Postfazione
Afterword
L'analisi quantitativa nella ricerca linguistica

Arjuna Tuzzi

Università di Padova

In linguistica il termine "analisi del testo" rimanda a una tradizione di studi qualitativi per l'analisi critica, retorica e stilistica che è parte della 'cassetta degli attrezzi' in dote alla disciplina. In altre aree, invece, si attribuiscono a questo termine significati diversi, con riferimento a tradizioni di studi diverse ma che fanno ugualmente parte del bagaglio culturale e del sapere condiviso di ciascuna disciplina.

Con una prospettiva interdisciplinare si può osservare che il termine "analisi del testo" è molto generico e può fare riferimento a un ampio ventaglio di approcci diversi: qualitativi, quantitativi e misti (Russell & Ryan 2010 e 1998; Tuzzi 2003). Con il termine "analisi dei dati testuali" si fa più precisamente riferimento a un processo di rilevazione, classificazione, codifica, sintesi, analisi e interpretazione delle informazioni contenute in un testo, ponendo l'attenzione sulla qualità dei dati, dalla loro raccolta all'esposizione dei risultati. Quando in questo processo entrano in gioco strumenti statistici (o, più in generale, quantitativi), si può parlare di "analisi statistica (o quantitativa) dei dati testuali" (Lebart, Salem & Berry 1998).

Nel corso degli anni, l'analisi dei dati testuali (ADT) è diventata un oggetto di ricerca in alcuni settori della linguistica, dell'informatica e della statistica (Roberts 2000) e uno strumento di ricerca per numerose altre discipline, come per esempio psicologia, semiotica, sociologia, sociolinguistica, storia, studi politici, della comunicazione, dell'educazione, ecc. I campi di applicazione sono tanti quanti i contesti di ricerca in cui è opportuno usare documenti scritti (o trascritti) e studiosi di diverse discipline fanno un uso costante e intensivo di

corpora testuali nella ricerca applicata, come raccolte di articoli di giornale, messaggi pubblicitari, opere letterarie, scambi epistolari, discorsi pubblici e istituzionali, documenti di interesse storico, atti pubblici, interviste, storie di vita, messaggi postati in rete, ecc.

Nata negli Stati Uniti d'America come *content analysis* all'inizio del secolo scorso (Lasswell 1927), nelle scienze sociali l'ADT gode di una lunga tradizione di studi e ha sperimentato ciclicamente periodi di espansione e periodi di ripensamento (Pool 1959). In linguistica i metodi quantitativi si possono far risalire a illustri matematici del passato come Eustop, Mandelbrot, Marcov, Shannon, Zipf e ai grandi padri della linguistica quantitativa (Reed 1949, Yule 1939, 1944, Herdan 1956, Guiraud 1960, Muller 1968) ma, fino agli anni '70, hanno costituito una nicchia rispetto al *mainstream* degli studi linguistici. Grazie allo sviluppo dell'informatica e delle *information technologies* (IT), nel corso del tempo si è sviluppata una vera e propria costellazione di aree di ricerca: linguistica computazionale, analisi statistica dei dati testuali, linguistica dei corpora, linguistica matematica, *information retrieval, natural language processing, text mining*, ecc., per citarne solo alcune più vicine agli interessi della linguistica.

L'importanza dei corpora, dei dati testuali e delle metodologie per la loro analisi è cresciuta e continua a crescere in ragione della disponibilità accresciuta di testi in formato elettronico e di software per la loro elaborazione. L'area è marcatamente interdisciplinare e i diversi settori di ricerca si qualificano e distinguono, oltre che per i diversi approcci teorici di riferimento, anche in base al tipo di dati che vengono analizzati, agli obiettivi perseguiti e agli strumenti implementati.

Il processo che conduce dall'osservazione di una collezione di testi a un insieme strutturato di dati elaborabili, cioè la cosiddetta fase di "costruzione del dato", è complesso perché, a monte, esiste il problema della costituzione di corpora in possesso di caratteristiche di ampiezza, coerenza, omogeneità e, in alcuni casi, di esaustività che li rendano idonei agli scopi perseguiti dalla ricerca. Per valutare l'insieme dei testi che compongono un corpus, i criteri da tenere in considerazione sono numerosi; si possono citare la lingua, il tema, lo stile, l'autore (Muller & Brunet 1988) e, solo per quello che riguarda la lingua, il problema della variazione può essere articolato in ulteriori cinque dimensioni (Berruto 1987): diacronica, diatopica, diafasica, diastratica, diamesica. Alla luce delle tante variabili da tenere sotto controllo, raramente si perviene a un risultato del tutto soddisfacente (Popescu, Mačutek & Altmann 2009).

