Il corpus di prosa letteraria del progetto RIND (1830-1930). Assunti teorici e vincoli pratici

Stefano Ondelli[1], Pietro Mazzarisi[2]
[1] Università degli studi di Trieste, Italia - sondelli@units.it
[2] Università degli studi di Trieste, Italia - pietro.mazzarisi@units.it

ABSTRACT[1]
Questo contributo descrive la componente prodotta in Italia di un corpus di 1.000 testi comprendenti romanzi e raccolte di racconti scritti da autori italiani e tradotti da altre lingue e pubblicati tra il 1830 e il 1930. In particolare, si sofferma sui criteri ipotizzati per il bilanciamento del corpus atti a garantirne la rappresentatività e la possibilità di indagarne le caratteristiche con metodi quantitativi: per es. distribuzione temporale, lunghezza, genere e grado di appartenenza al canone letterario dei testi da una parte, genere e provenienza degli autori dall'altra. Si illustrano anche le difficoltà derivanti dalla scarsa conoscenza della popolazione di riferimento e i problemi pratici legati al reperimento dei testi.
Il corpus qui illustrato è oggetto di studio del progetto di ricerca "Reading the Italian Novel at a Distance (1830-1930)" (RIND). Scopo finale sarà la verifica empirica della periodizzazione tradizionale delle fasi dello sviluppo della narrativa in Italia (dal romanzo storico e verista al modernismo) e la valutazione del contributo a tale sviluppo apportato dai modelli esogeni accessibili tramite le traduzioni.
Parole chiave: discorso riportato; linguistica dei corpora; romanzo italiano; storia della letteratura italiana; traduzioni.

ABSTRACT
The corpus of literary prose of the RIND project (1830-1930). Theoretical assumptions and practical constraints. This paper describes the non-translated component of a corpus of 1,000 texts comprising novels and short story collections written by Italian authors and translated from other languages, published between 1830 and 1930. In particular, the paper discusses possible criteria for balancing the corpus in order to ensure it is representative and can be investigated through quantitative methods: e.g. the chronological distribution, size, genre and canonicity of texts on the one hand, and the gender and origin of the authors on the other. The difficulties arising from the lack of knowledge of the overall reference population, compounded by practical problems in finding the texts, are also illustrated.
The corpus illustrated here provides the material for the research project 'Reading the Italian Novel at a Distance (1830-1930)' (RIND). Its final objectives include the empirical verification of the traditional periodization of literary prose in Italy (from historical and realist novels to Modernism) and the evaluation of the contribution provided by foreign models made accessible to the general public through translations.
Keywords: reported speech; corpus linguistics; Italian novel; history of Italian literature; translations.

1. IL PROGETTO RIND (1830-1930)

Nel progetto interdisciplinare "Reading the Italian Novel at a Distance (1830-1930) - RIND" (PRIN 2022JAYFJH) confluiscono studi linguistici computazionali e dei corpora, teoria letteraria e narratologia cognitiva, scienza dei dati e statistica. La cornice metodologica è fornita dal concetto di "distant reading" (Moretti, 2013), atto a valutare la presenza, la distribuzione e il ciclo di vita dei fenomeni letterari attraverso l'analisi quantitativa di grandi collezioni di testi. I metodi quantitativi di analisi dei dati testuali (Lebart et al. 1998) sono stati applicati a numerosi campi, dalla storiografia alla giurisprudenza. Negli studi letterari, questi approcci hanno preso le mosse da stilometria classica e attribuzione dell'autorialità (Burrows, 1987; Hoover, 2013), per poi ampliarne il campo di applicazione (Jockers, 2013; Piper, 2018; Underwood, 2019).
Il progetto RIND analizza un corpus di 1.000 romanzi e raccolte di racconti di autori italiani o tradotti da originali stranieri e pubblicati in Italia tra il 1830 e il 1930, ovvero un periodo che, approssimativamente, va dalla nascita e affermazione del genere romanzesco (storico e non), passa dal Realismo del XIX secolo e giunge fino al Modernismo del primo Novecento. Le fasi e i compiti, distribuiti tra le unità di ricerca partecipanti, comprendono: la creazione del corpus; la definizione formale del Discorso Riportato (DR) a livello linguistico; la compilazione di repertori lessicali relativi all'ambientazione delle narrazioni; lo sviluppo

[1] Pur essendo questo contributo il risultato di uno sforzo congiunto, specifichiamo che Ondelli è responsabile dei §§ 1 e 2 e Mazzarisi dei §§ 3 e 4.

