Sul distant reading: una visione critica

di Fabio Ciotti

Accade spesso, nel mondo vasto e proteiforme delle *Digital Humanities*, che la comparsa di una innovazione tecnica o metodologica produca una sorta di «effetto domino» intellettuale. Il «nuovo gioco in città», sia esso un metodo o uno strumento inizialmente sviluppato per specifiche e settoriali esigenze di ricerca, viene entusiasticamente adottato e adattato in ambiti più o meno contigui a quello di origine. Questa propagazione non sempre è accompagnata da una adeguata riflessione metodologica ed epistemologica e da una sufficiente consapevolezza circa i fondamenti teorici e concettuali che ne giustificano l'uso. Rientrano senza dubbio in questa fattispecie l'insieme di approcci analitici rubricati sotto l'etichetta di *distant reading*, a loro volta inseriti nel contesto della cosiddetta «rivoluzione dei Big Data» da cui ereditano metodi e strumenti[1].

Il principale luogo in cui sono state condotte le prime sperimentali applicazioni di tali tecniche nell'ambito degli studi letterari è lo *Stanford Literary Lab* fondato e diretto da Franco Moretti, con il supporto di Matthew Jockers, nel 2010[2]. Allo stesso Moretti si debbono i tentativi più brillanti di fornire una base teorica ed epistemologica a queste sperimentazioni con l'introduzione della nozione di *distant reading*[3], coniata in opposizione a quella tradizionale di *close reading*, intesa non solo nel senso stretto che identifica il metodo formalista prediletto dai *New Critics* nord americani, ma come metodo di accesso all'opera letteraria basato sulla lettura profonda e sull'analisi di dettaglio comune a gran parte degli studi letterari del '900.

Accanto alla naturale attrazione per la moda tecnoculturale del momento, il fascino esercitato dal nuovo paradigma del distant reading nella comunità degli studi digitali, e in particolare in quelli di ambito letterario, è dovuto al fatto che esso sembra avere la capacità di far uscire finalmente l'analisi computazionale dei testi dalla condizione di minorità – se non irrilevanza – rispetto al *mainstream* degli studi letterari nella quale è stata a lungo relegata.

Questa limitata influenza è stata tematizzata e discussa sin dagli anni '80 del secolo scorso, quando l'analisi testuale informatica aveva già diversi decenni di sperimentazione e riflessione metodologica alle spalle[4]. Nonostante pochi anni dopo la diffusione delle tecnologie digitali e della rete Internet abbiano innescato una grande trasformazione socioculturale, che ha investito anche il mondo della ricerca umanistica, la considerazione dei metodi computazionali negli studi letterari non è mutata molto. All'affacciarsi del nuovo millennio Jerome McGann nella prefazione al suo bel libro *Radiant Textuality* individua con grande chiarezza la causa di questa perdurante condizione nel fatto che i metodi e gli strumenti digitali non sono riusciti a incidere sul reale specifico degli studi letterari e umanistici (intesi nel senso più ampio), il lavoro dell'interpretazione[5]:

> Digital technology has remained instrumental in serving the technical and precritical occupations of librarians and archivists and editors. But the general field of humanities education and scholarship will not take the use of digital technology seriously until one demonstrates how its tools improve the ways we explore and explain aesthetic works – until, that is, they expand our interpretational procedures.

Pochi anni dopo Willard McCarty, altra figura prominente nelle Digital Humanities – specie per il suo acume teorico ed epistemologico – ha dedicato diversi lavori al tema delle difficoltà metodologiche e disciplinari dell'informatica letteraria, analizzandolo sia dal punto di vista storico sia da quello teorico e arrivando a conclusioni sostanzialmente concordi con l'analisi di McGann[6]:

> The point […] is that literary computing has thereby served only as mutely obedient handmaiden, and so done nothing much to rescue itself from its position of weakness, from which it can hardly deliver the benefits claimed for it by the faithful. It has done little to educate scholars methodologically.

L'elenco di citazioni sul tema potrebbe proseguire, ma esse non farebbero altro che convergere sulle medesime conclusioni: gli strumenti e i metodi computazionali "tradizionali", nonostante abbiano portato a risultati critici rilevanti (ci riferiamo ai successi conseguiti negli studi di attribuzione o ad alcuni lavori di analisi stilometrica), non sono stati capaci di interagire in modo soddisfacente con l'orizzonte teorico e metodologico degli studi letterari[7].

Se questa è la situazione, per tornare al tema centrale di questo lavoro, dobbiamo chiederci se e in che modo i cosiddetti approcci basati sul distant reading possano integrarsi con la tradizione teorica e con l'apparato interpretativo degli studi letterari, o possano persino ambire a costruire un nuovo paradigma negli studi letterari. Questione che può essere riformulata in accordo all'inquadramento metodologico che lo stesso Moretti ha proposto mutuando dall'epistemologia di P.W. Bridgman il concetto di *operazionalizzazione*[8]:

> Forget programs and visions; the operational approach refers specifically to concepts, and in a very specific way: it describes the process whereby concepts are transformed into a series of operation — which, in their turn, allow to measure all sorts of objects. Operationalizing means building a bridge from concepts to measurement, and then to the world. In our case: from the concepts of literary theory, through some form of quantification, to literary texts.

