Rivista Internazionale di Tecnica della Traduzione / International Journal of Translation 2021 (23), 213-218
DOI: 10.13137/2421-6763/33467 ISSN 1722-5906 (print) ISSN 2421-6763 (online)
https://www.openstarts.units.it/dspace/handle/10077/2503

Premessa agli articoli di Sciandra, Trevisani e Tuzzi e di Sciumbata, Nadalutti e Tringali

Floriana Carlotta Sciumbata[1]
Università di Trieste
fsciumbata@units.it

Gli studi illustrati nei due articoli che seguono sono basati sullo stesso corpus di 100 opere di narrativa, scritte originariamente in italiano e pubblicati tra il 1825 e il 1923, per un totale di quasi 7.500.000 *token*, di cui illustreremo i dettagli qui di seguito. Si tratta di una raccolta preparatoria che abbiamo utilizzato per testare metodologie e strumenti da applicare prossimamente nell'ambito del progetto *Distant Reading for European Literary History* (COST Action CA16204). Questo stabilisce precisi criteri[2] per la creazione del corpus da analizzare – denominato *ELTeC* (*European Literary Text Collection*) –, come la percentuale di autrici (inclusa tra il 10 e il 50%), la presenza di massimo tre titoli per ogni autore preso in analisi, la distribuzione di testi di diversa dimensione, l'inserimento di prime edizioni e l'impiego solo di opere non tradotte. Il corpus deve inoltre essere vario, bilanciato e rappresentativo del periodo preso in analisi; deve contenere opere letterarie e paraletterarie; e deve includere solo opere che non sono più protette dal diritto d'autore. Il nostro corpus preparatorio segue solo alcuni di questi criteri ed è da considerarsi come una raccolta ancora in fase di lavorazione, che in futuro sarà adattata alle richieste per i corpora del progetto COST Action.

1 Hanno contribuito all'estrazione dei dati usati nella premessa e alla sua redazione anche Michele Cortelazzo, Paolo Nadalutti, Stefano Ondelli, Andrea Sciandra, Matilde Trevisani, Luca Tringali e Arjuna Tuzzi.

2 Cfr. https://distantreading.github.io/sampling__proposal.html (consultato il 10/12/2021).

Le statistiche che illustreremo qui di seguito sono state calcolate dopo una fase preliminare di pulizia automatica del corpus. I file, di pubblico dominio, raccolti da database online, sono probabilmente stati realizzati con l'ausilio di tecnologie OCR (*optical character recognition*), cioè strumenti che permettono la digitalizzazione di documenti cartacei e immagini. Tuttavia, l'uso di OCR può determinare errori, in particolare dovuti alla presenza di invii a capo che spezzano parole o frasi; all'indicazione del numero pagina; alla presenza di intestazioni e piè di pagina. Si tratta di elementi che falsano le statistiche relative al testo (per esempio, al numero totale di *token*); rendono difficile l'utilizzo di software di estrazione automatica dei dati, come lemmatizzatori e *tagger* di parti del discorso; non permettono l'applicazione di strumenti come gli indici di leggibilità. Per la pulizia, è stato creato uno *script* nel linguaggio di programmazione *Python* che corregge in automatico gli errori frequenti tramite espressioni regolari.[3]

Dopo la pulizia, il corpus è stato sottoposto a tokenizzazione per determinare il numero esatto di *token* totali. Sono stati perciò esclusi tutti gli spazi e i segni di punteggiatura per distinguere le parole vere e proprie.

Il corpus che abbiamo utilizzato contiene i titoli elencati qui di seguito, pubblicati tra il 1825 e il 1923, per i quali specifichiamo anche autore e anno di pubblicazione.

autore	**titolo**	**anno**
Aleramo, Sibilla	Il passaggio	1919
Aleramo, Sibilla	Una donna	1907
Baccini, Ida	Il principino	1883
Barbieri, Ulisse	Lucifero	1871
Battaglia, Giacinto	La lega lombarda	1834
Bersezio, Vittorio	La carità del prossimo	1868
Bersezio, Vittorio	Povera Giovanna	1869
Boito, Camillo	Il maestro di Setticlavio	1891
Boito, Camillo	Senso, Nuove storielle varie	1883
Capuana, Luigi	Giacinta	1879
Capuana, Luigi	Il marchese di Roccaverdina	1901
Capuana, Luigi	Profumo	1892
Caracciolo, Enrichetta	I misteri del chiostro napoletano	1864
Carcano, Giulio	Damiano	1850
Carcano, Giulio	La nunziata	1849

3 Lo script, a cura di Luca Tringali e Floriana Sciumbata, è disponibile all'indirizzo https://github.com/flometis/ExSTRA/blob/master/pulisci.py (consultato il 10/12/2021).

