzando il CMS Omeka Classic, popolato con le riproduzioni fotografiche e le trascrizioni HTML e XML-TEI delle lettere, i relativi metadati (secondo lo standard Dublin Core) e l'apparato di note e schede descrittive a commento dei testi.

[2] Le lettere sono conservate negli archivi della Fondazione di Studi Storici "Filippo Turati" di Firenze (165 documenti) e della Fondazione Arnoldo e Alberto Mondadori di Milano (231 documenti).

di tangenza tra l'andamento diacronico emerso da una lettura *close* e gli *output* di una "macroanalysis" ([10]).

Tra le varie tecniche possibili, si è scelto di utilizzare il *topic modeling*, per la rilevazione dei principali "temi", e la *sentiment analysis*, per il calcolo della polarità e dell'andamento emozionale. Entrambe le analisi sono state eseguite, tra i tre livelli proposti da Bing Liu per la *sentiment analysis* ma estendibili anche al *topic modeling*, al "document level", considerando cioè "each document as a whole and does not study entities or aspects inside the document or determine sentiments expressed about them" ([12]: 47).[3]

Tra i molteplici studi che hanno applicato, con più o meno successo, *tool* per la *text analysis* in ambito letterario e storico-culturale, non molti sono quelli che ne hanno sperimentato le possibilità su *corpora* epistolari. Tra questi, ne elenchiamo alcuni tra i più recenti e significativi:

- *Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic (CKCC) – ePistolarium*[4] (2013)[5]: il progetto CKCC – nato da una collaborazione tra il Descartes Centre for the History and Philosophy of the Sciences and the Humanities dell'Università di Utrecht e l'Istituto Huygens – ha come obiettivo quello di rendere accessibile un *database* di lettere di eruditi attivi nella Repubblica Olandese del Diciassettesimo secolo. Il lavoro sul *corpus*, che annovera quasi 20.000 lettere, ha dato origine a un *virtual research environment* denominato ePistolarium, "which allows researchers to explore and analyze the corpus in innovative ways". Il *topic modeling* è stato utilizzato per ampliare le possibilità di esplorazione del *corpus* epistolare offerte da ePistolarium: in particolare, dopo una fase di pre-elaborazione, sono stati applicati tre diversi modelli di *topic modeling* (LDA – Latent Dirichlet Allocation –, LSA – Latent Semantic Analysis – e RI – Random Indexing) per calcolare un indice di similarità tra i vari pezzi epistolari e consentire una ricerca per testi affini o per proporre suggerimenti rispetto alle *keywords* delle *query*. Come illustrato nel capitolo III.4 di *Reassembling the Republic of Letters in the Digital Age* (Hotson, Wallnig 2019), nonostante dal 2013 ePistolarium abbia visto pochi aggiornamenti, il progetto è ancora oggi passibile di ampliamenti, sia per quanto concerne nuovi *corpus* testuali che per il patrimonio di metadati, soprattutto alla luce della sua inclusione nell'imponente progetto

[3] Per la *sentiment analysis*, Bing Liu ha infatti distinto tra "document level" – che assume come unità di riferimento i singoli documenti –, "sentence level" – che agisce invece al livello delle singole frasi –, e "aspect level" – che analizza il sentiment distinguendo tra le entità cui esso è riferito. A questi si può aggiungere il "concept level", individuato da Erik Cambria e Amir Hussein ([1]), che focalizza l'analisi sul livello concettuale non limitandosi alla superficie testuale del *corpus*.

[4] <http://ckcc.huygens.knaw.nl/> (04/2021).

[5] Data di prima pubblicazione.

di mappatura di carteggi del Diciassettesimo e Diciottesimo secolo *Mapping the Republic of Letters* dell'Università di Stanford.[6]

- *Epistolario De Gasperi*[7] (2016)[8]: l'edizione nazionale delle lettere di Alcide De Gasperi rappresenta un progetto di grande portata sia per il numero di fondi archivistici coinvolti (a oggi 226) che per il gruppo di ricerca (39 ricercatori coinvolti, in una sinergia tra la Fondazione Trentina Alcide De Gasperi, la Fondazione Bruno Kessler di Trento e l'Istituto Luigi Sturzo di Roma). Allo stato attuale, sono state pubblicate 3915 lettere, un numero destinato ancora a crescere. Il lavoro è perfetta testimonianza infatti di come le edizioni digitali siano prodotti fluidi, passibili di una integrazioni dei materiali coinvolti e di aggiornamenti delle modalità di fruizione. Come è stato recentemente dichiarato da Tonelli, Sprugnoli, Moretti *et al.* ([24]), tra i responsabili del progetto, le prospettive future di un progetto già completamente fruibile sul web sono molteplici, sia per quanto riguarda l'incremento delle risorse archivistiche proposte che per l'affinamento delle tecnologie utilizzate (sono infatti previste nuove modalità di ricerca all'interno del database, il raggruppamento delle lettere per aree tematiche, la possibilità di accedere ai file xml-tei e, soprattutto, il rilascio con licenza Creative Commons 4.0 dell'infrastruttura di trascrizione utilizzata). Anche in questa edizione si sono sperimentate tecnologie di analisi automatica dei testi – estrazioni di *keyphrase*, *topic detection*.[9] Dello stato attuale del progetto Epistolario De Gasperi ha recentemente scritto Stefano Malfatti ([13]).

- *Epistolario Svevo*[10] (2016)[11]: il progetto diretto da Cristina Fenu per la Biblioteca Attilio Hortis di Trieste in collaborazione con il MaLeLab del Dipartimento di Ingegneria e Architettura dell'Università di Trieste, il Master in Digital Humanities dell'Università di Ca' Foscari e Gabriele Sarti (cfr. [4]) rappresenta senza dubbio un progetto pionieristico almeno per quanto riguarda il panorama italiano. L'epistolario di Italo Svevo – di cui è proposta la sola, parziale digitalizzazione (sono visualizzabili soltanto i carteggi con Eugenio Montale e con James Joyce, mentre i testi sono stati ricavati dall'epistolario dello scrittore curato da Bruno Maier) senza trascrizioni né apparati critici – è stato infatti processato con tecnologie di analisi automatica del linguaggio, nello specifico *topic modeling* e *sentiment*

6 <http://republicofletters.stanford.edu/> (04/2021).

7 <https://www.epistolariodegasperi.it/> (04/2021).

8 Data di inizio del progetto.

9 Grazie all'utilizzo della piattaforma ALCIDE ([13]; [15]) e del *tool* KD ([15]), mentre la pubblicazione dei documenti e dei loro metadati è stata realizzata con l'utilizzo del *software* LETTERE (LETters Transcription Environment for REsearch; cfr. [17]): sia ALCIDE che LETTERE sono risorse *tailor made*, a oggi non accessibili *open source*.

10 <http://svevo-ar.online.trieste.it/progetto/archivio-digitale/> (04/2021).

11 Data di inizio del progetto.

> *analysis*. Per il *topic modeling* le lettere sono state pre-processate per rimuovere le *stopwords* (parole grammaticali e parole ad alto tasso di occorrenza in un epistolario, come le formule di saluti o le intestazioni) e poi analizzate con la libreria Python Gensim: l'interrogazione ha prodotto cinque temi fondamentali – famiglia, lavoro, letteratura, salute, viaggio – la cui distribuzione all'interno del *corpus* è stata rappresentata con un diagramma alluvionale. La *sentiment analysis*, invece, è stata realizzata utilizzando il pacchetto Syuzhet in ambiente R e gli output sono stati rappresentati sia con diagrammi a torta che con istogrammi orizzontali.

Nonostante l'intento meramente esemplificativo, questa rassegna sembra tuttavia comprendere la maggior parte (se non la totalità) delle sperimentazioni di *text analysis* su carteggi letterari e storico-culturali. Un numero molto esiguo, se si considera da un lato la costante crescita di sperimentazioni di simile tecniche su testi letterari e, soprattutto, la grande disponibilità di edizioni elettroniche di epistolari, come documentato dal *Catalog of Digital Scholarly Editions* a cura di Patrick Sahle,[12] dove i *corpora* epistolari rappresentano ben il 17% del totale dei lavori censiti. Questa scarsità di sperimentazioni può dipendere, anzitutto, da un naturale maggiore interesse della comunità umanistica per le opere letterarie, in linea, nel digitale, con quanto da sempre accade anche negli studi "tradizionali", in cui i lavori sui carteggi rappresentano un sotto-ambito di grande interesse ma "di nicchia", specie perché i *corpora* epistolari risultano essere fonti di secondo grado in supporto allo studio dei testi letterari. Inoltre, è probabile che l'applicazione delle nuove metodologie di analisi computazionale si concentri laddove le esigenze ermeneutiche risultino maggiori, ossia nella lettura interpretativa di testi letterari, di contro alle edizioni di epistolari dove, di prassi, il livello di approfondimento critico e di contestualizzazione dei testi è graduato a seconda delle intenzioni del curatore, senza che il risultato finale ne risenta: numerose, infatti, soprattutto nel digitale, sono le edizioni di carteggi proposte con un apparato di commento ridotto – spesso limitato a cenni chiarificatori riguardo a passaggi o riferimenti "oscuri" –, quando non del tutto assente e sostituito da schede descrittive e notizie archivistiche.

