Enthymema XXX 2022

Macchine per leggere: promuovere la lettura con il *distant reading*

Fabio Ciotti e Alberto Baldi
Università di Roma "Tor Vergata"

Abstract – Questo articolo presenta il progetto *Macchine per leggere*, nato dalla collaborazione tra il Dipartimento di Studi letterari, filosofici e di storia dell'arte dell'Università di Roma "Tor Vergata" e il Centro per il Libro e la Lettura del MiC. Scopo del progetto è la realizzazione di un ambiente digitale (*desktop* e *mobile*) che introduca gli studenti della scuola secondaria di secondo grado alla conoscenza e all'utilizzo delle tecniche di analisi computazionale dei testi, qui proposte come spunto per accostarsi alla lettura dei classici della letteratura italiana (tra gli altri che saranno resi disponibili sul sito, *Il piacere*, *I Malavoglia*, *Il fu Mattia Pascal*…). L'approccio del *distant reading* e gli strumenti informatici per l'analisi e la rappresentazione dei *corpora* e per il *text mining* (*word cloud*, indici di frequenza di termini e sintagmi, *topic modeling*, *sentiment* e *network analysis*), uniti ad altre risorse come la geolocalizzazione su mappe digitali dei luoghi letterari, saranno presentati sia come metodo per simulare in ambiente virtuale *reading strategies*, che come modelli per integrare le prassi ermeneutiche tradizionali – *close reading*. In una sezione del sito dedicata, gli algoritmi di cui si darà dimostrazione saranno a disposizione degli studenti in forma di *web app*, così che possano sperimentare in autonomia e su altri testi in loro possesso l'approccio *distant* proposto nell'ambito del progetto.

Parole chiave – Reading strategies; Promozione della lettura; Text analysis; Distant reading; Classici della letteratura italiana.

Abstract – We present *Macchine per leggere* project, a collaboration between the Department of Literary, Philosophical and Art History Studies of the university of Rome "Tor Vergata" and the MiC Center for Books and Reading. The project aims at creating a digital environment (desktop and mobile) that introduces secondary school students to the knowledge and use of computational text analysis as a cue for approaching the reading of classics of Italian literature (among others that will be made available on the site, Il piacere, I Malavoglia, Il fu Mattia Pascal...). The distant reading approach and a set of computer tools for corpus analysis and text mining (word cloud, term and syntagma frequency indices, topic modeling, sentiment, and network analysis), combined with other resources such as geolocation on digital maps of literary places, will be presented both as a method for simulating reading strategies in a virtual environment and as models for integrating traditional hermeneutic

practices-close reading. In a dedicated section of the site, the various tools will be available to students in the form of a web application so that they can experiment independently and on other texts in their possession with the distant approach proposed as part of the project.

Keywords – Reading strategies; Reading Promotion; Text Analysis; Distant Reading; Classics of Italian Literature.

Ciotti, Fabio, e Alberto Baldi. "*Macchine per leggere*: promuovere la lettura con il *distant reading*". *Enthymema*, n. XXX, 2022, *La letteratura e la rete. Alleanze, antagonismi, strategie*, a cura di Stefano Ballerio e Marco Tognini, pp. 173-92.

http://dx.doi.org/10.54103/2037-2426/19557

https://riviste.unimi.it/index.php/enthymema

ISSN 2037-2426

Macchine per leggere: promuovere la lettura con il *distant reading*

Fabio Ciotti e Alberto Baldi
Università di Roma "Tor Vergata"

1. Introduzione

Lo studio e la trasmissione pedagogica delle competenze testuali e delle tecniche della lettura, orientate a stimolare nei giovani l'interesse per testi culturalmente complessi e ad aiutarli nell'attuare un approccio critico che consenta loro di fruire a pieno dell'esperienza di lettura, è un ambito ampiamente approfondito dalle scienze pedagogico-didattiche. Tuttavia, soprattutto alla luce dell'impatto dei nuovi *media* digitali sul rapporto tra giovani generazioni e testo scritto – per opinione diffusa assai negativo, sebbene manchino sostanziali evidenze sperimentali a conferma –, anche gli studiosi di letteratura e storia culturale sono ormai chiamati a dare il loro contributo teorico e pratico alla definizione di nuove strategie per promuovere la lettura di testi letterari complessi, soprattutto di quei classici che, pur essendo ancora centrali nei percorsi di studio scolastico e universitario, sono così distanti dall'esperienza quotidiana dei più giovani. Questa necessità ha dato stimolo a un filone di studi innovativi, che indagano le possibilità di una didattica digitale della letteratura italiana (Riva; Magherini; Giusti) o che studiano i margini di coesistenza tra i classici e i nuovi linguaggi (emblematica, in tal senso, la rassegna compiuta da Luca Sarti sulle riscritture di *Pinocchio* in epoca digitale, tra *tweet* ed *emoji* – Sarti). D'altronde sul ruolo pedagogico dei nuovi *media* digitali nella didattica della letteratura verte una intera corrente dei *Digital Literary Studies*, quella legata allo studio e all'applicazione delle tecniche ipertestuali, a far data dagli storici esperimenti di George Landow negli anni 80 alla Brown University.

L'idea di una testualità decentrata, plurale e aperta al contributo dei lettori/autori che è alla base delle tecnologie ipertestuali sembrerebbe prestarsi in modo quasi naturale a una pratica didattica che aiuti gli studenti, soprattutto universitari, a entrare dentro un testo, e decostruirlo, esplorandone le numerose riserve di senso e di contestualizzazione. Tuttavia, dopo gli iniziali entusiasmi degli anni Novanta questa prospettiva è stata progressivamente abbandonata, mentre la comunità scientifica degli umanisti digitali iniziava a rivolgere la sua attenzione verso altre tecnologie e ambiti di ricerca e sperimentazione, come le biblioteche e gli archivi digitali, le campagne di digitalizzazione massive, l'edizione scientifica digitale dei testi, le *spatial humanities* e soprattutto l'esplorazione di tutte le possibilità scientifiche, critiche e comunicative offerte dalla rete e soprattutto dal Web. Non è questa la sede per esplorare il territorio delle *Digital Humanities*, ormai vastissimo, e rimandiamo per una visione d'insieme alle due ponderose edizioni del *Companion to Digital Humanities* (Schreibman et al.).

Ci concentriamo qui, piuttosto, su un'ampia provincia di tale territorio che ha molto a che fare con i testi, la letteratura e la cultura: l'analisi computazionale dei testi e dei fenomeni culturali. Questo campo ha una lunga tradizione, ma ha visto una forte espansione negli ultimi due decenni, spinto dall'emergenza del paradigma metodologico del *distant reading*, formulato da Franco Moretti per la prima volta nel saggio "Conjectures for World Literature", e successivamente divenuto quadro di riferimento metodologico per una ingente quantità di studi e sperimentazioni (Underwood; Jockers; Gavin). Il successo di questo paradigma negli studi letterari digitali è stato propiziato non solo dalla base teorica 'interna' ma anche da due impetuosi

processi, per così dire, 'esterni': la disponibilità di vasti archivi di testi digitali e di metadati, prodotti nelle campagne di digitalizzazione su ampia scala condotte nell'ultimo decennio del secolo scorso; lo sviluppo e la diffusione, anche in ambito umanistico, di tecniche e strumenti di analisi dei dati (meglio se *big data*), quali il *machine learning*, che fino a inizio millennio erano ancora in uno stadio prototipale. Queste tecniche, applicate su collezioni testuali, permettono di fare emergere strutture, regolarità e *pattern* che giocano un importante ruolo esplicativo nella comprensione dei processi letterari come l'evoluzione dei generi, la diffusione di uno stile, l'intertestualità, la presenza di temi e contenuti ricorrenti in un dato periodo storico-letterario. La critica e la storiografia letteraria, dunque, possono 'giustificare' le loro spiegazioni e generalizzazioni sulla base dell'evidenza empirica, fornita dall'analisi dei dati, e non solo delle capacità soggettive di intuizione e argomentazione dello studioso. Fin qui, la teoria e la metodologia, oltre alla storiografia e alla critica letteraria. Ma ci siamo chiesti: queste tecniche e questi metodi possono trovare spazio nella didattica della letteratura e, magari, anche nella promozione della lettura?

La rilevanza dei vari metodi di *text analysis* nella didattica della letteratura, specie al livello dell'istruzione superiore, appare più che plausibile, come testimonia la recente letteratura dedicata a questo tema, nel quadro di una crescente attenzione verso la dimensione pedagogica delle *Digital Humanities* (Hirsch; Croxall, Jakacki). Molto più controversa, se non provocatoria, l'idea che un approccio che considera il testo letterario come un set di dati quantitativi da elaborare mediante processi statistico-matematici, implementati in programmi per computer, possa aiutare a promuovere la lettura (per non dire che Moretti, teorizzando il *distant reading*, afferma testualmente che dovremmo "imparare a non leggere…"). Ma, forse, facendo avvicinare i giovani ai testi da una prospettiva così straniante, dove dalle pagine dei romanzi derivano grafici, misure, reti e modelli predittivi, non potrebbe scattare un meccanismo di attenzione in chi vive la dimensione digitale, la visualità e la sintesi come condizione antropologica primaria?

