COME SCRIVEVA GRAMSCI?

METODI MATEMATICI PER RICONOSCERE SCRITTI GRAMSCIANI ANONIMI

di Maurizio Lana

Il giovane Antonio Gramsci nei suoi primi anni di attività collaborò a vari quotidiani sui quali i suoi articoli, come quelli degli altri collaboratori, non sempre uscivano firmati. Tali giornali contengono quindi articoli di Gramsci per così dire nascosti nella massa degli articoli anonimi. L'articolo descrive l'attività di un gruppo di ricerca che ha messo a punto due metodi di tipo matematico (e non statistico) per il riconoscimento dei testi gramsciani anonimi. La qualità di tali metodi è stata verificata in due fasi (test in chiaro e test cieco) ottenendo risultati interessanti ed affidabili (più dell'80% di riconoscimenti correnti, e nessun falso positivo), prima di iniziare il lavoro di attribuzione sui testi anonimi. L'analisi venne richiesta dagli studiosi curatori dell'edizione nazionale delle opere di Antonio Gramsci, diretta e curata dal Ministero per i Beni e le Attività Culturali e dalla Fondazione Gramsci di Roma. Il fatto che con metodi matematici molto lontani dai tradizionali criteri linguistici, stilistici, grammaticali, di segmentazione del testo si possano ottenere risultati significativi apre questioni e pone interrogativi di grande portata.

Young Antonio Gramsci started his career as a journalist and he wrote for newspaper with a clear left-winged collocation: «Il grido del popolo», «Avanti!», and some other. At those times it was usual not to sign the articles, so those newspaper published a big amount of unsigned articles, and among them are hidden and anonymous articles written by Gramsci. In the last years, the «Edizione nazionale delle opere di Antonio Gramsci» has started, by the Italian ministry for the cultural heritage (MiBAC) and the «Fondazione Gramsci» of Rome. As it is a long time that the anonymous newspaper articles of the young Gramsci are studied, searched and investigated, the direction of the National Edition asked for new light about the matter, activating a long term research project aimed to discover hidden Gramsci articles by means of quantitative attribution methods (one based on n-grams counting, and one based on information entropy measurement). The research is ongoing, with interesting results and confirmations in every situation where defined infos and data are available and many more interesting open questions.

———

A Questo articolo prende spunto da un lavoro di ricerca che ha per oggetto l'individuazione, all'interno di un *corpus* di articoli di giornale anonimi, di articoli scritti da Antonio Gramsci. La ricerca, che si svolge utilizzando metodi di analisi quantitativi, è condotta da chi scrive insieme a tre colleghi fisici matematici, Dario Benedetto ed Emanuele Caglioti dell'Università La Sapienza di Roma, e Mirko Degli Esposti dell'Università di Bologna con C. Basile, dottoranda in matematica, ed è realizzata su richiesta della Commissione per l'edizione nazionale delle opere di A. Gramsci, nella persona del presidente G. Vacca. Dell'edizione nazionale delle opere fanno parte anche gli scritti giornalistici del periodo 1913-26, che come era abituale in quegli anni non sempre venivano firmati. Il *corpus* di articoli anonimi pubblicati dai giornali ai quali Gramsci collaborava è studiato dagli storici curatori dell'edizione nazionale allo scopo di valutare caso per caso la *gramscianità* di ogni singolo scritto che esso contiene. L'operazione è più semplice e lineare quando si hanno dati, o almeno indizi precisi, collegati al contenuto di un determinato articolo (per esempio un testimone che afferma: *ricordo che nel gennaio del 1920 Gramsci usò il tal argomento nella polemica con il sindaco di Torino sul Grido del Popolo*, e 2 articoli del gennaio 1920 di polemica contro il sindaco di Torino effettivamente contengono l'argomento ricordato dal testimone); molto meno quando tutto ciò manchi e lo storico sia costretto a ragionare sulla sola base della analisi e valutazione delle idee espresse. Così il presidente della Commissione volle portare un differente punto di vista nella valutazione degli scritti, il punto di vista dell'attribuzione con metodi quantitativi.

Riconoscimento di testi con metodi matematici? Come funziona?

L'idea di base dell'attribuzione con metodi quantitativi adottata dal gruppo di ricerca appena descritto è che i testi siano investigabili con metodi di analisi di tipo matematico che ne analizzano gli aspetti quantificabili, proprio come avviene normalmente per altri fenomeni del mondo fisico e biologico (l'analisi dei suoni, o delle immagini delle TAC, o del DNA, o degli andamenti del mercato borsistico, sono alcuni

esempi). Si tratta di un'idea non nuova nell'ambito degli studi letterari: si possono infatti studiare le tradizioni manoscritte trattandole come trasmissioni di caratteristiche genetiche all'interno di una popolazione, utilizzando quindi i metodi e i software cladistici che a tale scopo impiegano i biologi [1]. Ciò che caratterizza questo tipo di situazione è che i vari testimoni sono tutti sostanzialmente simili e differiscono solo in alcuni punti.

Gli scritti giornalistici anonimi tra i quali si vogliono individuare quelli gramsciani, sono dotati di alcune caratteristiche che li rendono molto ostici in linea di principio a un facile riconoscimento:

- sono per lo più brevi (da qualche decina a qualche centinaio di parole; sono rari quelli che superano il migliaio di parole);
- non si differenziano tra loro per il contenuto;
- non si differenziano tra loro per il lessico.

Testi lunghi sono più adatti a un lavoro di analisi di tipo quantitativo perché contengono e mostrano più caratteristiche misurabili e analizzabili; ma anche testi che si caratterizzano gli uni rispetto agli altri per il contenuto perché ciò si esplicita tra l'altro nella scelta delle combinazioni di parole utilizzate; e anche – ed è ovvio – testi che si differenziano gli uni dagli altri per il lessico. Basti un esempio semplice: se si ha un *corpus* formato dai singoli capitoli dei principali romanzi inglesi dell' 800, e in alcuni di questi – e non in altri – ritornano i nomi David e Copperfield, sarà semplice riconoscere questi capitoli come parte di una medesima opera.

