

Fabio Ciotti

What's in a Topic Model?

Critica teorica di un metodo computazionale
per l'analisi del testo



Testo & Senso

n.18, 2017

www.testoesenso.it

Introduzione

Nel 1975 William A. Woods, all'epoca ricercatore alla Bolt Beranek and Newman company (uno dei luoghi natali di Internet), pubblicò un articolo dal titolo "What's in a link: Foundations for Semantic Networks".¹ Il lavoro, che costituì un vero e proprio punto di svolta nella storia dell'intelligenza artificiale, consisteva in una serrata e argomentata critica del concetto di "rete semantica", considerato in tutte le sue diverse incarnazioni, dalla iniziale e per molti versi semplicistica versione di Ross Quillian fino alle più sofisticate e complesse nozioni di *dipendenza concettuale*, *script* e *frame* sviluppate da Roger Shank, Marvin Minsky e altri.²

Occorre ricordare come il concetto di rete semantica fosse alla base di gran parte delle teorie semantiche formali e computazionali dell'epoca. Il limite principale di tale nozione, secondo Woods, era la carenza di una soddisfacente definizione teorica: «there is currently no "theory" of semantic networks. The notion of semantic networks is for the most part an attractive notion which has yet to be proven. Even the question of what networks have to do with semantics is one which takes some answering». ³ Vorrei qui mettere in evidenza questo punto: il problema non consisteva solo e tanto nella consistenza logica, espressività e calcolabilità, nelle proprietà formali insomma, delle diverse implementazioni del concetto di rete semantica; bensì, cosa più rilevante, nella indeterminazione del suo ruolo concettuale per la spiegazione dei fenomeni semantici nel contesto di una *teoria* della semantica.

Questo breve excursus in funzione di introduzione ha lo scopo di spiegare la ragione della immodesta allusione all'articolo di Woods nel titolo di questo articolo: è mia intenzione, infatti, misurarmi criticamente con i fondamenti teorico/letterari di una nozione (e di un metodo analitico) attualmente assai diffuso negli studi letterari computazionali: la nozione di *topic model*.⁴

Il contesto: distant reading

Il *topic modeling*, ovvero l'estrazione di cluster lessicali co-occorrenti in una collezione di documenti sulla base di calcoli statistico-probabilistici, è uno tra i più innovativi e diffusi metodi analitici computazionali tra quelli che vengono comunemente ricompresi nell'insieme dei metodi del *distant reading*. La definizione di questo approccio metodologico all'analisi dei testi (letterari e non), come noto, è dovuta al critico e storico della letteratura Franco Moretti, che ha coniato il termine stesso, giocando sulla

¹ WILLIAM A. WOODS, *What's in a Link: Foundations for Semantic Networks*, in *Readings in Cognitive Science*, Elsevier, 1988, p. 102–25, <http://linkinghub.elsevier.com/retrieve/pii/B9781483214467500145>.

² M. ROSS QUILLIAN, "Semantic Networks," in *Semantic Information Processing*, ed. Marvin L. Minsky, Cambridge, MA, MIT Press, 1968; M. MINSKY, *A Framework for Representing Knowledge*, Cambridge, MA, MIT AI Lab. Memo 306, 1974; ROGER CARL SCHANK e ROBERT P. ABELSON, *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*, Lawrence Erlbaum Associates, 1977.

³ W.A. WOODS, *op. cit.*, p.102.

⁴ Le locuzioni inglesi *topic model/modeling* non sono traducibili in italiano in modo sensato ed essendo ormai entrate nell'uso comune abbiamo preferito adottarle in lingua originale. Per una introduzione: DAVID M. BLEI, *Introduction to probabilistic topic models*, «Communication of the ACM», 2011; MATTHEW L. JOCKERS, *Macroanalysis: Digital Methods and Literary History*, University of Illinois Press, 2013; TED UNDERWOOD, *Topic modeling made just simple enough*, «The Stone and the Shell» (blog), 2012, <https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/>.

antinomia con la tradizione del *close reading* – termine critico/metodologico introdotto nella critica letteraria da I.A. Richards e poi ripreso e divulgato dai *New Critics* nord-americani negli anni '30 del secolo scorso –⁵, che qui viene inteso in un senso ampio e generico, come il metodo critico che si basa sulla lettura analitica e approfondita dei testi, da cui consegue l'interpretazione e il giudizio. A questo approccio “teologico” Moretti oppone, provocatoriamente, l'idea di abbandonare l'atto della lettura come metodo di conoscenza della letteratura:⁶

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.