A nostro avviso, tuttavia, i carteggi di letterati – ma, in generale, le loro carte d'archivio, come diari, taccuini di lavoro, appunti… – rappresentano un campo di applicazione della *text analysis* con grandi potenzialità, perché si tratta di testi che, pur non necessitando di un intervento interpretativo "sul" testo, richiedono una contestualizzazione "per" i testi, che li inquadri in una più ampia prospettiva di ricostruzione storico-culturale. Inoltre, non presentano, a livello linguistico e semantico, le stesse criticità dei testi letterari, soprattutto per quanto attiene alla "figuratività" del linguaggio di questi ultimi, che spesso ingenera fraintendimenti o risultati "opachi" negli *output* delle analisi, sebbene, come ha scritto Lisa Rhody, anche le apparenti opacità, nel tentativo di chiarificarle, possano rivelare "rich deposits of hermeneutic possibility" ([19]: 32).

L'applicazione della *text analysis* su questo tipo di carte, allora, se da un lato non avrà intento strettamente critico – nella misura in cui il fabbisogno interpretativo, per questi testi, è per ovvie ragioni inferiore rispetto a quello dei testi letterari –, ciò nonostante il contributo che se ne potrà

[12] <http://www.digitale-edition.de/> (04/2021).

derivare, specie se integrandone i risultati in funzione editoriale, avrà un valore "orientativo" per l'utente-lettore, che beneficerà di questi strumenti per ampliare la propria esperienza di fruizione, con prospettive di lettura e informazioni che altrimenti potrebbero restare precluse, come hanno scritto Shawn Graham e Ian Milligan, riguardo all'utilizzo del pacchetto MALLET (cfr. *infra*) per il *topic modeling* su carte d'archivio: "Historians can use it to take a large archival collection with robust OCR, run it through the system, and begin to see the overall contours and shape of the material. While it does not replace in-depth close reading, it does provide invaluable context and pointers towards issues that might have otherwise been missed" ([5]: 72).

2.1. Topic modeling

Per predisporre il *corpus* di 396 lettere al processo di *topic modeling* si è dapprima eseguita una serie di operazioni di pre-elaborazione dei testi al fine di ottimizzare l'efficacia dell'analisi. Nonostante gli algoritmi di *topic modeling* prevedano fasi di pre-elaborazione dell'*input*, si è scelto comunque di integrare nel nostro *workflow* un ulteriore passaggio finalizzato alla riduzione degli elementi di possibile disturbo e all'ottenimento di una base di dati testuali ancor più adeguata agli *output* auspicati.

Fra le varie risorse per la pre-elaborazione del *corpus* disponibili *open-source* – sotto forma di librerie per il NLP (Spacy e NLTK, ad esempio) o di *software* (Carrot2, GATE, Orange...) – si è scelto di utilizzare Lexos,[13] sviluppato dal Wheaton College. Si tratta di un *tool*, giunto nel 2019 alla *release* 4.0, scritto in Python e con un'interfaccia realizzata con jQuery. È utilizzabile sia in modalità *web based* che, alla necessità di processare *corpora* di dimensioni ingenti, in locale. Lexos è pensato per riunire tutti gli strumenti necessari a compiere una serie di operazioni su basi di dati testuali, ad esempio la produzione di statistiche e indici di frequenza, il calcolo della similarità tra lemmi o sintagmi, l'analisi dei contenuti. Tra le sue funzionalità, tuttavia, spicca la sezione dedicata alla fase di pre-elaborazione dei testi, suddivisa in tre sotto-ambiti: *scrub*, *cut* e *tokenize*. Di questi, ci siamo avvalsi soprattutto degli strumenti proposti come *scrub*, che offrono la possibilità di uniformare tutte le lettere alla forma minuscola (*lowercase*), rimuovere cifre, spazi, interruzioni di riga, punteggiatura e tabulazioni, lemmatizzare il *corpus* e filtrarlo tramite un elenco di *stopwords* (cfr. *infra*). Inoltre, è prevista la possibilità di specificare la presenza nei testi di *tag* (HTML, XML, SGML), di modo da raffinare ulteriormente gli *output*.

Nella fattispecie della preparazione delle 396 lettere del nostro carteggio in funzione dell'analisi dei *topic*, si è utilizzato Lexos per uniformare le lettere alla forma minuscola e per rimuovere cifre e punteggiatura. Inoltre, si è utilizzato lo strumento "Consolidation" – che consente la sostituzione di specifici caratteri – per isolare le parole apostrofate (soprattutto gli articoli e le preposizioni articolate). Una volta verificato il risultato dell'operazione (grazie alla funzione *preview*, che consente di visualizzare un'anteprima dell'*output*), si è proceduto all'esportazione dei nuovi file TXT. Si è scelto invece di non procedere – né con Lexos né con altre risorse (ad esempio la libreria Snowball) – né allo *stemming* né alla lemmatizzazione del *corpus*. Di contro a quanto positivamente riscontrato da altri studiosi nell'ambito di sperimentazioni del *topic*

[13] <http://lexos.wheatoncollege.edu/upload> (04/2021).

modeling su testi umanistici, si è infatti potuto osservare che entrambe le operazioni sarebbero risultate controproducenti alla migliore riuscita dell'analisi finale, giacché gli *output* così prodotti sono apparsi forieri di numerosi fraintendimenti: a titolo d'esempio, un termine tecnico come "contratto", assai frequente tra i documenti del carteggio, sarebbe stato ridotto – in caso di lemmatizzazione – alla forma "contrarre", disperdendo il portato semantico originario.

Fra le varie operazione di pre-elaborazione, l'utilizzo di un'efficace lista di *stopwords* è spesso determinante per il buon esito delle operazioni di *text analysis*, abbattendo i livelli di "rumore" statistico e favorendo l'emersione dei termini realmente significativi per l'obiettivo dell'analisi. Ciò vale, in particolar modo, per il *topic modeling*, laddove il modello non prevede un approccio *lexicon based* che stabilisca un diverso valore tra le possibili occorrenze di parola: nel *topic modeling*, in specie se condotto con algoritmi LDA, i termini hanno di *default* il medesimo valore ai fini del sottostante calcolo statistico, al punto che l'uso delle *stopwords* costituisce – assieme alla scelta dei parametri d'analisi – il principale campo di possibile intervento nei procedimenti che conducono agli *output* (tanto che le loro possibili modalità di utilizzo hanno suscitato importanti riflessioni di studiosi, cfr. ad esempio [20]).

Esistono numerose liste pre-configurate, anche per la lingua italiana, spesso già incluse nelle librerie per il NLP utilizzabili per l'operazione di filtraggio, rese disponibili su portali di condivisione per risorse informatiche (ad esempio GitHub), o derivabili, se *open source*, da altri progetti. Tuttavia, data la particolare natura testuale del nostro *corpus*, dettata soprattutto dal formulario tipico della scrittura epistolare (i saluti e le formule di chiusura, gli appellativi di cortesia, le indicazioni delle località di spedizione o di destinazione) e dall'ampio numero di nomi propri e titoli di opere o collane editoriali, si è scelto di predisporre una lista *ad hoc*.

Come base di partenza si è utilizzata la lista di *stopwords* per l'italiano presente in NLTK, un totale di 279 termini, per lo più parole grammaticali oltre alle coniugazioni complete dei verbi "avere", "essere", "fare" e "stare". Per ampliare la lista, si è messo in pratica quanto suggerito di Brian Kokensparger nel capitolo di *Guide to Programming for the Digital Humanities* dedicato alle *stopwords*, ossia:

> [...] counting the frequencies of all of the unique words in the text, and writing out to a text file the words which should be considered "stop words," or those words which should be ignored when looking at the word frequency in a text. If you were able to process a corpus (more than one text in a collection of chosen texts), then the words of highest frequency over the entire corpus would become your list of stop words. ([11]: 49)

Per generare un indice di frequenza e al contempo filtrare le parole inferiori alle quattro lettere si è utilizzato la *suite* di risorse per l'analisi dei testi Voyant Tools, creata da Stéfan Sinclair e Geoffrey Rockwell e tradotta in italiano da un *team* della Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD) guidato da Fabio Ciotti. Il risultato finale, comprensivo anche di lessemi polirematici come alcuni titoli di libri siloniani o nomi di associazioni, collane editoriali e altri enti – inseriti manualmente –, ha contato 2.177 termini.

Una volta predisposto il *corpus* tramite le operazioni di pre-elaborazione si è passati alla fase di analisi dei dati. Per eseguirla, si è scelto di utilizzare il pacchetto MALLET, un *tool* Java sviluppato da Andrew McCallum insieme ai suoi colleghi e agli studenti della University of Massachusetts Amherst.[14]

Di MALLET esistono più GUI, come il Topic Modeling Tool[15] sviluppato da Jonathan Scott Enderle dell'University of Pennsylvania, che ha aggiornato una prima versione[16] creata da David Newman e Arun Balagopalan. L'interfaccia consente di settare i parametri dell'algoritmo in modo analogo a quanto possibile operando direttamente da codice, oltre a permettere di indicare i metadati di uno dei fogli CSV in cui confluiranno gli *output*, presentando di fatto delle significative migliorie rispetto alla precedente versione che permetteva soltanto di impostare il numero delle iterazioni e la soglia minima di distribuzione dei *topic*. Ciò nonostante, si è scelto di procedere alle analisi senza farvi ricorso e accedendo agli algoritmi direttamente da codice. Topic Modeling Tool, infatti, pur consentendo un elevato grado di parametrizzazione, incontra difficoltà nell'eseguire un elevato numero di iterazioni del modello.