I metodi del distant reading si possono configurare come un modo adeguato di operazionalizzare, almeno in parte, quella tradizione teoretica e analitica per riadattarla al contesto digitale negli studi letterari? Prima di formulare una risposta occorre vedere più da vicino in cosa consistano questi metodi, compatibilmente con i limiti di questo lavoro.

L'idea fondamentale è che esistono fatti e fenomeni letterari e culturali, sia sincronici sia diacronici, che non sono accessibili ai tradizionali metodi di close reading[9] – basati sulla lettura profonda di una o di poche grandi (o supposte tali) opere e sulla interpretazione puntuale delle loro caratteristiche formali o del loro contenuto – ma che richiedono l'analisi di massa di centinaia o migliaia di testi e la loro considerazione come totalità, non come individui. Questo, sostiene Moretti provocatoriamente, è possibile solo assumendo un punto di vista esterno e distante verso tale totalità e rinunciando all'incontro con il testo, all'atto della lettura[10]:

> the trouble with close reading (in all of its incarnations, from the new criticism to deconstruction) is that it necessarily depends on an extremely small canon […] At bottom, it's a theological exercise – very solemn treatment of very few texts taken very seriously – whereas what we really need is a little pact with the devil: we know how to read texts, now let's learn how not to read them. Distant reading: where distance, let me repeat it, is a condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text: devices, themes, tropes – or genres and systems. And if, between the very small and the very large, the text itself disappears, well, it is one of those cases when one can justifiably say, Less is more.

Il distant reading, dunque, si basa sulla adozione di metodi computazionali che permettono di analizzare grandi insiemi di opere focalizzandosi su precise caratteristiche quantificabili del testo. Le grandi campa-

gne di digitalizzazione degli ultimi decenni hanno reso disponibili vasti archivi testuali. Sono questi i depositi di Big Data testuali che rendono possibile applicare negli studi letterari tecniche e strumenti di analisi dei dati finora appannaggio di ambiti assai distanti come le scienze fisiche, biomediche e parti di quelle sociali ed economiche. Solo queste tecniche possono far emergere pattern, schemi, correlazioni, altrimenti non conoscibili, che hanno un rilevante ruolo esplicativo nella comprensione dei processi letterari quali l'evoluzione dei generi, l'affermazione di un stile, la presenza di temi e contenuti ricorrenti in un dato periodo storico-letterario. Si determina così un vero e proprio salto di paradigma negli studi letterari come afferma Matt Jockers[11]:

> Close reading is not only impractical as a means of evidence gathering in the digital library, but big data render it totally inappropriate as a method of studying literary history […] massive digital corpora offer us unprecedented access to the literary record and invite, even demand, a new type of evidence gathering and meaning making.

L'insieme di tecniche e strumenti computazionali adottati nel distant reading sono stati sviluppati in ambiti disciplinari diversi dalle scienze umane, sebbene sin dalle origini siano stati il prodotto di ricerche con una spiccata impronta multidisciplinare. Rientrano nell'area informatica i metodi e tecniche di *text mining* (a sua volta una specializzazione del più generale campo del *data mining*) e *machine learning*[12]. Si tratta in generale di un insieme di sistemi e metodologie di analisi dei dati orientate alla ricerca di pattern ricorrenti o schemi impliciti all'interno di grandi o grandissimi masse di dati (da cui il termine Big Data) scarsamente o del tutto non strutturati[13]. La ricerca di questi pattern si basa su complessi algoritmi statistici, che derivano dall'analisi multivariata e dalla teoria probabilistica bayesiana.

In generale i metodi di text mining si dividono in due classi: metodi supervisionati (*classificazione*); metodi non supervisionati (*clustering* e *topic modeling*). Nel primo caso le categorie nelle quali vanno suddivisi i dati sono note a priori. Pertanto il programma può essere 'addestrato', mediante un *data set* preventivamente categorizzato da esperti umani, a individuare le distribuzione di valori delle variabili quantitative che identificano ciascuna categoria. Tali distribuzioni sono poi utilizzate per categorizzare i dati non classificati forniti in input al sistema. I metodi non supervisionati invece producono una organizzazione del data set senza disporre di alcun criterio a priori, usando diverse tecniche di inferenza statistica. Per questo sono i metodi preferiti nella cosiddetta "ricerca esplorativa", quando ci si trova di fronte a enormi masse di dati osservativi senza disporre di chiare ipotesi sulle regolarità e correlazioni in essi soggiacenti.