I metodi di analisi utilizzati nella ricerca qui delineata sono di tipo matematico e non statistico. La statistica è una tecnica di analisi utile di fronte a grandi moli di dati quantitativi dei quali viene costruita una visione d'insieme sintetica (un esempio banale: a fronte dell'amplissima varietà delle misure della statura degli italiani, per dare conto delle variazioni nel corso del tempo si ricorre all'altezza media) che inevitabilmente semplifica la varietà iniziale dei dati eliminandone

[1] Si vedano a titolo di esempio B. Ehrman, *The text of the New Testament: its transmission, corruption and restoration*, Oxford University Press, USA, New York, 2005, pp. 207 sgg. e P. Robinson, R. O'Hara, *Cladistic analysis of an Old Norse manuscript tradition*, in "Research in Humanities Computing" 4 (1996), pp. 115-137 (http://rjohara.net/cv/1996-rhc).

o ignorandone una parte (le tecniche di analisi statistica multivariata collocano gli oggetti studiati dentro uno spazio multidimensionale, le cui dimensioni vengono poi ridotte così da conservare solo quelle che spiegano una quantità rilevante di informazione contenuta nei dati analizzati). Non bisogna dimenticare che in statistica tabelle di dati molto sparsi (cioè tabelle che contengono degli zero in metà o più delle caselle) in genere non sono reputate valide per ricavarne delle conclusioni. In questa ricerca invece i dati sono tutti utilizzati per giungere alle conclusioni, anche quando si presentano dispersi, perché i dati sono prodotti a partire dai testi in modo da rappresentare *misure di distanza* fra un testo e un altro.

I testi a un approccio matematico si presentano come una sequenza di simboli [2]. Sono simboli non solo le lettere dell'alfabeto ma anche i numeri, i segni di interpunzione, lo spazio. La sequenza di simboli chiamata testo è prodotta da un'emittente convenzionalmente chiamata autore. L'autore sceglie i simboli secondo regole probabilistiche che gli sono proprie e che differiscono da quelle di un altro autore. L'assunto è che dall'analisi dei simboli dei testi è possibile ricostruire le regole probabilistiche che li hanno originati e quindi è possibile distinguere le emittenti. Poiché l'emittente è un essere umano e non una macchina le sequenze potranno contenere delle variazioni rispetto al criterio generativo di base (per questo si parla di regole probabilistiche).

A partire da questi principi, per lo studio del *corpus* anonimo di scritti giornalistici si stanno utilizzando due metodi di analisi differenti che misurano altrettanti fenomeni testuali: gli *n*-grammi e l'entropia informativa. Vediamo brevemente di che cosa si tratta.

N-GRAMMI

Un *n*-gramma è una sequenza di *n* segni alfanumerici presenti in un testo. Poniamo che si voglia riprodurre un testo di un autore, per ap-

[2] In questa esposizione riprendo elementi provenienti da uno scritto non pubblicato dei colleghi D. Benedetto, E. Caglioti, M. Degli Esposti.

prossimazioni successive; qui prenderemo a riferimento un mini-*corpus* di scritti gramsciani. Si potrebbe iniziare con un'approssimazione *di ordine 0* scegliendo a caso, con pari probabilità, all'interno dell'insieme dei simboli disponibili, uno dopo l'altro, i simboli da utilizzare e si potrebbero ottenere testi di questo tipo:

mZmJMux,1UrsN.u 13HEpf7.hy-!

Se i simboli venissero estratti a uno a uno con probabilità *uguali a quelle che si hanno in corpus di riferimento* (per esempio 100 articoli gramsciani firmati) con un'approssimazione *al primo ordine* si otterrebbero testi di questo tipo:

illfmbaoaocnn e aai,sfrmrta eeoiddmaoo

Ma i simboli potrebbero anche essere estratti tenendo conto del carattere che li precede nel *corpus* di riferimento che si vuole approssimare. Poniamo che il primo carattere estratto sia una *c*; a quel punto si vedrà quali sono le lettere che nel *corpus* di riferimento seguono la *c*: si tratta di *a* nel 9% dei casi, di *e* nel 13% dei casi, di *h* nel 21% dei casi, e così via. Il carattere successivo a *c* viene scelto in modo che nell'insieme del testo prodotto le combinazioni di 2 caratteri (bigrammi, *n*-grammi di lunghezza 2) abbiano le stesse frequenze del *corpus* di riferimento. Con tale approssimazione *del secondo ordine* si otterrebbe un frammento di testo di questo genere:

Loncueresono astantà chedali co le prora Lafra Seoccoro

Un'approssimazione *del terzo ordine* terrà conto della probabilità della presenza di un simbolo in relazione ai due che lo precedono. Se i primi 2 simboli sono *c* e *h* (*ch*) le probabilità che essi siano seguiti da *a*, oppure *o* oppure *u* sono pari a 0; mentre sono pari al 74% per *e* e al 16% per *i*. Ne risulterà un frammento testuale di questo tipo:

La pietra fondamentale nel contegno delle due alleate, quando si è convertito,

Si potrebbe procedere vincolando la scelta di un simbolo alla sequenza dei 3 che lo precedono, e così via [3]. È chiaro che alla fine, con il crescere

[3] Quando la scelta di un carattere è vincolata alle sue probabilità di presenza in un testo di riferimento in relazione ai caratteri che lo precedono, la sequenza di carat-

dell'ordine ci si avvicinerebbe sempre più al testo di partenza (non a caso si era parlato qualche riga sopra di *approssimazioni* a un testo di riferimento). Quel che interessa è che le sequenze di 3 simboli (3-grammi) sono già capaci di riprodurre con un discreto grado di precisione le regole di produzione del testo adottate da una emittente. Tanto più le sequenze di maggiore lunghezza. Quindi si può ipotizzare che scomponendo un testo sconosciuto in sequenze di *n*-grammi, e confrontando le frequenze di tali *n*-grammi con le frequenze degli *n*-grammi di una serie di testi di autori noti, si possa individuare come autore del testo sconosciuto l'autore del testo noto i cui *n*-grammi sono più simili per tipologia e frequenza a quelli del testo sconosciuto.

Come si individuano gli *n*-grammi di un testo? Si può immaginare che sul testo da analizzare venga fatta scorrere una finestra di lunghezza *n* segni che si sposta di 1 segno alla volta (nel riquadro è evidenziata la sequenza alfanumerica che costituisce un 8-gramma):

viene individuato facendo scorrere sul testo da analizzare
viene individuato facendo scorrere sul testo da analizzare
viene individuato facendo scorrere sul testo da analizzare
viene individuato facendo scorrere sul testo da analizzare

Un esempio di 8-gramma può essere

: un n-g

Si tratta di un caso un po' estremo, che rende chiaro il concetto che un *n*-gramma non risponde a nessun criterio tradizionale di segmentazione del testo. Ugualmente sono 8-grammi anche

segmenta
ma rende

cioè *n*-grammi piuttosto lunghi corrispondono a parti sensatamente riconoscibili del testo, in base a criteri di tipo linguistico, grammaticale, sintattico, e così via. Ma un *n*-gramma può anche avere lunghezza 2 o 3 o 4. La lunghezza dell'*n*-gramma viene scelta empiricamente, sperimentalmente, non esistono a oggi criteri astratti che guidino lo studioso. Il criterio empirico è che, ad esempio per gli articoli giornalistici di

---

teri prende il nome di *catena di Markov*.

cui parliamo, *n*-grammi di lunghezza inferiore a 8 non riuscivano a distinguere correttamente i testi di un autore da quelli di un altro.