All'atto dell'analisi, per ogni parametro si sono prese in considerazione diverse possibili configurazioni, confrontandole tra loro e valutandone i risultati. Le variabili prese in considerazione – incrociandole nei vari, possibili tentativi – sono state: `--num-topics` (5; 10; 15), `--num-iterations` (5000; 10000; 15000), `--alpha` (100; 50; 20), `--beta` (0,001; 0,0001; 0,00001), mentre gli altri settaggi sono stati mantenuti invariati rispetto alla configurazione di *default*. La scelta di sperimentare un ridotto numero di combinazioni tra parametri è stata dettata dalla volontà di non intervenire oltre i necessari aggiustamenti, per evitare il rischio di etero-dirigere i risultati dell'analisi e forzarli nella direzione delle aspettative. Tra le varie combinazioni, la più appropriata è risultata essere quella espressa nella seguente stringa di codice:

```
bin/mallet train-topics --input topic_input.mallet --num-topics 5 --num-iterations 15000 --optimize-interval 10 --optimize-burn-in 20 --use-symmetric-alpha false --alpha 20 --beta 0.00001 --num-threads 4 --num-top-words 15 --output-state ~/OUTPUT_MALLET/ topic-state_num5_al20_bet00001-keepsequence.gz --output-doc-topics ~/OUTPUT_MALLET/num5_al20_bet00001.txt --output-topic-keys ~/OUTPUT_MALLET/topic-key_num5_al20_ bet00001.csv
```

Un ridotto numero di *topic* (cinque) è infatti apparso più coerente nell'analizzare un epistolario autore-editore, dove i temi che ricorrono sono limitati alla sfera professionale e, nella fattispecie, editoriale. Quanto al numero di iterazioni, si è optato per il più alto (15000) dal momento che l'unica controindicazione prevista è di prolungare l'elaborazione. I parametri *alpha* e *beta*, invece,

[14] <http://mallet.cs.umass.edu/index.php> (04/2021).

[15] <https://github.com/senderle/topic-modeling-tool> (04/2021).

[16] <https://github.com/arunbg/Topic-Modeling-Tool> (04/2021).

hanno richiesto una riflessione più approfondita. Il parametro *alpha* stabilisce la densità dei *topic* nei documenti: un parametro più elevato (partendo dal 100 di *default*) induce l'algoritmo a presumere un più alto grado di compresenza tra *topic* all'interno di uno stesso documento; il *beta*, invece, stabilisce il numero di parole che afferiscono a un determinato *topic*. Pertanto, si è scelto di ridurre entrambi i parametri (20 per l'*alpha*, 0,0001 per il *beta*), considerando la natura dei nostri documenti: com'è tipico della comunicazione commerciale, infatti, difficilmente in una missiva si toccano più argomenti diversi tra loro, preferendo magari l'invio di più documenti in contemporanea; quanto al numero di parole identificative dei *topic*, si è preferito ridurre il valore dal momento che la scrittura epistolare è ricca di formule ricorrenti che non riguardano l'effettivo contenuto del messaggio: un basso valore al parametro *beta* – coordinato a un'efficace lista di *stopwords* – consente così di individuare i termini più significativi e di enucleare con più precisione gli argomenti di cui si costituisce lo scambio.

MALLET produce tre diverse tipologie di *output*: un file compresso in archivio gzip (`--output-state`) che risulta ininfluente ai fini dell'analisi dei *topic* ma che costituisce una rappresentazione del modello generato: in esso, viene offerto uno schema che riassume l'assegnazione di ogni parola di ogni documento a un determinato *topic*, permettendo così l'eventuale riutilizzo. Gli altri due file concorrono invece alla rappresentazione del modello distributivo dei *topic* all'interno del *corpus* documentario, fornendo da un lato (`--output-topic-keys`) un elenco delle parole chiave (15, secondo il parametro `–num- top-words`) di ciascun *topic*, e dall'altro (`--output-doc-topics`) la distribuzione percentuale dei *topic* all'interno di ogni documento. Dal file `--output-topic-keys` si è proceduto alla definizione singoli *topic*, assegnando – in seguito a una valutazione degli elenchi di parole – un'etichetta di sintesi relativa al contenuto. Nella fattispecie, le liste di parole derivanti dall'analisi sono state le seguenti:

1. `0  0,262 ristampa edizione bozze auguri correzioni risposta notizia programma pubblicazione esaurito vendita nota precedente ritorno copertina;`
2. `1  0,04785 pubblicità italiana poesie interesse film forma giudizio librai impressione proposito soggetto rimanere situazione leggere distribuzione;`
3. `2  0,07152 racconto traduzione dattiloscritto scritto rivista prefazione progetto pubblicazione idea saggio interesse decisione tema tratta interessa;`
4. `3  0,01556 settore diritti ritenuta pagamento legge letteratura responsabilità cultura fattura esperienza importo serie compenso regime somme;`
5. `4  0,11316 contratto diritti edizione accordo ufficio contratti cessione traduzione edizioni editoriali pubblicazione condizioni richiesta paesi firma.`

Dei cinque *topic*, i numeri 0, 2, 3 e 4 sono apparsi da subito più coerenti e definiti. Rispettivamente, si sono assegnate le seguenti etichette: "Questioni redazionali", "Proposte di

pubblicazione e altri progetti", "Questioni economiche", "Questioni contrattuali". Si tratta, com'è evidente, di dimensioni concettuali volutamente estese, e che tuttavia sono apparse da subito efficaci nella definizione degli argomenti diffusi tra le lettere pur mantenendosi coerenti con l'*output* generato dall'algoritmo. Nonostante la co-occorrenza di alcuni termini in due o più liste, giustificata dalla loro riconducibilità a più ambiti (ad esempio, il termine "diritti" è presente, coerentemente, in lettere "contrattuali" che discutono degli aspetti legali degli accordi tra scrittore ed editore così come in lettere "economiche" che notificano il pagamento dei diritti maturati), la presenza di alcuni termini "spia" permette di etichettare con efficacia i singoli *topic*: nella fattispecie, "ristampa", "correzioni", "bozze", "edizione", "programma" per le "Questioni redazionali"; "progetto", "dattiloscritto", "scritto", "idea", "interesse", "interessa", "decisione" per le "Proposte di pubblicazione e altri progetti"; "diritti", "ritenuta", "pagamento", "fattura", "importo", "compenso", "somme" per le "Questioni economiche"; "contratto", "diritti", "accordo", "contratti", "cessione", "condizioni", "firma" per le "Questioni contrattuali". A questi, se ne aggiungono altri la cui occorrenza è comprensibile se contestualizzata: è il caso ad esempio del termine "tratta", nella lista riferibile alle "Proposte di pubblicazione…", spesso utilizzato da Silone o dagli esponenti della Mondadori nel presentare i contenuti dei loro progetti editoriali, o del termine "paesi" nella lista "Questioni contrattuali", che è giustificato dai colloqui intorno alla cessione dei diritti dei libri di Silone all'estero.

La conformità di queste etichette è stata poi verificata tramite un riscontro manuale tra i documenti del *corpus*. Quanto al *topic* 1, la sua valutazione è risultata più complessa, anche in seguito a un controllo nelle lettere: se infatti la sua prevalenza sembra coincidere con argomenti che attengono alle dinamiche di distribuzione editoriale (carenza dei libri nelle librerie, ritardi da parte del distributore, esaurimento delle copie nei magazzini dell'editore...), la sua presenza è registrabile tuttavia anche in lettere incentrate su questioni relative alla promozione e alla ricezione delle nuove uscite e agli adattamenti televisivi o teatrali delle opere di Silone. Pertanto, si è preferito ricorrere a una generica etichetta "Altro", al fine di non forzarne l'attribuzione, pur consapevoli che questa etichetta risulti ininfluente ai fini dell'apporto conoscitivo sul complesso testuale fermo restando il suo valore nella complementarità rispetto alla distribuzione complessiva degli altri *topic*.

Dopo aver selezionato e assegnato un'etichetta per ciascuno dei cinque argomenti prodotti dall'analisi, si è proceduto alla resa grafica dell'*output* numerico.

Nell'impossibilità di fornire i grafici completi, si è scelto di suddividere l'*output* grafico in sottoinsiemi di lettere corrispondenti a momenti particolarmente significativi dello scambio Silone-Mondadori e al contempo sufficientemente esemplificativi dei risultati, più o meno validi, dell'applicazione del *topic modeling* alla nostra base di dati testuali. Come si è già illustrato in precedenza, cinque sono i *topic* derivanti dalla nostra analisi, etichettati come "Proposte di pubblicazione e altri progetti" (1), "Questioni contrattuali" (2), "Questioni economiche" (3), "Questioni redazionali" (4), "Altro" (5).