La diffusione di testi accessibili, trattati in modo uniforme e facilmente interrogabili in formato elettronico rappresenta un obiettivo irrinunciabile per la ricerca linguistica ma, in questo volume, il lavoro di Manuel Barbera e Cristina Onesti sottolinea come la disponibilità di corpora di ampie dimensioni e di strumenti specifici per la loro analisi risulti ancora troppo limitata. Lo stesso lavoro offre una riflessione metodologica e un insieme di linee guida per l'annotazione di varietà di linguaggio giornalistico che rendono evidenti il ruolo e l'importanza di un modello di trattamento dei testi standardizzato e condiviso (*tokenization* e *markup*).

Per quanto riguarda il tipo di dati impiegati nell'ADT, le informazioni raccolte possono essere collocate a livello fonetico, grammaticale (morfologico e sintatti-

co) e lessicale. Tra gli studi pionieristici di area statistica si trovano esempi di corrispondenza fonetica; a livello grammaticale sono stati sviluppati modelli (anche matematici) per rappresentare il processo di formazione di strutture complesse a partire da unità testuali elementari (fonemi, morfemi, lettere, sillabe, parole, ecc.); la ricerca lessicale si è concentrata su unità testuali elementari e sequenze di unità testuali elementari (segmenti ripetuti, *multi-words*, *n-grams*, ecc.) per cogliere i tratti lessicali significativi di un testo (Lebart, Salem & Berry 1998).

Tra i saggi raccolti in questo volume, il lavoro di Anna Macedoni è un esempio di ricerca sistematica per la classificazione di informazioni a livello testuale, morfosintattico e lessicale. Dal punto di vista della raccolta dei dati, lo studio segue il percorso classico di analisi linguistica di un *corpus* e vi aggiunge il problema specifico dell'interferenza. Sulla stessa linea, anche se scende maggiormente nel dettaglio, il lavoro di Stefano Ondelli e Matteo Viale, mentre si concentrano su fenomeni più specifici il contributo di José Francisco Medina Montero (marcatori del discorso) e quello di Giuseppe Palumbo e Maria Teresa Musacchio (connettori interfrasali). Gli ultimi tre contributi citati si spingono, inoltre, fino alla quantificazione dei fenomeni oggetto di studio.

L'elaborazione delle informazioni necessita di metodi quantitativi soprattutto quando i corpora sono di grandi dimensioni e i dati devono essere estratti in maniera automatica. L'idea di fondo dell'ADT è di rappresentare alcuni tratti caratteristici di un testo attraverso semplici misure quantitative e di utilizzarle a fini comparativi. È il caso della *weirdness* che, grazie alla mediazione di un modello di riferimento (CORIS), viene usata da Palumbo e Musacchio per identificare i connettivi caratteristici di due corpora specialistici (articoli di argomento economico scritti originariamente in italiano e tradotti dall'inglese). Un modello di riferimento (il *Vocabolario di Base*) viene chiamato in causa anche da Ondelli e Viale per confrontare il lessico di due corpora di italiano giornalistico (italiano originale e delle traduzioni nei quotidiani) e, nel corso del loro contributo, vengono introdotti vari indicatori, come la densità di forestierismi non adattati per 1.000 occorrenze o la distribuzione percentuale delle occorrenze per *parts-of-speech*.

Esistono intere famiglie di misure (Köhler & Altmann 2009, Popescu 2009, Strauss, Fan & Altmann 2008) per valutare la ricchezza lessicale, l'articolazione in categorie grammaticali (*part-of-speech distribution*), la concentrazione tematica, la co-occorrenza di parole, la leggibilità di un testo, ecc., ma in letteratura risulta nota l'esistenza di problemi di comparazione quando i testi sono di lunghezza significativamente diversa. In pratica questi "parametri" non sono mai veri e propri parametri in senso statistico (quantità fisse che si possono considerare come il vero valore di una variabile) perché restano dipendenti dalla dimensione dei testi (Strauss Fan & Altmann 2008, Cortelazzo & Tuzzi 2008, Tweedie & Baayen 1998). Per tentare di neutralizzare l'effetto della dimensione, la letteratura propone varie soluzioni, sia a livello algebrico che a livello di procedura di calcolo, e in questo contesto si inserisce il tentativo di stima del rapporto tipi-occorrenze basato su campioni di uguale dimensione del lavoro di Ondelli e Viale.