ENTROPIA INFORMATIVA RELATIVA

Il concetto di entropia nell'ambito dell'informazione venne introdotto da un fondamentale articolo di C. E. Shannon del 1948 [4]. In tale scritto Shannon si propone di definire la quantità di informazione contenuta in un messaggio definendola come il minimo numero di bit necessari per veicolare il messaggio. Se il messaggio fosse

TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

esso avrebbe un contenuto informativo pari a 320 bit se scritto in un alfabeto occidentale (in tal caso sono infatti necessari 8 bit per carattere per poter rappresentare tutti i caratteri), ma in un linguaggio di programmazione esso potrebbe essere ridotto a un'istruzione corrispondente a qualcosa come *scrivi 40 T*. L'entropia è il numero di bit per carattere necessari per codificare il messaggio. Se il messaggio è una sequenza di DNA, per la quale si utilizzano 4 caratteri (C G A T) l'entropia è di 2 bit (2 bit bastano a codificare 4 caratteri).

L'esperienza quotidiana di ricerca della minima sequenza di bit necessaria per codificare un messaggio è quella dell'uso dei programmi di compressione come *winzip*, o *winrar*, o *stuffit*. Il rapporto di compressione tra file originale e file compresso dà una stima abbastanza precisa dell'entropia del testo: maggiore la compressione, minore l'entropia [5]. Per ottenere questo risultato il programma di compressione costruisce un dizionario del testo da comprimere, in cui – semplificando un po' – le sequenze di caratteri più frequenti sono sostituite con altre più brevi. La misurazione dell'entropia può essere utilizzata per confrontare 2 testi differenti, ottenendo così una misurazione dell'entropia relativa,

[4] C. E. Shannon, *A Mathematical Theory of Communication*, in "The Bell System Technical Journal" 27 (1948), pp. 379-423, 623-656.

[5] *WinRAR* riduce a 100 byte un documento di testo contente 800 occorrenze della medesima lettera; mentre riduce a 540 byte un testo reale lungo 800 caratteri.

come hanno scritto Benedetto, Caglioti e Loreto [6]. Poniamo di comprimere un canto della Divina Commedia e un articolo di giornale. Poi comprimiamo l'articolo una seconda volta utilizzando il dizionario che il programma aveva creato per comprimere il canto della Divina Commedia. Infine misuriamo la differenza di compressione dell'articolo nei due casi: otteniamo così una misura della differenza (distanza) tra il canto della Divina Commedia e l'articolo di giornale [7] basata sull'entropia informativa.

L'APPLICAZIONE AGLI SCRITTI GRAMSCIANI ANONIMI

I metodi di attribuzione delineati nelle righe precedenti per quanto abbiano pochi anni di vita sono stati valutati e verificati attraverso il consueto percorso delle pubblicazioni scientifiche. Non sono però mai stati in passato utilizzati in modo continuativo nel contesto di attività di ricerca reali, cioè attività di ricerca che utilizzano degli esiti dell'attribuzione. Ci si è perciò posti il problema di verificare l'affidabilità reale di questi metodi non accontentandosi della loro validità teorica. A tale scopo i metodi hanno passato due fasi di verifica che sono state chiamate *test aperto* e *test cieco*.

Nella fase di test aperto il gruppo di ricerca ha lavorato su 50 articoli giornalistici sicuramente gramsciani e su 50 articoli giornalistici di autori (Bianchi, Bordiga, Carena, Tasca, Togliatti e altri) che collaboravano ai medesimi giornali nei medesimi anni [8]. Il primo era il gruppo dei testi da attribuire, il secondo era il gruppo dei testi di con-

[6] D. Benedetto, E. Caglioti, V. Loreto, *Language Trees and Zipping*, in "Physical Review Letters" 88/4 (2002).

[7] Il caso estremo dall'altra parte sarebbe di ricomprimere il medesimo file una seconda volta: la differenza di compressione tra i due file sarebbe nulla, perché i due file sono identici. Se anziché con un articolo di giornale si lavorasse con testo poetico in volgare di area toscana la differenza di compressione sarebbe minore perché i due testi sono più simili.

[8] I testi furono forniti (e sono tutt'ora forniti nel corso della ricerca) in formato digitale dalla Fondazione Gramsci, con uno scrupoloso lavoro di archivio su originali e microfilm.

trollo. Ogni testo del gruppo da attribuire era confrontato con ogni testo del gruppo di autori certi (il gruppo di controllo) ottenendo per ogni coppia di testi una misura della distanza dai testi gramsciani e dai testi non gramsciani. L'attività consisteva nel regolare i metodi di riconoscimento delle somiglianze e differenze in modo da massimizzare le attribuzioni corrette a Gramsci e ridurre al minimo (e possibilmente evitare completamente) le attribuzioni erronee a Gramsci. La questione è complessa. Il rischio naturalmente era che in tal modo si perdesse un numero rilevante di scritti effettivamente gramsciani [9] e dunque la fase di messa a punto era estremamente importante. Ma era necessario ridurre al minimo, o anzi eliminare del tutto, se possibile, l'eventualità che uno o più scritti non gramsciani fossero attribuiti a Gramsci, perché ciò avrebbe minato l'affidabilità dell'intero lavoro agli occhi degli storici curatori dell'edizione nazionale.

I metodi messi a punto durante questa fase preliminare costituivano un'evoluzione a opera di Degli Esposti del metodo degli *n*-grammi (con cui V. Keselj aveva vinto nel 2006 la gara internazionale di attribuzione di testi *aaac*) e del metodo dell'entropia relativa, a opera di Benedetto e Caglioti. I due metodi confluirono nella strategia complessiva di attribuzione, che li utilizzò entrambi con lo scopo di diminuire il numero di falsi positivi: in sostanza si decise di attribuire a Gramsci i soli testi che entrambi i metodi attribuivano a Gramsci. Furono dunque attribuiti a Gramsci i soli testi che *tutti i metodi attribuivano a Gramsci*, cioè 43 su 50. In Figura 1 l'ascissa (orizzontale) di ogni punto rappresenta l'indice di gramscianità fornito dal metodo degli n-grammi, in ordinata (verticale) c'è il valore dell'analogo indice fornito dal metodo dell'entropia relativa; nel quadrante in altro a destra ci sono dunque i testi che entrambi i metodi attribuiscono a Gramsci (rappresentati dai punti grigi). I punti grigi negli altri quadranti rappresentano le attribuzioni non riuscite. Nel quadrante in alto a destra non ci sono testi non gramsciani (punti neri), cioè il sistema non genera dei falsi positivi (testi attribuiti a Gramsci ma che non sono gramsciani).

