Andrea Salvini

"Estrarre" o "generare" conoscenza? Dilemmi epistemologici e virtù metodologiche emergenti da un'indagine empirica sulla religiosità in Italia.

G. Venturi, A. Cimino, F. Dell'Orletta, *La fede dichiarata. Un'analisi linguistico-computazionale*, FrancoAngeli, Milano, 2022, p. 188.

Parole chiave
Natural language processing, de-contestualizzazione epistemologica, religiosità

Andrea Salvini è professore ordinario di Sociologia Generale presso il Dipartimento di Scienze Politiche dell'Università di Pisa. I suoi interessi scientifici riguardano la metodologia della ricerca sociale, con specifico riferimento ai metodi qualitativi e alla Grounded Theory (andrea.salvini@unipi.it)

Il volume si inserisce nella cospicua serie di testi che accompagnano e arricchiscono la diffusione dei risultati della recente indagine sulla religiosità in Italia, curata da Roberto Cipriani (Cipriani 2020), serie che persegue il meritorio obiettivo di mostrare come l'enorme messe di dati raccolti siano il frutto dell'adozione di metodi differenziati di analisi, e come tale opzione consenta di assegnare alle generalizzazioni empiriche una legittimazione molto estesa e profonda basata proprio

ISSN: 2784-8272 (print) | DOI: 10.53145/indiscipline.v4i2
www.morlacchilibri.com/universitypress/indiscipline

sulla pluralità metodologica – aspetto non consueto nel panorama delle ricerche empiriche nazionali e internazionali. Alcuni di questi testi – come quello considerato in questa recensione -, dunque, hanno una vocazione specificamente metodologica e illustrano con efficacia i diversi modi attraverso cui una medesima base empirica possa essere ispezionata con efficacia per generare conoscenze utilizzabili e comparabili ai fini dell'avanzamento teorico (cfr. Cipriani, Faggiano, Piccini 2020; Punziano 2020; Quagliata 2020).

Nel volume che stiamo considerando, la base empirica è costituita da una vera e propria miniera di materiale testuale, esito della effettuazione di un ampio numero di interviste (164, per la precisione), in parte non strutturate, in parte semi-strutturate. Procedendo ancora con la metafora della miniera (tra l'altro piuttosto apprezzata in letteratura, stando alla diffusione della pratica del *text mining*), nel volume è frequente l'utilizzo dell'espressione "estrarre conoscenza", che esprime non soltanto l'esito desiderato del processo analitico, ma anche la cornice epistemologica (per quanto non particolarmente esplicitata e discussa) cui quel processo è connesso. Come ben chiarito nelle *Conclusioni*, il titolo ne descrive la vocazione programmatica: la prima parte (*La fede dichiarata*) serve a perimetrare sinteticamente sia le coordinate sostantive di riferimento (la posizione e gli atteggiamenti della popolazione italiana verso la religiosità), sia il modo in cui quegli atteggiamenti vengono espressi (mediante l'uso del linguaggio naturale in dinamiche comunicative – le interviste). Il sottotitolo, a sua volta, specifica esplicitamente la prospettiva analitica con cui sono stati esplorati i resoconti testuali delle interviste, cioè l'analisi linguistico-computazionale.

L'operazione conoscitiva che viene perseguita e presentata nel volume è innovativa e, nel contempo, ambiziosa: innovativa, poiché è finalizzata a promuovere l'incontro tra prospettive metodologiche e disciplinari differenti; ambiziosa, perché le condizioni di possibilità di questa convergenza prevedono la de-contestualizzazione epistemologica dei metodi adottati, una circostanza che deve essere attentamente valutata e discussa alla luce degli esiti empirici e conoscitivi che vengono generati – su questo punto ritorneremo al termine della recensione.