Le tecniche di text mining non supervisionate più diffuse in ambito umanistico sono di due tipi:

- *text clustering*: applicazione di algoritmi probabilistici di confronto testuale al fine di suddividere un insieme di testi in sotto-gruppi (cluster);
- *topic modeling*: individuazione dei cluster di parole che caratterizzano un insieme di testi e analisi delle loro distribuzioni nei vari testi.

Il topic modeling, in particolare, suscita in questo momento grande interesse negli studi testuali e letterari, poiché viene comunemente considerato la migliore approssimazione disponibile di una analisi "semantica" del testo. In realtà il termine topic modeling denota una intera classe di algoritmi con proprietà e caratteristiche matematiche e computazionali assai diverse. Attualmente il più diffuso (anche per la buona disponibilità di software open source che lo implementa[14]) è quello noto come *Latent Dirichlet Allocation* (LDA), che si basa su un approccio probabilistico bayesiano[15].

Per descrivere in modo intuitivo l'algoritmo LDA possiamo dire che esso parte da un assunto semplicistico ma efficace sul modo in cui un testo viene generato: quando un autore scrive un testo in prima battuta sceglie l'insieme degli argomenti (topic) di cui vuole parlare e poi determina la proporzione con cui ciascun argomento sarà presente. Ammettiamo ora che ogni possibile topic possa essere caratterizzato come un insieme di parole con una data distribuzione: una specie di sacchetto di parole dove le singole parole possono essere ripetute in ragione diversa a seconda delle loro rilevanza rispetto all'argomento. Il nostro autore dunque non dovrà far altro che pescare in modo casuale dai vari sacchetti che corrispondono agli argomenti di cui intende scrivere ed estrarre da ciascuno un numero di parole proporzionale al peso che intende assegnare all'argomento stesso. Alla fine non dovrà far altro che mettere in sequenza il suo

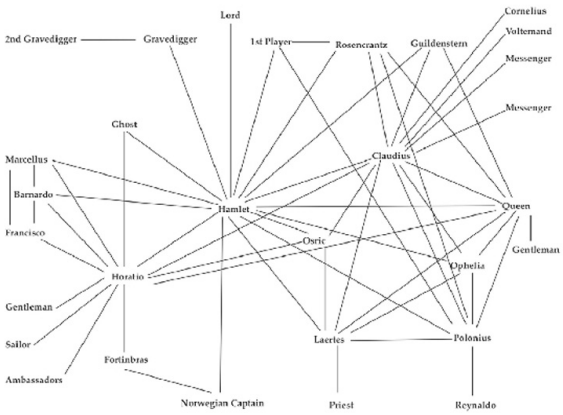
mucchietto di parole ed ecco che avrà ottenuto il suo testo, in cui ovviamente le parole appartenenti al topic più rilevante saranno presenti in misura maggiore rispetto a quelle del topic di secondo piano e così via. In termini tecnici si dice che in LDA un testo è una distribuzione di probabilità su un insieme di topic e un topic una distribuzione di probabilità su un insieme di parole. La cosa interessante di questo semplice modello generativo del testo è che esso può essere invertito: è possibile cioè definire un algoritmo che è in grado di estrapolare i topic presenti in un insieme di documenti, sottoforma di una serie di liste di parole che co-occorrono con frequenza notevole, corredata della loro distribuzione di probabilità[16].

Di recente hanno riscontrato un certo interesse in ambito letterario anche le tecniche di *sentiment analysis*[17]; si tratta di sistemi che cercano di individuare l'atteggiamento emotivo veicolato da una frase o da un testo, mediante la costruzione di una metrica emozionale assegnata alle parole. A questo fine vengono usati metodi di text mining supervisionati con il supporto di strumenti di elaborazione del linguaggio naturale (NLP) per pre-processare i dati (i diversi ruoli grammaticali e sintattici delle parola, infatti, giocano in questo caso un ruolo rilevante). Il caso più interessante e controverso di utilizzo delle tecniche di sentiment analysis in ambito letterario è legato allo sviluppo di una applicazione da parte di Matt Jockers (suggestivamente denominata *Syuzhet*) che traccia l'andamento emozionale all'interno dei testi sotto l'assunzione che essa sia una approssimazione dello sviluppo dell'intreccio; lo stesso Jockers ha utilizzato questo programma per una sperimentale analisi di testi narrativi inglesi che vanno dal 18° al 20° secolo, arrivando alla conclusione che esistono al massimo sei o sette tipologie archetipiche di intrecci[18].

Accanto ai sistemi di analisi testuale esaminati qui sopra, rientrano nel novero dei metodi del distant reading le tecniche di *network analysis*, sviluppate nell'ambito delle scienze sociali quantitative[19]. Si tratta di modelli formali a rete, basati sulla teoria matematica dei grafi, che rappresentano le relazioni tra entità atomiche (siano esso soggetti individuali, collettivi o istituzioni) all'interno di un gruppo sociale. Ogni individuo costituisce un nodo (o vertice) e ogni relazione un arco (o spigolo). La rete risultante è un modello astratto della struttura relazionale interna al gruppo. Alcune proprietà matematiche della rete possono essere calcolate e adottate come succedanei (*proxy*) delle caratteristiche qualitative del dominio modellizzato: le centralità di un nodo, la distanza di un (sottoinsieme di) nodi rispetto a un altro, l'individuazione di sottoinsiemi di nodi prossimali e così via.