2. Il progetto *Macchine per leggere*

Macchine per leggere rappresenta un tentativo esplorativo e altamente sperimentale di rispondere a questa domanda. Il progetto, dal punto di vista istituzionale, nasce come una collaborazione tra il Dipartimento di Studi Letterari, Filosofici e di Storia dell'Arte dell'Università di Roma "Tor Vergata" e il Centro per il Libro e la Lettura del MiC (CEPELL, https://cepell.it),[1] e muove dall'idea che anche le tecnologie per l'analisi automatica dei testi possano ricoprire un ruolo di rilievo in questo tentativo di promuovere la lettura dei classici della letteratura in epoca digitale.

Tecniche e algoritmi per l'analisi statistica dei *corpora* (elaboratori di indici di frequenza, di concordanze, di collocazioni…) o per più raffinate operazioni di *text mining* e *information extraction* (*topic modeling*, *sentiment analysis*, analisi stilometrica computazionale…) sono ormai piuttosto diffuse anche nell'ambito dell'informatica umanistica italiana, con successi critici e ricezione variabili. In parallelo, stanno iniziando a svilupparsi dei modelli didattici che utilizzano queste tecniche come dispositivi utili a mediare l'incontro tra i giovani e i testi letterari, da un lato agevolando gli aspetti più meccanici della lettura, dall'altro presentandosi come risorse 'ludiche' per accostarsi ai libri in modo diverso, inusuale. È il caso, ad esempio, dell'esperimento compiuto in una classe del liceo Liceo Scientifico Galileo Galilei di Trento (Valitutti e Dalla Torre), in cui gli studenti sono stati guidati nell'analizzare il *sentiment* di *Io non ho paura* di Niccolò

[1] In particolare, promotore del progetto presso il CEPELL è stato il Dott. Angelo Piero Cappello, direttore dello stesso e acuto studioso dell'opera dannunziana. La recente edizione de *Il piacere* a sua cura per i tipi della Rizzoli (D'Annunzio) ospita una postfazione di Marcello Esposito e Marco Dotti in cui il romanzo viene analizzato con metodologie quantitative e tecniche di network analysis.

Ammaniti, o della proposta di Alessandro Iannella di creare con Google Dialogflow un *chatbot* che simuli un dialogo con la poetessa Saffo.

Anche l'idea di un assistente virtuale per aiutare i giovani nell'approccio alla lettura ha dei precedenti, come, ad esempio, il *tool* Sobek, sviluppato «to support educational applications […] from assisting teachers to review student's work to helping kids in reading and writing activities»[2] da due ricercatori dell'Università federale del Rio Grande do Sul (cfr. Retaegui et al.) e finalizzato all'estrazione di una rete di parole chiave da un testo, o, soprattutto, Readerbench,[3] una piattaforma digitale (o «Personal Learning Environment») realizzata dall'Università Politecnica di Bucarest (cfr. Dascalu et al.) che si rivolge a studenti e insegnanti mettendo a disposizione una serie di strumenti per analizzare *corpora* testuali (in lingua inglese o francese), compiendo analisi come l'estrazione di *keywords*, il calcolo dell'indice di leggibilità di un testo, la *sentiment analysis*. Tuttavia, né Sobek né Readerbench – così come la maggior parte dei *tool* per la *text analysis* disponibili *online*, anche se non espressamente indirizzati agli studenti e spesso a pagamento (Monkeylearn, Meaning Cloud, Aylien…) – sono destinati al supporto della lettura e dello studio di opere letterarie, tantomeno italiane, laddove invece, soprattutto per alcune delle tecniche coinvolte – in particolare la *sentiment analysis* – la lingua di riferimento è fondamentale.

Il nostro progetto mira allo sviluppo di un ambiente digitale *web based*, utilizzabile anche su piattaforma *mobile*, che introduca i lettori ad alcune delle principali tecniche di analisi testuale e *text mining*. Insieme ai testi integrali di dieci classici della letteratura italiana, si propongono esempi applicativi di alcune tra le più note tecniche di analisi computazionale dei testi, qui utilizzate come strumenti per replicare digitalmente i procedimenti delle *reading skills* fondamentali (analisi del lessico, dei temi e dei *plot*, studio delle interazioni tra personaggi…), permettendo agli studenti di conoscere nuove metodologie di approccio ai testi, sia esplorando gli *output* dei libri pre-analizzati nell'ambiente della piattaforma sia provando a riprodurre le analisi in autonomia su testi a loro piacimento grazie a una sezione che mette a disposizione gli algoritmi in forma di *web app*.

La scelta delle opere da proporre in versione integrale e da analizzare tramite le tecniche di analisi computazionale selezionate per il progetto è stata vincolata dalla necessità di optare, tra i testi riconducibili al canone delle opere affrontate nel percorso dell'educazione secondaria di secondo grado,[4] per autori 'fuori diritti'.[5] Inoltre, si è preferito per il momento limitare il nostro *corpus* di riferimento alla narrativa, per ragioni prettamente tecniche: nonostante – con le rispettive peculiarità – le tecniche siano oggi applicate in uguale misura a opere in prosa e a opere in versi, quest'ultime richiedono un lavoro di pre-elaborazione dei testi maggiore, dettato per lo più dalla scansione metrica, oltre a pretendere una maggiore capacità di analisi degli *output*, spesso meno chiari a causa della figuratività del linguaggio poetico (ma non per questo meno proficui, anche nella loro 'oscurità' – cfr. Rhody o, per una recente rassegna di esempi applicativi, il volume *Tackling the Toolkit*, Plecháč et al.). Questi due vincoli hanno determinato l'esclusione di alcuni tra gli autori più letti e studiati nelle classi di licei e istituti tecnici e professionali (Giuseppe Ungaretti, Eugenio Montale, Italo Calvino…), ma hanno comunque consentito di allestire una rosa di dieci titoli (uno per autore) che rappresenta, a nostro giudizio,

[2] http://sobek.ufrgs.br/index-en.html.
[3] https://readerbench.com/.
[4] Cfr. https://www.indire.it/lucabas/lkmw_file/licei2010/indicazioni_nuovo_impaginato/_decreto_indicazioni_naz-ionali.pdf; https://www.indire.it/lucabas/lkmw_file/nuovi_tecnici/INDIC/_LINEE_GUIDA_TECNICI_.pdf.
[5] Condizione determinata dal fatto che siano trascorsi 70 o più anni dalla loro morte, come specificato dal d.lgs 22/2014, che ha esteso questo termine rispetto ai 50 anni previsti dalla legge sul diritto d'autore 633/1941, tuttora vigente in Italia.

un soddisfacente catalogo della narrativa italiana ottocentesca e primo-novecentesca. Nello specifico, si sono incluse le seguenti opere canoniche:

- Alessandro Manzoni, *I promessi sposi*;
- Giovanni Verga, *I Malavoglia*;
- Carlo Collodi, *Le avventure di Pinocchio*;
- Gabriele D'Annunzio, *Il piacere*;
- Federico De Roberto, *I Viceré*;
- Luigi Capuana, *Il marchese di Roccaverdina*;
- Luigi Pirandello, *Il fu Mattia Pascal*;
- Grazia Deledda, *Canne al vento*;
- Federigo Tozzi, *Tre croci*;
- Italo Svevo, *La coscienza di Zeno*.

Le trascrizioni digitali dei testi provengono in gran parte dalla biblioteca *online* Progetto Manuzio, curata dalla comunità di Liber Liber.[6] Tra le varie opzioni possibili (Biblioteca italiana,[7] Biblioteca della letteratura italiana,[8] Wikisource,[9] ecc.), si è privilegiato Liber Liber per l'opportunità di scaricare i testi in formato EPUB, che ha agevolato la conversione dei libri in pagine HTML (con l'eccezione di *Tre croci*, non disponibile in EPUB e pertanto riconvertito in HTML partendo da una versione 'solo testo'). Come standard per valutare l'affidabilità delle proprie trascrizioni, Liber Liber ha coniato un indice che va da 0 (affidabilità bassa) a 3 (affidabilità ottima). I testi da noi utilizzati sono tutti contrassegnati da un indice di affidabilità 1 ma, come si legge nella guida per i collaboratori al progetto,[10] si tratta di un'indicazione di *default*, riservando a valutazioni di esperti l'assegnazione di punteggi superiori. Tuttavia, in fase di conversione e di caricamento sulla piattaforma, si sono comunque revisionati i testi, al fine di individuare eventuali refusi o problemi di formattazione.

3. Gli strumenti per l'analisi testuale

Il primo gruppo di strumenti che abbiamo selezionato per la piattaforma rientra nell'ambito dei metodi classici della *text analysis* e permette di applicare semplici analisi di statistica linguistica descrittiva, oltre a generare concordanze KWIC. A questo fine abbiamo adottato, personalizzandola, la notissima *suite Voyant Tools*,[11] un'applicazione Web client-server *open source* ideata e sviluppata da Stéfan Sinclair (purtroppo prematuramente scomparso nell'agosto del 2020) e Geoffrey Rockwell. Rilasciato in prima versione nel 2003 e attualmente giunto alla *release* 2.6.0 (agosto 2022), Voyant mette a disposizione degli utenti quasi 30 strumenti per l'elaborazione, l'analisi quantitativa e la relativa visualizzazione di *corpora* testuali e costituisce un *unicum* nel panorama degli applicativi per l'analisi testuale: nonostante alcuni di questi strumenti risultino molto simili tra loro per funzionamento e risultati e quindi quasi sovrapponibili, non si ha notizia di piattaforme altrettanto longeve che offrano un'analoga varietà di tecniche (forse soltanto la serie di *software* "Ant",[12] sviluppata a partire dai primi anni 2000[13] da Laurence Anthony, che non è però *web based*), rese *user friendly* grazie a un'interfaccia molto intuitiva (e senza, pertanto, necessità di usare stringhe di codice) e, per di più, completamente gratuite.