Castellani Fantoni Benaglio, Ines	Mia	1884
Collodi, Carlo	Le avventure di Pinocchio	1883
D'Annunzio, Gabriele	Il fuoco	1900
D'Annunzio, Gabriele	Il piacere	1889
D'Annunzio, Gabriele	Le vergini delle rocce	1895
De Amicis, Edomondo	Cuore	1886
Deledda, Grazia	Canne al vento	1913
Deledda, Grazia	La madre	1920
Deledda, Grazia	La via del male	1896
De Marchi, Emilio	Demetrio Pianelli	1890
De Marchi, Emilio	Giacomo l'idealista	1897
De Marchi, Emilio	Il cappello del prete	1888
De Roberto, Federico	I viceré	1894
De Roberto, Federico	Processi verbali	1890
Di Luanto, Regina	Nuovissimo amore	1903
Di Luanto, Regina	Per il lusso	1912
Dossi, Carlo	La colonia felice	1874
Dossi, Carlo	La desinenza in A	1873
Farina, Salvatore	Fante di picche	1874
Fogazzaro, Antonio	Daniele Cortis	1885
Fogazzaro, Antonio	Malombra	1881
Fogazzaro, Antonio	Piccolo mondo moderno	1901
Garibaldi, Giuseppe	Cantoni il volontario	1870
Garibaldi, Giuseppe	Clelia	1870
Guerrazzi Domenico	L'assedio di Firenze	1863
Guerrazzi Domenico	Pasquale Paoli	1860
Guglielminetti, Amalia	Anime allo specchio	1915
Guglielminetti, Amalia	La porta della gioia	1920
Guglielminetti, Amalia	Le ore inutili	1919
Imbriani, Vittorio	Dio ne scampi dagli Orsenigo	1876
Invernizio, Carolina	I misteri delle soffitte	1901
Invernizio, Carolina	Il bacio di una morta	1886
Invernizio, Carolina	La trovatella di Milano	1889
Manzoni, Alessandro	I promessi sposi	1825-1827[4]
Manzoni, Alessandro	Storia della colonna infame	1840
Marchesa Colombi	Prima morire	1881
Melegari, Dora	Caterina Spadaro	1908
Neera	Lydia	1888
Neera	Un nido	1903

4 La datazione fa riferimento alla prima datazione dell'opera, pubblicata tra il 1825 e il 1827. Nelle rappresentazioni dei due articoli che seguono l'opera sarà indicata con la data intermedia (1826).

Neera	Una passione	1880
Negri, Ada	Le solitarie	1917
Nievo, Ippolito	Le confessioni di un italiano	1860
Nievo, Ippolito	Il barone di Nicastro	1867
Oriani, Alfredo	Al di là	1877
Oriani, Alfredo	No	1881
Perodi, Emma	Il principe della marsiliana	1891
Petruccelli, Ferdinando	Il sorbetto della regina	1890
Petruccelli, Ferdinando	Memorie di Giuda	1867
Petruccelli, Ferdinando	I moribondi del palazzo Carignano	1862
Pirandello, Luigi	I vecchi e i giovani	1904
Pirandello, Luigi	Il fu Mattia Pascal	1913
Pirandello, Luigi	L'esclusa	1901
Praga, Emilio	Memorie del presbiterio	1881
Righetti, Carlo	La scapigliatura	1862
Rovani, Giuseppe	Cent'anni	1859
Rovetta, Gerolamo	Casta Diva	1903
Rovetta, Gerolamo	Mater dolorosa	1882
Ruffini, Giovanni	Il dottor Antonio	1856
Ruffini, Giovanni	Lorenzo Benoni	1853
Salgari, Emilio	Alla conquista di un impero	1907
Salgari, Emilio	I misteri della giungla nera	1887
Salgari, Emilio	I pirati della Malesia	1896
Serao, Matilde	Cuore inferno	1881
Serao, Matilde	Dal vero	1879
Serao, Matilde	La mano tagliata	1912
Silvola, Rodolfo Giuseppe	Alboino in Italia	1840
Slataper, Scipio	Il mio Carso	1912
Svevo, Italo	La coscienza di Zeno	1923
Svevo, Italo	Senilità	1898
Svevo, Italo	Una vita	1892
Tarchetti, Iginio Ugo	Fosca	1869
Tartufari. Clarice	Il miracolo	1909
Tartufari. Clarice	L'albero della morte	1912
Tartufari. Clarice	Maestra	1887
Tommaseo, Niccolò	Fede e bellezza	1840
Torricelli, Giovan Battista	Oddantonio Feltre	1860
Tozzi, Federigo	Con gli occhi chiusi	1919
Tozzi, Federigo	Tre croci	1920
Vamba	Giamburrasca	1912
Verga, Giovanni	Il marito di Elena	1882
Verga, Giovanni	Tigre Reale	1875

Verga, Giovanni	Una peccatrice	1866
Vertua Gentile, Anna	Il romanzo di una signora perbene	1897
Zuccoli, Luciano	La freccia nel fianco	1913
Zuccoli, Luciano	Roberta	1897

Il corpus presenta le seguenti statistiche:

	numero di autori
donne	16 (29,6%)
uomini	38 (70,4)
totale	**54**

	numero di opere
donne	30
uomini	70
totale	**100**

anno	**numero di opere**
1826	1
1834	1
1840	3
1849	1
1850	1
1853	1
1856	1
1859	1
1860	3
1862	2
1863	1
1864	1
1866	1
1867	2
1868	1
1869	2
1870	2
1871	1
1873	1
1874	2
1875	1
1876	1
1877	1
1879	2
1880	1
1881	5

1882	2
1883	3
1884	1
1885	1
1886	2
1887	2
1888	2
1889	2
1890	3
1891	2
1892	2
1894	1
1895	1
1896	2
1897	3
1898	1
1900	1
1901	4
1903	3
1904	1
1907	2
1908	1
1909	1
1912	5
1913	3
1915	1
1917	1
1919	3
1920	3
1923	1
totale	**100**

decennio	**numero di opere**
1820_1829	1
1830_1839	1
1840_1849	4
1850_1859	4
1860_1869	13
1870_1879	11
1880_1889	21
1890_1899	15
1900_1909	13
1910_1920	17
totale	**100**