La network analysis esercita una notevole attrazione anche in virtù del fatto che essa può essere facilmente espressa in forma di visualizzazioni grafiche (anche dinamiche) alquanto esplicative. Le sue applicazioni nell'ambito delle Digital Humanities, sia in ambito storico sia in ambito letterario, sono numerose. In ambito letterario, in particolare, si deve riconoscere nuovamente a Franco Moretti il ruolo di pioniere nella sperimentazione e nella riflessione teorica e metodologica. Nei suoi studi su Amleto e Dickens, ad esempio, ha usato queste tecniche analitiche per studiare i rapporti tra i personaggi[20]:

> A network is made of vertices and edges; a plot, of characters and actions: characters will be the vertices of the network, interactions the edges, and here is what the Hamlet network looks like:



Più avanti nel medesimo saggio Moretti delinea la natura modellistica di queste astrazioni e ne individua con chiarezza le potenzialità euristiche[21]:

> […] once you make a network of a play, you stop working on the play proper, and work on a model instead. You reduce the text to characters and interactions, abstract them from everything else, and this process of reduction and abstraction makes the model obviously much less than the original object […] but also, in another sense,

> much more than it, because a model allows you to see the underlying structures of a complex object.

Ovviamente sono possibili numerose intersezioni tra metodi di text mining e network analysis: per esempio i dati prodotti da una applicazione di topic modeling o di clustering possono essere forniti come input a un programma per la visualizzazione e l'analisi di reti e grafi.

Questa veloce rassegna ci permette di tornare con maggiore consapevolezza a riflettere sulla adeguatezza teorica ed epistemologica dei metodi e delle tecniche analitiche quantitative usate nel distant reading, in relazione al dominio degli oggetti e dei metodi interpretativi letterari. Anticipo che la mia è una posizione critica, dove però il termine 'critica' va assunto nel suo senso più ampio e pieno: la critica non è mai un giudizio aprioristico di rifiuto e di condanna, ma un processo dialettico che si confronta con il suo oggetto, ne individua i possibili limiti ma ne evidenzia anche le potenzialità, e se possibile suggerisce una sintesi superiore. Ecco dunque alcune criticità teoriche e metodologiche che ritengo di avere individuato in questi metodi e nella loro applicazione (e teorizzazione) ingenua (atteggiamento assai diverso dalla posizione metodologica sofisticata e supportata da una solidissima competenza letteraria di Moretti).

1) Gli algoritmi di data mining in generale sono completamente indipendenti dal contesto (ovvero essi possono essere applicati indifferentemente alle transazioni della borsa così come a grandi incorpora testuali). Essi individuano similarità e pattern ricorrenti in modo indipendente dalla semantica dei dati. Nelle scienze umane (in quelle letterarie in particolare) invece, i dati sono sempre semanticamente e pragmaticamente relativi al contesto di emissione e di ricezione del testo. Simmetricamente, ogni algoritmo di text mining può produrre risultati diversi a partire dallo stesso set di dati, e non esiste nessun metodo automatico per scegliere quale di tali interpretazioni dei dati sia migliore. Ne consegue che non si può assumere ingenuamente che i risultati di una data analisi siano la migliore spiegazione o giustificazione di una ipotesi interpretativa[22].

2) La nozione di dato non è innocente. In primo luogo la dimensione del data set su cui si applicano le tecniche di analisi è fortemente significativa. L'efficacia e l'adeguatezza degli algoritmi probabilistici degrada notevolmente se tale dimensione è limitata, come avviene spesso in ambito letterario. In secondo luogo se un data set testuale è composto da documenti distribuiti cronologicamente su un lungo periodo di tempo, la variazione diacronica nell'uso del linguaggio (sia a livello sintattico sia semantico) possono invalidare misure puramente quantitative e statistiche. In terzo luogo molti dei fenomeni culturali ad ampio raggio e di lunga durata che i cultori del distanti reading asseriscono essere i veri oggetti di questi nuovi metodi sono intrinsecamente multilinguistici, mentre tutti i metodi standard di topic modeling e clustering lavorano su stringhe di caratteri codificati e non possono essere applicati su corpora multilinguistici. Infine i risultati di un metodo analitico sono condizionati dalla rappresentazione dei dati, ovvero dalla modalità con cui l'insieme di caratteristiche testuali viene selezionato e tradotto in termini quantitativi. Poiché esistono diversi modi di attuare questa rappresentazione, sia in relazione alla selezione (parole, n-grammi casuali di caratteri o di parole, caratteristiche morfosintattiche, …) sia alla quantizzazione (frequenza assoluta, binary scoring, TF-IDF score, …), occorre una estrema consapevolezza metodologica nella loro scelta e una grande cautela nell'interpretare i risultati dell'analisi.