[6] https://www.liberliber.it/online/.
[7] http://www.bibliotecaitaliana.it/.
[8] http://www.letteraturaitaliana.net/.
[9] https://it.wikisource.org/wiki/Pagina_principale.
[10] https://www.liberliber.it/online/aiuta/progetti/manuzio/collaborare/.
[11] https://voyant-tools.org/.
[12] http://www.laurenceanthony.net/software.html.
[13] La *release* iniziale di AntConc, il primo *software* della serie, è del 2002.

Inoltre, Voyant ha una interfaccia localizzata in tredici lingue (tra cui la traduzione italiana, a cura di Fabio Ciotti e altri membri dell'Associazione per l'Informatica Umanistica e la Cultura Digitale). Sinclair e Rockwell – e molti altri studiosi dopo di loro – hanno dato ampia dimostrazione delle possibilità di Voyant in campo umanistico sia per la ricerca sia, soprattutto, per la didattica (cfr. Sinclair e Rockwell), ma la versatilità del sistema consente di utilizzarlo su testi di qualsiasi ambito o disciplina. Inoltre, tra i principi che hanno guidato lo sviluppo di Voyant, la sua architettura modulare consente di integrarlo (*in toto* o limitatamente ad alcuni moduli) in altri siti web.

Passando in rassegna i vari strumenti disponibili, Voyant sembra privilegiare – per numero ed efficacia – i metodi destinati all'analisi lessicale e statistica dei *corpora*, come indici di frequenza, concordanze, distribuzione dei termini e *collocation*. Ciò nonostante, non mancano risorse per analisi più avanzate: ad esempio, l'integrazione – definita dagli stessi sviluppatori «rudimentary» –[14] di una implementazione in Javascript dell'algoritmo di *topic modeling* LDA (Latent Dirichlet Allocation), [15] uno strumento per la rappresentazione geo-spaziale dei testi, addirittura un chatbot – Veliza, mutuato dal celebre Eliza –[16] con cui interagire dialogando sul *corpus*. Da questo ampio insieme di strumenti si è scelto di estrapolare cinque *tool* da integrare sulla nostra piattaforma:

- *Cirrus*: tra i cinque *tool* selezionati, Cirrus è certo il più elementare, con scarso impatto ermeneutico, ma utile a introdurre – visivamente, soprattutto – l'idea di una testualità alternativa, rimodellata grazie a uno strumento informatico. Si tratta essenzialmente della versione Voyant di una *wordcloud*, ossia una rappresentazione 'pesata', utilizzando cioè caratteri di corpi, colori e dimensioni diverse per rappresentare sinotticamente le 50 parole più frequenti nel *corpus*. Voyant consente di variare il numero di parole incluse nella nuvola, partendo da un minimo di 25.
- *Termini*: Termini elabora un indice di frequenza dei termini del *corpus*. La visualizzazione in colonne consente di ordinare la lista in ordine alfabetico o per numero di occorrenze, oltre che di conoscere il dato di frequenza assoluta (espresso in interi) e il dato di frequenza relativa, ossia quanto incida la frequenza di un termine rispetto al totale delle parole che costituiscono il *corpus* (espresso in percentuale). Oltre alla lista, Termini mette a disposizione una barra di ricerca che, oltre al singolo termine, consente la ricerca con caratteri jolly per ovviare alle variabili morfologiche (ad esempio cercando per «cappott*») e la possibilità di selezionare due o più termini, per compararne la frequenza.
- *Contesti*: lo strumento Contesti equivale a un compilatore di concordanze, ossia mostra ogni singola occorrenza delle parole più ricorrenti all'interno del rispettivo contesto sintattico. Permette di ampliare o diminuire l'intervallo testuale mostrato e, dai valori della colonna "Posizione", è possibile desumere da che parte del *corpus* provenga il frammento. In modo analogo a Termini, oltre alla visualizzazione gerarchica anche Contesti permette di eseguire una ricerca testuale per uno o più termini.
- *Microricerca*: Microricerca offre una rappresentazione del 'tessuto' di un testo e mostra la distribuzione delle parole più ricorrenti al suo interno come punti su righe, giocando con il colore più o meno vivido a seconda del grado di diffusione.
- *Sintagmi*: Sintagmi, infine, mostra una lista dei sintagmi più frequenti all'interno del *corpus* (e dunque non interessato, ai fini dell'analisi, dal filtro delle *stopwords*). È possibile

[14] https://voyant-tools.org/docs/#!/guide/topics.
[15] I primi ad applicare un algoritmo di LDA al *machine learning* e in particolare all'analisi dei *topic* sono stati David Blei, Andrew Ng e Michael I. Jordan nel 2003. Si tratta di una soluzione statistica fondata su un approccio generativo bayesiano di distribuzione probabilistica delle parole all'interno dei documenti.
[16] Assistente virtuale progettato tra il 1964 e il 1966 da Joseph Weizenbaum e ottimizzato – nella sua versione DOCTOR – per parodiare il colloquio con uno psicoterapista rogersiano. Se ne può trovare una versione più recente all'indirizzo http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm.

definire la lunghezza minima e massima del singolo sintagma e visualizzarli in ordine di lunghezza o di frequenza.

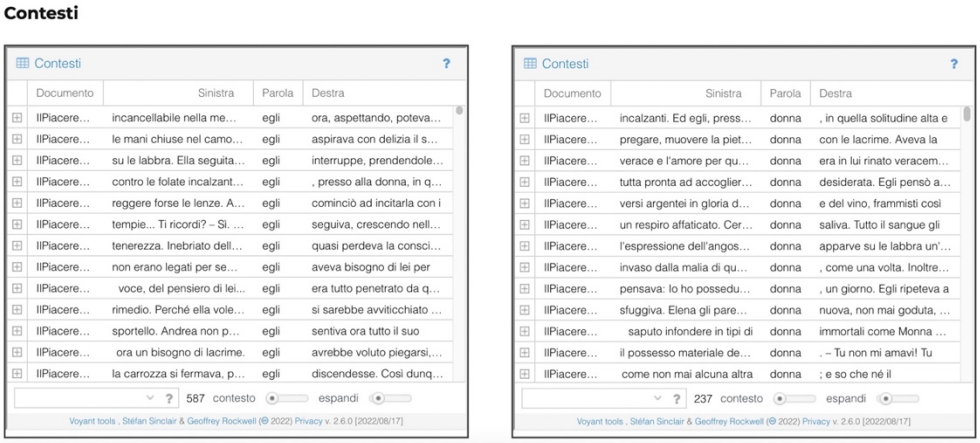


Fig. 1 – Esempio della visualizzazione affiancata (con e senza *stopwords*) del *tool* Contesti per il *Piacere*.

Di tutti questi strumenti, a eccezione dell'ultimo in ordine di visualizzazione ("Sintagmi"), si presentano due visualizzazioni affiancate: la prima derivante dall'analisi del testo nella sua totalità, la seconda dopo l'eliminazione delle principali *stowpords*, un elenco di nomi propri e parole semanticamente vuote che Voyant consente di filtrare. Si tratta di un insieme di 700 parole grammaticali o ad alta frequenza (come le forme, usate in funzione di ausiliare, dei verbi "essere" e "avere"), allestito tramite lo spoglio degli elenchi di frequenza dei primi romanzi analizzati. A questo gruppo, per ogni testo si è aggiunta una lista integrativa composta dai nomi propri di persona o di luogo, estratta in automatico grazie all'algoritmo per la *Named Entity Recognition* della *suite* per il *Natural Language Processing* Spacy[17] e poi revisionata a mano. Spacy offre uno dei migliori algoritmi per l'individuazione di entità in lingua italiana, grazie soprattutto al modello linguistico "it_core_news_lg"[18] che garantisce un tasso di precisione dichiarato dell'88%. La lista di entità, limitata ai nomi propri di persona, è stata poi riutilizzata in un secondo momento per la *network analysis* delle interazioni tra i personaggi (cfr. *infra*).

4. Alla ricerca dei temi del testo: il *topic modeling*

L'insieme di metodi di *text mining* che vengono comunemente rubricati sotto l'etichetta di *topic modeling* hanno avuto un notevole successo in ambito umanistico e letterario in particolare (Ciotti). Si tratta di tecniche statistico-probabilistiche volte a estrarre automaticamente l'insieme dei temi che caratterizza una collezione di documenti testuali. L'assunto di base che le contraddistingue è che ogni documento di una collezione è generato da una distribuzione di *topic*, che a loro volta sono rappresentati come una distribuzione di parole. L'algoritmo, procedendo a ritroso a partire dalla distribuzione delle parole nella collezione, desume tali distribuzioni. Come detto Voyant include una versione del più noto e diffuso degli algoritmi di *topic modeling*, LDA. Nella piattaforma di *Macchine per leggere* abbiamo tuttavia preferito integrare una

[17] https://spacy.io/.
[18] https://github.com/explosion/spacy-models/releases/tag/it_core_news_lg-3.4.0.

soluzione più evoluta, l'algoritmo *BERTopic*, che integra anche i metodi di *word embedding* e dei *transformer*, alla base dei più recenti e potenti modelli linguistici basati su reti neurali.[19]

BERTopic è uno strumento di *machine learning open access* sviluppato da Maarten Grootendorst nel 2020[20] che combina il modello linguistico BERT (*Bidirectional Encoder Representations from Transformers*),[21] implementato da Google e reso disponibile tra il 2019 e il 2020, con un l'algoritmo c-TF-IDF.[22] Questo è una variante del classico metodo statistico per l'*information retrieval* TF-IDF (*term frequency-inverse document frequency*), finalizzato alla misurazione della rilevanza dei termini occorrenti in un documento all'interno di un *corpus*, con la sola variante che i documenti, nel caso di c-TF-IDF, vengono riuniti per classi.[23]

BERTopic è molto flessibile e consente di analizzare documenti partendo da formato solo testo o da tabelle CSV. Soprattutto, caratteristica tipica di BERT, non richiede nessun lavoro preliminare di pre-etichettatura del testo o di 'addestramento' su uno specifico *corpus*: una volta completato l'*upload* e aggiunta – se si desidera – una lista di *stopwords*, l'algoritmo è pronto per produrre gli *output* desiderati.