Sulla scorta dei lavori di Moretti e del suo gruppo di ricerca allo Stanford Literary Lab,⁷ molti studiosi hanno adottato in modo entusiastico questo nuovo paradigma critico e metodologico, e la ricerca in questo ambito è cresciuta in modo rilevante. A ben vedere, la tradizione degli studi quantitativi e statistici nella ricerca letteraria e umanistica in generale è molto più antica del nuovo corso morettiano⁸ (in effetti, è coeva con la fondazione stessa di quel campo di studi che oggi chiamiamo Digital Humanities⁹). Tuttavia prima dell'introduzione dell'espressione “distant reading” da parte di Moretti questi studi erano considerati altamente specialistici, se non esoterici, e avevano un impatto assai ridotto sugli studi letterari genericamente intesi. Al contrario oggi sono numerosi gli studiosi impegnati in progetti di ricerca basati sull'elaborazione quantitativa e statistica di collezioni testuali, sulla network analysis e altre tecniche e metodi a essi correlati e, almeno nel panorama nord americano, i loro lavori appaiono ormai nelle maggiori riviste di ambito teorico e critico-letterario.

Tra gli studiosi che hanno entusiasticamente adottato questo approccio emerge talvolta una sorta di atteggiamento riduzionista nei confronti dei fenomeni letterari, una implicita credenza nel fatto che essi possano essere ridotti senza alcun residuo a un insieme di fatti puramente quantitativi e dunque misurabili e analizzabili con metodi numerici o statistici.

⁵ La nozione di *close reading* appare per la prima volta in I. A. RICHARDS, *Practical Criticism*, K. Paul, Trench, Trubner, London, 1929. Viene successivamente adottata e rifunzionalizzata dai *New Critics* statunitensi; su questo si veda J. NORTH, *What's 'New Critical' about 'Close Reading'?: IA Richards and His New Critical Reception*, «New Literary History», 44, 1, 2013.

⁶ FRANCO MORETTI, *Conjectures on World Literature*, «The New Left Review», 2000, <http://newleftreview.org/A2094>. Poi in: FRANCO MORETTI, *Distant Reading*, London, Verso, 2013.

⁷ Per informazioni sullo *Stanford Literary Lab* e le sue attività si veda il sito Web <https://litlab.stanford.edu>, che contiene tre le altre cose gli ormai famosi *pamphlet*, pubblicazioni digitali, spesso dallo stile colloquiale, realizzati da membri del gruppo (in genere uno o due studiosi esperti e un gruppo di studenti o dottorandi). Sono stati recentemente raccolti nel volume FRANCO MORETTI, ed., *Canon/Archive: Studies in Quantitative Formalism*, New York, N+1, 2017.

⁸ Su questo si veda il recente T. UNDERWOOD, *Genealogy of Distant Reading*, «Digital Humanities Quarterly», 11, 2 2017.

⁹ DAVID L. HOOVER, *Quantitative Analysis and Literary Studies*, in *A Companion to Digital Literary Studies*, in *Blackwell Companions to Literature and Culture*, Oxford, Blackwell, 2008), p. 517–33, http://www.digitalhumanities.org/companion/view?docId=blackwell/9781405148641/9781405148641.xml&chunk.id=ss1-6-9&toc.depth=1&toc.id=ss1-6-9&brand=9781405148641_brand; DAVID L. HOOVER, *Textual Analysis*, in KENNETH M. PRICE, RAY SIEMENS, *Literary Studies in the Digital Age*, Modern Language Association of America, 2013, <http://dlsanthology.commons.mla.org/textual-analysis/>.

Matt Jockers, uno dei primi collaboratori di Moretti, epitomizza in modo esemplare questo atteggiamento:¹⁰

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.