di metodi per estrarre i dati testuali, identificarne e rappresentarne l'evoluzione diacronica (serie temporali) e analizzarne il ciclo di vita; la valutazione delle tradizionali periodizzazioni dei movimenti letterari e la classificazione di generi e autori attraverso approcci quantitativi che incrocino l'illustrazione degli sviluppi sociali (l'ambientazione) e lo sviluppo delle tecniche narrative (DR); lo studio dell'interazione tra modelli stranieri e produzione italiana.
I dati linguistici che il progetto intende estrarre dal corpus appartengono a due ambiti principali: il DR e l'ambientazione delle vicende. Nel primo caso, l'obiettivo è rilevare i cambiamenti diacronici nella resa della "coscienza" dei personaggi di finzione, la cui istanziazione formale è identificabile nel DR in quanto insieme di strutture morfo-sintattiche che eccedono i confini frastici e investono il livello pragmatico-testuale. L'ipotesi di partenza è che l'evoluzione delle strategie di resa del DR rifletta la transizione tra diversi periodi letterari, come tradizionalmente riconosciuto dalla critica. A livello teorico, il DR può essere considerato parte della "polifonia" narrativa, introdotta da Bachtin (1997 [1974]) come dialogismo e ampiamente adottata nella teoria letteraria (Ducrot, 1984; Roulet, 2011). Finora la ricerca linguistica sul DR si è concentrata principalmente sulla tipologia, portando alla classificazione delle strutture dirette rispetto a quelle indirette e all'identificazione dei marcatori formali (Mortara Garavelli, 1995; Calaresu, 2004). Sono però pochi i tentativi di rilevare il DR con metodi quantitativi (Byszuk et al., 2020).
Anche nel caso dell'"ambientazione narrativa", il dato linguistico contribuisce a cogliere quantitativamente la nozione di *setting* come definita dalla geo-critica e dalle *spatial humanities* (per es., Cooper et al., 2016; Bushell, 2020). Pur riconoscendo il valore della vasta letteratura sull'impiego del *Named Entity Recognition* (NER) nella ricerca letteraria (Nouvel et al., 2016; Archana et al., 2018; Frontini et al., 2020), occorre notare che non è possibile limitarsi ai nomi propri (in particolare toponimi), ma è necessario includere qualsiasi luogo (naturale e architettonico, reale e fittizio) che contribuisce all'ambientazione in senso lato: città vs. campagna, interno vs. esterno, pubblico vs. privato ecc., con tutte le possibili sottocategorizzazioni (per es. edifici residenziali, commerciali, industriali, privati, pubblici) e sovrapposizioni (per es. commerciali vs. pubblici).
Alle ambientazioni contribuiscono inoltre i ruoli sociali dei personaggi, come figure professionali, gradi militari, cariche istituzionali e titoli religiosi. Naturalmente, i repertori lessicali si scontrano con la difficoltà di delimitare elenchi potenzialmente aperti e diversificati in base al periodo storico, mentre l'estrazione dei dati deve affrontare la questione della polisemia lessicale e della variabilità ortografica. Tuttavia, anche in questo caso, l'ipotesi di ricerca prevede che i cambiamenti nelle ambientazioni (per es. dalla campagna alla città) e nei ruoli sociali (per es. dai contadini agli operai) siano in grado di riflettere l'evoluzione delle tematiche attraverso le diverse fasi dello sviluppo sociale, nonché il passaggio tra diversi periodi letterari in Italia e all'estero (nelle traduzioni).
Infine, nel confronto tra le opere di autori italiani e le traduzioni di autori stranieri, si ipotizza che l'evoluzione diacronica dei dati testuali relativi ai due ambiti di indagine illustrati sopra getti luce anche sul contributo all'evoluzione del sistema letterario italiano da parte di modelli esogeni accessibili al grande pubblico (Even-Zohar, 1990; Toury, 1995).

2. IL CORPUS: PREMESSE TEORICHE

L'arco temporale (1830-1930) è stato scelto poiché copre il periodo che parte all'incirca dai primi esempi di romanzo in Italia fino all'ascesa del Modernismo e perché i testi possono essere liberamente consultati, manipolati e diffusi senza violare i diritti d'autore (Legge 22 aprile 1941, n. 633). Il corpus, una volta completato e bilanciato, sarà reso liberamente accessibile, scaricabile e consultabile per ulteriori ricerche sul sito del progetto (https://rind.units.it/home/) o su *Github*, con licenza CC BY-NC-ND 4.0.
La compilazione del corpus, atta a garantirne autenticità, rappresentatività e bilanciamento, è una delle fasi critiche dal punto di vista teorico e pratico (Bode, 2018; Ondelli, 2018; Odebrecht, Burnard & Schöch, 2021). Il progetto prevede il reperimento di 1.000 romanzi e raccolte di racconti e novelle, suddivisi in due subcorpora di 500 opere di autori italiani e 500 traduzioni da lingue straniere. I criteri di bilanciamento ipotizzati sono i seguenti: per gli autori, genere, origine geografica e numero di opere; per i testi, anno di pubblicazione, lunghezza, appartenenza al canone letterario o alla paraletteratura; nel caso delle traduzioni si aggiunge la lingua di partenza. I criteri applicati agli autori rendono conto delle variazioni diastratiche e diatopiche e sono mirati a evitare la sovrarappresentazione di singole personalità. I criteri relativi ai testi riguardano la copertura in diacronia e la presenza di generi letterari diversi, oltre a evitare la sovrarappresentazione di determinate opere a causa delle loro dimensioni.
Il problema principale ai fini del bilanciamento risiede nella mancata conoscenza della popolazione di riferimento, non esistendo un repertorio completo delle opere letterarie in prosa pubblicate in Italia (come