[9] È la nota questione dell'impossibilità di massimizzare sia il *richiamo* sia la *precisione* quando si effettuano ricerche complesse all'interno di raccolte di testi.

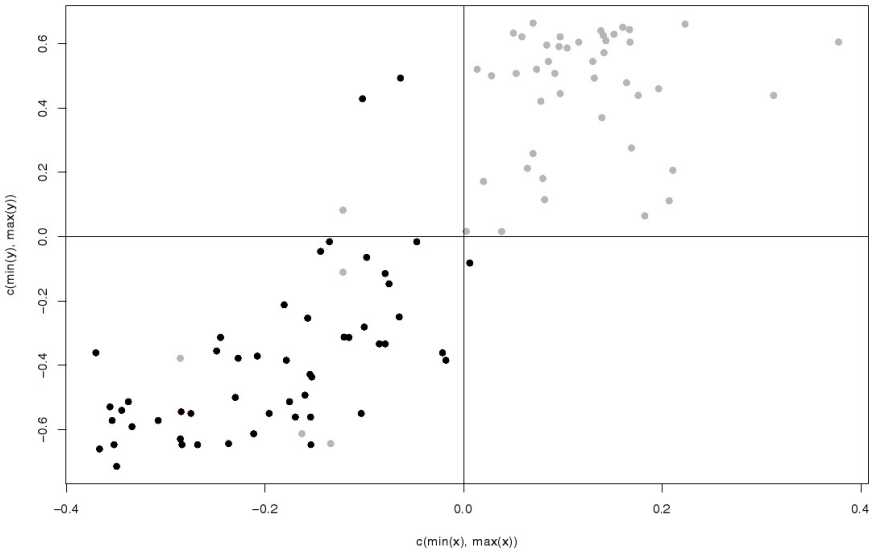

Figura 1. Attribuzioni degli scritti anonimi al termine della fase di test in chiaro

A questo punto si poteva affrontare il test cieco, che fu effettuato su 40 testi dell'elenco, consegnati anonimi al gruppo di ricerca alla fine di giugno 2006.

1. Gramsci, *La rievocazione di Gelindo*, «Il Grido del Popolo», 25 dicembre 1915.
2. Leo Galetto, *In tema di guerra*, «Il Grido del Popolo», 8 novembre 1915.
3. Gramsci, *Maurizio Barrès e il nazionalismo sensuale*, «Il Grido del Popolo», 2 marzo 1918.
4. Gramsci, *Disciplina*, «La Città futura», 11 febbraio 1917.
5. B.B. [Bruno Buozzi], *La Conferenza del lavoro e il Convegno di Zimmerwald*, «Il Grido del Popolo», 7 gennaio 1916.
6. Gramsci, *Il socialismo e l'Italia*, «Il Grido del Popolo», 22 settembre 1917.
7. Gramsci, *Stenterello*, «Avanti!», 10 marzo 1917.
8. G.B. [Giuseppe Bianchi], *Una volta per sempre*, «Il Grido del Popolo», 15 gennaio 1916 [19 15:240].
9. Gramsci, *Il Cottolengo e i clericali*, «Avanti!», 30 aprile 1917.
10. A.T. [Angelo Tasca], *Sempre più chiaramente*, «Il Grido del Popolo», 7 novembre 1914.

11. O.P. [Ottavio Pastore], *Il Papa al congresso della pace*, «Il Grido del Popolo», 15 aprile 1916.
12. Gramsci, *Una verità che sembra un paradosso*, «Avanti!», 3 aprile 1917.
13. G.M.S. [Giacinto Menotti Serrati], *Il più gran terremoto*, «Il Grido del Popolo»,12 agosto 1916.
14. Gramsci, *Con mani di vetro ...*, «Il Grido del Popolo», 13 aprile 1918.
15. Alfonso Leonetti, *Evoluzione e rivoluzione*, «Il Grido del Popolo», 3 agosto 1918.
16. Gramsci, *La lingua unica e l'esperanto*, «Il Grido del Popolo», 16 febbraio 1918.
17. Decio Pettoello, *La dottrina di Norman Angell*, «Il Grido del Popolo», 10 agosto 1918.
18. Gramsci, *Repubblica e proletariato in Francia*, «Il Grido del Popolo», 20 aprile 1918.
19. Zino Zini, *Marx nel pensiero di un cattolico*, «Il Grido del Popolo», 31 agosto 1918.
20. Gramsci, *Due inviti alla meditazione*, «La Città futura», 11 febbraio 1917.
21. A.V. [Andrea Viglongo], *La Costituzione parlamentare inglese*, «Il Grido del Popolo», 5 ottobre 1918.
22. Pietro Gavosto, *Le opinioni dei compagni. Guerra, patria e proletariato*, «Il Grido del Popolo», 9 gennaio 1915.
23. A.T. [Angelo Tasca], *Noterelle di guerra*, «Il Grido del Popolo», 16 gennaio 1915.
24. Gramsci, *Il privilegio dell'ignoranza*, «Il Grido del Popolo», 13 ottobre 1917.
25. Gramsci, *I monaci di Pascal*, «Avanti!», 26 febbraio 1917.
26. Gino [Gino Castagno], *Cinismo*, «Il Grido del Popolo», 20 febbraio 1915.
27. Gramsci, *Disciplina e libertà*, «La Città futura», 11 febbraio 1917.
28. Leo Galetto, *Il proletariato deve servire da «materia anatomica»*, «Il Grido del Popolo», 20 marzo 1915.
29. Gramsci, *Modello e realtà*, «La Città futura», 11 febbraio 1917.
30. Cincali, *Luci ed ombre*, «Il Grido del Popolo», 23 ottobre 1915.
31. Corso Bovio, *ll problema del Mezzogiorno*, «Avanti!», 27luglio 1917.
32. Gramsci, *La Giustizia*, «Il Grido del Popolo», 13 ottobre 1917.
33. Omero Concetto, *Diagnosi interessata*, «Avanti!», 10 agosto 1917.
34. Gramsci, *Letteratura italica: La prosa*, «Avanti!», 17 aprile 1917.
35. Egidio Gennari, *Nazionalisti od internazionalisti?*, «Avanti!», 27 agosto 1917.
36. Gramsci, *Rispondiamo a Crispolti*, «Avanti!», 19 giugno 1917.
37. Francesco Ciccotti, *Il reazionario democratico*, «Avanti!», 2 settembre 1917.
38. O.B., *Problemi presenti e futuri*, «Avanti!», 12 settembre 1917.
39. Gramsci, *Spezzatino d'asino e contorno*, «Il Grido del Popolo», 29 aprile 1917.
40. Gramsci, *Analogie e metafore*, «Il Grido del Popolo», 15 settembre 1917.