Per *Macchine per leggere*, nonostante i testi presi in esame differissero tra loro per voluminosità e struttura, si è tentato di elaborare un unico modello di analisi che potesse essere – con ottimizzazioni minime – replicato per ciascun libro, senza quindi 'etero-dirigere' eccessivamente l'elaborazione ma, al contempo, cercando di ottenere dei risultati proficui. Nello specifico, quindi, oltre a integrare di volta in volta le liste di *stopwords* predisposte per ogni singolo libro, si sono richiesti all'algoritmo l'estrazione delle prime dieci parole più indicative per ogni *topic*, il calcolo della probabilità di ogni *topic* di essere presente in un dato documento, l'inclusione come elenco di *stopwords* della lista aggiunta sotto il nome "vectorizer_model", e, soprattutto, l'individuazione di *topic* non troppo 'vasti' (3 il valore impostato, contro il 10 previsto di *default*) e la riduzione dell'*output* finale a un massimo di dieci *topic*. Quest'ultima scelta è stata indotta dai primi tentativi di applicazione dell'algoritmo su alcuni dei nostri titoli (soprattutto *I Malavoglia*) che, coi valori di *default*, vedevano risultare un unico, onnicomprensivo *topic*, largamente più diffuso rispetto agli altri, presenti ma ignorati. Senza aver raggiunto conclusioni certe (non si dà notizia, nel pur frequentatissimo *forum* dedicato a BERTopic, di problematiche analoghe), è probabile che ciò fosse dettato da una ridotta varietà linguistica interna all'opera, che impediva all'algoritmo di produrre risultati sufficientemente significativi. Con questa soluzione, nonostante l'incidenza degli altri *topic* risulti di fatto minore, si è comunque riusciti a ottenere un'analisi coerente ed esaustiva, con dieci *topic* per romanzo.

BERTopic consente di estrarre i *topic* da un *corpus* e di compiere delle operazioni sugli *output*, come ad esempio quella di individuare in quali porzioni di testo ("sentences", che possono corrispondere a paragrafi o singole frasi, a seconda della suddivisione a cui si sottopone il testo in fase di importazione) sia più presente uno tra i *topic* individuati o, viceversa. Tuttavia, abbiamo preferito limitarci a presentare in forma grafica i risultati delle analisi dei *topic* nei sei

[19] A loro volta fondati sugli assunti della semantica distribuzionale (Firth).
[20] https://github.com/MaartenGr/BERTopic.
[21] https://github.com/google-research/bert. Tra le varie versioni del modello linguistico BERT, per analisi su testi in italiano (e per più di 50 altre lingue oltre all'inglese) è disponibile il modello "paraphrase-multilingual-MiniLM-L12-v2", parte del *framework* Sentence Transformers, https://www.sbert.net/.
[22] https://maartengr.github.io/BERTopic/api/ctfidf.html#bertopic._ctfidf.ClassTFIDF.
[23] Se la funzione di peso TF-IDF aumenta in proporzione al numero di occorrenze di un termine in un dato documento, ma diminuisce proporzionalmente alla sua presenza nel *corpus*, presumendone così una particolare significatività nell'economia del singolo documento, la variante c-TF-IDF si comporta in modo analogo ma riunisce i documenti che hanno in comune un *topic* (ossia un insieme di parole giudicato coerente) e li tratta come un unico elemento.

possibili formati di *output* previsti dal modello e prodotti con la libreria grafica Python Plotly.[24] Nello specifico:

- *Intertopic Distance Map*: un grafico cartesiano dove i *topic* sono rappresentati come cerchi, di diametro variabile a seconda del numero di parole comprese. La vicinanza tra i cerchi rappresenta il grado di tangenza tra gli insiemi lessicali che costituiscono i singoli *topic*.
- *Topic Probability Distribution*: istogramma orizzontale che presenta la probabilità di occorrenza di ogni *topic* all'interno del *corpus*.
- *Hierarchical Clustering*: un grafico che suddivide i *topic* estratti in *cluster* gerarchici.
- *Topic Word Scores*: le cinque parole più rilevanti per ciascuno dei dieci *topic* individuati, proposte in istogrammi orizzontali che ne denotano la frequenza.
- *Similarity Matrix*: una rappresentazione visiva in cui a diversi colori sono associati i possibili gradi di similarità tra l'uno e l'altro *topic*.
- *Term score decline per Topic*: indica il grado di distintività per l'algoritmo di un *topic* all'interno del *corpus*, riportando il progressivo diminuire del numero di termini necessari all'identificazione. Più che un *output* con valori per la consultazione, si tratta quindi di un resoconto "tecnico", che denota l'evidenza dei singoli *topic* nel progredire dell'analisi.

Il codice dell'algoritmo eseguito, nella sua versione completa, è accessibile nella *repository* GitHub del progetto *Macchine per leggere*.[25]

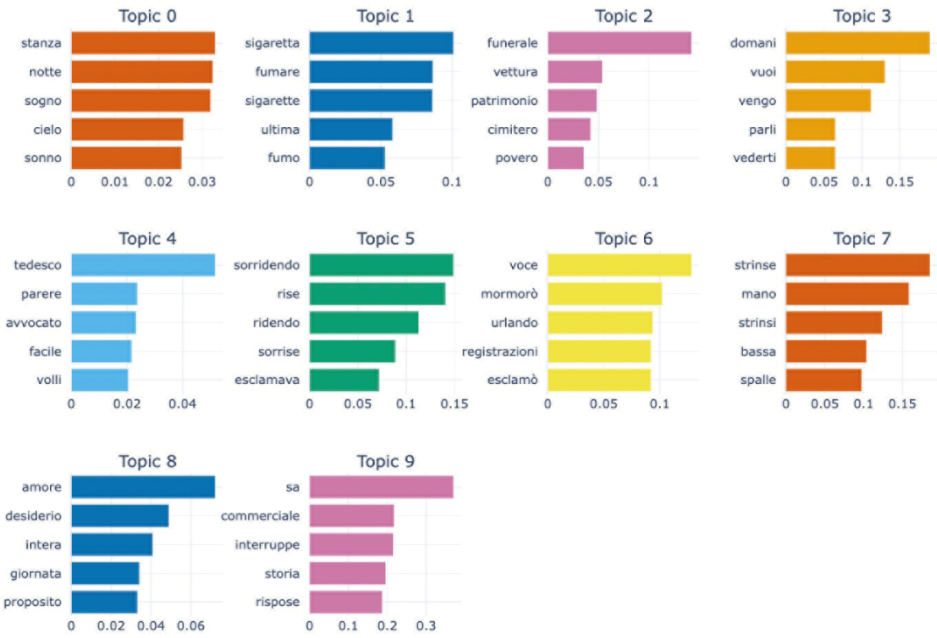


Fig. 2 – *Topic Word Scores* di *La coscienza di Zeno*. Tra temi connessi a elementi lessicali che denotano interazioni tra personaggi, come ad esempio il numero 5, il numero 6 e il numero 7, l'algoritmo ha isolato insiemi di parole semanticamente significativi, di cui si può individuare con una certa facilità il grado di coerenza, come ad esempio il numero 1, che si ricollega al rapporto di Zeno col vizio del fumo, o il numero 8, legato alla tematica amorosa. Il *topic* numero 0, invece, sembra denotare i momenti legati al sogno, una serie di episodi diffusa in modo piuttosto uniforme nell'arco di tutto il romanzo.

[24] https://plotly.com/python/.
[25] https://github.com/Albertobaldi/Macchine-per-leggere.

5. Le emozioni nel testo: la *sentiment analysis*

La *sentiment analysis* è un insieme di metodi volti a determinare in modo automatico la polarità (positiva o negativa) e l'andamento o *score* 'emozionale' (basato sulle varie tipologie di emozioni primarie elaborato dalle teorie psicologiche) di un testo o di un *corpus* testuale (Liu). Si tratta di un approccio analitico utilizzato prevalentemente in ambito sociologico, commerciale e politico, soprattutto in relazione al marketing e allo studio dei *social media*, nonostante negli ultimi anni abbia trovato applicazione anche negli studi culturali e letterari, anche in relazione al cosiddetto "affective turn" (Hogan et al.).

