Rivista Internazionale di Tecnica della Traduzione / International Journal of Translation 2021 (23), 179-186
DOI: 10.13137/2421-6763/33465 ISSN 1722-5906 (print) ISSN 2421-6763 (online)
https://www.openstarts.units.it/dspace/handle/10077/2503

Corpora e storia della lingua

Michele A. Cortelazzo
Università di Padova
cortmic@unipd.it

Abstract

During the last twenty years, in particular at the lexical level, the study of the history of the Italian language has benefited greatly from corpora including literary texts, such as "Letteratura Italiana Zanichelli" or "Biblioteca Italiana" (and, perspectively, the Italian section of the ELTeC - European Literary Text Collection). However, those corpora were compiled mainly according to the literary canon, and they include a small number of texts compared to the time span they are supposed to cover. Thus, they can provide a limited account of the evolution of literary Italian in time.
Today, a large number of texts are available online (to begin with Google Books), covering a large time span and broad variety of genres. The analysis of such large corpora calls for suitable tools and approaches to ensure their effective distant reading.
Consequently, the interdisciplinary co-operation of linguists, statisticians and academics of literature may lead to an innovative approach to the study of corpora to assess the evolution of literary language.

Keywords

Distant reading, corpus linguistics, diachronic corpora, chronological textual data, history of the italian language

Luca Serianni ha pubblicato, nel 2006, un contributo dal titolo *Gli archivi elettronici e la lessicografia storica*, che riprende un intervento tenuto qualche anno prima a Saarbrücken al convegno tenutosi in occasione del settantesimo compleanno di Max Pfister. Questo contributo rappresenta bene l'entusiasmo che la disponibilità dei primi *corpora* digitalizzati aveva generato negli studiosi, ma al tempo stesso la consapevolezza della loro insufficienza per uno studio pieno della storia della lingua di uno strumento costituito da una selezione di testi che, pur con qualche scelta originale, si basava, necessariamente direi, sul canone tradizionale della storia letteraria. Serianni sottolinea la rilevanza di quello che allora era lo strumento principe per lo studio della lingua italiana, la LIZ, *Letteratura italiana Zanichelli*, per lo studio prima di tutto della lingua poetica, "data la tradizionale vischiosità della tradizione lirica italiana e il rilievo esemplare che vi assumono i grandi classici" (Serianni 2006: 42). Ne rileva anche i limiti, consistenti principalmente nella quasi esclusiva letterarietà del canone di riferimento. Non solo sono totalmente esclusi testi di natura non letteraria (come i testi di natura tecnica, con il risultato che "è scarso il contributo alla conoscenza dei linguaggi settoriali, se non in quanto suscettibili di rifluire nella lingua letteraria, come avviene in qualche misura per il lessico giuridico o per quello medico", Serianni 2006: 42), ma anche nella scelta delle opere di autori inclusi nel repertorio (nota, ad esempio, che di un letterato, ma anche scienziato, come Lorenzo Magalotti è raccolta la *Relazione sulla China* e non i *Saggi di naturali esperienze*, di uno storico come Pietro Giannone la *Vita* e non la *Istoria civile del Regno di Napoli*, di un economista come Antonio Genovesi le *Memorie autobiografiche* e non le *Lezioni di commercio o sia di economia civile*. Queste scelte sono ben comprensibili in un'ottica concentrata sulla storia letteraria, ma riducono l'utilità dello strumento per la storia della lingua.

Nonostante questi limiti, la LIZ, nelle sue diverse edizioni, sempre più accresciute, ha costituito un serbatoio enorme anche per chi intendeva studiare la storia della lingua e, soprattutto, il suo lessico. Più di un risultato interessante è proposto proprio da Serianni (2006).