Il volume, come detto, ha una natura essenzialmente metodologica, e si pone l'obiettivo di descrivere l'applicazione dei metodi linguistico-computazionali al corpus testuale delle interviste raccolte durante l'indagine sulla religiosità. Esso illustra con efficacia le potenzialità e la fecondità della convergenza tra metodi di analisi dei dati qualitativi consolidati nella tradizione sociologica e metodi di *natural language processing*, il cui sviluppo si è accresciuto negli ultimi anni grazie agli avanzamenti della linguistica computazionale – in molta parte orientata all'esplorazione dell'enorme universo di dati reperibili sui social networks.

L'intento programmatico del volume è reso esplicito fin dall'inizio, cioè capire se e in che misura il ricorso a tecnologie di trattamento automatico della lingua (TAL) nell'analisi di un corpus testuale come le trascrizioni di interviste possa contribuire in modo significativo a "estrarre nuova conoscenza", ad esempio in termini di "generalizzazioni che fanno astrazione dalle stringhe di caratteri che si susseguono nel testo, oppure di relazioni tra le entità identificate come rilevanti" (p. 70). In una qualche misura, questo intento è condiviso con le prospettive analitiche qualitative incorporate nei software che supportano la CAQDA (Computer Assisted Qualitative Data Analysis) – non a caso, infatti, il volume si apre con un ampio capitolo dedicato alla descrizione delle virtù analitiche di NVIVO, uno dei pacchetti più noti per l'analisi dei dati qualitativi. Ciò che invece rischia di costituire un ambito di incompatibilità tra CAQDAS e strumenti TAL è proprio il riferimento al trattamento automatico dei dati testuali, rispetto al quale sussiste una significativa diffidenza da parte degli studiosi di orientamento qualitativo, non soltanto nelle scienze sociali (cfr., ad esempio, van Peer 1989; Altheide, Schneider 2013).

Tuttavia, c'è un altro punto che dev'essere considerato cruciale per comprendere il senso complessivo dell'operazione metodologica condotta nel volume: l'analisi dei testi condotta con NVIVO e descritta nel primo capitolo si è basata su un set predefinito di 219 categorie concettuali, stabilite a priori dal gruppo di ricercatori, che ha costituito la griglia attraverso cui sono stati filtrati i dati testuali. Si tratta di una

scelta ovviamente legittima, ma che limita la possibilità di valorizzare l'approccio abduttivo (e induttivo) nell'analisi dei dati, che consente di rintracciare nei testi gli elementi informativi sorprendenti e inattesi, non previsti entro le cornici dei concetti sensibilizzanti del pur ampio insieme di categorie predefinite in partenza. Al fine di ovviare ai limiti di questa scelta, che nel volume viene definita di tipo top-down, si è affidato al trattamento automatico della lingua il compito di esplorare il corpus testuale, valorizzando l'approccio di tipo bottom-up; questo compito è possibile, secondo gli autori, partendo dalla premessa che "la conoscenza contenuta in un documento è convogliata attraverso le sue strutture linguistiche", in particolare "sotto forma di unità terminologiche che è dunque possibile identificare ed estrarre in modo automatico" (p. 70).

Il volume, dunque, si articola in modo da accompagnare il lettore nel processo di graduale comprensione dei metodi analitici del TAL, per fare apprezzare – anche mediante le numerose applicazioni esemplificative – le potenzialità insite nella linguistica computazionale per l'analisi dei testi. I due corposi capitoli centrali del volume, il n. 3 e il n. 4, sono espressamente dedicati alla illustrazione e discussione di tali potenzialità: nel primo, significativamente intitolato *Dal testo alla conoscenza*, si presentano le caratteristiche essenziali degli strumenti informatici adottati e della logica che presiede al loro utilizzo. In particolare, la metodologia di estrazione della conoscenza è affidata alla piattaforma Text to Knowledge (T2K) progettata e implementata dall'Istituto di Linguistica Computazionale del CNR di Pisa (cfr. Dell'Orletta, Venturi, Cimino, Montemagni 2014); il software combina strumenti statistici e di TAL in modo tale da "trasformare la conoscenza implicitamente codificata all'interno di un corpus di documenti in conoscenza esplicitamente strutturata" (p. 89). Mediante tali strumenti, esso è in grado di identificare le entità informative più rilevanti nei testi e le relazioni che le legano. Non entreremo nel dettaglio delle procedure tecniche che caratterizzano le procedure estrattive, tuttavia vale la pena sottolineare come i "nuclei informativi di base" siano individuati ricercando "iterativamente sequenze ripetute di unità linguistiche (morfo-sintattiche) corrispondenti a unità