Il primo blocco che proponiamo corrisponde alle prime 29 lettere del carteggio Silone-Mondadori (1946-1948), che vanno dagli inizi del rapporto tra lo scrittore e il suo editore alla prima edizione Mondadori di *Fontamara*:

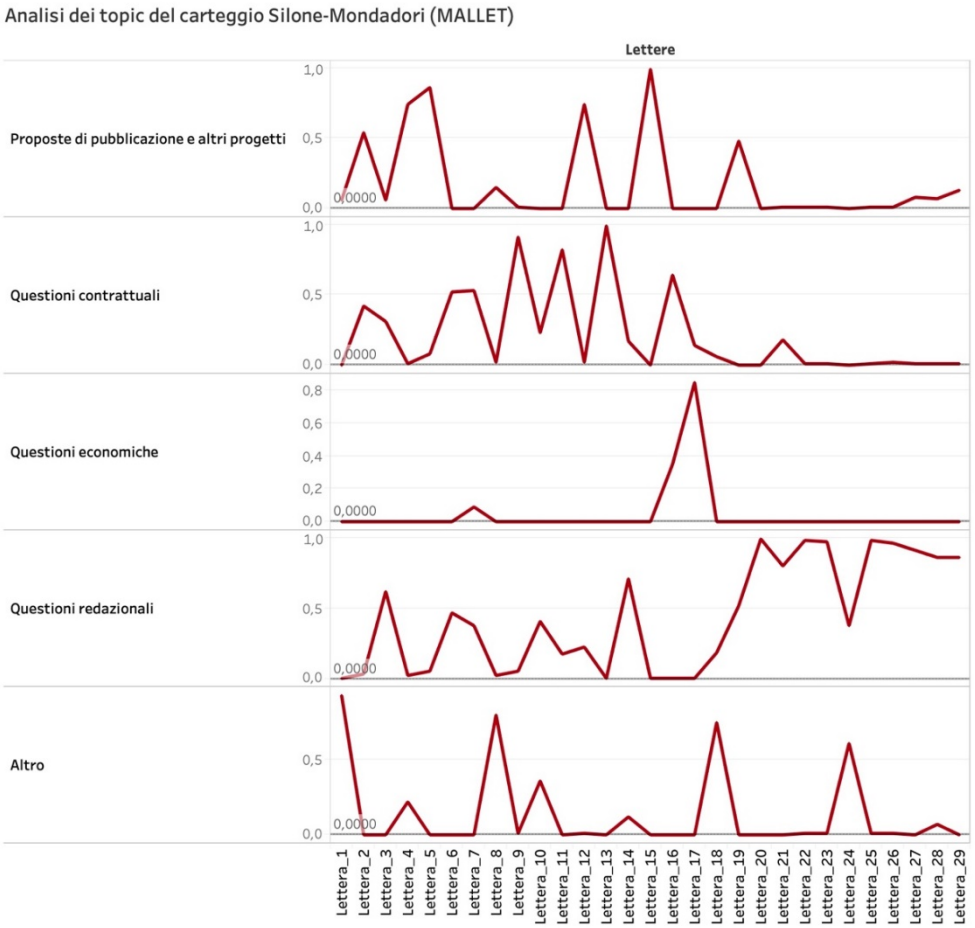

Figura 1: Analisi dei topic

Come risulta dai grafici, in questo periodo iniziale la corrispondenza è, soprattutto nelle prime venti lettere, incentrata sui *topic* "Proposte di pubblicazione e altri progetti" e "Questioni contrattuali". Ciò è facilmente verificabile e riconducibile al fatto che i primi momenti del carteggio vertano attorno ai progetti di pubblicazione dei primi libri di Silone in Mondadori e alle relative vicende contrattuali. Tuttavia, dei picchi del *topic* "Proposte di pubblicazione e altri progetti", soltanto 4, 5 e 15 sembrano avere un'effettiva corrispondenza col contenuto delle lettere: nella lettera 4 (12 novembre 1946), Silone presenta a Mondadori Giulio Macchi, che svolgerà ruolo di intermediario per le proposte di pubblicazione dei suoi libri e riassumerà gli ultimi sviluppi contrattuali con il suo primo editore in Italia, Faro Editrice; Arnoldo Mondadori, nella lettera 5 (19 novembre 1946), riassume i punti fondamentali del suo colloquio con Macchi; nella lettera 15 (18 ottobre 1947), invece, minimizza le possibili conflittualità tra due diversi progetti di pubblicazione, la rivista di Guido Tonella *Pane e vino* e l'omonimo romanzo di Silone.

Tutte e tre le lettere presentano termini riconducibili al *topic* di riferimento ("pubblicazione", "progetto", "uscita"…). La lettera 12 (25 luglio 1947), invece, è un telegramma in cui Arnoldo invita Silone a un ricevimento in onore di Thomas Mann presso villa Mondadori a Meina, sul lago Maggiore, dove non si fa nessun accenno a possibili progetti editoriali: l'equivoco, probabilmente, è stato generato dalla testualità ridotta tipica della forma-telegramma che, ulteriormente minimizzata dalla pre-elaborazione del *corpus*, riduce drasticamente la base di dati utile a un'analisi accurata e induce l'algoritmo al fraintendimento, dettato dalla presenza del lemma "promessa" in un contesto verbale ridotto a soli cinque termini. Un simile errore si è verificato anche in merito al *topic* "Questioni contrattuali", dove, dei quattro picchi (lettere 9, 11, 13 e 16), il primo, che è da ritenersi inesatto – trattandosi di un messaggio di Alberto Mondadori in cui si allude a un servizio giornalistico del Pen Club, della cui sezione italiana Silone è stato presidente dal 1945 al 1959 –, corrisponde ancora una volta a un telegramma (in cui, di quattro parole risultanti, il solo termine "proposta" può giustificare questa rilevazione), mentre le altre tre lettere – soprattutto per la presenza di termini quali "contratto", "accordo", "diritti" –, sono effettivamente etichettabili sotto il *topic* "Questioni contrattuali", nella misura in cui Silone e Arnoldo Mondadori, soprattutto nella 11 e nella 13, si accordano sugli estremi contrattuali per la pubblicazione delle prime cinque opere di Silone con Mondadori (*Fontamara*; *Pane e vino*; *Seme sotto la neve*; *Scuola dei dittatori*; *Ed egli si nascose*). In un *corpus* testuale così strutturato, dove l'analisi è compiuta al "document level", a incidere nell'accuratezza del *topic modeling* non è dunque la dimensione della base di dati nel suo complesso (l'intero carteggio), quanto la lunghezza dei singoli documenti, che varia a seconda del contenuto del messaggio e della forma epistolare prescelta dal mittente. Quanto ai telegrammi, la loro incidenza nel totale del *corpus* (45 su 396 documenti), sebbene significativa, non può giudicarsi tale da compromettere l'efficacia dell'analisi nella sua totalità, considerando inoltre che non per tutte le occorrenze di questa tipologia di missiva l'*output* è da valutarsi nullo o errato.

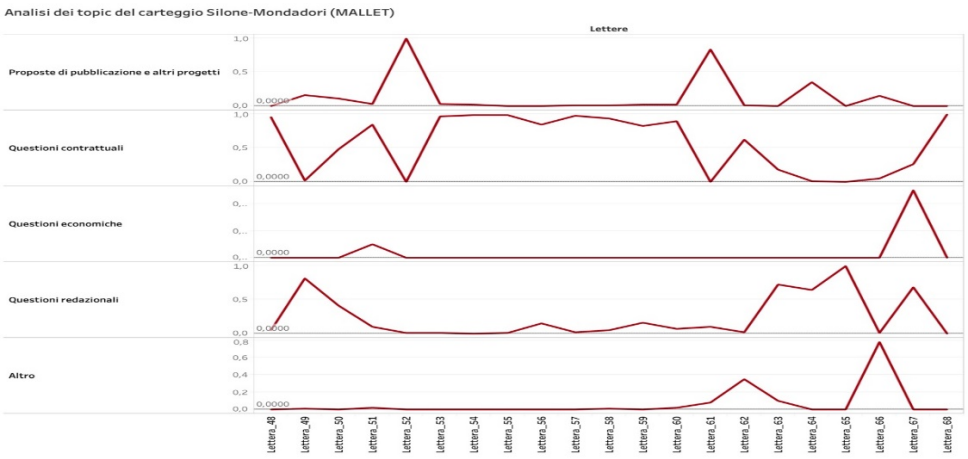

Figura 2: Analisi dei topic

In questo caso, invece, osserviamo un blocco di corrispondenza in cui, nell'arco dalla lettera 53 (27 febbraio 1952) alla 60 (25 giugno 1952), si segnala una preminenza del *topic* "Questioni contrattuali", a discapito degli altri che risultano del tutto o quasi assenti. Questa rilevazione risulta essere particolarmente accurata rispetto al reale contenuto delle lettere, giacché all'epoca Silone e Alberto Mondadori (insieme agli esponenti della segreteria della casa editrice) erano in stretta trattativa per trovare un accordo relativo alla pubblicazione di *Una manciata di more*. In questo caso, tuttavia, sarà utile osservare in parallelo gli sviluppi dei singoli grafici, per notare come la nostra scelta di applicare ai *topic* delle etichette volutamente "estensive", se da un lato riduce il rischio di produrre *output* inesatti, dall'altro aumenta la necessità di ricontestualizzarli a posteriori. In questo caso, ad esempio, il picco del *topic* "Questioni redazionali" delle lettere 63-65, pur corretto, riguarda *Una manciata di more* solo relativamente alla prima (19 febbraio 1953) e, soprattutto, alla terza lettera (23 febbraio 1953), con la quale Alberto Mondadori ragguaglia circa le vendite del romanzo pur allontanando per il momento l'eventualità di una ristampa, mentre nella lettera 64 (21 febbraio 1953), sempre a firma di Alberto Mondadori, il *focus* del discorso passa sulle valutazioni intorno a un libro segnalato da Silone (*Hexensabbat* di Alexander Weißberg-Cybulski), che non verrà poi pubblicato perché giudicato troppo voluminoso.