Certo, leggere questo contributo circa un ventennio dopo la sua prima stesura ingenera un senso di nostalgia e tenerezza, pensando alla soddisfazione che provavamo nell'avere a disposizione un repertorio che, nella sua ultima edizione, comprendeva circa mille testi, interrogabili in tempo reale. Sembrava un'enormità. Ma mille testi per quasi mille anni di storia letteraria significa una media di un testo per anno. Troppo poco per studiare in dettaglio l'evoluzione della lingua che, si sa, è caratterizzata, anche nelle sue realizzazioni scritte, da un alto grado di varietà non solo cronologica (quella che è il centro di interesse nel caso di indagini diacroniche), ma anche di genere testuale, di funzione, di provenienza regionale.

Io stesso, in quegli anni, mi ero basato sul corpus della prima LIZ (quindi su poco meno di 400 testi), per ricostruire la protostoria e la storia della perifrasi

progressiva (Cortelazzo 2007) e mi pareva da una parte di annegare in mezzo ai dati, dall'altra di aver fondato la storia del costrutto su una base testuale ineguagliabile. Oggi, la base testuale utilizzata solo una quindicina di anni fa appare ben modesta e certamente da incrementare soprattutto per le fasi più recenti (ottonovecentesche) della storia della nostra lingua.

Dal punto di vista quantitativo, i *corpora* nella piena disponibilità degli utenti non sono aumentati in misura considerevole. La più ampia raccolta aperta e filologicamente attendibile di testi della letteratura italiana, la Biblioteca Italiana (sigla BiBit, sito http://www.bibliotecaitaliana.it), comprende più di 1600 opere: si tratta certamente di un notevole incremento rispetto alla LIZ, che però raffina solo parzialmente la densità cronologica dei testi. Del resto, l'obiettivo della BiBit (erede di imprese pionieristiche, come quella avviata dal CIBIT, Centro Interuniversitario Biblioteca Italiana Telematica, fin dal 1996) è pur sempre quella di allestire una biblioteca digitale formata, come si legge nel sito, da "testi rappresentativi della tradizione culturale e letteraria italiana dal Medioevo al Novecento". Cioè anche questa iniziativa prevede, di necessità, l'individuazione di un canone (quello che permette di riconoscere la rappresentatività di un testo rispetto a un altro). Lo stesso corpus otto-novecentesco ELTeC (European Literary Text Collection), costituito nell'ambito del progetto europeo (COST Action) *Distant Reading for European Literary History* (Burnard, Schöch, Odebrecht 2021), prevede per ogni lingua la raccolta di 100 opere, collocate cronologicamente di norma tra il 1840 e il 1920 (per l'italiano ne è stata finora costituita una versione preliminare, di testi tra il 1825 e il 1923), con una densità, dunque, del tutto analoga a quella della LIZ.

Indubbiamente *corpora* come quelli che ho ricordato hanno il vantaggio di essere codificati secondo sistemi standard e, almeno in parte, annotati (per es. in relazione alle parti del discorso o ad alcune categorie di nomi: Schoch et alii 2021; ed altri *corpora* sono più finemente annotati, sotto il profilo sintattico o sotto quello semantico o tematico). Sul piano qualitativo si tratta, quindi, di *corpora* più ricchi di informazioni dei semplici testi digitalizzati; ma la codificazione richiede un impegno che rallenta la predisposizione dei *corpora*, obbliga a trovare stretti criteri di selezione dei testi e ci riporta, quindi, alla prefigurazione di un canone.

Ancor più selettivi sono *corpora* nati con finalità specifiche, come il "PIC, Padua Italian Corpus" (così denominato da Savoy 2018), costituito per identificare le similarità tra i testi di Elena Ferrante e quelli di altri autori italiani a lei contemporanei.

2. *Corpora* elettronici per la storia della lingua italiana

I *corpora* letterari finora raccolti sono stati di grandissimo aiuto anche per la storia della lingua italiana. Il fatto di avere a disposizione sul proprio computer dai mille ai milleseicento testi dell'intera storia letteraria italiana, e poterli interro-

gare in tempo reale (senza la necessità di consultarli uno per uno alla ricerca di una parola o di un costrutto) ha modificato profondamente il modo di raccogliere i dati per ricostruire pezzi di storia della nostra lingua. Il vantaggio maggiore riguarda la ricostruzione della storia delle parole, ma anche i costrutti grammaticali più facilmente identificabili per mezzo di definiti elementi formali (come è il caso della citata perifrasi progressiva) hanno avuto una più precisa illustrazione storica.