Asserzioni così radicali (anche se a onore del vero va detto che l'estremismo delle formulazioni da manifesto teorico non sempre corrisponde alla reale pratica critica che ne consegue) hanno ovviamente determinato reazioni critiche che si articolano su un ampio spettro. Ci sono ovviamente le posizioni "tradizionaliste" che rivendicano la radicale irriducibilità del letterario a ogni approccio formale e quantitativo, scientifico in senso stretto insomma, la sua natura "qualitativa" se non ineffabilmente spirituale (*mutatis mutandis* gli stessi argomenti furono usati contro lo strutturalismo e la semiotica del testo negli anni 70). Poi ci sono le posizioni di ascendenza ermeneutica o *reader oriented*, che ricordano come ogni lettura sia (debba essere) una operazione di disvelamento e costruzione del senso da parte del lettore e che dunque non possa mai basarsi su dati oggettivi (*a fortiori* non su dati empirici di carattere quantitativo); o quelle di ascendenza neo-storicista e contestualista, che rifiutano la natura rigorosamente formalista e linguistica degli studi quantitativi.¹¹ Anche dall'interno del campo delle Digital Humanities sono emerse posizioni "dialettiche", che propongono approcci metodologici di mediazione come l'idea di *scalable reading* di Martin Muller¹², o l'idea di una *computational hermeneutics* proposta di Andrew Piper¹³; queste posizioni in generale richiamano alla necessità di articolare l'analisi quantitativa su vasta scala con quella qualitativa, in profondità, almeno per alcune specifiche tipologie testuali (in particolare quelle poetiche e letterarie).

Vanno infine menzionati i numerosi lavori che, partendo da un punto di vista eminentemente tecnico e matematico/statistico, avanzano critiche relative alla applicabilità di specifici metodi e strumenti su dati umanistici, o che evidenziano come un loro uso "poco" consapevole possa condurre a risultati fallaci o a fraintendimenti. Uno dei più efficaci e generali è l'ottimo articolo di Sculley e Pasanek che evidenzia come alcune assunzioni implicite nei vari sistemi di *text mining* possano avere conseguenze problematiche in ambito umanistico¹⁴: la presupposizione che la distribuzione delle caratteristiche analizzate nel *dataset* sia indifferenziata e non motivata (che nel caso dei testi letterari è ovviamente assai problematica, se è vero che, per dirla con Jakobson, un testo poetico proietta l'asse della selezione su quello dell'associatività lineare); l'influenza dei metodi di rappresentazione dei dati sui risultati di un algoritmo; il cosiddetto *No free lunch theorem*: nessun algoritmo è ottimale e non è possibile costruire un meta-algoritmo che ne minimizzi le criticità in ogni data applicazione. Ogni metodo produce risultati che

¹⁰ M. JOCKERS, *Macroanalysis: Digital Methods and Literary History*, Urbana, Chicago and Springfield, Un. of Illinois Press, 2013.

¹¹ Moretti stesso ha proposto la formula "formalismo quantitativo" per caratterizzare il suo metodo. S. D. ALLISON, R. HEUSER, M. L. JOCKERS, F. MORETTI, *Quantitative Formalism: An Experiment*, Stanford Literary Lab, 2011.

¹² MARTIN MUELLER, *Scalable Reading*, «Scalable Reading» (blog), 2012, https://scalablereading.northwestern.edu/?page_id=22.

¹³ A. PIPER, *Novel Devotions: Conversional Reading, Computational Modeling, and the Modern Novel*, «New Literary History», 46, 1, 2015, p. 63–98, <https://doi.org/10.1353/nlh.2015.0008>.

¹⁴ D. SCULLEY, BRADLEY M. PASANEK, *Meaning and Mining: the Impact of Implicit Assumptions in Data Mining for the Humanities*, «Literary and Linguistic Computing», vol. 23 / 4, 2008, p. 409-424.

sono inevitabilmente influenzati dal metodo stesso e ciò ha delle conseguenze rilevanti soprattutto se i metodi sono utilizzati in contesti esplicativi e non esplorativi come avviene nelle scienze dure.