3) Una delle assunzioni dei metodi di text mining (soprattutto di quelli non supervisionati) è l'approccio esplorativo del procedimento analitico. L'idea soggiacente a questo approccio è che l'analisi non debba presupporre alcuna ipotesi, teoria regolativa o modellizzazione a priori e non richieda alcun trattamento complesso dei dati. Il problema è che in ambito umanistico sin dal livello della costruzione dei dati è implicata una grande quantità di interpretazione e di teoria. La natura intenzionale degli oggetti letterari rende assai difficile individuare fenomeni rilevanti senza avere un modello ipotetico a priori. Altrettanto problematica in ambito umanistico (ma non solo) è l'idea che nell'analisi dei Big Data si debba rinunciare alla ricerca di una vera e propria relazione causale tra i fenomeni, per sostituirla con quella di correlazione. Il fatto è che dall'analisi di un data set si possono derivare numerose correlazioni senza che esista alcun criterio formale per decidere quali di queste possa avere un reale valore esplicativo.

4) Il significato nei testi letterari è articolato su molteplici livelli e alcuni di questi livelli non hanno

una lessicalizzazione diretta o ne hanno una molto complessa e dispersa (si pensi ad aspetti di un testo narrativo a differenti livelli di astrazione come l'anafora, i temi e motivi, l'intreccio, la fabula, le isotopie). Molti interessanti lavori basati su metodi quantitativi statistici o su network analysis non hanno a che fare con le proprietà intrinseche dei testi letterari quanto piuttosto su aspetti sociologici relativi alla loro produzione, diffusione e recezione come artefatti. Questo è un campo di studi molto promettente che attiene alla letteratura come fenomeno socioculturale (anche se la sociologia della letteratura non è un propriamente una novità), ma di sicuro non ci dice molto riguardo l'interpretazione dei testi.

5) I testi, come ogni altro oggetto culturale, sono fondamentalmente oggetti intenzionali: per citare il filosofo Daniel Dennet essi sono il prodotto dell'atteggiamento intenzionale dei loro creatori e fruitori[23]. Il significato di una parola, l'uso di una metafora, la scelta di una soluzione metrica e ritmica in una poesia sono determinati dall'attribuzione di senso da parte dell'autore e del lettore (non discuto qui se il primo sia più o meno rilevante del secondo). Essi possono essere, e spesso sono, idiolettali o persino idiosincratici. Una analisi puramente quantitativa e su vasta scala non riesce a catturare questo genere di fenomeni, poiché si pone a un livello affatto diverso. E tuttavia l'interpretazione dei testi non può prescinderne.

Questo ultimo punto è in effetti il cuore della mia argomentazione, poiché esso implica un fatto che la maggior parte dei lettori e, tra essi, la maggior parte degli studiosi e dei critici letterari, crede fermamente: l'interpretazione di un testo è un processo intenzionale; ogni discorso intorno ai testi letterari utilizza un insieme di nozioni e termini intenzionali (personaggio, influenza, punto di vista, autorialità, agency e così via) al fine di spiegare cosa un testo significhi e come. Nel distant reading queste nozioni vengono ignorate o sottoposte a una sorta di "riduzionismo quantitativo".

Per comprendere meglio questa argomentazione mi pare possa essere utile ragionare per analogia con il dibattito teorico ed epistemologico che ha caratterizzato le scienze cognitive negli ultimi decenni. In estrema sintesi, possiamo dire che nelle scienze cognitive e nella filosofia della mente, si sono manifestati e, in parte, avvicendati due grandi orientamenti: quello funzionalista computazionale e quello neurale, con le sue varianti computazionali, il connessionismo, e biologiche, le neuroscienze, alleate a geometrie variabile[24].

Il punto di discrimine tra questi due paradigmi che ci interessa in questa sede, e che presento in forma molto schematica, è il seguente. L'approccio funzionalista ha tra le sue preoccupazioni teoriche la salvaguardia delle nozioni della psicologia di senso comune. Questo ovviamente non significa che esse vengono accolte così come sono: tuttavia in ambito funzionalista si cerca di spiegare nozioni come quello di credenza, intenzionalità e semantica attraverso la loro riduzione a processi computazionali astratti, relativamente indipendenti dal sostrato biologico della cognizione. Nell'orientamento neurale invece l'assunto è che ogni fenomeno mentale è in realtà un processo neurofisiologico, e dunque la spiegazione della mente prima o poi sarà una riduzione del mentale al biologico; di conseguenza le nozioni intenzionali potranno essere eliminate dal discorso scientifico.