I criteri che (seguo ancora Serianni 2006) hanno portato alla costituzione dei *corpora* citati, sono ragionevoli, ma non rappresentano sempre la soluzione migliore per lo storico della lingua. Il primo criterio è quello della rappresentatività storico-letteraria o, in alternativa quello della costruzione di un corpus bilanciato, in base a criteri che non sempre corrispondono ai criteri che sono ugualmente validi per tutte le epoche storiche (basti pensare al bilanciamento in base al genere dell'autore, in una tradizione letteraria che è stata prevalentemente maschile). Il secondo criterio è quello della disponibilità di edizioni affidabili, meglio se recenti, in quanto facilmente digitalizzabili (criterio che oggi è caduto in secondo piano, grazie al perfezionamento degli strumenti di lettura automatica dei testi a stampa originali).

Oggi abbiamo la possibilità materiale di uscire dalle strette della rappresentatività esclusivamente letteraria grazie alla enorme disponibilità in rete di volumi scansionati e resi disponibili almeno in formato pdf (in Google Libri, o nel preziosissimo Internet Archive, o in altre biblioteche digitali a libero accesso, come Hathi Trust o ulteriori raccolte di altre Università di oltreoceano, tutte ricche anche di libri italiani). Fondamentali anche, per l'italiano postunitario, gli archivi storici dei giornali, anche se non tutti pienamente utilizzabili ai nostri fini: è il caso della raccolta del "Corriere della sera", che non permette la ricerca per significanti, dal momento che elabora le richieste con criteri semantici. Ma utilissimi sono gli archivi della "Stampa", del "Piccolo", dell'"Avanti!", dell'"Unità", oltre che di "Repubblica" (se lo studio riguarda la storia dell'italiano più recente). Preziosi anche gli archivi storici di Camera e Senato.

In primo luogo risultano comunque fondamentali le raccolte digitali di volumi, che, per la maggior parte dei testi non più protetti dal diritto d'autore, consentono anche di scaricare la versione pdf del testo. Si tratta di una risorsa di grandissima utilità per chi voglia descrivere con maggiore precisione la configurazione dell'italiano scritto degli ultimi secoli e le dinamiche che si sono realizzate, uscendo dai canoni di matrice letteraria che, ancor più da quando sono risultati disponibili i *corpora* di prima generazione come la LIZ, hanno forzatamente guidato la ricerca sistematica sui testi, basata su risorse automatiche di interrogazione, con un riferimento alla lingua letteraria più forte di quello che si verifica nei più tradizionali studi qualitativi.

Ma le grandi raccolte di testi digitalizzati ci danno immediato accesso all'insieme globale di volumi pubblicati in italiano, o alla sua gran parte, con la possibilità di attingere anche alla letteratura non canonica (quella letteratura di con-

sumo dal limitato valore artistico, ma che ha avuto una circolazione di estrema ampiezza tra i lettori), alla letteratura tecnica e scientifica, alle traduzioni. In questi mesi sto sperimentando quanto questa disponibilità di testi ci permetta di meglio datare il lessico degli ultimi due secoli. La lezione che traggo da questi spogli (che si basano su due presupposti di fondo: l'attingibilità di raccolte vaste di testi finora non studiati e l'esistenza di appropriati strumenti di estrazione dei dati) è che questo enorme giacimento di testi digitalizzati ci permette di studiare in maniera più approfondita, e con una grana più fine, la lingua almeno degli ultimi due secoli. Anzi, questa nuova disponibilità di risorse ci impone di farlo. Ma per utilizzare appieno il patrimonio disponibile è necessario, a mio avviso, mutare la prospettiva con cui si guarda ai *corpora*.