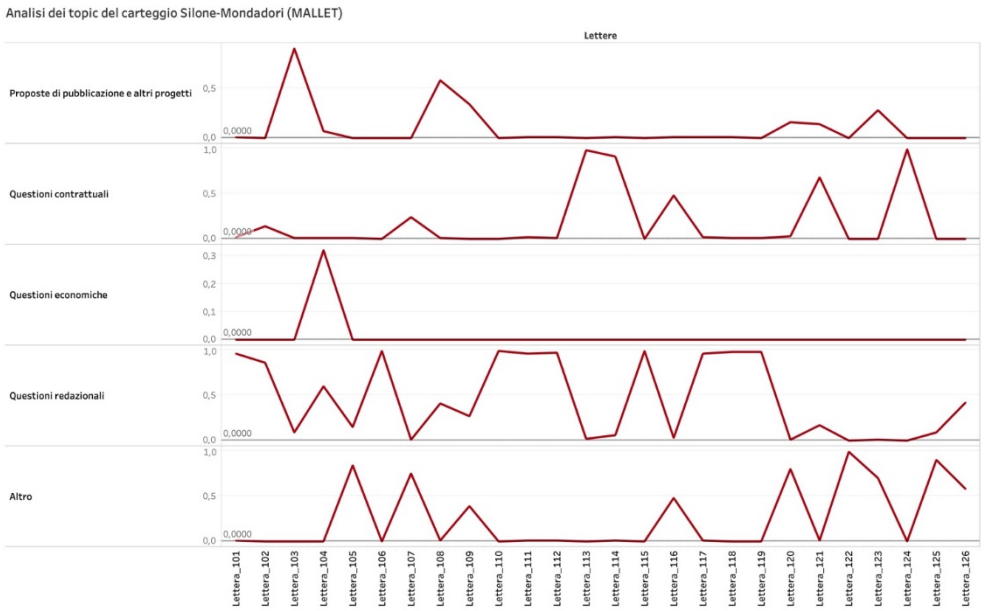

Figura 3: Analisi dei topic

In questo terzo grafico, invece, osserviamo l'andamento dei *topic* nel periodo di corrispondenza (11 novembre 1955-30 aprile 1957) incentrato sulla pubblicazione del *Segreto di Luca*. Anche in questo caso, è da considerarsi pienamente coerente la prevalenza del *topic* "Questioni redazionali", e i picchi registrati (soprattutto le lettere 110-112, 115 e 117-119) confermano la

correttezza dell'analisi, essendo comunicazioni tutte riferibili, per motivazioni differenti, al lavoro redazionale che precede la pubblicazione di un volume: nello specifico, da 110 a 112 Silone e Alberto Mondadori concordano le modalità di invio e di lavorazione delle bozze, cercando di coordinarle rispetto agli impegni di Silone; in 115, invece, è lo scrittore che illustra le principali correzioni che ha apportato alle bozze, segnalando punto per punto i suoi interventi; infine, da 117 a 119, Alberto Mondadori comunica a Silone l'avanzamento dei lavori al volume, che, dopo la fase di correzione, è passato alla confezione. È peraltro interessante mettere a paragone l'andamento del *topic* "Questioni redazionali" con quello dei *topic* "Proposte di pubblicazione e altri progetti" e "Questioni contrattuali": si ottiene, infatti, un'efficace rappresentazione dell'evoluzione tematica attorno alle fasi di pubblicazione di un libro, introdotte dapprima da discussioni preliminari e manifestazioni d'interesse (*topic* "Proposte di pubblicazione e altri progetti", lettera 103) e seguite dal lavoro redazionale che si alterna alla definizione dei termini contrattuali (*topic* "Questioni contrattuali", lettere 113 e 114).

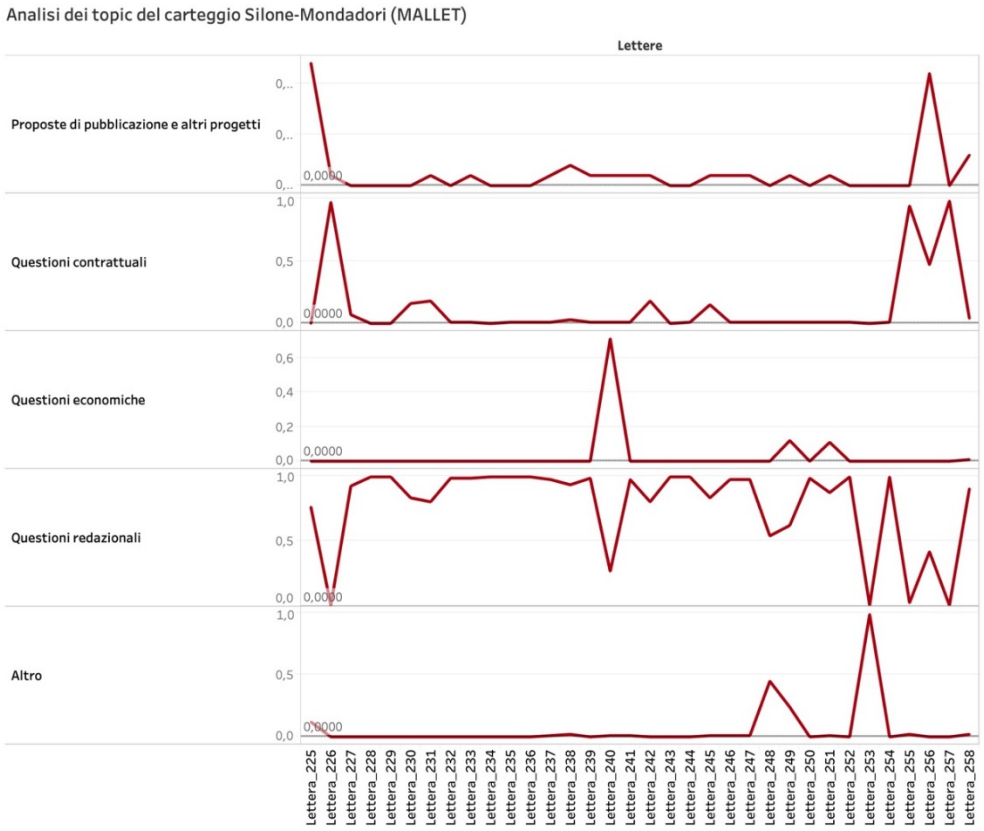

Figura 4: Analisi dei topic

In questo ultimo segmento, infine, la netta prevalenza del *topic* "Questioni redazionali" è da ascrivere non tanto ai lavori che precedono la prima pubblicazione di un volume quanto a una serie di riedizioni sollecitate da Silone, in particolare di *Una manciata di more* e di *Vino e pane*, e alle conseguenti revisioni per mano dello scrittore e della redazione Mondadori.

Come si evince già da questi estratti, è evidente come il *topic* "Questioni redazionali" sia, tendenzialmente, il più diffuso nell'intero carteggio. Ciò dipende essenzialmente da due fattori: *in primis*, la sua effettiva continuità, considerando che – mentre le questioni contrattuali e le proposte di pubblicazione, ad esempio, tendono in genere a esaurirsi in brevi scambi di una o due lettere (con alcune eccezioni, come si è osservato nel caso della Figura 2) – i processi redazionali si caratterizzano per una maggior durata nel tempo, che, di fatto, li rende l'argomento principale dello scambio tra scrittore ed editore; d'altro lato, l'etichetta "Questioni redazionali", rispetto alle altre a caratterizzazione esplicita (escludendo la generica "Altro"), rappresenta volutamente un dominio tematico più estensivo: in essa, infatti, sono confluiti non soltanto i pezzi epistolari incentrati sui lavori di impaginazione e correzione dei volumi, ma anche quelli in cui sono presenti riferimenti alla messa in produzione e alla calendarizzazione dei libri, o, ancora, ad altri progetti (come nel citato caso di *Hexensabbat* o riguardo alla pubblicazione delle poesie di Alberto Mondadori su *Tempo presente*) che comunque afferiscono all'elaborazione di prodotti editoriali.