Pur con la dovuta cautela con cui va preso ogni ragionamento per analogia, mi sembra che in molte caratterizzazioni del distant reading si rinvenga un atteggiamento simile: in fondo i fenomeni letterari possono essere ridotti senza residui a fenomeni quantitativi misurabili e analizzabili in virtù di metodi puramente numerici e statistici. In ultima analisi potremmo dire che un approccio puramente quantitativo, per quanto sofisticato, sia di fatto una forma di riduzionismo o persino di eliminativismo rispetto alle nozioni intenzionali della critica di senso comune e anche delle nozioni più formali della tradizione critica strutturalista e semiotica.

Al contrario per la tradizione teorico-critica letteraria, pur nella estrema variabilità degli approcci e delle teorie, i testi sono oggetti intenzionali, e l'interpretazione consiste nella elaborazione e applicazione (più o meno consapevole a seconda del livello di lettura) di una serie di nozioni intenzionali al fine di spiegarne il funzionamento: nozioni come quella di storia, personaggio, autore e lettore implicito, descrizione. Possiamo dire che la tradizione semiotico strutturalista ha adottato verso queste nozioni la stessa strategia del funzionalismo in teoria della mente: ne ha fornito (o ha tentato di fornire) una spiegazione in termini di concetti formali più generali, o astratti, come quelli di attante, funzione narrativa, intreccio o isotopia. Questi concetti tuttavia preservano la natura intenzionale dell'interpretazione e non giustificano né richiedono

l'abolizione dei termini della critica letteraria di senso comune.

Accade talvolta che per progredire in un campo del sapere sia opportuno volgere lo sguardo all'indietro (magari ai nonni e agli zii, tralasciando i padri, come suggeriva Victor Šklovskij). Occorre cioè avere la consapevolezza del patrimonio di idee e teorie con cui nel passato le nostre discipline hanno tentato di affrontare problemi simili a quelli che oggi ci troviamo ad affrontare. Il patrimonio teorico dell'analisi semiotica e strutturale del testo si potrebbe rivelare una miniera teorica e metodologica, se opportunamente ricontestualizzato e rifunzionalizzato. Sono convinto che i metodi computazionali faranno progredire la nostra conoscenza del campo letterario e dei testi che lo costituiscono nella misura in cui, abbandonando ogni tentazione riduzionista, sapremo trovare una sintesi con quella tradizione teorica.

Bibliografia

Aiden E., Michel J.-B., *Uncharted: Big Data as a Lens on Human Culture,* New York, Riverhead Books 2013.

Bessinger J.B., Parrish S.M., *Literary Data Processing Conference Proceedings*, White Plains, NY, IBM 1965.

Blei D., *Probabilistic topic models*, in «Communications of the ACM», 55.4(2012), pp. 77-84.

Blei D., *Topic modeling and digital humanities*, in «Journal of Digital Humanities», 2.1 (2013).

Borgman C.L., *Big Data, Little Data, No Data: Scholarship in the Networked World*. Cambridge, Massachusetts, The MIT Press 2015.

Busa R., *The Annals of Humanities Computing: The Index Thomisticus. Computers and the Humanities*, in «Computers and the Humanities» 14 (1980), pp. 83-90.

Caldarelli G., Catanzaro M., *Networks: A Very Short Introduction*, Oxford, Oxford University Press 2012.

Dennett, Daniel C., *The Interpretation of Texts, People and Other Artifacts*, «Philosophy and Phenomenological Research», 50.S (1990), pp. 177-94.

Di Francesco M., *Realismo mentale, naturalismo e scienza cognitiva* in *Ritorno alla realtà*, ed. M. Ferraris e M. De Caro. Torino, Einaudi 2012.

Feldman R., Sanger J., *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge, Cambridge University Press 2007.

Hoover, D., Culpeper L.J., O'Halloran K., *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*, New York, Routledge 2014.

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

Jockers M.L., *Text Analysis with R for Students of Literature*, New York, Springer 2014.

Kadushin C., *Understanding Social Networks: Theories, Concepts, and Findings*, New York, Oxford University Press 2012.

Liu B., *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, New York, NY, Cambridge University Press 2015.

Mayer-Schönberger V., Cukier K., *Big Data: A Revolution That Will Transform How We Live, Work, and Think,* Boston, Houghton Mifflin Harcourt 2013.

McCarty W., «Literary enquiry and experimental method: What has happened? What might?», in *Storia della Scienza e Linguistica Computazionale: Sconfinamenti Possibili*, ed. Liborio Dibattista, Milano, Franco Angeli 2009.

McGann J., *Radiant Textuality: Literature After the World Wide Web,* New York, Palgrave 2001.

Morando S., *Almanacco letterario Bompiani 1962*, Milano, V. Bompiani & C. 1962.

Moretti F., *Distant Reading*, London, Verso 2013.

Moretti F., *Graphs, Maps, Trees: Abstract Models for a Literary History*, London, Verso 2005.

Moretti F., *Operationalizing: Or, the Function of Measurement in Literary Theory,* in «New Left Review» 84 (Nov/Dec 2013), pp. 103-19.