La consuetudine di agire su *corpora* ampi, ma ristretti rispetto alla realtà di riferimento, scelti in base a criteri di rappresentatività (e quindi in base a un canone generale o a un canone definito *ad hoc*) e annotati è scarsamente compatibile con l'attuale ampia disponibilità di testi digitalizzati. È anche scarsamente coerente con gli sviluppi del trattamento automatico dei dati. Certamente le raccolte sempre più estese di testi non possano essere definite in senso stretto *big data*, in quanto consistono sì in un volume di dati fino ad oggi impensabile e sono caratterizzate anche da grande varietà, ma sono dati statici, che non variano con la velocità e l'imprevedibilità tipica dei *big data*. Ma la tendenza che si sta imponendo di elaborare grandi quantità di dati non può non estendersi, con le dovute finezze metodologiche, anche allo studio della storia della lingua, quando viene sviluppato non attraverso l'individuale capacità di identificazione delle dinamiche culturali che stanno alla base dei movimenti della lingua, ma attraverso lo spoglio o lo studio quantitativo di *corpora*.

3. Strumenti per una visione da distante della storia della lingua italiana

Ci troviamo, quindi, davanti a uno scenario nuovo, che ci permette di indirizzarci verso la possibilità di confrontarci con un numero di testi finora impensabile e di farlo nell'ottica del *distant reading* (che si oppone alla classica prospettiva del *close reading*), cioè attraverso quell'analisi indiretta dei testi che avviene attraverso elaborazioni di vario genere (per es. grafici, carte, alberi). Un'ottica che "fa vedere meno dettagli, vero: ma fa capire meglio i rapporti, i *pattern*, le forme" (Moretti 2005: 3) ed è certamente l'unica possibile se vogliamo utilizzare *corpora* di grandi dimensioni. Ma, in una sorta di circolo vizioso, la disponibilità di ampie raccolte di testi non giustifica più l'utilizzo di campioni ridotti, di insieme di testi rappresentativi, scelti sulla base di canoni più o meno condivisi. Questo è particolarmente vero quando l'obiettivo è quello di individuare andamenti diacronici su ampi archi cronologici, per i quali non sono sufficienti testimonianze uniche per ogni porzione temporale (come abbiamo visto accadere un po' in tutti i *corpora* ora disponibili); ed è ancor più valido in relazione a fasi storiche che hanno

visto verificarsi rilevanti processi di evoluzione linguistica (preciso che, anche in coerenza con gli altri interventi pubblicati in questa sede, penso principalmente, all'Otto-Novecento).

Si tratta, allora, di trovare gli strumenti che ci permettano di procedere a una lettura da distante delle ampie raccolte di testi digitalizzati ora disponibili, di individuare modelli e tendenze, in un'ottica che per uno storico della lingua è principalmente esplorativa e cioè tale da suggerire ipotesi di lavoro da verificare poi con analisi di tipo più tradizionale, basate su spogli da vicino dei testi appartenenti alle raccolte disponibili.

A questo fine possiamo affiancare alla ormai consolidata tradizione di costituire *corpora* di testi forniti di annotazioni, più o meno elaborate, una metodologia di analisi quantitativa di dati testuali basati su *corpora* testuali se non grezzi, certamente debolmente normalizzati, ma di dimensioni ben più ampie di quelle tradizionali (ciò significa anche, nel campo della ricostruzione dello sviluppo della lingua, caratterizzati da una rappresentatività dell'evoluzione cronologica molto più fine).