In questa sede, tuttavia, non intendo affrontare il problema generale dell'ammissibilità o plausibilità del *distant reading* e dei metodi quantitativi negli studi letterari. Piuttosto voglio discutere se sia possibile attribuire uno statuto teorico e metodologico in ambito critico letterario a uno specifico metodo computazionale quantitativo, il *topic modeling*. Quando parlo di statuto teorico mi riferisco al ruolo che la nozione di *topic modeling* gioca nel contesto della teoria del testo letterario o della metodologia della critica letteraria, e non ai suoi aspetti puramente matematici, che sono ovviamente saldamente basati sulla statistica e sulla teoria della probabilità bayesiana (Blei, 2012).

Chiaramente questa domanda implica una precisa tesi metodologica: un metodo di analisi quantitativo o computazionale riveste interesse in ambito critico letterario nella misura in cui può essere messo in correlazione (anche fornendo dati osservativi) con termini o concetti propri di una teoria del testo letterario o di una metodologia critica¹⁵. Esso dunque è accettabile se può essere assunto come un modello formale e computazionale (non necessariamente solo quantitativo/numerico) isomorfo al dominio critico letterario. Da questo punto di vista la modellizzazione computazionale è una variante della interpretazione critico/semiotica, intesa come spiegazione del funzionamento del testo, teorizzata da Umberto Eco in opposizione alla interpretazione semantica (che invece si limita ad attribuire una interpretazione specifica)¹⁶:

L'interpretazione semantica o semiosica è il risultato del processo per cui il destinatario, di fronte alla manifestazione lineare del testo, la riempie del significato. L'interpretazione critica o semiotica è invece quella per cui si cerca di spiegare per quali ragioni strutturali il testo possa produrre quelle (o altre alternative) interpretazioni semantiche.¹⁷

Topic model: cosa è tecnicamente

L'analisi del ruolo teorico e metodologico del *topic modeling* nello studio dei testi letterari richiede tanto un'adeguata conoscenza teorico-letteraria, quanto una appropriata comprensione delle basi matematiche e dell'algoritmo che ne determina il funzionamento. Uno dei rischi più evidenti e dibattuti dell'applicazione di metodi di *text*

¹⁵ D'altra parte lo stesso Moretti nel suo lavoro più eminentemente metodologico si pone il medesimo problema, e lo risolve adottando la nozione di "operazionalizzazione" dalla epistemologia operazionalista di Percy Bridgeman (*The Logic of Modern Physics*, New York, Macmillan, 1927): «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». Cfr. F. MORETTI, *Operationalizing: Or, the Function of Measurement in Literary Theory*, «New Left Review», vol. 84, 2013, p. 10

¹⁶ UMBERTO ECO, *I limiti dell'interpretazione*, Milano, Bompiani, 1990, p. 29.

¹⁷ Sono convinto che le teorie semiotiche del testo letterario elaborate a partire dalla metà degli anni 60 del secolo scorso, prima della svolta post-stutturalista, rappresentino una fonte di modelli e teorie di enorme utilità in ambito computazionale. L'interpretazione di un testo letterario consiste nella creazione ed elaborazione di una serie di nozioni intenzionali adottate per spiegarne il funzionamento: nozioni come quella di storia, personaggio, eroe, autore e lettore, descrizione. Possiamo dire che la scuola semiotica nella teoria e nella critica letteraria ha adottato verso queste nozioni la stessa strategia che la corrente funzionalista nella teoria della mente ha avuto verso i termini della psicologia del senso comune: ne ha fornito una spiegazione in termini di concetti (come quella di attante, funzione narrativa, fabula, intreccio, isotopia) e modelli formali più rigorosi. Questi concetti non negano la natura intenzionale dell'interpretazione, ma non la intendono come un fenomeno irriducibile.

mining e *machine learning* negli studi umanistici è appunto che essi siano adottati come *black box*, procedure incapsulate di cui non si conoscono i fondamenti e che sono usati senza la necessaria consapevolezza. Benjamin Schmidt, ad esempio, ha efficacemente dimostrato come l'applicazione ingenua degli algoritmi di *topic modeling* probabilistici possa generare seri malintesi interpretativi, in particolare a causa di due comuni assunzioni intuitive circa la natura delle "topic" che si rivelano fallaci¹⁸:

First, it is coherent: a topic is a set of words that all tend to appear together, and will therefore have a number of things in common. Second, it is stable: if a topic appears at the same rate in two different types of documents, it means essentially the same thing in both.