I risultati del test cieco, presentati a Roma alla Fondazione Istituto Gramsci nel luglio 2006, sono mostrati in Figura 2 costruita con lo stesso procedimento usato per l'analoga figura per i 100 articoli del test aperto: in ascissa l'indice di *gramscianità* fornito dal metodo degli *n*-

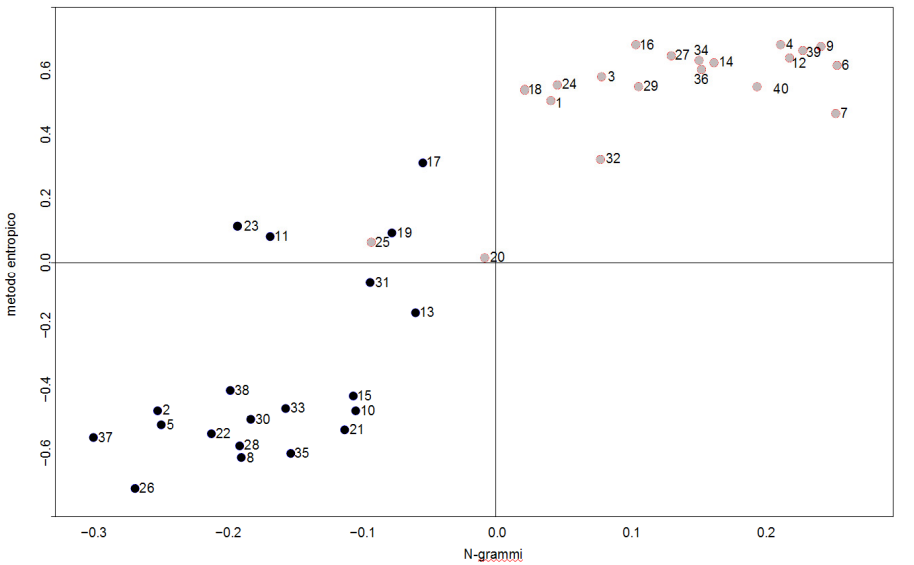

Figura 2. L'esito del test cieco

Come si può osservare vennero correttamente individuati e attribuiti 18 testi gramsciani su 20 (punti grigi), pari al 90%, senza falsi positivi. I due testi gramsciani non riconosciuti sono il n. 20 (Gramsci, *Due inviti alla meditazione*, «La Città futura», 11 febbraio 1917: un testo breve, di poche righe, quindi piuttosto difficile da attribuire) e il n. 25 (Gramsci, *I monaci di Pascal*, «Avanti!», 26 febbraio 1917, testo che non presenta caratteristiche che aiutino a capire perché l'attribuzione non è risultata corretta). Non sembrava possibile ottenere un risultato così chiaro e significativo perché era chiara la complessità del dominio in cui si operava e il permanere di alcune questioni aperte di tipo scientifico e metodologico che imponevano e impongono cautela.

Soprattutto se la classificazione dei testi di un *corpus* per attribuirli agli autori è spesso indistinguibile in realtà da una classificazione di

contenuti [1], l'attribuzione dei testi giornalistici anonimi di Gramsci si trova a operare su testi indistinguibili dal punto di vista del contenuto, cioè degli argomenti trattati e quindi dei termini utilizzati; e gli autori a cui si possono attribuire i testi hanno una medesima (o molto simile) visione e descrizione del mondo. Di qui la difficoltà per gli storici di attribuire i testi sulla base di criteri di tipo semantico, ideologico e simili. Con questi testi quindi non v'è speranza che gli autori possano essere distinti e classificati sfruttando *marche lessicali* caratteristiche dell'uno o dell'altro. Infine i testi dei singoli autori sono molto brevi, quando invece è noto che le grandi dimensioni del testo sono condizione importante per concludere con successo un lavoro di attribuzione con metodi quantitativi [2].

Il risultato finale convinse la Fondazione ad affidare al gruppo di ricerca il lavoro di attribuzione, che è tutt'ora in corso, lavoro di attribuzione vero, senza rete, per così dire, cioè l'analisi ed eventualmente l'attribuzione di testi giunti anonimi. Le attribuzioni (e le non attribuzioni) effettuate su gruppi di articoli omogenei per anno di pubblicazione vengono analizzate e valutate, e viene scritto un report inviato allo storico curatore del volume a cui quegli scritti eventualmente appartengono, storico che liberamente deciderà in base alle sue valutazioni critiche se le attribuzioni e non attribuzioni proposte siano da tenere in conto o siano da rigettare.

E se già una serie di cautele e di interrogativi si era posta mentre i metodi di analisi venivano valutati, messi a punto, applicati, un'altra ancora deve essere presa in considerazione. Il primo pensiero riguarda l'importanza del gruppo di testi di riferimento. I testi di riferimento devono essere della stessa epoca dei testi da riconoscere perché si deve

---

[1] Clement e Sharp scrivono in termini chiarissimi: *It was conjectured that there may well be a correlation between the authors of the books, and the subjects of those books. i.e. that accuracy of authorship attribution for these data sets might not be due to the detection of the* authorship *signal in the text, but the detection of a correlated* topic *signal* (R. Clement and D. Sharp, *Ngram and Bayesian Classification of Documents for Topic and Authorship*, in "Literary and Linguistic Computing", 18/4 (2003), p. 433.

[2] Sempre Clement e Sharp scrivono, commentando la buona percentuale di attribuzioni ottenuta in una loro ricerca: *This level of accuracy is high, but this is due to having large amounts of material (whole books) for each author both for training and classification (authorship attribution)*. (*Ibidem*).

supporre, in assenza di acquisizioni definitive su questo argomento, che il contenuto e il modo di scrivere di un autore evolva nel corso del tempo e che quindi testi da attribuire e testi di riferimento debbano essere coevi. Che cosa significhi *coevi* non si può stabilire a priori ma tocca agli studiosi dell'autore dirlo: nel caso di Gramsci, secondo gli storici curatori dell'edizione nazionale, una prima fase compiuta del pensiero e della scrittura è quella che va dagli inizi dell'attività giornalistica fino al 1918.