Insomma, se abbiamo la possibilità (e forse anche il dovere) di utilizzare ampie raccolte di testi, siamo costretti a ridurre al più basso livello possibile l'elaborazione finalizzata alla costruzione del dato. I testi oggetto di analisi vengono, sostanzialmente, a corrispondere a testi digitalizzati in forma basica, quindi in semplice formato di testo; i testi vanno certamente ripuliti degli errori che qualsiasi sistema di lettura e digitalizzazione automatica comporta e di eventuali tratti "sporchi" che possono venire introdotti nella transcodificazione da un formato all'altro (tipicamente da pdf a txt); inoltre vanno normalizzate le consuetudini grafiche (a meno che non siano esse stesse oggetto di analisi), per esempio unificando la scelta del tipo di segnaccento nelle parole tronche (è noto, ad esempio, che Einaudi utilizza l'accento acuto per tutte le vocali chiuse, comprese *i* ed *u*; pur riconoscendo che, alla base di questa scelta, c'è una motivazione non priva di buone ragioni, mantenerla nei testi significherebbe rilevare a livello quantitativo e attribuire quindi ai testi pubblicati da un editore una diversità che si situa solamente sul piano delle scelte grafiche, ma che finirebbe con l'allontanare tali testi da testi nella sostanza molto simili).

La diminuzione di informazione rispetto a testi annotati deve essere, però, compensata dalla raffinatezza degli strumenti di analisi quantitativa. Da statistici o matematici lo storico della lingua, capace di offrire temi di indagine e insiemi di testi validi per affrontare tali temi, deve aspettarsi di ricevere strumenti di misurazione e di classificazione che permettano quella visione da distante che è tanto più necessaria quanto più ampio è l'insieme dei testi sottoponibile ad analisi.

Dalla statistica ci possono giungere strumenti raffinati per riconoscere e raffigurare l'andamento temporale di forme o lemmi presenti in *corpora* diacronici, cioè in *corpora* di testi capaci di rappresentare l'evoluzione storica di un tipo testuale (Trevisani e Tuzzi 2013) o di uno specifico dominio nozionale, come, ad

esempio, la letteratura specialistica relativa a un preciso ambito scientifico (Tuzzi 2018, Trevisani e Tuzzi 2018, Trevisani e Tuzzi 2015).

In questo numero della rivista Sciandra, Trevisani e Tuzzi hanno messo alla prova gli strumenti statistici su un corpus di narrativa italiana; ma l'obiettivo più ambizioso è quello di poter tratteggiare l'evoluzione dell'uso lessicale almeno scritto di un'intera comunità linguistica, guardando da distante una quantità rilevante di testi di varia natura, capaci di rappresentare con ricchezza di dati ciascuna porzione (per esempio annuale) dell'arco temporale indagato. Da un'altra prospettiva, con tecniche ascrivibili al *distant reading* e basate sul *machine learning*, si è cercato di valutare l'efficacia delle periodizzazioni proposte per la rappresentazione della storia della lingua italiana (Cortelazzo et alii in print).

L'utilizzo di *corpora* di ampie dimensioni, costituiti dal maggior numero di testi possibili, indipendentemente da una scelta legati a canoni letterari (per quanto larghi), capaci di collocarsi in una fitta linea del tempo, implica un vero lavoro interdisciplinare tra studiosi di lingua italiana ed esperti di metodi quantitativi di analisi. Ai primi spetta, comunque, il recupero dei testi e la loro pulizia, l'individuazione dei dati più significativi estraibili (almeno potenzialmente) da quei testi, l'interpretazione sul piano della storia culturale dei risultati che emergono dall'elaborazione dei dati; ai secondi di individuare i modi più raffinati, e se possibile più innovativi, per rappresentare l'evoluzione nel tempo dei fenomeni rappresentati dai dati oggetto di analisi (meglio se permettono di individuare tendenze non immediatamente percepibili con la lettura da vicino).

Mi è già capitato (Tuzzi, Cortelazzo 2018) di utilizzare un adattamento di una frase di Baricco (1994) per indicare di cosa è necessario disporre quando ci si appresta a svolgere indagini testuali con metodi quantitativi. I ricercatori orientati verso una considerazione quantitativa dei testi possono dirsi: "non sei fregato veramente finché hai da parte un buon corpus e un metodo con cui analizzarlo". Ciò mi appare particolarmente vero quando si vuole operare in chiave storica, su una grande quantità di testi (e allora occorre ricorrere a metodi particolarmente raffinati); oppure quando si hanno disposizione metodi particolarmente raffinati (e allora servono *corpora* ampi, che permettano di rappresentare nella maniera più densa e fine l'intero arco cronologico oggetto di analisi).