A maggior ragione, se ci proponiamo di discutere la validità di questi metodi nell'ambito della teoria e critica letteraria, occorre chiarire cosa è tecnicamente un topic model. In generale si tratta di una classe di tecniche di *text mining* non supervisionate; esistono infatti diversi algoritmi di topic modeling, tra i quali il più diffuso e applicato è quello noto come *Latent Dirichlet Allocation* (LDA). Esso è stato sviluppato da David Blei ed è basato su un approccio probabilistico bayesiano. Le cui proprietà sono così riassunte dal suo stesso ideatore:

- Topic models are algorithms for discovering the main themes that pervade a large and otherwise unstructured collection of documents. Topic models can organize the collection according to the discovered themes.
- Topic modeling algorithms can be applied to massive collections of documents.
- Topic modeling algorithms can be adapted to many kinds of data.

In modo intuitivo possiamo dire che alla base di LDA vi è un semplicistico modello generativo del testo: 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 nel testo stesso. 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 parole possono essere ripetute in ragione diversa a seconda delle loro rilevanza rispetto all'argomento. Il nostro autore dunque potrà estrarre parole in modo casuale dai vari sacchetti che corrispondono agli argomenti di cui intende scrivere: in particolare estrarrà da ciascuno di essi un numero di parole proporzionale al rilievo che intende assegnare a ciascun argomento. Alla fine non dovrà far altro che ordinare sequenzialmente l'insieme di parole ottenuto ecco che avrà generato il suo testo. Ovviamente le parole del testo avranno una distribuzione determinata dalla rilevanza degli argomenti/sacchetti da cui sono state estratte e dalla rilevanza delle parole nei singoli sacchetti. Se il tema *amore* conta nel testo per il 50% e la parola "amore" ha la il 90% di probabilità di apparire in quella topic, ne consegue che essa avrà una alta frequenza relativa nel testo. In termini tecnici si dice che in LDA

- un testo è una distribuzione di probabilità su un insieme di topic;
- una topic è una distribuzione di probabilità su un insieme di parole.

La cosa interessante di questo modello generativo è che può essere invertito e grazie all'aggiunta di alcune assunzioni *ad hoc* se ne può dedurre un algoritmo che è in grado di estrapolare automaticamente le *topic* presenti in un insieme di documenti e la loro

¹⁸ BENJAMIN M. SCHMIDT, *Words Alone: Dismantling Topic Models in the Humanities*, «Journal of Digital Humanities», n. 5, 2013, <http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/>.

distribuzione statistica nel corpus. Intuitivamente possiamo descrivere l'algoritmo in questo modo¹⁹:

1. in primo luogo i documenti vengono rappresentati sotto forma di una struttura dati detta *bag of word*. A questo fine prima viene generato un dizionario delle n forme distinte presenti nel corpus; poi ogni documento viene rappresentato come un vettore a n dimensioni in cui sono indicate, secondo l'ordine del dizionario, le frequenze assolute delle forme verbali effettivamente riscontrate nel testo o 0 se la forma è assente;
2. a questo punto occorre specificare quante topic possono essere presenti nel corpus (tralasciamo la questione del modo in cui si possa determinare questo numero). In prima istanza l'algoritmo assegnerà ogni parola a una o più topic temporanee. Queste assegnazioni vengono effettuate modo semi-casuale (secondo una distribuzione di Dirichlet, per essere esatti). Ciò significa che se una parola appare n volte, essa può essere assegnata a diverse topic;
3. dopo l'assegnazione iniziale l'algoritmo in modo iterativo verifica e aggiorna le assegnazioni delle parole alle topic, analizzando ogni parola in ogni documento. Per ciascuna parola W , l'assegnazione alla topic T nel documento D (ovvero il calcolo della probabilità che essa vi appartenga nel documento dato) viene aggiornata in base a due criteri:
 - a. la frequenza relativa con cui W appare in T ;
 - b. la frequenza assoluta dei restanti termini di T in D .