in altri paesi) nel periodo considerato, tantomeno corredato dei metadati qui presi in considerazione. Oltre a ciò, sorgono difficoltà teoriche in relazione ai seguenti aspetti: definizione dei fattori di possibile impatto linguistico in diatopia (luogo di nascita dell'autore vs. luogo di residenza vs. luogo di pubblicazione del testo); attribuzione delle singole opere al canone o alla paraletteratura; definizione delle fasce e dei limiti massimi e minimi di lunghezza dei testi; trattamento dei racconti pubblicati in momenti diversi e raccolti successivamente in volume; certezza della lingua di partenza della traduzioni. Poiché tali criteri sono subordinati alla fattibilità del corpus in relazione alla disponibilità di tempo e risorse, la loro valutazione in dettaglio può essere condotta più agevolmente dopo una prima ricognizione della disponibilità e reperibilità effettiva dei testi a partire da archivi digitali e biblioteche on-line come *Biblioteca Italiana, Progetto Manuzio, Project Gutenberg, Wikisource, Internet Archive, Hathi Trust* ecc., oltre a corpora costituiti in occasione di studi precedenti e messi gentilmente a disposizione da altri gruppi di ricerca.
Infine va sottolineato che i criteri adottati contraddistinguono il corpus RIND da altri corpora di (o con) prosa narrativa in lingua italiana. Per es., l'ELTeC (Odebrecht, Burnard & Schöch, 2021), nella sua sezione italiana, conta 100 romanzi, mentre il corpus RIND dieci volte tanti (se consideriamo anche le traduzioni); il periodo storico coperto dall'ELTeC è minore (1840-1920); ogni sua sezione è destinata a una singola nazione e non prevede la presenza di traduzioni. Altre differenze salienti riguardano il bilanciamento, che nell'ELTeC condiziona la compilazione *ab origine*. Infatti il progetto europeo prefissava per ciascun subcorpus nazionale almeno il 10%-50% di donne; la presenza di almeno 9-11 autori con 3 romanzi; almeno il 30% di opere appartenenti al canone e almeno il 30% di opere non canoniche (con la discriminante "canone/non canone" stabilita in base al numero di ristampe realizzate tra il 1970 e il 2009); almeno il 20% costituito da novelle (identificate come testi di lunghezza compresa tra le 10.000 e le 50.000 parole) e almeno il 20% composto di romanzi di ampio respiro (intendendo i testi di 200.000 e oltre parole); per ultimo, la lunghezza dei testi veniva distribuita in tre fasce: "brevi" (tra le 10.000 e le 50.000 parole), "medi" (tra le 50.000 e le 100.000 parole) e "lunghi" (oltre le 100.000 parole). Questi criteri hanno consentito al progetto europeo di ottenere risultati tendenzialmente omogenei nella creazione dei vari subcorpora nazionali, ma hanno limitato, e sicuramente condizionato, la rappresentatività del campione in relazione alla popolazione di riferimento (peraltro non nota).
Anche il progetto RIND ha affrontato il problema della insufficiente conoscenza della popolazione di riferimento, soprattutto nel caso della componente relativa alla paraletteratura e alle traduzioni. Tuttavia, piuttosto che impostare il bilanciamento del corpus secondo criteri predeterminati e arbitrari, si è optato per un approccio più spiccatamente – per così dire – *bottom up* che, tra l'altro, tenesse in considerazione gli aspetti legati alla fattibilità del progetto. In altre parole, dopo aver stabilito un numero congruo ma realistico di testi che renda possibile l'indagine statistica delle serie temporali dei fenomeni linguistici presi in considerazione (i 500 testi di ciascun subcorpus si concretizzano in poco meno di 5 testi per ognuno dei 101 anni considerati, con una media presunta di circa 250.000 parole per anno, stante la lunghezza media dei romanzi moderni), si è proceduto a un sondaggio preliminare teso a recuperare il maggior numero possibile di testi adatti disponibili in formato digitale e caratterizzati da un grado di "pulizia" sufficiente a ipotizzarne l'impiego nelle ricerche previste. Solo al termine di questa prima raccolta sono state prese decisioni ponderate alla luce della distribuzione interna dei testi secondo i criteri già considerati (distribuzione temporale, lunghezza, genere letterario, genere dell'autorialità, canonicità ecc.), basandosi sull'esistente ed eventualmente prevedendo integrazioni ad hoc per risolvere gli squilibri più evidenti. In altre parole, per quanto la raccolta dei testi già disponibili in formato elettronico non possa in alcun modo essere considerata un'immagine della popolazione esistente, alcune distribuzioni (per es. la maggiore o minore numerosità dei testi in alcune decadi, la minore o maggiore presenza delle autrici nel corso del tempo, l'andamento della lunghezza media dei testi, la distribuzione dei generi letterari ecc.) possono fornire indicazioni utili per il bilanciamento, soprattutto alla luce di eventuali conferme provenienti dalla storia della letteratura e dell'editoria italiana (e dei loro contatti con l'estero).