Un secondo pensiero riguarda il fatto che gli n-grammi risultati più efficaci nel discriminare correttamente i testi nella fase iniziale della produzione gramsciana sono di lunghezza 8, e quindi le sequenze di simboli così rilevate sono piuttosto specifiche e poco frequenti, e non sono molto numerose quelle comuni ai testi analizzati; eppure questi dati che con termine statistico si chiamerebbero *sparsi* sono risultati efficaci per discriminare e riconoscere i testi gramsciani nelle due fasi di test e nelle fasi successive di attribuzione *reale* [3].

Un terzo pensiero riguarda il fatto che l'analisi non considera oggetti di studio tradizionali e compiuti nell'ambito delle scienze che si occupano dei testi (critica letteraria, storia della letteratura, linguistica: parole, prefissi, suffissi, lemmi, forme, sintagmi, desinenze, fonemi, …) bensì oggetti nuovi quali gli *n*-grammi e l'entropia che sono ben noti e studiati ma in altri ambiti scientifici. Ciò fa sì che in generale gli interlocutori di area umanistica rimangano spiazzati e non siano in grado di argomentare pro o contro aspetti e caratteristiche di questi metodi. Un effetto collaterale è che lo studioso di formazione umanistica si senta in un certo modo espulso a opera di una *macchina* da un territorio che era abituato a considerare suo (suo esclusivo).

Si tratta peraltro di un processo che si verifica o si è verificato in molti campi: gli architetti delle cattedrali non utilizzavano *Autocad* e tutto ciò che un architetto oggi demanda al programma di progettazione era in altri tempi un sapere specialistico dell'architetto; analogamen-

[3] È significativo che gli storici curatori delle annate dell'edizione nazionale non hanno espresso fino a ora sfiducia nelle attribuzioni (nelle proposte di attribuzione) derivanti dall'analisi quantitativa, segno che tali attribuzioni, anche nel caso non vengano accolte dallo storico, risultano comunque credibili e non danno luogo a clamorosi risultati palesemente insostenibili.

te è accaduto in medicina ad opera degli esami clinici di laboratorio. Tutto ciò si colloca a mio avviso nel quadro di quello che si potrebbe chiamare *il paradigma Licklider-Engelbart* che vede negli strumenti informatici un mezzo di potenziamento dell'intelletto umano, liberato dai *clerical tasks* e quindi in grado di dedicarsi di più e meglio ai compiti che gli sono propri e che il computer non può svolgere:

> *intendiamo accrescere la capacità di un uomo di affrontare una situazione che presenta un problema complesso, di acquisire comprensione adatta alle sue specifiche necessità, e derivare soluzioni per i problemi. Capacità accresciuta, in questo senso, è usato per indicare una miscela dei seguenti elementi: comprensione più rapida, comprensione migliore, possibilità di acquisire un livello di comprensione utile in una situazione che prima era troppo complessa, soluzioni più rapide, soluzioni migliori, e la possibilità di trovare soluzioni per problemi che prima sembravano irrisolvibili. E con situazioni complesse … non parliamo di trucchi intelligenti e isolati che aiutano in una particolare situazione. Ci riferiamo invece ad un modo di vivere in un dominio integrato dove intuizioni, tentativi, beni immateriali, e la qualità specificamente umana della percezione della situazione coesistono in modo utile con concetti potenti, terminologia e notazione efficienti, metodi sofisticati e aiuti elettronici di alto livello.* [4]

Ma su questa linea di pensiero (di grande importanza perché propone un modo di interpretare, e di dare risposta agli interrogativi che sorgono laddove i computer iniziano a svolgere compiti che prima erano svolti dalle persone), non è il caso di inoltrarsi per non allontanarsi troppo dal tema del riconoscimento degli articoli gramsciani anonimi.

[4] D. Engelbart, *Augmenting Human Intellect: A Conceptual Framework, Summary Report*, Stanford Research Institute, on Contract AF 49(638)-1024, October 1962, accesso online full text all'URL http://www.invisiblerevolution.net/engelbart/full_62_paper_augm_hum_int.html.

QUALCUNO DIRÀ: 'MA QUESTA È STILOMETRIA! '

Chi è arrivato fino a questo punto nella lettura avrà notato, forse, che non è stata mai utilizzata la parola *stile*. Stile è l'insieme delle caratteristiche distintive della scrittura di un autore. Affermazione, concetto, che intuitivamente si riconosce per vero ma che nella pratica della ricerca rimane sfuggente. In questo articolo si è preferito parlare di regole probabilistiche per la produzione di testi da parte di un emittente. Un'espressione molto meno affascinante, ma riconducibile, come si è visto, a elementi verificabili e misurabili, senza per questo ritenere che sia risolto (per altra via, o per … eliminazione diretta) il problema dello stile. Perché rimane aperto l'interrogativo di fondo, se lo stile rimanga uguale a se stesso nel corso del tempo o se invece esso evolva; e parallelamente se lo stile sia un prodotto non cosciente da parte dell'autore o se invece l'autore sia in grado di governare coscientemente e finalisticamente le caratteristiche profonde del suo stile. E quali siano gli elementi di superficie in cui si manifestano queste scelte non consce dell'autore costituisce ancora un altro aspetto del problema, cioè: che cosa si esamina per individuare nello scritto del testo l'espressione delle scelte stilistiche?

Da una parte l'idea che lo stile sia una sorta di carattere originario della scrittura di un autore ha generato l'espressione *stylistic fingerprint*: l'impronta digitale rimane identica dalla nascita ed è assolutamente unica. Da un'altra abbiamo un termine come *stilometria* (usato per la prima volta da W. Lutosławski nel 1897 [5]) che implica l'idea che lo stile si misuri, contandone le caratteristiche di superficie dalle quali traspare il carattere nascosto, sepolto, inconscio. L'approccio della ricerca sugli scritti gramsciani anonimi si colloca sostanzialmente nella prospettiva stilometrica, anche se in modo deviante, in quanto le caratteristiche di stile misurate e contate sono state, nel corso del tempo, le costruzioni sintattiche [6], le parole più frequenti, le parole grammaticali (anche det-

[5] W. Lutosławski, *On Stylometry. Abstract of a paper read at the Oxford Philological Society on May 21st by Dr. W. Lutosławski, of Drozdowo, near Lomza, Poland*, in "Classical Review" 11 (1897), pp. 284-286; W. Lutosławski, *Principes de stylométrie*, in "Revue des études grecques" 41 (1898), pp. 61-81.

[6] È l'approccio che adottò W. Lutosławski nel suo studio della cronologia di

te parole vuote), parole contenenti una specifica lettera [7], ma nessuno di questi oggetti, come si è visto, è oggetto dei conteggi operati per l'attribuzione degli scritti gramsciani.