Alla fine del ciclo di iterazioni la distribuzione si assesta e di conseguenza è possibile calcolare la presenza relativa delle topic in ciascun documento.

Da questa descrizione possiamo enucleare alcune caratteristiche fondamentali del concetto di *topic model*: in primo luogo dalla iniziale rappresentazione come *bag of words* sappiamo che la struttura sintagmatica del testo viene completamente ignorata. In secondo luogo, la generazione del *topic model* assume che la distribuzione delle topic e delle parole nei testi sia stocastica e non motivata. Infine una topic è una distribuzione statistica di parole cooccorrenti in un insieme di documenti, e di conseguenza essa va *interpretata* al fine di assegnarle una interpretazione e una denominazione coerente.

Topic model: cosa sono teoricamente?

Ora che abbiamo visto in modo sintetico come funziona l'algoritmo LDA per la generazione di topic model da un insieme di documenti possiamo tornare alla questione che ci eravamo posti in apertura: quale nozione teorica della teoria letteraria corrisponde al concetto statistico/probabilistico di topic (e di topic model)? O, per usare la terminologia comune nell'ambito del *text mining*: di quale fatto o caratteristica testuale funge da "proxy" (sostituto prossimale) un topic model?

¹⁹ Tralasciamo qui alcuni dettagli tecnici, come il ruolo degli iper-parametri (o *priors*) per cui si rimanda alla letteratura specializzata. Una buona introduzione discorsiva è in T. UNDERWOOD, *Topic Modeling Made Just Simple Enough*, cit.

Un elemento discriminante ai fini di questa individuazione è il fatto che il concetto di *topic model* (anche in virtù del significato lessicale del sostantivo *topic* in inglese) viene considerato un fenomeno di carattere semantico che *emerge* dalla manifestazione lessicale lineare (o livello discorsivo) del testo, il livello in cui un testo consiste di una sequenza di unità atomiche linguistiche dotate di significato. In particolare si deve trattare di un fenomeno semantico distintivo del testo, qualcosa che abbia a che fare con la sua *aboutness*, cioè sui cui verte o di cui parla. Nella tradizione teorica e critica letteraria possiamo individuare almeno i seguenti concetti che possono essere candidati a svolgere il ruolo di correlati di *topic* (model):

- il concetto di *topic* (o *thema*) nella coppia *topic/comment* della linguistica funzionalista e testuale;
- i concetti di *tema* e *motivo* della critica tematica;
- la nozione di *isotopia* della tradizione strutturalista e semiotica;
- la nozione foucaultiana di discorso.

La coppia di concetti *topic/comment* è stata sviluppata nell'ambito della linguistica testuale a partire dalla sinonima coppia *thema/rema* dell'analisi funzionale dell'enunciato di Mathesius e dei Neo praghensi.²⁰ Come ogni altro concetto teorico della nostra lista la sua definizione non è univoca nella letteratura scientifica. In prima approssimazione il *topic* è il contenuto di un testo o di un suo sottoinsieme che ne determina una interpretazione coerente e compiuta da parte del fruitore. Esso dunque da una parte fa appello ai lessemi che sono stati letti fino a un dato momento, dall'altra alla capacità del lettore di selezionare e attualizzare il significato più adeguato di quei lessemi. Tuttavia da questo punto di vista un *topic* è un fenomeno prevalentemente pragmatico (la sua individuazione si basa su selezioni circostanziali), e non una proprietà intrinseca o apriori del testo.

I concetti di *tema* e *motivo* sono due fra i più antichi e controversi termini teorici della scienza letteraria. Non è possibile in questo contesto nemmeno accennare alla vasta serie di possibili interpretazioni (anche contraddittorie) che essi hanno avuto. Temi e motivi sono entità semantiche di diverso livello: i temi sono macro contenuti regolativi di un testo, i motivi sono unità più ristrette, che possono essere costituenti delle unità di livello tematico ma anche avere manifestazioni autonome, al limite altamente stereotipiche fino a divenire dei veri e propri topoi codificati (per esempio quello del *locus amoenus*). Entrambi possono manifestarsi linguisticamente in specifici sintagmi, enunciati o macro-enunciati, anche se non necessariamente un motivo o un tema lessicalizzato presenta una ricorrenza elevata; ma possono altresì essere veicolati da vaste porzioni discorsive (a limite un testo nella sua interezza) senza avere nessun correlato linguistico immediato.²¹ Sono identificati in quanto temi e non generici contenuti concettuali in virtù della loro natura di contenuti culturali stereotipici, che risiedono nella memoria culturale collettiva a cui autori e lettori attingono, pur mutandone nel tempo e nello spazio i valori semantici