3. IL SUBCORPUS DI OPERE DI AUTORI ITALIANI: UN CENSIMENTO PRELIMINARE

Per quanto riguarda il subcorpus di opere di autori italiani, il sondaggio delle fonti già disponibili ha portato alla raccolta di 1.198 testi caratterizzati da un grado di "pulizia" accettabile ai fini delle analisi previste dal progetto. In assenza di dati certi sulla popolazione reale di riferimento (cioè le opere effettivamente pubblicate nel periodo considerato), si può assumere come riferimento l'andamento cronologico dei testi reperiti, illustrato in Fig. 1, possibilmente messo in relazione ai principali snodi della storia della letteratura e dell'editoria in Italia (Cadioli & Vigini, 2018; Ragone, 1999; Turi, 1997).

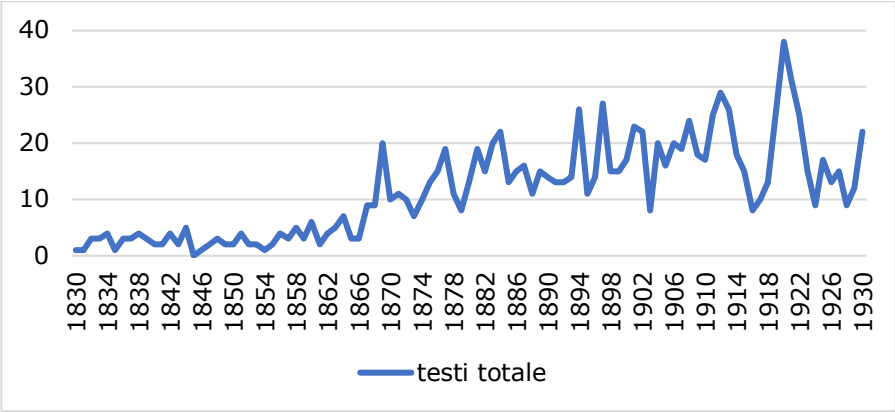

Figura 1. Distribuzione cronologica delle opere di autori italiani reperite

Per esemplificare gli assunti illustrati sopra (par. 2) per quanto concerne la significatività della distribuzione delle opere reperite, l'evidente penuria nel periodo precedente l'Unità può essere messa in relazione alla scarsa reattività del panorama nazionale alle novità letterarie almeno fino alla Scapigliatura e al Verismo, come testimoniano il notissimo saggio di Madame de Staël (1816) all'origine del dibattito che segna l'inizio del Romanticismo nella penisola a scapito di Arcadia e Neoclassicismo, ma anche il successivo contributo a firma di Niccolò Tommaseo e Giuseppe Bianchetti (1832) che, proprio nell'anno della morte di Walter Scott, illustra opinioni critiche e favorevoli in relazione all'introduzione del romanzo storico nella letteratura italiana. Un altro fattore determinante è da ricercarsi nell'iniziale mancanza di un'editoria attiva a livello nazionale, quando invece, nel decennio successivo all'Unità, l'ampliamento del pubblico dei lettori e delle attività editoriali determina un netto incremento delle pubblicazioni che si dimostrerà altalenante nei primi decenni del XX secolo, con battute d'arresto (probabilmente dovute a questioni di tipo economico e censorio) in corrispondenza della Grande Guerra e dell'avvento del Fascismo.
Prima di affrontare vari aspetti legati alle variabili considerate per il bilanciamento del corpus, occorre notare che la decisione di dimensionare il corpus in base al numero di testi è arbitraria: perché 1.000 testi e non 100 o 10.000? Perché non considerare piuttosto il numero di tokens? Benché gli esperimenti di Distant Reading previsti dal progetto non risentano particolarmente dell'accuratezza filologica, sulla scorta di un'ampia messe di studi precedenti (Jockers, 