Ai fini dell'inclusione nella nostra piattaforma di uno strumento semplice da usare e allo stesso tempo adatto allo scopo di supportare la lettura e la comprensione di un testo narrativo, sono stato presi in esame diversi tra le risorse liberamente disponibili *online*. Si è dovuto rilevare come gran parte di tali strumenti fossero inadeguati, o per incompatibilità linguistica (nella sua declinazione *lexicon based*, infatti, questo tipo di analisi risente molto della lingua di riferimento) o per *output* scarsamente rappresentativi ai fini del nostro progetto. Ad esempio, FEEL-IT,[26] uno dei migliori *tool* di *sentiment analysis* per l'italiano – basato su una versione italiana di BERT, UmBERTo,[27] – (Bianchi et al.) per quanto molto accurato nell'analisi è purtroppo limitato a brevi stringhe di testo (essendo ottimizzato per analizzare i tweet) e non dà *output* rilevanti se applicato, come nel nostro caso, a testi di grandi dimensioni come i romanzi. Si è quindi deciso, cercando di tenere fede al proposito di proporre strumenti che introducano a queste tecniche – a livello concettuale e applicativo – anche e soprattutto attraverso possibili visualizzazioni 'alternative' del testo letterario, di riproporre uno dei primi e più celebri modelli di *sentiment analysis* applicata ai testi letterari, ossia la libreria Syuzhet sviluppata da Matthew L. Jockers nel 2014.

Syuzhet è stata creata da Jockers seguendo la suggestione di Kurt Vonnegut riguardo alle *shapes of stories*. Lo scrittore statunitense aveva infatti teorizzato, in un intervento di cui è ancora oggi possibile rintracciare il filmato *online*,[28] che le trame dei romanzi della letteratura mondiale possano essere ridotte a un massimo di otto strutture base, rappresentabili su diagramma cartesiano con "bene/male" (o "positività" e "negatività") come estremi dell'asse y e con inizio e fine narrazione sull'asse della x.[29] Su tale base, Jockers ha sviluppato un algoritmo capace di suddividere un testo in sequenze e di assegnare un indice di positività o di negatività a ciascuna di esse, così da riprodurre in digitale quanto realizzato da Vonnegut in forma analogica. L'assegnazione del punteggio alle sottounità avviene in base alla presenza, all'interno di esse, di parole chiave che, secondo i dizionari emozionali integrati nella libreria, esprimono polarità negativa o positiva. Questo metodo, piuttosto rudimentale, non è stato esente da critiche, come le puntuali rilevazioni di Annie Swafford, soprattutto circa l'assegnazione dei punteggi con metodo *lexicon based*, che giocoforza ignora il ruolo semantico dei modificatori (ad esempio le negazioni), di parole annoverate come positive anche se ricorrenti in forma neutrale (caso tipico dell'inglese, con termini come "like"), di espressioni ricercate o desuete (perché i dizionari integrati sono dizionari dell'uso). Tuttavia, ancora oggi, Syuzhet non manca di suscitare interesse nella comunità degli umanisti digitali, come dimostra la recente e puntuale disamina di

[26] https://github.com/MilaNLProc/feel-it.
[27] https://github.com/musixmatchresearch/umberto.
[28] https://www.youtube.com/watch?v=oP3c1h8v2ZQ; https://www.youtube.com/watch?v=j9Qsiu8qqvA.
[29] Jockers è stato il primo ma non l'unico umanista digitale a sperimentare un modello informatico per la teoria di Vonnegut. Di particolare rilievo è infatti anche il contributo di Andrew J. Reagan *et al.*, che hanno verificato l'assunto di Vonnegut su un *corpus* di circa 1300 romanzi in lingua inglese, investigato con il *tool* Hedonometer (da loro sviluppato nel 2013, insieme a un ampio *team*, nell'ambito del Computational Story Lab dell'Università del Vermont – https://hedonometer.org/, 08/2022). La loro ricerca ha individuato un numero pari a sei di archi emozionali fondamentali.

Kim che, muovendo dalle critiche di Annie Swafford, torna a evidenziarne i punti deboli pur constatandone l'indubitabile successo rispetto ad altri strumenti affini.

Per questo motivo, sempre considerando lo scopo più dimostrativo che ermeneutico della nostra piattaforma e soprattutto l'indubbia risonanza che questo strumento ha riscosso nel tempo, si è scelto di includerlo come esemplificazione della *sentiment analysis*, nonostante il suo *output* non sia – almeno nella forma del grafico cartesiano – agevolmente riconducibile a questo tipo di analisi.

Syuzhet è inoltre uno strumento molto ergonomico. Si tratta di una libreria in linguaggio R (un linguaggio di programmazione ottimizzato per il calcolo statistico) facilmente accessibile (R dispone di un ottimo ambiente desktop, RStudio) e semplice da utilizzare: con poco più di dieci righe di codice è possibile processare il testo e ottenere un grafico come quello relativo a *Pinocchio* in Fig. 3.

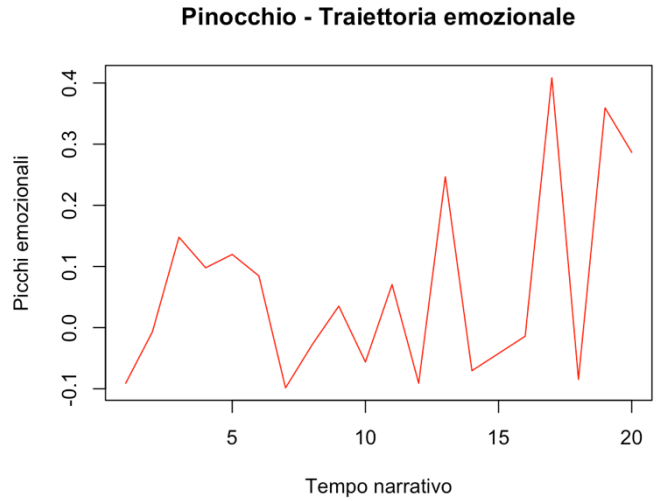


Fig. 3 – Traiettoria emozionale di *Pinocchio*. Nel plot possiamo notare tre picchi al negativo: il primo, intorno ai 7/20 della trama, va dall'episodio dell'impiccagione di Pinocchio da parte del Gatto e della Volpe ai primi momenti di Pinocchio febbricitante dalla Fata turchina; il secondo, intorno ai 12/20, dal mancato arresto di Pinocchio al rischio di finire fritto in padella dal pescatore; il terzo, infine, intorno ai 18/20, che segue la "cuccagna" nel Paese dei Balocchi (il picco più alto dell'intera trama), riguarda la trasformazione di Pinocchio in ciuchino, la sofferenza al circo, l'inghiottimento da parte del pescecane e termina col ritrovamento di Geppetto, che prelude al lieto fine.

6. Le reti dei personaggi: la *network analysis*

La *network analysis* è un metodo matematico che permette di studiare le strutture relazionali tra gruppi di individui (dove gli individui possono essere persone, enti istituzioni, oggetti o concetti astratti). Ogni individuo costituisce un nodo (o vertice) e ogni relazione un arco (o spigolo). Le proprietà matematiche della rete possono essere adottate come succedanei delle caratteristiche qualitative del dominio: le misure di centralità di un nodo permettono di identificare l'individuo più importante nella rete (ad esempio il protagonista di una storia); l'individuazione di sottoinsiemi di nodi prossimali la presenza di gruppi omogenei nella popolazione (ad esempio gruppi di personaggi che interagiscono tra loro); e così via. È stato ancora una volta Franco Moretti, assieme ai membri dello *Stanford Literary Lab* da lui fondato, il primo a proporre l'applicazione della *network analysis*, molto diffusa nelle scienze sociali, in ambito

letterario. La sua proposta (applicata, ad esempio, all'*Amleto*, cfr. "Network Theory, Plot Analysis" – anche se senza l'utilizzo di risorse informatiche per l'estrazione e la rappresentazione delle informazioni) è stata quella di considerare i personaggi come nodi e le loro interazioni – intese come scambi di battute nei drammi, ma anche come prossimità sintattica nei romanzi – come archi relazionali, e così indagare, ad esempio, il grado di centralità di un personaggio rispetto all'opera.

Seguendo il modello proposto da Moretti, poi ripreso e, soprattutto, declinato al digitale da diversi studiosi (cfr. Jannidis et al.; John et al.; Smeets), si è deciso di includere una rappresentazione a nodi delle relazioni tra i personaggi di ciascuno dei romanzi proposti sulla nostra piattaforma. Per farlo, si è utilizzato un *workflow* Python proposto da Milán Janosov, che ha compiuto un'operazione analoga sulla serie di romanzi *fantasy The Witcher* (Jasonov). Si tratta di un flusso di lavoro semi-automatico, perché prevede comunque che l'utente predisponga una lista di riferimento con i nomi dei personaggi del romanzo. Uno volta in possesso della lista, strutturata in un file CSV, l'algoritmo scompone in frasi il *corpus* testuale e utilizza la già citata libreria per la *Named Entity Recognition* di Spacy per identificare in esse la presenza di una o più delle entità presenti nella lista fornita dall'utente. Per computare le relazioni, si suddividono poi le entità co-occorrenti nel medesimo contesto come "source" e "target", attribuendo le due etichette sulla base di un ordine sintattico. Infine, si raggruppano le relazioni individuate e si procede a sommarle in un intero, che costituisce il 'peso' delle interazioni tra due personaggi. Si esporta il risultato in formato GEXF grazie alla libreria Python NetworkX e si procede alla personalizzazione del grafico di *output* utilizzando Gephi,[30] uno ddei più noti *software open source*per la network analysis. Per la pubblicazione *online* sulla piattaforma, il grafico ottenuto con Gephi è riprocessato con VOS-viewer,[31] un *tool* che consente l'esportazione di grafici in formati interattivi per il *web*, come mostra la Fig. 4 per *I Malavoglia*.