²⁰ VILÉM MATHESIUS, JOSEF VACHEK, *A functional analysis of present day English on a general linguistic basis*, Praga: Academia, 1975.

²¹ Si veda il bellissimo capitolo *Tema/motivo* in CESARE SEGRE, *Avviamento All'analisi Del Testo Letterario*, Torino, Einaudi, 1985.

connotativi. Anche in questo caso, dunque, le varie accezioni che possiamo assegnare ai concetti di tema e motivo paiono avere poco a che fare con i risultati di un algoritmo di *topic modeling* come LDA

La nozione di isotopia è altrettanto problematica. La prima formulazione del concetto è stata avanzata come noto da Greimas:²² “un insieme di categorie semantiche ridondanti che rendono possibile la lettura uniforme di una storia”. Esso è stato successivamente ripreso più volte dallo stesso semiologo francese e da numerosi altri teorici e semiologi del testo, che lo hanno adottato per indicare classi di fenomeni completamente diversi a livello fonetico, prosodico, stilistico, retorico, sintattico, oltre che ovviamente semantico e narrativo.

In ogni caso, anche limitandosi alla nozione eminentemente semantica greimasiana, una isotopia è classe paradigmatica di tratti semantici (*semi* è il termine usato da Greimas) che presentano una omogeneità semantica e che si manifestano in modo ricorrente nel testo attraverso lessemi, morfemi, sintagmi ma anche enunciati o strutture metaforiche. Tuttavia l'isotopia non è un fenomeno unicamente lessicale e soprattutto non è interamente immanente: esso infatti è anche il prodotto della cooperazione interpretativa del lettore che attribuisce coerenza semantica a una serie di elementi testuali in virtù della selezione di specifici tratti attualizzati nell'atto della lettura e dell'interpretazione. In conclusione appare assai dubbio che l'isotopia possa essere in generale ipostatizzata in una semplice lista di lessemi.

Consapevole della difficoltà di assegnare un plausibile valore semantico letterario alle “topic” prodotte da un algoritmo statistico quantitativo come LDA, Ted Underwood ha proposto di interpretarle come “discorsi”, ovvero “kinds of language that tend to occur in the same discursive contexts”²³. Mi sembra tuttavia che questa soluzione sia solo un modo per spostare il problema, poiché richiede che lo statuto del concetto di discorso sia meglio definito. Qualcuno è stato tentato di applicare in questo contesto la nozione di formazione discorsiva di Foucault; il fatto è che il filosofo francese ha usato il concetto di discorso in molti modi non sempre coerenti e comunque mai limitandolo al livello puramente linguistico:²⁴

Una simile analisi [...] dovrebbe descrivere dei sistemi di dispersione. Nel caso in cui, tra un certo numero di enunciati si possa descrivere un simile sistema di dispersione, nel caso in cui, tra gli oggetti, i tipi di enunciazione, i concetti, le scelte tematiche, si possa definire una regolarità (un ordine, delle correlazioni, delle posizioni e dei funzionamenti, delle trasformazioni) si dirà convenzionalmente che ci si trova di fronte a una formazione discorsiva.

Senza contare che per Foucault il concetto stesso di enunciato non sembra poter essere inteso come datità linguistica, ma come “funzione di esistenza” o atto che dà luogo tale datità.²⁵ D'altra parte nella proposta di Underwood il concetto di discorso appare più simile a una nozione di sistema di dizione linguistico verbale. Ma anche questa interpretazione è problematica, poiché

²² A.J. GREIMAS, *Del Senso*, Milano, Bompiani, 1974, p. 188.