North J., *What's 'New Critical' about 'Close Reading'?: IA Richards and His New Critical Reception*, in «*New Literary History*» 44.1 (2013), pp. 141-57.

Piper A., *Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel*, in «New Literary History» 46.1 (2015), pp. 63-98.

Ramsay S., *Reading Machines: Toward an Algorithmic Criticism*. Urbana, University of Illinois Press 2011.

Richards I.A., *Practical Criticism: A Study of Literary Judgment*, London, K. Paul, Trench, Trubner 1929.

Sculley D., Pasanek B.M., *Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities*, in «Literary and Linguistic Computing» 23.4 (2008), pp. 409-24.

Underwood T., *Topic modeling made just simple enough*, in «The Stone and the Shell», 2012, https://tedunderwood.wordpress.com/2012/04/07/topic-modeling-made-just-simple-enough/.

Weingart S., *Topic Modeling and Network Analysis*, 2011, http://www.scottbot.net/HIAL/?p=221.

Note

1 Sul tema la letteratura è ormai sterminata, ci limitiamo dunque a fornire alcuni suggerimenti di lettura: V. Mayer-Schönberger and K. Cukier, *Big Data: A Revolution That Will Transform How We Live, Work, and Think*, Boston, Houghton Mifflin Harcourt, 2013; E. Aiden and J.-B. Michel, *Uncharted: Big Data as a Lens on Human Culture*, New York, Riverhead Books 2013; C.L. Borgman, *Big Data, Little Data, No Data: Scholarship in the Networked World*, Cambridge, Massachusetts, The MIT Press 2015.

2 Il Lab, ancora sotto la direzione di Moretti, coordina oggi numerosi ricercatori e studiosi che lavorano a diversi progetti di ricerca e organizzano iniziative scientifiche e didattiche. Cfr. http://litlab.stanford.edu.

3 F. Moretti, *Graphs, Maps, Trees: Abstract Models for a Literary History*, London, Verso 2005; F. Moretti, *Distant Reading*, London, Verso 2013.

4 Come noto l'origine del campo di studi che oggi chiamiamo Digital Humanities viene unanimemente indicato nel progetto di padre Roberto Busa per la lemmatizzazione

e indicizzazione del corpus delle opere di Tommaso D'Aquino che, secondo le testimonianze dello stesso Busa, inizia sin dagli anni quaranta (R. Busa, *The Annals of Humanities Computing: The Index Thomisticus. Computers and the Humanities*, in «Computers and the Humanities» 14 (1980), pp. 83-90). Le sperimentazioni per la produzione di concordanze con l'ausilio del computer proseguono poi in modo sporadico fino agli inizi degli anni 60 quando queste esperienze convergono in una serie di conferenze "fondative", di cui la prima probabilmente fu la Literary Data Processing Conference organizzata nel 1962 dalla IBM a Yorktown Heights (J.B Bessinger and S. M. Parrish, *Literary Data Processing Conference Proceedings*, White Plains, NY, IBM 1965). Due anni dopo Joseph Raben avvia la pubblicazione della prima rivista specializzata, *Computer in the Humanities*. Ma dobbiamo ricordare qui che già nel 1962 il prestigioso annuale *Almanacco Bompiani* dedica il numero a "Le Applicazioni dei Calcolatori Elettronici alle Scienze Morali e alla Letteratura", testimoniando come anche in Italia già allora vi fossero ricerche di avanguardia (S. Morando, *Almanacco letterario Bompiani1962*, Milano, V. Bompiani & C. 1962).

[5] J. McGann, *Radiant Textuality: Literature After the World Wide Web*, New York, Palgrave 2001, p. xii.

[6] W. McCarty, «Literary enquiry and experimental method: What has happened? What might?», in *Storia della Scienza e Linguistica Computazionale: Sconfinamenti Possibili*, ed. Liborio Dibattista, Milano, Franco Angeli 2009, p. 41.

[7] Senza dubbio una delle cause principali di questa difficoltà consiste nella notevole distanza tra gli assunti fondamentali delle più influenti correnti teoriche e critiche degli ultimi decenni, dal decostruzionismo agli studi culturali e in genere a tutto l'universo dei "post-strutturalismi", e i metodi formali e quantitativi tipici della critica letteraria computazionale.

[8] F. Moretti, *Operationalizing: Or, the Function of Measurement in Literary Theory*, in «New Left Review» 84 (Nov/Dec 2013), pp. 103-19.

[9] La nozione di close reading è stata coniata dal teorico e critico letterario inglese I.A. Richards nel libro *Practical Criticism* (London, K. Paul, Trench, Trubner 1929) e successivamente è stata adottata dai New Critics nordamericani come Ransom, Brooks, Warren, Wimsatt e Bearsdley, che la trasformarono in un approccio puramente formalista; su questo si veda J. North, *What's 'New Critical' about 'Close Reading'?: I.A. Richards and His New Critical Reception*, in «New Literary History» 44.1 (2013). Come detto, tuttavia, in questo contesto il termine indica in generale il metodo di lettura e analisi dei testi che ha caratterizzato gran parte della critica letteraria del 20º secolo, dal formalismo alle varie correnti post-strutturaliste, nonché la critica militante eclettica e la stessa didattica della letteratura.