2013; Piper, 2018; Underwood, 2019) è senza dubbio preferibile includere lo stesso numero di testi completi nei due subcorpora, nonostante la possibilità che emergano differenze in termini di tokens. Considerate sia la scarsa disponibilità di testi in italiano digitalizzati con una certa cura sia le risorse economiche del progetto eventualmente destinabili a ulteriori reperimenti e digitalizzazioni, la soglia di 1.000 testi totali rappresenta un traguardo realistico e comunque ben superiore a progetti analoghi giù realizzati (cfr. Odebrecht, Burnard & Schöch, 2021).
La scelta di raccolte di racconti e novelle – oltre ai romanzi – quale format narrativo ammesso discende dall'ipotesi secondo cui le raccolte presentino una certa unitarietà di stile (DR) e contenuti (agentivi e ambientazioni), mentre la brevità dei singoli racconti condurrebbe a squilibri nel bilanciamento del corpus. Sempre per evitare tali squilibri, le soglie minima e massima di lunghezza dei testi sono state ipotizzate tra le 50.000 e 150.000 parole. All'atto pratico, la distinzione tra "romanzo", "racconto" e "novella" in base a semplici considerazioni dimensionali non risulta agevole; per es. è definita novella *Maestra* di Clarice Tartufari (1887) con 15.392 parole, ma anche *Gli americani di Rabbato* di Luigi Capuana (1912) con 31.366 parole. È invece definito "racconto" *Storia della Tonia* di Giovanni Battista Cereseto (1856) con 17.046 parole, ma anche *Catene* di Cordelia o Virginia Tedeschi Treves (1882) con 46.387 parole. Inoltre, l'etichetta "racconto" viene talvolta attribuita anche a testi divisi in più volumi come *La montanara* di Anton Giulio Barrili (1886), con 117.700 parole. Anche la letteratura per l'infanzia presenta grandi oscillazioni: *Il principino* di Ida Baccini (1881) con 16.996 parole è definito "racconto", così come *Uccelletto nero* di Guido Fabiani (1908) con il triplo delle parole (50.571). D'altra parte, sono chiamati "romanzi" testi assai brevi come *Amori moderni* di Grazia Deledda (1907) con 12.709 parole o *La colonia felice* di Carlo Dossi (1874) con 16.228 parole. Infine, per quanto riguarda le raccolte, bisogna prestare attenzione ristampe o riedizioni per non includere nel corpus due copie dello stesso testo. Capuana è un caso emblematico: il testo intitolato *Il dottor Ficicchia* compare con questo titolo nella raccolta *Nostra gente* del 1915, mentre una precedente versione – stavolta intitolata *Il dottore dei poveri* – era già stata inclusa nel 1894 nella raccolta di racconti dal titolo *Le paesane*.