lessicali e terminologiche"; questa analisi, come si capisce, va ben oltre il computo delle co-occorrenze delle unità lessicali, in quanto consente di estrarre anche le "azioni nelle quali tali entità sono inserite" (p. 95). Le procedure di indicizzazione, inoltre, consentono di rintracciare i contesti (gli atti illocutori, cioè i brani di testo) in cui le unità terminologiche occorrono (allo stesso modo delle tecniche di retrieving nell'analisi qualitativa), così come di ricostruire le connessioni con le altre unità, che possono essere visualizzate come strutture reticolari di tipo semantico (anche in questo caso, come accade per le funzioni di mapping dei codici nei CAQDAS). In effetti, durante la lettura del testo, può risultare del tutto naturale comparare mentalmente le funzioni di T2K – almeno quelle presentate nel volume – e le procedure adottate da un qualsiasi software di analisi qualitativa dei dati. Questi ultimi, infatti – al di là del modo in cui si è deciso di procedere nella specifica indagine sulla religiosità –, sono stati progettati per svolgere attività analitiche di dati testuali (e non solo testuali, ma anche di tipo audio e video) perseguendo un approccio di tipo bottom-up – attraverso la costruzione di codici e categorie concettuali il cui livello di astrazione, sul piano concettuale, è gradatamente crescente. A parità di approccio bottom-up (si potrebbe semplicemente dire induttivo), la differenza fondamentale che sembra istituirsi tra T2K e i CAQDAS (dai più complessi, come NVIVO o Atlas.ti, ai più semplici, ma non necessariamente meno efficaci, come Qualcoder), risiede nella circostanza per cui il processo di estrazione della conoscenza, nel primo caso, è essenzialmente affidata ad algoritmi statistici e procedure linguistico-computazionali; nel secondo, alle virtù interpretative del ricercatore. L'attribuzione di rilevanza alle espressioni presenti nelle interviste dipende infatti, nel primo caso, dalla ricostruzione di glossari terminologici compiuta mediante meccanismi di co-occorrenza dei lemmi nei contesti testuali (le co-occorrenze possono prevedere anche espressioni combinate di lemmi), nel secondo caso dalla salienza che quelle espressioni acquistano rispetto al dominio esplorato (l'oggetto di indagine, nel vocabolario metodologico). Questa salienza deriva da un atto interpretativo del ricercatore, il quale ha il compito – non irrilevante – di render conto di tale attribuzione di rilevanza (utilizzando i cosiddetti

memos) mediante l'esplicitazione del processo logico-argomentativo in base al quale ha assegnato significato alle espressioni, sintetizzando quel significato nei codici e nelle categorie.

Da quello che si può osservare dalle esemplificazioni presenti nel volume, c'è una singolare similitudine tra i corrispondenti lessicali dei lemmi estratti da T2K e la forma con cui normalmente si definiscono i codici nei CAQDAS: nel primo caso, tuttavia, le espressioni concettuali assumono significato in quanto classificate sulla base del loro peso quantitativo nei contesti testuali, sia in termini assoluti, sia in termini relativi, cioè disaggregando quel peso e articolandolo secondo le variabili indipendenti prese in considerazione nell'analisi (come per esempio il genere, l'età, la ripartizione territoriale dei soggetti intervistati). Nel secondo caso, i codici e le categorie sono ricostruzioni interpretative operate dal ricercatore sulla base dei suoi sistemi di rilevanza e della costante comparazione tra i diversi elementi della base empirica. Questi brevissimi cenni sul raffronto tra le procedure di analisi quantitativa (operata da strumenti come T2K) e qualitativa (operata dai CAQDAS) delle basi empiriche testuali possono essere utili per apprezzare le potenzialità dei diversi strumenti informatici, ma anche e soprattutto le differenti logiche che sottendono al loro utilizzo (su questo punto, ovviamente si rinvia alla letteratura specifica, come ad esempio Krippendorf 2004; Bolasco 2021).