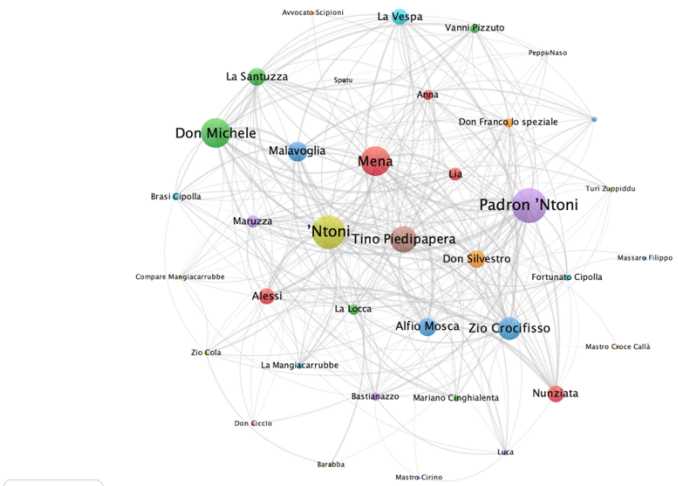


Fig. 4 – *Network analysis* dei *Malavoglia*. I nodi, che corrispondono ai personaggi del romanzo, variano in dimensione a seconda della loro centralità nel complesso delle relazioni. Lo spessore delle linee, invece, varia in base al numero di interazioni tra i due personaggi collegati.

7. Il design dell'interfaccia e l'architettura del sito

La comunicazione visiva e l'architettura informativa di una *web application*, tanto più se deve coinvolgere i fruitori nell'esecuzione di compiti complessi e cognitivamente impegnativi, va studiata con molta cura e deve consentire un accesso multicanale. Nella scelta del *layout* di base

[30] https://gephi.org/.
[31] https://www.vosviewer.com/.

per il sito di *Macchine per leggere*, pertanto, si è privilegiata un'impostazione grafica che permettesse un'elevata scalabilità su dispositivi *mobile* (tablet e smartphone). L'*homepage* è stata strutturata in riquadri di grandezza variabile su sfondo bianco, limitando i contenuti di testo e dando maggior risalto alle immagini, con collegamenti rapidi alle analisi digitali di alcune delle opere letterarie proposte, alla pagina di presentazione del progetto e alla guida introduttiva alle tecniche di analisi testuale digitale. Per le intestazioni e per il corpo dei testi, si è scelto di utilizzare per ciascuna pagina della piattaforma il Montserrat, set di caratteri *open source* progettato da Julieta Ulanovsky, pubblicato nel 2011 e aggiornato nel 2017. Un font *sans serif*, che mantiene un elevato indice di leggibilità in corpi minori (tanto da risultare il quarto set di caratteri più scaricato tra quelli disponibili nella libreria Google Fonts[32] e attualmente in uso su circa 15 milioni di siti web) e che favorisce l'accessibilità digitale.

Poiché lo scopo del progetto[33] è promuovere la riutilizzazione delle tecniche di analisi da parte degli utenti, soprattutto in contesti didattici, il sito offre una sezione dedicata alla guida alle tecniche di analisi computazionale dei testi, suddivisa in quattro sottosezioni

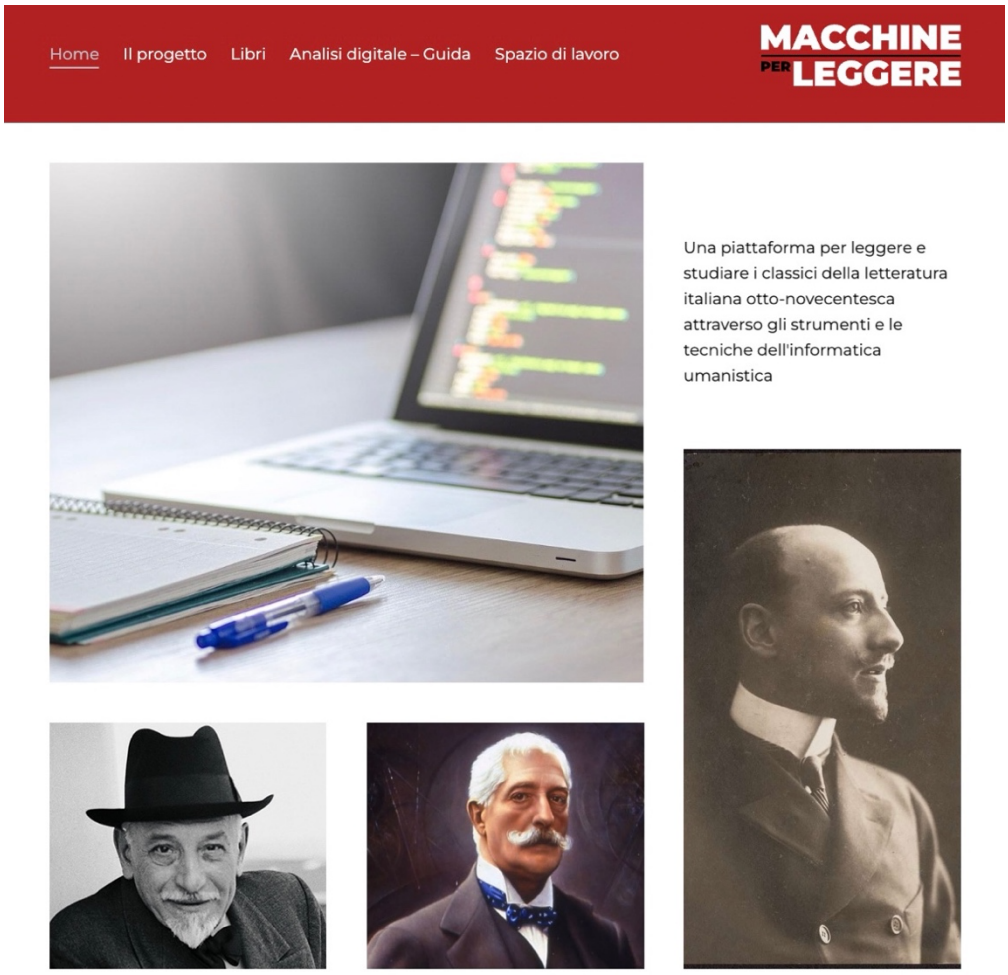
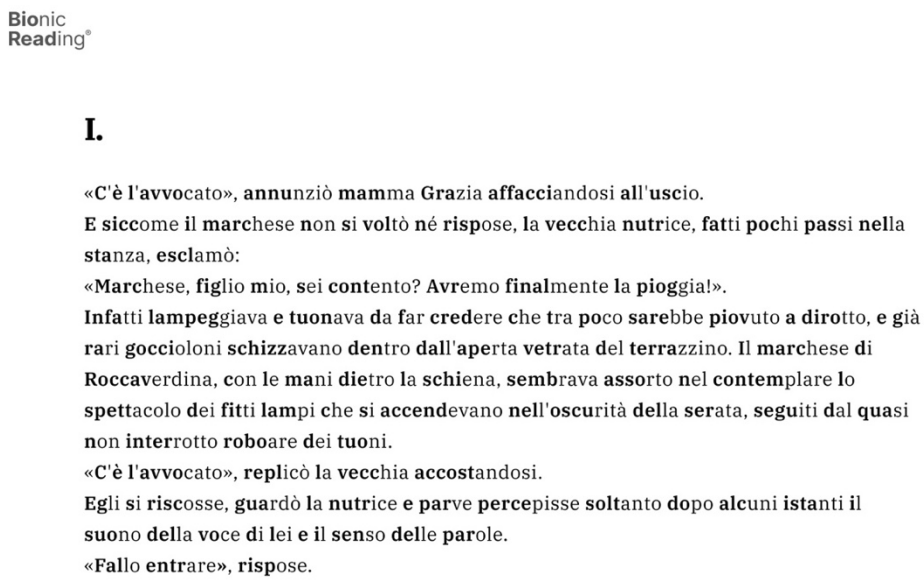


Fig. 5 – Sezione della homepage di *Macchine per leggere*.

[32] https://fonts.google.com/.
[33] Descritto con dettaglio nella sezione "Progetto" che dà anche notizia della fonte da cui sono state tratte le opere letterarie proposte e rinvia ai siti o alle pagine Github degli strumenti utilizzati nelle sezioni dedicate all'analisi computazionale dei testi, in un'ottica che aderisce ai principi dell'*open science*.


Bionic Reading®

I.

«C'è l'avvocato», annunziò mamma Grazia affacciandosi all'uscio.
E siccome il marchese non si voltò né rispose, la vecchia nutrice, fatti pochi passi nella stanza, esclamò:
«Marchese, figlio mio, sei contento? Avremo finalmente la pioggia!».
Infatti lampeggiava e tuonava da far credere che tra poco sarebbe piovuto a dirotto, e già rari gocciloni schizzavano dentro dall'aperta vetrata del terrazzino. Il marchese di Roccaverdina, con le mani dietro la schiena, sembrava assorto nel contemplare lo spettacolo dei fitti lampi che si accendevano nell'oscurità della serata, seguiti dal quasi non interrotto roboare dei tuoni.
«C'è l'avvocato», replicò la vecchia accostandosi.
Egli si riscosse, guardò la nutrice e parve percepisse soltanto dopo alcuni istanti il suono della voce di lei e il senso delle parole.
«Fallo entrare», rispose.