Fece scuola, come la scoperta di un filone d'oro che tutti sperano di poter anche loro trovare in altri territori ricorrendo ai medesimi strumenti di lavoro, lo studio di attribuzione sui *Federalist Papers*, per 12 dei quali, anonimi, l'attribuzione è incerta tra i due principali autori della raccolta, Hamilton e Madison. Mosteller e Wallace [8] individuarono che nei 14 papers di Madison *while* non ricorre mai, mentre ci sono 8 occorrenze di *whilst*; che in 14 dei 48 papers di Hamilton ricorre *while*, mentre *whilst* è assente; e quindi la presenza di *whilst* in 5 dei *papers* disputati portò verso l'attribuzione a Madison. Poiché però le parole capaci di discriminare in modo così forte erano meno di quante speravano, ricorsero anche allo studio di parole che caratterizzassero i due candidati avendo occorrenze significativamente differenti: *by*, per esempio, era raro in Hamilton e frequente in Madison; con *to* le posizioni si scambiavano. Dopo aver esaminato molte parole, Mosteller e Wallace si fermarono su 28 che avevano la più alta efficacia discriminante: *also*, *upon*, *by*, *of*, *on*, *there*, *this*, *to* e altre, tra le parole grammaticali; e *innovation*, *language*, *probability* e altre tra le parole piene. Confrontando l'uso di queste parole Mosteller e Wallace conclusero che i 12 papers dubbi erano da attribuire a Madison.

Platone (*The Origin and Growth of Plato's Logic*, London - New York - Bombay 1897, repr. Georg Olms, Hildesheim 1983). Egli conteggiò fra l'altro (come risulta da A. Kenny, *The Computation of Style*, Pergamon Press, Oxford 1982, p. 3):

> *Answers denoting subjective assent less than once in 60 answers.*
> *Superlatives in affirmative answers more than half as frequent as positives, but not prevailing over positives.*
> *Interrogations by means of* ara *between 15 and 24% of all interrogations.*
> Peri *placed after the word to which it belongs forming more than 20% of all occurrences of* peri.

[7] *1. words containing a specified letter; 2. words ending in a specified letter; 3. words with a specified letter as penultimate*, come scrive G. R. Ledger, *Re-counting Plato: A Computer Analysis of Plato's Style*, Clarendon Press, Oxford 1989, p. 6.

[8] F. Mosteller, D. L. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading MA 1964. Il loro lavoro è per lo più citato con il titolo della seconda edizione: *Applied Bayesian and Classical inference: The Case of the Federalist Papers*, Springer-Verlag, New York 1984.

Analogamente, alcuni anni prima Alvar Ellegård aveva studiato le *Lettere di Junius* [9] (una serie di 69 lettere anonime pubblicate in Inghilterra e scritte fra il 1769 e il 1772). Dalla metà dell'800 le ipotesi sull'autore si concentrarono su sir Philip Francis. Ellegård descrisse e studiò circa 500 parole ed espressioni che nelle lettere di Junius sono o molto più frequenti o molto meno frequenti che negli scritti dei contemporanei; a esse si aggiungevano circa 50 termini che Junius sceglieva all'interno di coppie o terne di termini che risultavano approssimativamente sinonimi, come *on* e *upon*; *kind* e *sort*; e così via. Le scelte espressive di Junius risultarono in grande sintonia con quelle di sir Philip Francis e poiché a ciò si aggiungeva il fatto che i suoi dati biografici erano compatibili con l'ipotesi dell'attribuzione su base puramente testuale, Ellegård concludeva che sir Philip Francis era Junius.

Ma si possono citare altri esempi di questa prospettiva basata sulla ricerca del giusto mix di indicatori derivati dalle tipologie di parole presenti nel testo: lo studio di Keim e Oelke e sui romanzi di J. London e M. Twain [10], o quello di A. M. García e J. C. Martín sulle traduzioni in *old english* dei vangeli sinottici [11], per menzionarne solo alcuni.

---

[9] A. Ellegård, *Who was Junius?*, Almquist & Wiksell, Stockholm 1962; Id., *A statistical method for determining authorship: The Junius Letters 1769-1772*, Acta Universitatis Gothenburgensis / Elanders Boktryckeri Aktiebolag, Göteborg, 1962.

[10] D. A. Keim, D. Oelke, *Literature Fingerprinting: A New Method for Visual Literary Analysis*, Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST 2007), Sacramento, October 30 - November 1, 2007, pp. 115-122. In riferimento a quanto scritto in apertura dell'articolo a proposito dell'impiego di strumenti matematici nello studio di fenomeni del mondo fisico e sociale, è interessante notare a puro titolo di esempio che Keim ha pubblicato fra l'altro anche articoli che studiano i movimenti finanziari (H. Ziegler, T. Nietzschmann, D. Keim, *Visual exploration and discovery of atypical behavior in financial time series data using two-dimensional colormaps*, Proceedings 11th International Conference on Information Visualization (IV 07), IV-VDM Symposium on Visualization and Data Mining, Zurich, Switzerland, July 2007, pp. 308-315) e il genoma virale (K. Neuhaus, D. Oelke, D. Fuerst, S. Scherer, D. Keim, *Towards automatic detecting of overlapping genes - clustered BLAST analysis of viral genomes*, Proceedings of 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBIO 2010), 2010, Istanbul, Turkey, in corso di pubblicazione), non essendo un economista né un biologo.

[11] A. M. García, J. C. Martín, *Function Words in Authorship Attribution Studies*, in

La questione in gioco, come si è detto, è quella che traspare attraverso i concetti di (*stylistic*) *fingerprint* e di (*stilo*)*metria*: il primo orientato verso una concezione deterministica dello stile, l'altro aperto a un approccio probabilistico, perché misurare è certo un'azione precisa e determinata, ma che cosa misurare e come, e con quali approssimazioni, è invece questione che deve ogni volta essere affrontata, valutata, risolta.