Ancora in relazione alla lunghezza, sono state scartate le opere reperite che eccedevano il limite massimo previsto (150.000 parole); tuttavia, quando permettevano di colmare una certa lacuna temporale nel corpus (specialmente nei primi quarant'anni considerati), si è preferito mantenerle, come nel caso di *Cent'anni* di Giuseppe Rovani (1859) con 413.000 parole e i due testi di Nievo *Le confessioni di un italiano* (1858) con 332.000 parole e *Le confessioni di un ottuagenario* (1867) con 329.000 parole.
Per quanto concerne la datazione delle opere, in precedenza Bertinetto (2003) si era basato sulle date di nascita e morte degli autori per risolvere il problema delle pubblicazioni a puntate, riedizioni, revisioni ecc. Ai nostri fini si è, invece, preferito optare per la data dell'edizione princeps in volume sia dei romanzi che delle raccolte di novelle e racconti, salvo nei casi noti di forti interventi anche sull'assetto linguistico, per es. *I promessi sposi* di Alessandro Manzoni, di cui è stata inclusa la "Quarantana". Degno di menzione *Il castello di Trezzo* di Giambattista Bazzoni, pubblicato per la prima volta nel 1828 (quindi al di fuori del periodo considerato nella ricerca) e, dopo revisioni e correzioni, riedito nel 1838, data con la quale appare nel nostro corpus. Inoltre, abbiamo deciso di includere edizioni successive quando non è stata reperita la princeps in formato digitale (per es. *Il diavolo al Pontelungo* di Riccardo Bacchelli, edizione 1951, princeps 1927), poiché si è ritenuto che eventuali interventi non riguardanti l'assetto linguistico del testo (impaginazione, grafia ecc.) non abbiano alcun impatto sulle analisi che intendiamo realizzare, dato lo scarso peso del rigore filologico.
L'appartenenza o meno di un'opera al canone letterario dipende non solo dal periodo storico (il canone odierno è diverso da quello di un secolo fa; cfr. Ricci, 2013) ma anche dal contesto culturale: la distinzione tra letteratura alta e di consumo è relativamente netta in Italia, molto più sfumata nei Paesi anglofoni, con le ovvie ricadute in termini di classificazione delle traduzioni. Tuttavia, per verificare se la tradizionale periodizzazione delle correnti letterarie sia in qualche modo collegata all'evoluzione del DR e delle ambientazioni, pare necessario includere opere di ampia circolazione, seppur limitata nel tempo, per evitare di riproporre una storia della letteratura basata esclusivamente sul canone. I problemi di attribuzione sono molteplici, a cominciare dai possibili incroci tra letteratura di genere e valore letterario. Analogamente, poiché si è deciso di includere la letteratura per l'infanzia, è ovvio che, accanto a Collodi, anche autori come Salgari hanno pieno diritto a comparire nel corpus, e considerazioni analoghe sono valide nel caso della letteratura rosa o poliziesca. Infine, un altro problema riguarda se assegnare o meno al canone l'intera opera di un autore oppure solo alcune opere (e quali i criteri di eventuali esclusioni?). Per es., se Luigi Capuana può essere considerato senza discussioni un autore canonico, lo è di conseguenza anche la sua novella per l'infanzia *Scurpiddu* (1898)? Per prendere decisioni dotate di una certa oggettività, per il subcorpus qui considerato si è deciso di adottare come riferimento quattro storie della letteratura pubblicate negli ultimi 50 anni (Asor Rosa, 1985; Contini, 1992; Guglielmino & Grosser, 2001; Ferroni, 2012). Definiamo quindi canonici titoli e autori che compaiono in almeno due di esse, fatta salva ovviamente l'eventuale classificazione degli stessi come esempi di paraletteratura (per es. Carolina Invernizio e Francesco Mastriani in Gugliemino & Grosser, 2001).
Anche il genere degli autori incide sulla distribuzione in base al canone. In una storia letteraria come quella italiana, in cui l'affermazione della scrittura femminile giunge tardi rispetto al panorama europeo, autrici come Virginia Tedeschi Treves (Cordelia) e Anna Gentile Vertua non sono considerate canoniche, ma la prima è imprescindibile nella letteratura per l'infanzia, la seconda, con oltre 150 romanzi di discreto successo, è fautrice di una maggiore indipendenza femminile malgrado il sentimentalismo.
A questo proposito, il genere degli autori è un'altra variabile interessante di cui ci sfugge l'effettiva incidenza nel periodo considerato, anche in considerazione dell'incrocio con i generi letterari. Ancor più problematica è la ricaduta sul bilanciamento del corpus della provenienza geografica di opere e autori, stante la forte influenza della diatopia nella storia dell'italiano. A questo proposito, data la difficoltà nello scegliere un criterio di selezione univoco (luogo di nascita, a prescindere dai confini nazionali? residenza ufficiale o vissuto personale? sede della casa editrice?), pur registrando il metadato del luogo di nascita degli autori, il bilanciamento del corpus non ha tenuto conto della diatopia.

4. IL SUBCORPUS DI OPERE DI AUTORI ITALIANI: UNA SELEZIONE PLAUSIBILE

Tenendo conto della compresenza simultanea e degli incroci di tutti i criteri esposti sopra, sulla base dei testi reperiti in questa prima fase di compilazione la cernita delle 500 opere componenti il subcorpus di narrativa scritta direttamente in italiano è stata mirata a conservare il maggior numero possibile di opere di scrittrici (per permettere eventuali campionamenti) e a distribuire in maniera bilanciata i testi in base all'autorialità, onde evitare l'eccessiva incidenza di singoli idioletti. Al momento (ma il corpus è ancora da considerarsi in fieri e perfezionabile), le 500 opere sono riconducibili a 214 autorialità, di cui 45 donne

(21%) e 171 uomini (79%). Le autrici sono riconducibili a 121 opere (24%) mentre gli autori a 379 opere (76%). La Fig. 2 ne illustra la distribuzione cronologica.