Vale la pena, a questo proposito, ricordare come alcuni CAQDAS (come lo stesso NVIVO), offrano funzioni basiche di analisi quantitativa dei testi, rispetto alle quali la piattaforma T2K offre un orizzonte di possibilità e di sviluppo analitico ben più ampio, strutturato e coerente, che permette al ricercatore di compiere esplorazioni approfondite e puntuali. Di particolare interesse, inoltre, è la possibilità, mediante T2K, di costruire grafi che consentono di visualizzare le relazioni semantiche che si istituiscono tra i lemmi estratti nel glossario, in modo da favorire l'analisi interpretativa dello studioso, che può esplorare la struttura delle relazioni seguendo i percorsi che collegano direttamente e indirettamente i lemmi.

Il quarto capitolo introduce la Sentiment Analysis come strumento che T2K rende disponibile per lo studio delle polarità di un testo

(positiva, negativa, neutra). Al di là delle finalità meramente descrittive, è interessante constatare come nelle esemplificazioni empiriche presentate, la polarità di un contesto (cioè di una proposizione, di un brano di testo) tenda a influenzare l'esistenza di specifiche relazioni tra concetti. Inoltre, la classificazione automatica delle polarità dei contesti in cui ricorrono i lemmi presenta coerenze significative con i macro-concetti nei quali è stato possibile organizzare i concetti nella precedente indagine compiuta con NVIVO. Quest'ultimo riferimento è rilevante per comprendere come sia continuamente presente, nella consapevolezza degli autori, la necessità di combinare gli automatismi connessi con le procedure di estrazione delle conoscenze con lo sguardo competente dell'esperto di dominio (in questo caso, dei sociologi della religione), in modo da attribuire senso sia alle scelte procedurali, sia alla interpretazione degli esiti di processo.

La continua interlocuzione tra percorso analitico e quadri concettuali costituisce il valore aggiunto dell'operazione metodologica prospettata nel volume: questa reciproca chiamata in causa, descritta nel volume come combinazione tra approccio top-down e approccio bottom-up, presuppone un continuo dialogo tra ambiti disciplinari solo apparentemente lontani, quello sociologico e quello linguistico-computazionale, che può offrire insights significativi su un tema così complesso come quello della religiosità degli individui. Come si è detto all'inizio di questo contributo, le condizioni di possibilità di tale dialogo prevedono la specificazione di alcuni aspetti epistemologici e metodologici che nel testo avrebbero potuto trovare maggiore attenzione. Infatti, l'applicazione della linguistica computazionale implica l'adozione di procedure standardizzate nell'indicizzazione delle unità lessicali e nei processi di estrazione della conoscenza – espressione che richiama un orientamento di tipo essenzialista nell'approccio ai dati testuali. Si tratta, dunque, di un insieme di tecniche e procedure riassumibili come analisi quantitativa-computazionale di dati qualitativi. Non mancano, nella letteratura sociologica, significative riflessioni che assegnano un posto di rilievo all'analisi quantitativa dei contenuti (cfr. ad esempio, Della Ratta 2009). Tuttavia, è opportuno valutare qui l'appropriatezza dei metodi