²³ TED UNDERWOOD, *What Kinds of 'topics' Does Topic Modeling Actually Produce?*, «The Stone and the Shell» (blog), 2012, <http://tedunderwood.com/2012/04/01/what-kinds-of-topics-does-topic-modeling-actually-produce/>.

²⁴ MICHEL FOUCAULT, *Le parole e le cose: un'archeologia delle scienze umane*, Milano, Rizzoli, 1967, p. 48.

²⁵ *Ibidem*, p. 101-102.

1) in questo modo i topic model escono del tutto dal campo del livello semantico e concettuale per essere considerati come fatti fondamentalmente sintattici;

2) la formulazione di Underwood sembra dare adito a circolarità, poiché è proprio la natura e la funzione letteraria di quel particolare insieme di parole che va spiegata.

D'altra parte una spiegazione puramente linguistica del concetto di *topic modeling* (a seconda dei casi si potrebbe fare appello a nozioni come quelle di *idioletto* o *socioletto*, che sono tuttavia concetti sistematici i quali presuppongono l'analisi della totalità del corpus di riferimento) ne ridurrebbe l'interesse in ambito letterario, visto che altri approcci statistici, assai più semplici e comprensibili, hanno dimostrato la loro efficacia nell'analisi di tipo stilistico.

Conclusioni

La discussione sulle possibili interpretazioni semiotico-letterarie della nozione di *topic model* e la constatazione della difficoltà teoriche che esse presentano ci porta ad affermare che in effetti non è possibile trovare un unico e soddisfacente correlato teorico-letterario dei risultati di questi metodi di analisi quantitativa. Alcuni dei termini presenti nelle descrizioni tecniche di tali metodi (come *topic* o *tema*) compaiono come termini teorici della teoria letteraria, tuttavia questa apparente congruenza si rivela una pura omonimia terminologica.

Questa difficoltà di definizione teorica, ovviamente, non comporta un rifiuto aprioristico e generalizzato dei metodi di *text mining* nello studio dei testi letterari. Tuttavia ci pare di poter dire che la loro utilizzazione è pienamente giustificata solo in un contesto di indagine esplorativa, dove una configurazione di unità linguistiche prodotta da LDA *topic modeling* (o derivati) può suggerire ulteriori percorsi di analisi, i potranno a loro volta basarsi su metodi computazionali o meno (fino a giungere al momento della interpretazione). Occorre tuttavia essere consapevoli che di volta in volta il ricercatore dovrà individuare, sulla base dei testi che ha sottoposto ad analisi, quale siano i fenomeni letterari che intende spiegare. La natura intenzionale degli oggetti letterari rende impossibile individuare fenomeni rilevanti senza avere un modello o una ipotesi critica che orienti e regoli l'indagine.

Assai più controversa, se non da scartare del tutto, la possibilità di usare il prodotto del *topic modeling* come *evidenza* empirica in un contesto esplicativo e meno ancora interpretativo. Come notava John Unsworth in una conferenza nel 2008 in cui presentava i risultati, non pienamente soddisfacenti, del progetto Monk (probabilmente il primo progetto di text mining in ambito umanistico che ha potuto disporre di ingenti finanziamenti):²⁶

I might be inclined to add one more, which is to avoid approaching the application of this technology as a matter of proving the truth of a hypothesis. For literary purposes [...] I think it makes more sense to think of text-mining tools as offering provocations, surfacing evidence, suggesting patterns and structures, or adumbrating trends. Whereas text-mining is usually about prediction, accuracy, and ground truth, in literary study, I think it is more about surprise, suggestion, and negative capability”

²⁶ JOHN UNSWORTH, *How Not To Read A Million Books*, 2008, <http://www3.isrl.illinois.edu/~unsworth/hownot2read.html>.

Questo non solo per il fatto che, come è stato osservato, in un contesto esplicativo occorre essere pienamente consapevoli del modo in cui i dati sono stati generati e di quale ruolo giochino come evidenze, ma anche perché, come abbiamo detto in precedenza, una spiegazione in ambito letterario deve rendere conto di nozioni e concetti teorici che traducono il linguaggio intenzionale con cui ogni lettore parla e interpreta i testi.