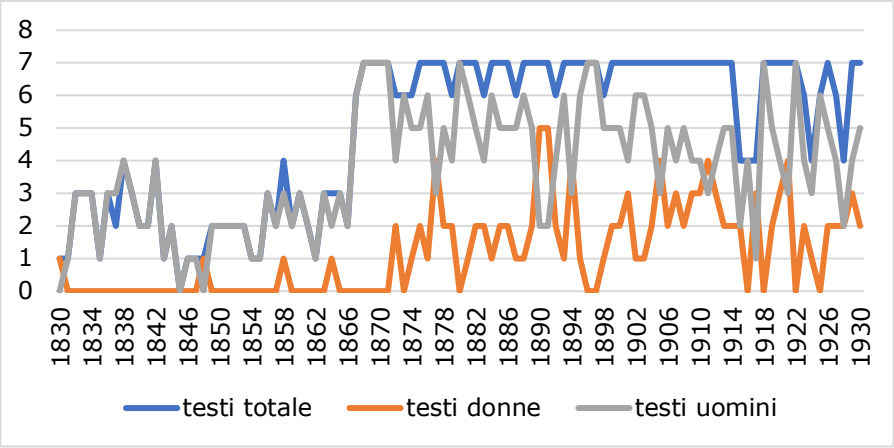

Figura 2. Distribuzione cronologica delle opere per anno, donne e uomini

Si tratta di una distribuzione realistica rispetto alla popolazione di riferimento? Sull'incremento del numero di opere edite nel tempo, valgono le considerazioni svolte sopra a commento della Fig. 1. Sul lato della distribuzione del genere, anche se in un periodo storico diverso, la selezione è confortata dal confronto con i vincitori dei maggiori premi letterari italiani: dal 1947 a oggi il premio *Strega* è stato attribuito a 13 donne (16,7%) e dal 1963 il premio *Campiello* è stato appannaggio di 17 donne (27%). Per quanto riguarda, invece, la lunghezza delle opere inserite nel subcorpus qui considerato, 10 non superano le 15.000, mentre 17 eccedono le 150.000 parole. La Fig. 3 illustra la distribuzione cronologica della lunghezza media e dimostra come, dopo una fase iniziale di definizione del genere "romanzo" e di penuria di testi disponibili, che registra differenze notevoli nel numero di opere selezionate per anno (fino alla completa assenza) e obbliga a mantenere testi molto brevi o molto lunghi, successivamente all'Unità si verifica una certa stabilizzazione.

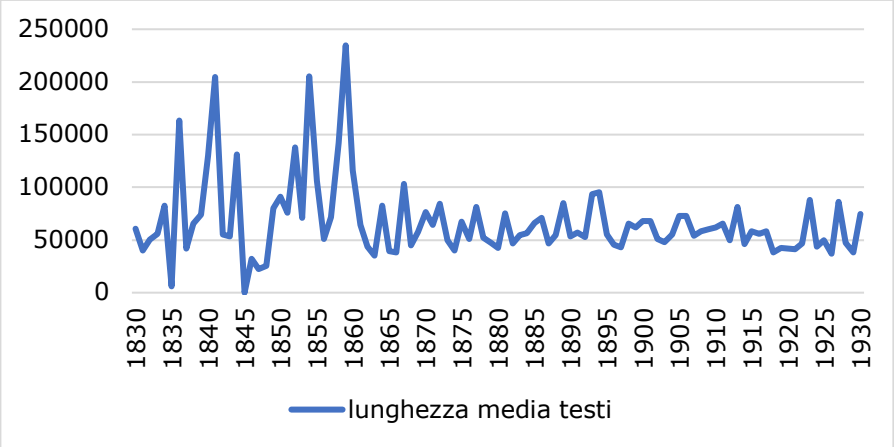

Figura 3. Distribuzione cronologica della lunghezza media dei testi in parole

Per motivi di spazio, i dati qui illustrati forniscono un'immagine solo parziale del subcorpus raccolto finora; peraltro una valutazione esauriente della sua rappresentatività dovrà tener conto della compresenza e dell'incrocio di tutti i criteri adottati, al fine di stabilire quali lacune colmare e che miglioramenti apportare in termini di bilanciamento e rappresentatività tramite il reperimento e la digitalizzazione ad hoc di testi non ancora disponibili in formato elettronico. Al momento quest'ultimo passaggio pare doversi indirizzare alla sostituzione di 27 opere per un migliore bilanciamento della distribuzione esistente. Pare invece più complessa, alla luce della scarsa "pulizia" dei testi disponibili e dell'onerosità di scansioni "a tappeto", l'eventuale integrazione delle opere che coprono le prime tre decadi considerate; non da ultimo, occorrerà attendere almeno una selezione parziale del subcorpus comprendente le traduzioni per ipotizzare interventi mirati a un bilanciamento complessivo ottimale.

BIBLIOGRAFIA

Archana, G., Gupta, V. & Kumar, M. (2018). Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review, 29, 21-43.

Asor Rosa, (1985). Storia della letteratura italiana. Firenze: La nuova Italia.

Bachtin, M. (1997). Estetica e romanzo. Torino: Einaudi.

Bertinetto, P. M. (2003). Tempi verbali e narrativa italiana dell'Otto/Novecento. Alessandria: Edizioni dell'Orso.