[10] F. Moretti, *Distant Reading*, cit., p. 48.

[11] M.L. Jockers, *Macroanalysis: Digital Methods and Literary History*, Urbana - Chicago - Springfield, University of Illinois Press 2013, Kindle edition.

[12] R. Feldman and J. Sanger, *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, Cambridge, Cambridge University Press 2007.

[13] Spesso nella letteratura tecnica di settore si usa il termine dati "non strutturati". Rileviamo tuttavia che a ben vedere questa nozione è scorretta sia perché nessun dato digitale può essere veramente non strutturato, sia perché le tecniche di mining si possono applicare anche a dati strutturati nel senso stretto; sarebbe pertanto opportuno parlare di diversi livelli o gradi di strutturazione dei dati.

[14] Segnalo qui solo la più diffusa delle applicazioni che implementano LDA, *Mallet* (sviluppata in Java) disponibile su http://mallet.cs.umass.edu/.

[15] Oltre al capitolo 8 di Jockers, *Macroanalysis*, una ottima introduzione al topic modeling e ai fondamenti matematici di LDA pensata per un lettore umanistico è il blog post di Ted Underwood "Topic modeling made just simple enough," 2012, https://tedunderwood.wordpress.com/2012/04/07/topic-modeling-made-just-simple-enough/. Segnalo anche il blog post di Scott Weingart "Topic Modeling and Network Analysis", 2011, http://www.scottbot.net/HIAL/?p=221. Dai link inclusi in questi articoli si possono raggiungere numerosi ulteriori articoli e blog sul tema dei medesimi autori e di altri. Infine un ottimo testo introduttivo alle varie tecniche e metodi di text mining in ambito umanistico con il linguaggio di programmazione R (esplicitamente progettato per effettuare analisi statistica dei dati) è di M. Jockers, *Text Analysis with R for Students of Literature*, New York, Springer 2014).

[16] Ovviamente i topic model così generati non sono altro che distribuzioni probabilistiche di parole. É compito del ricercatore interpretare la loro coerenza semantica ed assegnargli un senso (se ce ne è uno). Per una trattazione approfondita ma discorsiva dei principi matematici degli algoritmi di topic modeling si veda dall'ideatore di LDA D. Blei, *Topic modeling and digital humanities*, in «Journal of Digital Humanities», 2 (2013), http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/. Più tecnico, dello stesso Blei, *Probabilistic topic models*, «Communications of the ACM», 55(2012), pp. 77–84.

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

[18] Il primo post nel blog di Jockers che introduce la sua teoria è *A Novel Method for Detecting Plot*, giugno 2014, http://www.matthewjockers.net/2014/06/05/a-novel-method-for-detecting-plot/. Segue il post nel quale annuncia il completamento del software (basato sul linguaggio R): *Revealing Sentiment and Plot Arcs with the Syuzhet Package*, febbraio 2015, http://www.matthewjockers.net/2015/02/02/syuzhet/. Ne è conseguito un dibattito tecnico e teorico assai vivace che ha coinvolto lo stesso Jockers e numerosi altri studiosi molto attivi in questo campo (come Ted Underwood, Andrew Piper, Scott Weingart e Annie Swafford, cui si deve la più dettagliata critica tecnica del software). Non è possibile qui citare tutti i post, gli articoli e i tweet (!) che hanno animato questo dibattito, per cui rimandiamo alla ricostruzione su Storify che ne ha fatto Eileen Clancy, "A Fabula of Syuzhet", https://storify.

com/clancynewyork/contretemps-a-syuzhet.

19 C. Kadushin, *Understanding Social Networks: Theories, Concepts, and Findings*, New York, Oxford University Press 2012; G. Caldarelli and M. Catanzaro, *Networks: A Very Short Introduction*, Oxford, Oxford University Press 2012.

20 F. Moretti, *Distant Reading*, cit., p. 214.

21 F. Moretti, *Distant Reading*, cit., p. 219.

22 Su questo come sugli aspetti critici analizzati nel punto successivo si veda anche D. Sculley and B.M. Pasanek, *Meaning and Mining: The Impact of Implicit Assumptions in Data Mining for the Humanities*, in «Literary and Linguistic Computing» 23.4 (2008), pp. 409-24.

23 D.C. Dennett, *The Interpretation of Texts, People and Other Artifacts*, in «Philosophy and Phenomenological Research», 50.S (1990), pp. 177-94.

24 Un articolo introduttivo su questo dibattito è in M. Di Francesco, «Realismo mentale, naturalismo e scienza cognitiva», in *Ritorno alla realtà*, ed. M. Ferraris e M. De Caro, Torino, Einaudi 2012.