Fig. 6 – Primi paragrafi del *Marchese di Roccaverdina* in Bionic Reading.

("Introduzione", "Tecniche", "Glossario", "Bibliografia") che vuole costituire un breve compendio di questo approccio alla letteratura destinato agli studenti e ai loro docenti. Nella parte introduttiva, ci si sofferma sui punti fondamentali del dibattito teorico, dalla locuzione *distant reading* coniata come si è detto da Franco Moretti alle sue successive declinazioni, passando in rassegna le posizioni dei maggiori esperti in materia (Matthew Jockers, Ted Underwood, Andrew Piper…) e anche alcune tra le critiche più accese, come quelle mosse da Nan Z. Da su *Critical Inquiry* nel 2019. Nella sottosezione dedicata alle tecniche, si approfondiscono dapprima le metodologie utilizzate nell'ambito della nostra piattaforma (analisi statistiche, *topic modeling*, *sentiment analysis*, *network analysis*), per poi completare la panoramica fornendo ragguagli su altre possibili applicazioni del digitale all'analisi letteraria (stilometria computazionale, *text classification*, *word embedding*…). Chiudono la sezione del sito dedicata alla Guida un glossario con i termini più ricorrenti nell'ambito dell'analisi computazionale dei testi e una bibliografia con rimandi a testi teorici e a esperimenti applicativi.

Dalla pagina "Libri", in cui ogni titolo è accompagnato da un'immagine dell'autore, è possibile accedere al testo integrale dell'opera o a un singolo capitolo tramite un indice a comparsa, oltre che alla pagina che presenta gli *output* delle analisi digitali. L'utilizzo di moduli iFrame ha consentito un'agevole integrazione di tutti gli strumenti, rendendo le schermate risultanti consultabili e interattive e preservando la scalabilità della pagina. Inoltre, si è scelto di includere la possibilità di fornire agli utenti la possibilità di leggere il testo dei libri analizzati in modalità "bionic reading".[34] Si tratta di un formato tipografico-digitale di recente sviluppo, realizzato dal designer svizzero Renato Casutt. Attraverso la messa in rilievo in grassetto (o neretto) di parte delle parole, questa tecnologia facilita il mantenimento del *focus* dell'attenzione, favorendo la capacità di assorbimento del testo e la velocità di lettura:

> Bionic Reading revises texts so that the most concise parts of words are highlighted. This guides the eye over the text and the brain remembers previously learned words more quickly. […] Bionic Reading aims to play a supporting role in the absorption of volume text. We see technological progress as an opportunity for all those who want to increase the pleasure of reading in a noisy and hectic world in a focused way and without distraction. (Bionic Reading)

[34] https://bionic-reading.com/.

Bionic Reading, come si può notare dalla Fig. 6, si differenzia dall'utilizzo, convenzionale sebbene raro, del grassetto per l'enfatizzazione (concettuale prima che grafica) di parole o porzioni di testo tipico della tipografia tradizionale.[35] Nonostante si tratti di una tecnologia in via di sperimentazione e non riconducibile all'ambito delle tecniche di analisi computazionale dei testi, si è comunque deciso – d'accordo con lo sviluppatore, che ha peraltro dimostrato interesse nella possibilità di ricevere un *feedback* di utilizzo – di includerla nella nostra piattaforma, guidati dal proposito di non escludere nessuna risorsa digitale che possa, seppure in misura variabile, costituire un elemento di incentivo per la lettura dei classici letterari. A ciascuna delle opere letterarie proposte sulla piattaforma *Macchine per leggere* corrisponde una pagina dedicata all'analisi computazionale del testo e alla presentazione dei relativi *output*.

Come detto, il *focus* del nostro progetto è di sperimentare se le tecniche di *text analysis* possano fungere da veicolo per agevolare l'approccio alla lettura dei grandi classici: pertanto, la sezione "Spazio di lavoro" ne costituisce l'elemento fondamentale. Questa sezione è pensata per mettere a disposizione dell'utenza la possibilità di replicare le analisi proposte nelle pagine "Analisi computazionale" dei testi pre-processati, senza tuttavia uscire dalla piattaforma e soprattutto senza necessità di ricorrere a operazioni di programmazione e di configurazione del codice necessarie all'esecuzione degli algoritmi: sarà sufficiente possedere un testo codificato in formato solo testo per ottenere gli stessi *output* presentati per le opere pre-processate.

Per quanto riguarda gli strumenti di Voyant, la *suite*, nonostante sia molto flessibile, non consente di estrapolare singoli moduli senza prima aver processato un *corpus*: pertanto, in una cornice IFRAME, si è proposta la schermata iniziale, così come reperibile sull'h*omepage* del sito Voyant Tools, da cui è possibile effettuare l'*upload* di un *corpus* e procedere alle analisi. Per riprodurre il *workflow* di estrazione dei *topic* realizzato con BERTopic si è invece optato per il *framework open source* Streamlit,[36] che consente di sviluppare *web app* da codice Python, mettendo a disposizione una nutrita libreria di *widget* per personalizzare l'interfaccia grafica e renderla più ergonomica all'esperienza dell'utente. Streamlit, collegata a un repository GitHub che ospiti lo *script* dell'applicazione, inizializza la app sui suoi server, cui si accede con un semplice link, qui

Fig. 7 – La web app per l'analisi con BERTopic così come appare su *Macchine per leggere*: nella colonna sulla sinistra, la casella per inserire una lista di stopwords e l'uploader per i file di testo. Al centro, la schermata in cui vengono restituiti i risultati dell'analisi.

[35] Sull'efficacia di questa strategia tipografica come facilitazione alla lettura cfr. Sanocki e Dyson e soprattutto Macaya e Perea.
[36] https://streamlit.io/.

Fig. 8 – La web app per Syuzhet così come appare su *Macchine per leggere*.

riproposto sempre in una cornice IFRAME. Uno dei limiti principali di Streamlit è che prevede una logica di *reload* a ogni interazione dell'utente: questo non consente di mantenere memoria di variabili o *output* e costringe pertanto alla creazione di applicazioni che non abbiano *step* di analisi intermedia.

In modo analogo, per consentire di replicare in autonomia l'analisi presentata con Syuzhet, si è realizzata una *web app* che ne implementi il *workflow* essenziale, da file di testo a grafico cartesiano con l'arco emozionale del *plot* di un'opera. Per farlo, si è utilizzato Shiny,[37] un pacchetto per il linguaggio R (in cui è scritta Syuzhet) analogo a Streamlit per Python.

Queste soluzioni rappresentano a nostro avviso un buon compromesso tra usabilità, semplicità e flessibilità richieste perché i vari strumenti possano essere usati in contesti didattici, che costituiscono oggetto dei prossimi *workpackage* del progetto.

8. Conclusioni e sviluppi futuri

La predisposizione della piattaforma *Macchine per leggere*, come abbiamo detto più volte, rappresenta solo il passo iniziale del progetto complessivo. Dal punto di vista tecnico, che abbiamo privilegiato in questo articolo, il progetto prevede ulteriori sviluppi, in particolare l'aggiunta di un modulo di georeferenziazione del testo che potrebbe rappresentare un ottimo veicolo per stimolare la curiosità dei giovani lettori, soprattutto se il mondo finzionale del testo è connesso, almeno in parte, con il mondo dell'esperienza vissuta. Ma ancora più importante sarà la fase sperimentale. Sia la scelta dei metodi e degli strumenti di analisi, sia la progettazione grafica, sono state condotte nell'ottica di fare del sito una piattaforma di lavoro e di sperimentazione didattica che dovrà coinvolgere in primo luogo il contesto della formazione secondaria.

In sinergia col Centro per il Libro e la Lettura del MiC, è in corso di organizzazioneuna sperimentazione didattica (in presenza e a distanza) con alcune classi selezionate. Ai docenti sarà messa a disposizione una serie di brevi *tutorial* video che illustreranno i principali nodi teorici del *distant reading*, le singole risorse presenti sul sito e possibili percorsi di lettura digitale a partire dalle analisi presentate: dalla ricostruzione delle trame e dei relativi picchi, verificando l'esattezza degli *output* di Syuzhet, all'analisi dei temi emersi dal *topic modeling* assegnando etichette alle liste di parole e rintracciando nel corpo del testo la presenza degli argomenti corrispondenti. Una volta concluso il periodo di sperimentazione, saranno raccolti *feedback* e suggerimenti dai docenti e dagli studenti attraverso la somministrazione di questionari compilabili direttamente sulla piattaforma. Con i questionari si cercherà di rilevare sia l'ergonomia di utilizzo percepita della piattaforma – per quanto concerne la presentazione dei contenuti, l'accessibilità delle analisi digitali e l'efficacia della guida – sia l'impatto degli strumenti messi a disposizione sull'approccio alla lettura critica dei testi letterari.