Bode, K. (2018). A World of Fiction: Digital Collections and the Future of Literary History. Ann Arbor: University of Michigan Press. https://doi.org/10.3998/mpub.8784777.

Burrows, J. F. (1987). Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method. Oxford: Clarendon Press.

Bushell, S. (2020). Reading and Mapping Fiction: Spatialising the Literary Text. Cambridge: Cambridge University Press.

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A. & Eder M. (2020). Detecting Direct Speech in Multilingual Collection of 19th-century Novels. Proceedings of LT4HALA 2020. 1st Workshop on Language Technologies for Historical and Ancient Languages. Sprugnoli R. & Passarotti M. (Eds.). 100-104. Marseille: European Language Resources Association (ELRA).

Cadioli, A. & Vigini, G. (2018). Storia dell'editoria in Italia dall'Unità a oggi. Milano: Editrice Bibliografica.

Calaresu, E. (2004). Testuali parole. La dimensione pragmatica e testuale del discorso riportato. Milano: Franco Angeli.

Contini, G. (1992). La letteratura italiana Otto-Novecento. Milano: BUR.

Cooper, D., Donaldson, C. & Murrieta-Flores, P. (Eds.). 2016. Literary Mapping in the Digital Age. London: Routledge.

Ducrot, O. (1984). Le dire et le dit. Paris: Les éditions de Minuit.

Even-Zohar, I. (1990). The Position of Translated Literature within the Literary Polysystem. Poetics Today 11/1, 45-51.

Ferroni, G. (2012). Storia della letteratura italiana. 4 voll. Milano: Mondadori università.

Frontini, F., Brando, C., Byszuk, J., Galleron, I., Santos, D. & Stanković, R. (2020). Named Entity Recognition for Distant Reading in ELTeC. CLARIN Annual Conference 2020 (5-7 October). Navarretta, C. & Eskevich, M. (Eds.). 37-41. Madrid: Virtual edition. https://hal.science/hal-03160438/document.

Gugliemino, S. & Grosser, H. (2001). Il Sistema letterario 2000. 12 voll. Milano: Principato.

Hoover, D. L. (2013). Quantative analysis and literary studies. A Companion to Digital Literary Studies. Siemens R. & Schreibman S. (Eds.). 517-533. Hoboken: Wiley & Sons.

Jockers, M. L. (2013). Macroanalysis: Digital Methods and Literary History. Champaign: University of Illinois Press.

Lebart, L., Salem, A. & Lisette, B. (1998). Exploring Textual Data. Dordrecht: Kluwer Academic Publication.

Moretti, F. (2013). Distant Reading. London-New York: Verso Books.

Mortara Garavelli, B. (1995). Il discorso riportato. Grande grammatica italiana di consultazione. Vol. III. Renzi, L., Salvi, G. & Cardinaletti, A. (Eds.). 426-468. Bologna: Il Mulino.

Necker, A.-L.-G. (1816). Sulla maniera e la utilità delle Traduzioni. Biblioteca italiana, 1, Gennaio, pp. 9-18.

Nouvel, D., Ehrmann, M. & Rosset, S. (2016) Named Entities for Computational Linguistics. Hoboken: Wiley & Sons.

Odebrecht, C., Burnard, L. & Schöch, C. (Eds.). (2021). European Literary Text Collection (ELTeC), version 1.1.0, April 2021. COST Action (CA16204), Distant Reading for European Literary History. https://doi.org/10.5281/zenodo.4662444).

Ondelli, S. (2018). Treat Texts as Data but Remember They Are Made of Words: Compiling and Processing Corpora. Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. Tuzzi, A. (Ed.). 133-150. Cham: Springer.

Piper, A. (2018). Enumerations: Data and Literary Study. Chicago: The University of Chicago Press.

Ragone, G. (1999). Un secolo di libri. Storia dell'editoria dall'Unità d'Italia al post-moderno. Torino: Einaudi.

Ricci, L. (2013). Paraletteratura. Lingua e stile dei generi di consumo. Roma: Carocci.

Roulet, E. (2011). Polyphony. Discursive Pragmatics. Zienkowski, J., Östman, J-O. & Verschueren, J. (Eds.). 208-222. Amsterdam: John Benjamins.
Tommaseo, N. & Bianchetti, G. (1832). Discorsi critici intorno alla questione se giovi di ammettere o no nella letteratura italiana il romanzo storico. Treviso: Tipi di Gio. Paluello del fu Antonio.
Toury, G. (1995). Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins.
Turi, G. (Ed.). 1997. Storia dell'editoria nell'Italia contemporanea. Firenze: Giunti.
Underwood, T. (2019). Distant Horizons: Digital Evidence and Literary Change. Chicago: The University of Chicago Press.