della linguistica computazionale a quegli specifici *corpora* testuali costituiti dai resoconti di interviste nel quadro della ricerca sociologica, che, come segnalato anche nel volume, assumono caratteri diversi rispetto ad altri corpora (come i testi ricavabili dai social networks o i testi letterari). Attingendo alla letteratura più consolidata nella metodologia della ricerca sociale, si può legittimamente ritenere che coloro i quali decidono di svolgere la propria indagine (o parte di essa) mediante l'effettuazione di interviste libere o non strutturate perseguono l'obiettivo di cogliere i significati che le persone attribuiscono a eventi e situazioni rilevanti per l'oggetto di indagine, nonché i modi attraverso cui quei significati sono processualmente costruiti (adottando, dunque un orientamento costruzionista nella analisi e nei modi attraverso cui si genera conoscenza). Di conseguenza, può risultare problematico presupporre che tali significati siano convogliati attraverso le strutture linguistiche, intese come unità lessicali e terminologiche indicizzabili. Altrettanto controverso è il presupposto in base al quale quei contenuti possano essere estratti mediante un'analisi statistico-linguistica piuttosto che con un atto interpretativo dei riferimenti simbolici contenuti nei testi, e prima ancora, delle modalità comunicative con cui vengono espressi i resoconti linguistici (di cui la Sentiment Analysis non sembra ancora in grado di render conto pienamente).

Una condizione essenziale della applicabilità del trattamento automatico del linguaggio, dunque, è proprio quella della de-contestualizzazione epistemologica delle procedure metodologiche adottate per l'analisi dei dati qualitativi, e di assumere – come correttamente viene esplicitato nel testo, senza tuttavia compiere ulteriori specificazioni – che sia la terminologia nominale incardinata nelle strutture linguistiche a veicolare i concetti espressi dagli intervistati.

Del resto, è lo stesso Roberto Cipriani, nell'*Introduzione* al volume, ripercorrendo il percorso di legittimazione dei metodi qualitativi nel panorama sociologico, a ricordare che gli avanzamenti della conoscenza sui fenomeni sociali sono strettamente connessi con lo sviluppo del dialogo interdisciplinare e con la valorizzazione dell'incontro tra prospettive metodologiche ed epistemologiche differenziate. La convergenza tra analisi

linguistica-computazionale e analisi sociologica prospettata nel volume costituisce una tappa sicuramente significativa nella sperimentazione di questo percorso dialogico.

Riferimenti bibliografici

Altheide, D. L., Schneider, C. J.
2013, *Qualitative Media Analysis*, Sage Publications, Los Angeles.

Bolasco, S.
2021, *L'analisi automatica dei testi: Fare ricerca con il text mining*, Carocci, Roma.

Cipriani, R.
2020, *L'incerta fede. Un'indagine quanti-qualitativa in Italia*, FrancoAngeli, Milano.

Cipriani, R., Faggiano, M. P., Piccini, M. P.
2020, *La religione dei valori diffusi. Intervista qualitativa e approccio misto di analisi*, FrancoAngeli, Milano.

Dell'Orletta F., Venturi G., Cimino A., Montemagni S.
2014, *T2K². A System for Automatically Extracting and Organizing Knowledge from Texts*, Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014), European Language Resources Association (ELRA), Reykjavik, Islanda, 26-31 May 2014, pp. 2062-2070.

Della Ratta Rinaldi, F.
2009, *L'analisi testuale computerizzata*, in L. Cannavò, L. Frudà (a cura di), *Ricerca sociale: Tecniche speciali di rilevazione, trattamento e analisi*, pp. 133-152, Carocci, Roma.

Krippendorff, K.
2004, *Content Analysis: An Introduction to Its Methodology* (2nd ed.), Sage Publications, Thousand Oaks, CA.

Punziano, G.
2020, *Le parole della fede. Espressioni, forme e dimensioni della religiosità tra pratiche e sentire in Italia*, FrancoAngeli, Milano.

Quagliata, A. (a cura di)
2020, *Il dogma inconsapevole. Analisi del fenomeno religioso in Italia: il contributo qualitativo della Grounded Theory costruttivista*, FrancoAngeli, Milano.

van Peer, W.
1989, *Quantitative Studies of Literature. A Critique and an Outlook. Computers and the Humanities*, v. 23, n. 4/5, pp. 301-307.