2.2. Sentiment analysis

Per quanto riguarda l'analisi del *sentiment* del carteggio Silone-Mondadori, si è scelto di suddividerla tra il calcolo della polarità (positiva/negativa) e l'individuazione delle emozioni predominanti. Per compiere l'operazione si è utilizzata la libreria R *Syuzhet*,[17] sviluppata da Matthew Jockers a partire dal 2015. Come lessico "emozionale", la versione italiana dell'NRC Emo-Lex di Saif Mohammad, implementata di *default* nella libreria. La scelta di Syuzhet e dell'Emo-Lex è stata in parte condizionata dalla scarsità di strumenti per la *sentiment analysis* di testi in lingua italiana. Questo tipo di analisi computazionale, se ad approccio *lexicon based*, è infatti fortemente vincolata alla lingua per cui gli strumenti sono predisposti, e la gran parte di essi – almeno per quanto riguarda quelli disponibili *open source* – sono ottimizzati per la lingua inglese. Inoltre, rispetto a quanto osservato per il *topic modeling*, dove il pacchetto MALLET gode di ampia diffusione fra le varie discipline ed è utilizzato in progetti che esulano dall'ambito umanistico, relativamente alla *sentiment analysis* si è riscontrato un forte iato tra il proliferare di *tool* proprietari (MonkeyLearn, Meaning Cloud, Aylien...), spesso facenti parte di più ampie *software pipeline* per la *text analysis*, e soluzioni *tailor made* altamente specializzate, difficilmente riutilizzabili al di fuori del progetto originario e senza un elevato grado di competenza informatica.

Diversamente da quanto realizzato per predisporre il *corpus* di lettere al *topic modeling*, per la *sentiment analysis* non è stato richiesto un particolare lavoro preliminare sui testi. Se infatti MALLET agisce in modo inclusivo, ricomprendendo ogni termine presente nel *corpus* all'interno

[17] <https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html>(04/2021).

degli *output* e necessitando quindi una preparazione che ne migliori gli esiti, la *sentiment analysis*, e in particolare quella eseguita con approccio *lexicon based*, procede per esclusione, ossia individuando – tra l'insieme dei dati testuali – quelli che concorrono a determinare i risultati auspicati. Tuttavia, per agevolare il *matchmaking* tra i termini presenti nel *corpus* e quelli compresi nelle liste dei lessici emozionali, si è scelto comunque di sottoporre i testi a un processo di *lower casing*, riducendo tutte le lettere alla forma minuscola, e di eliminare cifre e punteggiatura. Il procedimento seguito, sempre utilizzando Lexos, ricalca quanto già illustrato per il *topic modeling*.

Successivamente, lo stesso *input* è stato convertito una tabella CSV priva di intestazioni per l'importazione in ambiente R.

Sebbene la libreria predisposta da Matthew Jockers si fondi su di un modello *keyword based*, offre la possibilità di processare i dati utilizzando più lessici emotivi, già integrati di *default* o importabili dall'utente. L'utilizzo di NRC EmoLex ha permesso di estrarre informazioni relative alla polarità dei documenti e alla diffusione in essi delle otto principali emozioni.[18]

Per eseguire Syuzhet si è utilizzato RStudio, un *software open source* che predispone un ambiente di sviluppo per il linguaggio R, fornendo degli strumenti per l'installazione e la gestione dei pacchetti e delle librerie e per l'*upload* dei dati.

L'*output* ha presentato gli *score* delle otto principali emozioni affiancati ai dati relativi alla polarità.

Per la polarità, a un confronto con la base di dati testuale, Syuzhet e, soprattutto, NRC EmoLex, pur con dei limiti nella gestione delle negazioni, hanno mostrato una discreta efficacia nell'individuazione dei testi con andamento neutrale o negativo, non incorrendo in particolari fraintendimenti. Non sono invece stati altrettanto efficaci nell'individuare i picchi di negatività – presenti, nel carteggio, soprattutto in relazione a lamentele da parte di Silone per i ritardi nelle uscite o nelle ristampe. A ciò è tuttavia risultato di efficace integrazione l'affiancamento dell'analisi delle otto singole emozioni, che è parsa in grado di offrire un quadro più accurato della reale distribuzione emozionale nei contenuti delle lettere.

In modo analogo a quanto fatto per il *topic modeling*, presentiamo alcuni grafici estrapolati dagli *output* della *sentiment analysis* del carteggio. In questo caso, come si è detto sopra, la lettura dei grafici deve essere condotta tenendo conto dei limiti intrinseci all'approccio *lexicon based*, in particolare nella gestione delle negazioni: queste tipologie di strumento, infatti, efficaci nell'individuazione della polarità positiva, risultano meno precise nella rappresentazione dei picchi di polarità negativa, che, quando presenti, non surclassano con evidenza l'alternativa ma, al massimo, la equiparano. Questo accorgimento, sebbene in apparenza indebolisca l'efficacia dello strumento nel suo complesso, è comunque sufficiente a renderne gli *output* pienamente fruibili e a valutarne la precisione nelle rilevazioni.

[18] Rabbia, paura, tristezza, disgusto, sorpresa, anticipazione, fiducia e gioia, secondo il modello della "wheel of emotions" di Robert Plutchik.

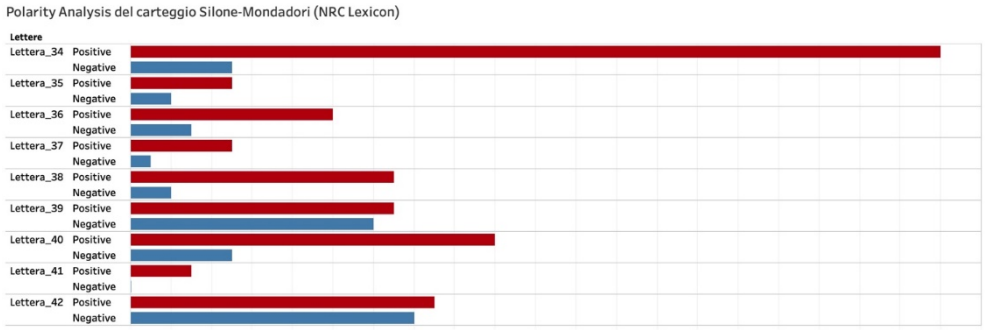

Figura 5: Polarity Analysis

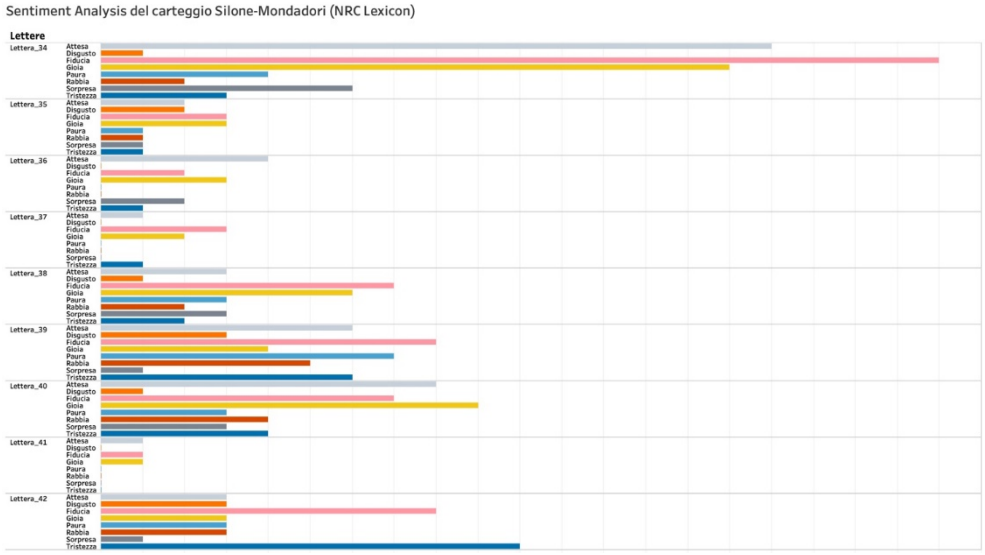

Figura 6: Sentiment Analysis

Ad esempio, il blocco di lettere dalla 34 alla 42 cui corrispondono i due grafici qui sopra è aperto da un documento fortemente positivo: si tratta, infatti, della lettera (26 agosto 1949) con cui Arnoldo Mondadori annuncia ai suoi autori il nuovo piano editoriale per le collane della casa editrice, affidate a grandi esperti dei vari settori. La positività della lettera è dovuta ai toni entusiastici della comunicazione di Arnoldo, e, pertanto, si registrano picchi di "fiducia" e "gioia", nonché di "attesa", in questo caso lieta, da leggersi come progettualità e fiduciosa speranza nel prossimo futuro dell'azienda. Nel blocco, tuttavia, si registrano anche due documenti a polarità negativa, se si considerano tali, come dicevamo, quegli *output* in cui i due valori si equivalgono. Le lettere 39 (17 febbraio 1950) e 42 (26 maggio 1950) ruotano, infatti, attorno all'insoddisfazione di Silone per i ritardi nella pubblicazione del *Seme sotto la neve*. Nella prima, è Silone che, in margine a delle riflessioni sull'opportunità di scrivere una prefazione a

The Way Out di Uys Krige, torna sul romanzo in sospeso, riguardo al quale aspetta notizie da più di sei mesi, al punto da dichiarare la propria delusione verso il trattamento subìto in Italia: "La mia presunzione non arriva fino al punto di sognare di avere con un editore italiano i rapporti di amicizia e confidenza che ho con i nominati editori esteri". La lettera 42 (26 maggio 1950) è invece la risposta di Alberto Mondadori che, infatti, registra un picco del valore "tristezza", giacché l'editore si dice molto dispiaciuto per quanto avvenuto, ma, al contempo, fiducioso che tutto potrà presto risolversi nel migliore dei modi (da cui deriva infatti un elevato valore di "fiducia").