Le possibilità di integrazione della piattaforma con il metodo didattico e la sua efficacia nella promozione dell'interesse verso il testo da parte dei giovani lettori, dei suoi eventuali benefici, sono pertanto ancora in fase di valutazione e andranno definendosi nella attuazione di questa seconda parte del progetto. Non avendo dati esaustivi e sperimentazioni consolidate su cui basarsi, le scelte progettuali fatte finora sono state guidate da assunti maturati nel corso della nostra esperienza di ricerca e di didattica, assunti che vanno misurati con l'esperienza.

[37] https://shiny.rstudio.com/.

Una ulteriore complicazione è rappresentata dal fatto che le generazioni cosiddette *born digital* (qualsiasi cosa si voglia intendere con questa formula) non sono affatto dotate di quelle competenze digitali avanzate previste dai teorici dell'*information literacy* (Lana), per non parlare della scarsa frequentazione del pensiero statistico e probabilistico, così centrale per capire e utilizzare proficuamente i metodi della cosiddetta *data science*, per non dire della loro applicazione ai testi letterari. Siamo dunque di fronte a una doppia scommessa: promuovere le competenze testuali e le capacità di lettura critica attraverso strumenti computazionali, ma prima ancora matematici, e allo stesso tempo promuovere un approccio sinergico e cooperativo tra i campi del sapere, superando la separazione tra saperi umanistici e saperi tecnico-scientifici che affligge il mondo della formazione e la nostra cultura più in generale. Questa nuova alleanza rappresenta a nostro parere il modo migliore per garantire un futuro alla cultura del testo e per rispondere alla sfida della complessità (Roncaglia).

9. Bibliografia

Bianchi, Federico, Debora Nozza e Dirk Hovy. "FEEL-IT: Emotion and Sentiment Classification for the Italian Language". *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, Association for Computational Linguistics, pp. 76-83.

Ciotti, Fabio. "What's in a Topic Model? Critica teorica di un metodo computazionale per l'analisi del testo". *Testo e Senso*, n. 18, 2017.

Croxall, Brian, e Diane Jakacki (a cura di). *Debates in Digital Humanities Pedagogy*. U of Minnesota P, *forthcoming*.

D'Annunzio, Gabriele. *Il Piacere*. A cura di Angelo Piero Cappello, BUR Rizzoli, 2021.

Dascalu, Mihai, et al. "*ReaderBench*, an Environment for Analyzing Text Complexity and Reading Strategies". *Artificial Intelligence in Education. AIED 2013. Lecture Notes in Computer Science*, a cura di H. Chad Lane, Kalina Yacef, Jack Mostow e Philip Pavlik, Springer, 2013, pp. 379-88.

Firth, John Rupert. *Papers in Linguistics 1934–1951*. Oxford UP, 1957.

Gavin, Michael. *Literary Mathematics: Quantitative Theory for Textual Studies*. Stanford UP, 2022.

Giusti, Simone. *Didattica della letteratura 2.0*. 2015. Carocci, 2020.

Hirsch, Brett D,.editor. *Digital Humanities Pedagogy: Practices, Principles and Politics*. Open Book Publishers, 2012.

Hogan, Patrick Colm, et al. (a cura di). *The Routledge Companion to Literature and Emotion*. Routledge, 2022.

Iannella, Alessandro. "'Ok Google, vorrei parlare con la poetessa Saffo': Intelligenza Artificiale, assistenti virtuali e didattica della letteratura". *Thamyris, nova series: Revista de Didáctica de Cultura Clásica, Griego y Latín*, vol. 10, 2019, pp. 81-104.

Jannidis, Fotis, et al. "Comparison of Methods for the Identification of Main Characters in German Novels". *Digital Humanities 2016: Conference Abstracts*, a cura di Maciej Eder, Jan Rybicki, Jagiellonian University & Pedagogical University, 2016, pp. 578-82.

Jasonov Milán. "A Network Map of The Witcher". *Nightingale. Journal of the Data Visualization Society*, 2021.

Jockers, Matthew L. *Macroanalysis: Digital Methods and Literary History*. U of Illinois P, 2013.

John, Markus, et al. "A Visual Approach for the Comparative Analysis of Character Networks in Narrative Texts". *2019 IEEE Pacific Visualization Symposium*, IEEE, 2019, pp. 247-56.

Kim, Hoyeol. "Sentiment Analysis: Limits and Progress of the Syuzhet Package and Its Lexicons". D*igital Humanities Quarterly*, vol. 16, n. 2, 2022.

Lana, Maurizio. *Introduzione all'information literacy: storia, modelli, pratiche*. Editrice Bibliografica, 2020.

Landow, George P. *Hypertext 2.0: The Convergence of Contemporary Critical Theory and Technology*. Johns Hopkins UP, 1997.

Liu, Bing. *Sentiment analysis: mining opinions, sentiments, and emotions*. Cambridge UP, 2015.

Macaya, María, e Manuel Perea. "Does Bold Emphasis Facilitate the Process of Visual-Word Recognition?". *The Spanish Journal of Psychology*, vol. 17, 2014.

Magherini, Simone. "Strumenti informatici per la letteratura italiana". *Didattica della letteratura italiana. Riflessioni e proposte educative*, a cura di Gino Ruozzi, Gino Tellini, Le Monnier Università-Mondadori Education, pp. 173-84.

Moretti, Franco. *A una certa distanza: leggere i testi letterari nel nuovo millennio.* Carocci, 2020.

—. "Conjectures on World Literature". *The New Left Review*, II, 1, 2000.

—. "Network Theory, Plot Analysis". *Stanford Literary Lab Pamhplets*, n. 2, 2011.

Plecháč, Petr, et al. (a cura di). *Tackling the Toolkit. Plotting Poetry through Computational Literary Studies*. Institute of Czech Literature of the Czech Academy of Sciences, 2021.

Reagan, Andrew J., et al. "The Emotional Arcs of Stories are Dominated by Six Basic Shapes". *EPJ Data Science*, vol. 5, 2016.

Retaegui, Eliseo, *et al.* "Can Text Mining Support Reading Comprehension?". *Methodologies and Intelligent Systems for Technology Enhanced Learning, 9th International Conference*, Rossella Gennari, Pierpaolo Vittorini, Fernando De la Prieta, et al. (a cura di), Springer, 2020, pp. 37-44.

Rhody, Lisa. "Topic Modeling and Figurative Language". *Journal of Digital Humanities*, vol. 2, no. 1.

Riva, Francesca. *Insegnare letteratura nell'era digitale*. Edizioni ETS, 2017.

Roncaglia, Gino. *L'età della frammentazione: cultura del libro e scuola digitale*. Laterza, 2020.

Sanocki, Thomas, e Mary C. Dyson. "Letter Processing and Font Information During Reading: Beyond Distinctiveness, where Vision Meets Design". *Atten Percept Psychophys*, vol. 74, 2012, pp. 132-45.

Sarti, Luca. "Narrare i classici nell'era digitale. Dai *tweet* alle *emoji*: il caso di Pinocchio". *Una/Κοινῇ*, n. 1, 2021, pp. 152-98.

Schreibman, Susan, Raymond George Siemens e John Unsworth (a cura di). *A Companion to Digital Humanities*. Blackwell Publishing, 2004.

—. *A New Companion to Digital Humanities*. Wiley-Blackwell, 2016.

Sinclair, Stéfan, e Geoffrey Rockwell. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT P, 2016.

Smeets, Roel. *Character Constellations. Representations of Social Groups in Present-Day Dutch Literary Fiction*. Leuven UP, 2021.

Swafford, Annie. "Problems with the Syuzhet Package". *Anglophile in Academia: Annie Swafford's Blog*.

Underwood, Ted. *Distant horizons: Digital Evidence and Literary Change*. The U of Chicago P, 2019.

Valitutti, Alessandro, e Cecilia Dalla Torre. "'Io non ho paura': Sentiment analysis nell'analisi di testi narrativi". *Proceedings of Didamatica-21*, AICA, 2021, pp. 166-69.

10. Sitografia

Biblioteca della letteratura italiana. http://www.letteraturaitaliana.net/.

Biblioteca italiana. http://www.bibliotecaitaliana.it/.

Bionic Reading. https://bionic-reading.com/.

Eliza. http://psych.fullerton.edu/mbirnbaum/psych101/eliza.htm.

FEEL-IT. https://github.com/MilaNLProc/feel-it.

Gephi. https://gephi.org/.

Google Fonts. https://fonts.google.com/.

Hedonometer. https://hedonometer.org/.

Indire. https://www.indire.it/.

Laurence Antony's Website – software. http://www.laurenceanthony.net/software.html.

Liber Liber. https://www.liberliber.it/online/.

Jockers, Matthew. *Introduction to the Syuzhet Package*. https://cran.r-project.org/web/packages/syuzhet/vignettes/syuzhet-vignette.html.

Plotly. https://plotly.com/python/.

Readerbench. https://readerbench.com/.

Shiny. https://shiny.rstudio.com/.

Sobek. http://sobek.ufrgs.br/index-en.html.

Spacy. https://spacy.io/.

Streamlit. https://streamlit.io/.

Syuzhet. https://github.com/mjockers/syuzhet.

UmBERTo. https://github.com/musixmatchresearch/umberto.

Vonnegut, Kurt. *Shapes of Stories*. https://www.youtube.com/watch?v=oP3c1h8v2ZQ; https://www.youtube.com/watch?v=j9Qsiu8qqvA.

Voyant Tools. https://voyant-tools.org/.

Wikisource. https://it.wikisource.org/wiki/Pagina_principale.