La gestione della diversa dimensione di due testi nell'ambito specifico delle misure di similarità tra testi viene affrontata in questo volume nel contributo di Dominique Labbé. La questione di come stabilire una misura di distanza tra due testi ha sempre affascinato gli studiosi e spesso acceso dispute appassionate, soprattutto quando si è intrecciata con la questione dell'attribuzione d'autore nel caso di opere anonime o di incerta autorialità. Il problema di stabilire la distanza tra due testi e, più in generale, di determinarne il grado di dipendenza è una tipica domanda di ricerca nell'ambito del *document clustering*: nel momento in cui è disponibile una misura di distanza, le tecniche di *cluster analysis* sono in grado di riconoscere in maniera automatica gruppi di testi 'vicini', cioè simili, senza utilizzare informazioni a priori sull'appartenenza a gruppi e senza interventi da parte del ricercatore.

Molte misure quantitative si basano su profili lessicali, cioè sulle liste di unità testuali elementari corredate dalle rispettive occorrenze nei testi, come i vocabolari di frequenza per forme o per lemmi (Baayen 2001, Popescu 2009). Le occorrenze rappresentano un insieme di indicatori semplici di presenza, assenza, frequenza di una parola e, a partire da questi indicatori semplici, si possono costruire indicatori composti di vario tipo, come per esempio la distanza intertestuale discussa da Labbé. Una fetta consistente dei metodi di ADT si basano sul *bag-of-words* allo scopo di rappresentare le relazioni esistenti tra i profili lessicali dei testi e di sintetizzarne i tratti principali attraverso grafici. In molti casi il *focus* delle analisi si sposta dallo studio delle strutture della lingua e delle leggi (universali) del linguaggio ai contenuti veicolati dai testi, con una prospettiva che è più vicina alla *content analysis* di matrice sociologica e psicosociale che a quella della linguistica quantitativa. È il caso dell'analisi delle corrispondenze (Greenacre 1984 e 2007; Murtagh 2005; Lebart, Morineau & Warwick 1984), una delle tecniche statistiche più diffuse in questo campo.

Gli strumenti informatici, linguistici e statistici offrono una gamma di possibilità molto ampia e a diversi livelli di complessità. Software sempre più potenti e sofisticati permettono l'elaborazione automatica dei dati testuali a partire da corpora di dimensioni così vaste da rendere impensabile un approccio di tipo qualitativo. Si tratta certamente di una grande opportunità per la ricerca, ma non bisogna dimenticare che i software riflettono gli approcci teorici delle scuole che li hanno sviluppati, i quali ne determinano gli obiettivi e le potenzialità così come i limiti e le rigidità. Nella scelta di un software devono essere chiari gli scopi perseguiti dalla ricerca, sia perché i prodotti disponibili sul mercato non possono soddisfare tutte le esigenze, sia perché è necessaria una discreta competenza metodologica per controllare i passaggi necessari allo svolgimento delle operazioni. Inoltre, alcuni software offrono soluzioni per l'analisi semiautomatica dei testi, cioè prevedono scelte discrezionali e richiedono l'intervento manuale del ricercatore.

L'ADT assistita da software garantisce una velocità e una sistematicità alle operazioni di ricerca, spoglio e sintesi delle informazioni di interesse che difficilmente possono essere garantite dalle analisi qualitative, soprattutto in presenza di corpora di grandi dimensioni. Permette, quindi, di superare quegli ostacoli che rappresentano i principali limiti dell'analisi qualitativa. Tuttavia, l'ADT non deve essere immaginata come alternativa ai tradizionali approcci

qualitativi: i metodi quantitativi offrono, da un lato, spunti per approfondimenti qualitativi e, dall'altro, strumenti per verificare su larga scala le intuizioni emerse da una prima analisi qualitativa.

È evidente che il passaggio dal testo al dato testuale sacrifica una parte della ricchezza espressa ma, se questa perdita è compensata dalla possibilità di elaborare grandi masse di dati in tempi contenuti, il ricorso all'elaborazione statistica risulta vantaggioso. Il fatto di poter lavorare con corpora di grandi dimensioni rappresenta, però, anche un limite di questi metodi perché i risultati migliorano al crescere delle dimensioni. Pertanto, l'ADT non è un approccio adatto a studiare piccoli testi che, tuttavia, possono essere la norma in alcune specifiche situazioni applicative.

I corpora sono entità complesse e possono essere valutati sulla base di criteri, anche soggettivi, che difficilmente possono essere trasformati in dati. L'ADT non è in grado di risolvere le ambiguità dell'espressione linguistica o di rilevare la 'bellezza' di un brano e sicuramente soffre della frammentazione in unità testuali semplici, che perdono il contesto d'uso e il significato dei componenti della frase. Per le stesse ragioni, questi metodi non riescono a cogliere la presenza di paradossi, ironia, figure retoriche, doppie negazioni, ecc. Forse sarebbe più corretto dire che non riescono *ancora* a cogliere certi aspetti, perché in questa direzione si sta muovendo a grandi passi tutta quell'area di ricerca che trova sbocco naturale nelle applicazioni di intelligenza artificiale. Il *text mining* e altre discipline come *knowledge discovery, machine learning, information extraction, natural language processing* si pongono obiettivi che, fino a pochi anni fa, sarebbero stati considerati fantascienza, come riconoscere relazioni lessico-semantiche, assegnare un peso ai testi su base tematica o rilevare in maniera automatica il *mood* di una comunicazione.