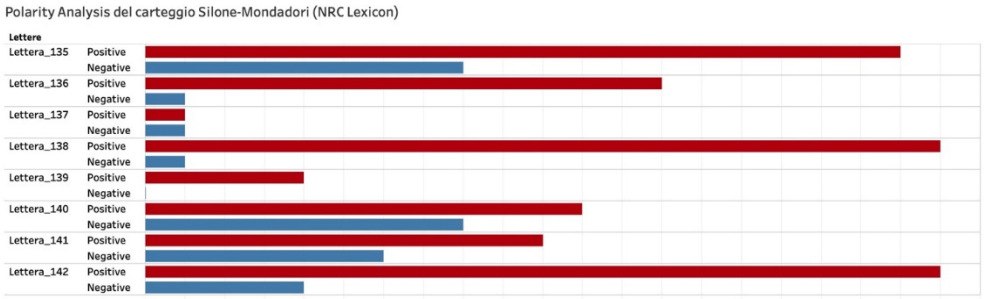

Figura 7: Polarity Analysis

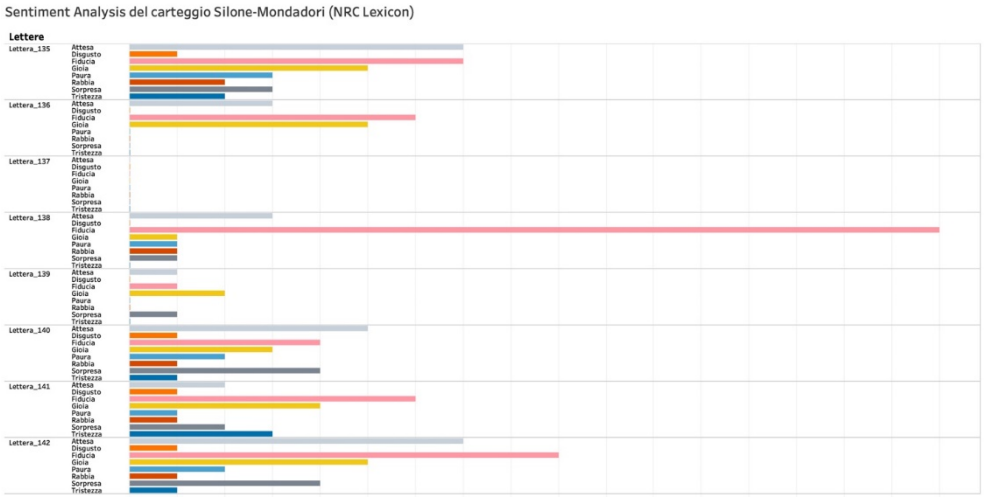

Figura 8: Sentiment Analysis

In questi secondi grafici (7-8) osserviamo invece un blocco di lettere ad andamento tendenzialmente positivo: la lettera 135 (8 maggio 1958), ad esempio, corrisponde a una proposta di collaborazione da parte di Alberto Mondadori a un progetto di almanacco letterario, i cui toni, come da consuetudine per i tentativi di coinvolgimento in nuove iniziative editoriali, sono ottimistici; la lettera 138 (30 ottobre 1958), invece, è ancora una comunicazione agli autori

firmata da Arnoldo Mondadori e tesa a riassumere le ultime modifiche ai quadri direttivi dell'azienda, contrassegnate da una generica fiducia per il futuro della casa editrice; nella 142 (11 gennaio 1959), infine, Silone comunica ad Arnoldo di aver proposto il suo nuovo racconto "La volpe" a suo figlio, sperando che non subentrino conflitti con i diritti di opzione di Mondadori, e che un giovane e capace regista (Amato Bottazzi) si appresta a realizzare una riduzione televisiva del *Segreto di Luca*: come si evince dalla sezione di grafico corrispondente (Figura 8), sebbene la lettera sia prevalentemente orientata a una diffusa positività per entrambi i progetti, non mancano elementi di "attesa" legati all'esito delle valutazioni in corso su "La volpe".

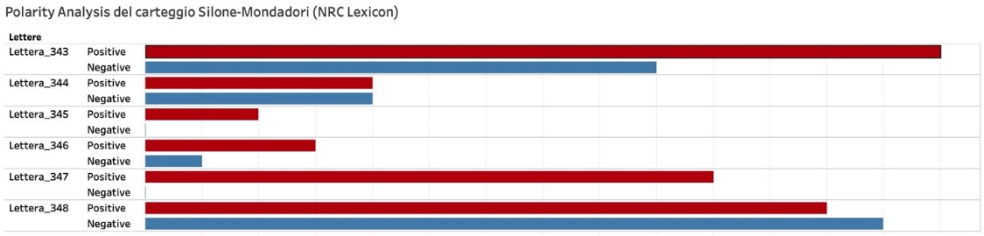

Figura 9: Polarity Analysis

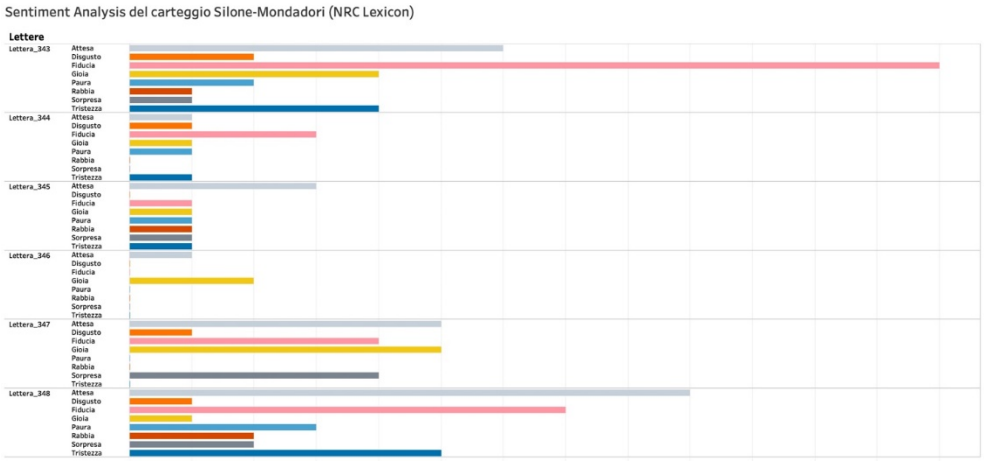

Figura 10: Sentiment Analysis

Analizzando la Figura 9, infine, troviamo due picchi di polarità negativa (lettere 343 e 348) che, al confronto con i rispettivi *output* nella Figura 10, sembrano necessitare un'ulteriore contestualizzazione. Si tratta di due lettere di argomento differente: nella 343 (9 giugno 1970), Silone scrive a Domenico Porzio per notificare un errore nella copertina dell'edizione Oscar di *Fontamara*; nella 348 (16 dicembre 1970), invece, la segreteria editoriale Mondadori motiva a Silone i ritardi nella chiusura degli accordi con la Minerva Italica per un'antologia scolastica dalle sue opere. Osservando i grafici delle singole emozioni, noteremo tuttavia – al netto di una forte

presenza del valore "tristezza", che, come si è visto, connota più genericamente testi in cui gli interlocutori esprimono il loro rammarico per qualche inconveniente – picchi di valori "positivi", soprattutto nella 343, dove questa discrepanza è da imputare alle due righe di apertura, fortemente positive ma estranee all'argomento centrale nella lettera, con cui Silone ringrazia Porzio per un intervento sulla rivista *Il Dramma*. Ciò dimostra, ancora una volta, come questi strumenti, se utilizzati su documenti a polarità apparentemente "ibrida", per quanto efficaci nel cogliere la presenza di una determinata sfumatura "emozionale", non sempre sono capaci di rappresentarne l'effettiva incidenza nel più ampio contesto dell'intero documento, presentando valori che necessitano comunque di un'ulteriore lettura per interpretare correttamente l'andamento del testo.

Text analysis ed edizioni di carteggi letterari: proposte di integrazione

Intorno alla *text analysis* applicata agli studi letterari *in toto* esiste un vivo dibattito. Pur con le criticità che sono state evidenziate, come ha notato Johanna Drucker, tuttavia, "text analysis has value" ([3]: 633), a patto di considerare i risultati conseguiti – siano essi semplici indici di frequenza o *output* che ambiscono a essere semanticamente più complessi – non come approdi ermeneutici ma come spunti, suggerimenti iniziali per tracciare nuove vie interpretative o per verificare la veridicità di ipotesi mediante modelli matematici. Sebbene la validità di certe critiche permanga, esse sembrano tendenzialmente fondarsi sull'erroneo presupposto di sopravvalutare – o, meglio – di fraintendere la funzione strumentale delle risorse informatiche, che invece

> […] are just that, tools, and their value is only as good as the models on which they are made, the protocols used to implement those models, and the qualifications that can be attached to the results. But these tools don't read, except in the most mechanical sense. ([3]: 633)

Si tratta di una posizione in parte affine a quanto affermato da Andrew Piper, secondo cui la "quantitative evidence can be a valuable *complement* to assist scholars in the process of generalization and help make our evidentiary claims more credible" ([18]: 10; corsivo nostro). In questa prospettiva, l'utilizzo di *tool* informatici per integrare le prassi di analisi testuale con ulteriori strumenti non può che essere ritenuto un valore aggiunto alla disciplina, nella misura in cui lo sia tutto ciò che favorisca (o che getti le basi per) un incremento conoscitivo.