Baricco A. (1994) *Novecento. Un monologo*, Milano, Feltrinelli.

Cortelazzo M. A. (2007) "La perifrasi progressiva in italiano è un anglicismo sintattico?", in *Studi in onore di Pier Vincenzo Mengaldo per i suoi settant'anni, a cura degli allievi padovani*, Firenze, SISMEL. Edizioni del Galluzzo, pp. 1753-1764.

Cortelazzo M. A., Gatti F. M. T., Mikros G. K. e Tuzzi A. (in stampa) "Does the Century matter? Machine learning methods to attribute historical periods in an Italian literary corpus", in *Quantitative Approaches to Universality and Individuality in Language. Selected papers of the 11th International Conference on Quantitative Linguistics (QUALICO)*. Ed. di M. Yamazaki *et alii*, Berlin, De Gruyter Mouton.

LIZ 1993 = *Letteratura italiana Zanichelli (= LIZ), CD-ROM dei testi della letteratura italiana*. A cura di P. Stoppelli e E. Picchi, Bologna, Zanichelli.

LIZ 1995 = *Letteratura italiana Zanichelli (= LIZ), CD-ROM dei testi della letteratura italiana*. A cura di P. Stoppelli e E. Picchi, Bologna, Zanichelli.

LIZ 1997 = *Letteratura italiana Zanichelli (= LIZ), CD-ROM dei testi della letteratura italiana*. A cura di P. Stoppelli e E. Picchi, Bologna, Zanichelli.

LIZ 2001 = *Letteratura italiana Zanichelli (= LIZ), CD-ROM dei testi della letteratura italiana*. A cura di P. Stoppelli e E. Picchi, Bologna, Zanichelli.

Moretti F. (2005) *La letteratura vista da lontano*, Torino, Einaudi.

Savoy J. (2018) "Is Starnone really the author behind Ferrante?", *Digital Scholarship in the Humanities*, 33(4), pp. 902-918.

Schöch C., Patras R., Erjavec T. e Santos D. (2021) "Creating the European Literary Text Collection (ELTeC). Challenges and Perspectives", *Modern Languages Open*, (1), p. 25.

Serianni L. (2006) "Gli archivi elettronici e la lessicografia storica", in *Nuovi media e lessicografia storica. Atti del colloquio in occasione del settantesimo compleanno di Max Pfister*. A cura di W. Schweickard, Tübingen, Niemeyer, pp. 41-58.

Trevisani M. e Tuzzi A. (2013) "Shaping the history of words", in *Methods and Applications of Quantitative Linguistics. Selected papers of the VIIIth International Conference on Quantitative Linguistics (QUALICO)*. Ed. by I. Obradović, E. Kelih & R. Köhler, Belgrade, Akademska Misao, pp. 84-95.

Trevisani M. e Tuzzi A. (2015) "A portrait of JASA: the History of Statistics through analysis of keyword counts in an early scientific journal", *Quality and Quantity*, 49, pp. 1287-1304.

Trevisani M. e Tuzzi A. (2018) "Learning the evolution of disciplines from scientific literature. A functional clustering approach to normalized keyword count trajectories", *Knowledge-based systems*, 146, pp. 129-141.

Tuzzi A. (2018) (ed.) *Tracing the Life Cycle of Ideas in the Humanities and Social Sciences*, Cham, Springer.

Tuzzi A. e Cortelazzo M. A. (2018) "What is Elena Ferrante? A comparative analysis of a secretive bestseling Italian writer", *Digital Scholarship in the Humanities*, 33(3), pp. 685-702.