L'idea che si possa estrarre da un testo senso e significato in maniera automatica non è unanimemente accettata da un punto di vista teorico e non è ancora praticabile nella ricerca applicata ma, ormai, la strada verso l'analisi automatica della semantica del testo sembra tracciata. Allo stato attuale, gli strumenti a disposizione sono ancora limitati e affetti da errori così grossolani da esporli a facili critiche. Tuttavia, prima di aderire alla schiera dei detrattori, vale la pena di riflettere su un punto cruciale: lo sviluppo di tutti questi strumenti è dettato dalla convinzione che anche l'intervento umano non è esente da errori; inoltre è soggettivo, non riproducibile e, cosa fondamentale per molte applicazioni, troppo costoso in termini di risorse e di tempo.

Riferimenti

Baayen H.R. (2001) *Word Frequency Distributions. Exploring Quantitative Aspects of Lexical Structure*, Dordrecht, Kluwer Ac. Pub.

Bernard H.R. & Ryan G.W. (1998) "Text Analysis: Qualitative and Quantitative Measures", in *Handbook of Methods in Cultural Anthropology*. Ed. by H. R. Bernard Walnut Creek, Altamira Press, pp. 595–646.

Bernard H.R. & Ryan G.W. (2010) *Analyzing Qualitative Data: Systematic Approaches*, Los Angeles, Sage.

Berruto G. (1987) *Sociolinguistica dell'italiano contemporaneo*,

Roma, La Nuova Italia Scientifica.

Cortelazzo M. & Tuzzi A. (2008) *Metodi statistici applicati all'italiano*, Bologna, Zanichelli.

Greenacre M.J. (1984) *Theory and Application of Correspondence Analysis*, London, Academic Press.

Greenacre M.J. (2007) *Correspondence Analysis in Practice*, London, Chapman & Hall.

Guiraud P. (1960) *Problèmes et méthodes de la statistique linguistique*, Paris, Presses Universitaires de France.

Herdan G. (1956) *Language as Choice and Chance*, Groningen, Noordhoff LTD.

Lasswell H.D. (1927) *Propaganda Technique in the World War*, New-York, Alfred A. Knopf.

Lebart L., Morineau A. & Warwick K.M. (1984) *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*, New-York, Wiley.

Lebart L., Salem A. & Berry L. (1998) *Exploring Textual Data*, Dordrecht, Kluwer Ac. Pub.

Muller C. (1968) *Initiation à la statistique linguistique*, Paris, Larousse.

Muller C. & Brunet E. (1988) "La statistique résout-elle les problèmes d'attribution?", *Strumenti critici*, n.s., 3, 3, pp. 367-387.

Murtagh F. (2005) *Correspondence Analysis and Data Coding with Java and R*, London, Chapman & Hall/CRC.

Pool I. de S. (1956) (ed.) *Trends in Content Analysis*, Illinois, University of Illinois Press.

Popescu I.-I. (2009) *Word Frequency Studies*, Berlin, Mouton de Gruyter.

Popescu I.-I., Mačutek J. & Altmann G. (2009) *Studies in Quantitative Linguistics 3*, Lüdenscheid, RAM-Verlag.

Reed D.W. (1949) "A Statistical Approach to Quantitative Linguistic Analysis", *Word* V, pp. 235-247.

Roberts C.W. (2000) "A Conceptual Framework for Quantitative Text Analysis", *Quality & Quantity* 34, pp. 259-274.

Strauss U., Fan F. & Altmann G. (2008) *Problems in Quantitative Linguistics 1*, Lüdenscheid, RAM-Verlag.

Tuzzi A. (2003) *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*, Roma, Carocci.

Tweedie F.J. & Baayen R.H. (1998) "How Variable May a Constant Be? Measures of Lexical Richness in Perspective", *Computers and the Humanities*, 32:5, pp. 323-352.

Yule G.U. (1944) *The Statistical Study of Literary Vocabulary*, Cambridge, Cambridge University Press.

Yule G.U. (1939) "On Sentence Length as a Statistical Characteristic of Style in Prose, with Applications to Two Cases of Disputed Authorship", *Biometrika* 30, pp. 363-390.