Quanto all'applicabilità su carte d'archivio, si è cercato di dimostrare nel commento ai grafici proposti che gli strumenti di *text analysis*, se utilizzati e interpretati, nei risultati, opportunamente, possono quindi fornire senza dubbio un primo, efficace spunto per una rappresentazione dell'andamento tematico ed "emozionale" di un *corpus* testuale, agevolando sia un approccio "panoramico" (*distant*), ossia consentendo uno sguardo orientativo sull'intero complesso epistolare, che una consultazione nel dettaglio dei singoli documenti, per constatare la singola distribuzione tematica ed "emozionale" nel più ampio contesto dell'intero carteggio.

Tuttavia, non è stato possibile ritenere gli *output* così ottenuti aprioristicamente validi, ma anzi è stato necessario, riandando ai testi delle lettere, sottoporli a operazioni di ricognizione, contestualizzazione e interpretazione, senza le quali avremmo rischiato di incorrere in fraintendimenti. Questa necessità, se da un lato ne ha limitato il potenziale "apodittico", dall'altro ci ha confermato quello che è da sempre uno degli assunti fondamentali delle *digital humanities*: le risorse informatiche "don't read, except in the most mechanical sense" ([3]: 633), e non possono sostituirsi all'intelligenza critica dell'umanista, ma, affiancando ai metodi tradizionali strumenti alternativi, possono supportarla per affinarla e ampliarne ulteriormente le prospettive di intervento. Ciò vale, altresì, per l'idea che queste tecniche possano produrre un contributo euristico, generando un senso aggiuntivo altrimenti inconoscibile: ancora una volta, si rischierebbe così una visione "magica" di questi strumenti, sminuendone l'effettivo contributo, che, se nel caso della loro applicazione a testi letterari può consistere nell'individuazione di spunti di lettura alternativi, favoriti dall'approccio macroscopico, anche nel nostro tentativo di utilizzo integrato all'edizione di materiali di archivio – dove le esigenze ermeneutiche sono, ovviamente, ridotte e gli *output* ricalcano la linearità di testi che non presentano particolari complessità – sembra comunque esistere.

Il *topic modeling* e la *sentiment analysis* – così come altri *tool* per la *text analysis* –, sulla base di quanto emerso dalla nostra lettura computazionale, possono avere infatti una finalità "descrittiva", se utilizzati per rappresentare l'andamento tematico ed emozionale di un *corpus* testuale, senza l'intenzione di sostituirli al discorso critico dell'esperto "umano", così da fornire strumenti per classificare e segmentare un'ampia mole testuale, dando al lettore-utente la possibilità di fruirne muovendo da una situazione discretizzata rispetto al flusso di informazioni continuo che deriva da un *corpus* non processato. In seconda istanza, il valore di questi *output* può essere ravvisato in una funzione "predittiva", permettendo, una volta costituito un modello di analisi, di ripetere l'operazione su altri *corpora* e, in base ai risultati, eseguire una comparazione e predire il contenuto in caso di andamenti più o meno similari. Utilizzi, tuttavia, che non devono avere pretesa di produrre un valore conoscitivo "assoluto", quanto di integrare l'edizione con una serie di informazioni macroscopiche affiancate, come strumenti di corredo, al commento "*close*", per rendere il più efficiente possibile l'esperienza di lettura, un "valuable complement" ([18]: 10) al prodotto editoriale complessivo.

References

[1] Cambria, Eric and Amir Hussain. 2015. *Sentic Computing. A Common-Sense-Based Framework for Concept-Level Sentiment Analysis*. Cham: Springer.

[2] Da, Nan Z. 2019. "The Computational Case against Computational Literary Studies." *Critical Inquiry* 45: 601-39.

[3] Drucker, Johanna. 2017. "Why Distant Reading Isn't." *PMLA* 132, 3: 628-35.

[4] Fenu, Cristina. 2017. "Sentiment Analysis d'autore: l'epistolario di Italo Svevo." In *AIUCD 2017 Conference. Il telescopio inverso: big data e distant reading nelle discipline umanistiche*, eds. Fabio Ciotti, Gianfranco Crupi, 133-9. Firenze: Associazione per l'Informatica Umanistica e la Cultura Digitale. http://doi.org/10.6092/unibo/amsacta/5885.

[5] Graham, Shawn, Ian Milligan. 2012. "Review of MALLET, Produced by Andrew Kachites McCallum." *Journal of Digital Humanities* 2, 1: 73-5.

[6] Hammond, Adam. 2017. "The Double Bind of Validation: Distant Reading and the Digital Humanities' 'trough of disillusionment'." *Literature Compass*. https://doi.org/10.1111/lic3.12402.

[7] Hotson, Howard, Wallnig, Thomas, eds. 2019. *Reassembling the Republic of Letters in the Digital Age. Standards, Systems, Scholarship*. Göttingen: Göttingen University Press. https://doi.org/10.17875/gup2019-1146.

[8] Jänicke Stefan, Franzini Greta, Cheema Muhammad Faisal, et al. 2015. "On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges." In *Eurographics Conference on Visualization. STARs – State of The Art Reports*, ed. by Rita Borgo, Fabio Ganovelli and Ivan Viola, 83-104. The Eurographics Association.

[9] Jannidis, Fotis. 2020. "On the Perceived Complexity of Literature. A Response to Nan Z. Da." *Journal of Cultural Analytics* 1. https://doi.org/10.22148/001c.11829.

[10] Jockers, Matthew L. Macroanalysis. 2013. *Digital Methods and Literary History*. Urbana-Chicago-Springfield: University of Illinois Press.

[11] Kokensparger, Brian. 2018. *Guide to Programming for the Digital Humanities: Lessons for Introductory Python*. Cham: Springer.

[12] Liu, Bing. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. New York: Cambridge University Press.

[13] Malfatti, Stefano. 2020. "The Digital Edition of Alcide De Gasperi's Correspondence. Some Reflections on an Ongoing Project." *Jlis* 11, 1: 89-105. http://dx.doi.org/10.4403/jlis.it-12599.

[14] Moretti, Franco. 2000. "Conjectures on World Literature." *New Left Review* 1: 54-68.

[15] Moretti, Giovanni, Rachele Sprugnoli, Stefano Menini, et al. 2016. "ALCIDE: Extracting and Visualising Content from Large Document Collections to Support Humanities Studies." *Knowledge-Based Systems* 111: 100-12.

[16] Moretti, Giovanni, Rachele Sprugnoli, Sara Tonelli. 2015. "Digging in the Dirt: Extracting Keyphrases from Texts with KD." In *Proceedings of the Second Italian*

Conference on Computational Linguistics CLiC-it 2015, a cura di Cristina Bosco, Sara Tonelli, Fabio Massimo Zanzotto, 198-203. Torino: Accademia UP.

[17] Moretti, Giovanni, Rachele Sprugnoli, Sara Tonelli. 2018. "LETTERE: LETters Transcription Environment for REsearch." In *Patrimoni culturali nell'era digitale. Memorie, culture umanistiche e tecnologia / Cultural Heritage in the Digital Age. Memory, Humanities and Technologies*, a cura di Daria Spampinato, 207-9. Bologna: Associazione per l'Informatica Umanistica e la Cultura Digitale.

[18] Piper, Andrew. 2019. "Do We Know What Are We Doing?." *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.11826.

[19] Rhody, Lisa M. 2012. "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2, 1: 19-35.

[20] Schofield, Alexandra, Måns Magnusson, David Mimno. 2017. "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*, vol. II, Short Papers, ed. by Mirella Lapata, Phil Blunsom and Alexander Koller, 432-6. Valencia: Association for Computational Linguistics.

[21] Shillingsburg, Peter L. 2006. *From Gutenberg To Google. Electronic Representations of Literary Texts*. New York: Cambridge University Press.

[22] Shillingsburg, Peter L. 2017. *Textuality and Knowledge. Essays*. University Park: The Pennsylvania State University Press.

[23] Sinclair, Stéfan, Geoffrey Rockwell. 2012. "Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies". In *Digital Humanities Pedagogy: Practices, Principles and Politics*, ed. by Brett D. Hirsch, 241-63. Cambridge: Open Book Publishers.

[24] Tonelli, Sara, Rachele Sprugnoli, Giovanni Moretti, et al. 2020. "Epistolario De Gasperi. National Edition of De Gasperi's Letters in Digital Format." In *La svolta inevitabile. Sfide e prospettive per l'informatica umanistica*, a cura di Cristina Marras, Marco Passarotti, Greta Franzini, et al., 253-59. Milano: Università Cattolica del Sacro Cuore.

[25] Underwood, Ted. 2017. "A Genealogy of Distant Reading." *Digital Humanities Quarterly* 11, 2.

[26] Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

[27] Underwood, Ted. 2019. Blog article. in *Computational Literary Studies: A Critical Inquiry Online Forum*. (<https://critinq.wordpress.com/2019/04/02/computational-literary-studies-participant-forum-responses-day-2-4/>).