Breve viaggio nel tempo

Gli studi stilometrici sulla cronologia delle opere di Platone meriterebbero da soli una ricostruzione storica dell'evoluzione dei metodi adottati e degli oggetti individuati e misurati, per il loro significato fondatore nell'area dell'attribuzione di testi e perché si individuerebbero e mostrerebbero linee di tendenza e orientamenti. Ma si può comunque, anche tralasciando tale ambito, individuare una tendenza dominante negli studi di attribuzione basati su metodi quantitativi, cioè l'evoluzione progressiva dell'individuazione degli oggetti da contare:

- da oggetti complessi, *sintattici* (le proposizioni che contava Lutosławski);
- a oggetti di tipo lessicale (le parole): nel 1880 Filippo Mariotti pubblicò uno studio sul lessico dantesco [12]; nel 1901 Th. C. Mendenhall pubblicò un articolo in cui venivano studiati e confrontati gli scritti di Shakespeare, Marlowe e Bacone [13]; Mosteller e Wallace e i Federalist Papers; Ellegård e le lettere di Junius [14];
- a oggetti parzialmente nuovi che costituiscono evoluzione degli oggetti tradizionali: le parole che contengono una data lettera in una data posizione, studiate da Ledger;
- fino a oggetti nuovi (almeno per l'ambito linguistico-letterario) quali per esempio gli n-grammi e l'entropia informativa.

"Literary and Linguistic Computing" 22/1 (2007), pp. 45-69.

[12] F. Mariotti, *Dante e la statistica delle lingue*, Barbera, Firenze 1880.

[13] T. C. Mendenhall, *A mechanical solution of a literary problem*, in "The Popular Science Monthly" 60/7 (1901), pp. 97-105.

[14] Per citare solo alcuni casi famosi.

Gli oggetti *nuovi*, quali sono quelli utilizzati nell'individuazione di testi gramsciani pervenuti anonimi, implicano l'ipotesi che i testi abbiano una struttura matematica latente, o forse ancora meglio che abbiano una struttura che può essere letta con strumenti matematici, struttura che potremmo chiamare latente perché sembra non apparire alla superficie del testo (ma non più latente della struttura grammaticale o sintattica di un testo scritto in una lingua ignota al lettore). Davanti a una serie di problemi nello studio di un testo, per esempio l'attribuzione, si può ricorrere ai consueti (per le scienze umane) strumenti di tipo semantico, linguistico, storico: si cerca di ampliare e approfondire le conoscenze sul contenuto del testo, sulla lingua in cui è scritto, sulla storia della sua composizione e trasmissione. Può succedere però che non si facciano progressi significativi e si può allora decidere di ricorrere allo studio di *altre* caratteristiche del testo perché ci si rende conto che il modello di studio del testo basato sul *contenuto* non risponde alle domande che lo studioso sta ponendo; e perché si ipotizza, con un atteggiamento esplorativo, che per cercare le risposte si debba costruire un differente modello (fa parte dell'atteggiamento esplorativo la consapevolezza che si potrebbe non trovare risposta) in cui contenuto, lingua, storia del testo *vengono messi in secondo piano*. Per far questo il primo nodo concettuale è quello del passaggio dagli oggetti ai numeri, dall'individuazione degli oggetti al loro conteggio, dell'assegnare numeri (reali) agli oggetti; o in termini più formalizzati, di trasformare un sistema di relazioni qualitative (il testo) in un sistema di relazioni quantitative (l'insieme dei dati che contiene le informazioni sugli oggetti dell'analisi) grazie a una o più operazioni di *classificazione* del testo (Doležel [15]). La scelta degli oggetti da studiare è, per ogni testo, veramente molto ampia: si potrebbe pensare, nella linea di Ledger, alle parole che iniziano con *a* e che hanno una *z* come quarta lettera, per contare e vedere come si distribuiscono nel testo da studiare e in altri vicini; oppure si può pensare alle ripetizioni di sillabe; e altro ancora, solo per indicare che le scelte possibili, se non si è più vincolati alle

[15] L. Doležel, *A note on quantification in text theory*, in S. Allén (a cura), *Text Processing. Text Analysis and Generation. Text Typology and Attribution*, Proceedings of Nobel Symposium 51, Almqvist & Wiksell International, Stockholm 1982, pp. 539-552; la parte a cui si fa qui riferimento sono le pp. 540-42.

parole, sono amplissime, quasi sconfinate.

Da un sistema qualitativo si passa dunque a un sistema quantitativo. Se i due sistemi corrispondono perfettamente (se sono isomorfi) la situazione risulta per certi versi insoddisfacente perché significa che il passaggio dal sistema qualitativo a quello quantitativo non ha messo in luce nessuna differenza, la costruzione di un nuovo sistema non ha fatto apparire nulla di nuovo. Ma se delle discrepanze tra i due sistemi appaiono, se si percepiscono differenze allora si è su una buona strada (l'informazione è una differenza che crea una differenza, la percezione della differenza permette di generare e acquisire informazione), perché diventa necessaria una riorganizzazione delle conoscenze allo scopo di capire come si rapportano i dati dei due sistemi e che cosa significano le differenze tra l'uno e l'altro. Alla base di questa riorganizzazione della conoscenza, che è nuova conoscenza, sta il riconoscimento che i dati quantitativi fungono da indicatori della presenza nel testo di proprietà qualitative che non appaiono in evidenza o che non sono state rilevate:

> *Sometimes when a structure is not objectively definable, there exists a series of objectively definable indicators which when taken collectively are almost co-extensive with the given structure.* [16]

L'interpretazione è quindi la ricerca dei fattori qualitativi, formali, da cui dipendono i valori dell'indicatore quantitativo; e la situazione si fa difficile quando si ha motivo di ritenere che uno specifico indicatore quantitativo sia controllato da molteplici fattori qualitativi (la lunghezza delle frasi di un testo dipende da molteplici fattori: idiosincrasie dell'autore, pubblico di riferimento, argomento, testi e stili di riferimento, e altro ancora), che non sono necessariamente sempre i medesimi in testi differenti.

Può essere interessante segnalare, senza pretesa di esaustività, alcune ricerche recenti che costituiscono espressioni significative della tendenza al riconoscimento dell'esistenza di strutture matematiche nei testi (o di strutture che si possono individuare e analizzare con stru-

[16] B. Brainerd, *Weighing Evidence in Language and Literature. A Statistical Approach*, University of Toronto Press, Toronto-Buffalo 1974, p. 218.

menti matematici) [17].

Nel 2001 Dmitri Khmelev (che aveva già pubblicato nel 2000 un articolo di argomento simile [18]) e Fiona Tweedie pubblicarono un articolo intitolato *Using Markov Chains for Identification of Writers* [19] in cui mostrarono i risultati che si potevano ottenere con una tecnica di attribuzione di testi con metodi quantitativi basata su catene di Markov (qualcosa di abbastanza simile a una valutazione dei possibili n-grammi). Nell'attribuzione di 387 testi differenti estratti dalla raccolta del Progetto Gutenberg, tutti di autore noto, effettuarono 288 attribuzioni corrette, pari al 74,4%; nell'attribuzione dei *Federalist Papers*, assumendo come riferimento i risultati di Mosteller e Wallace di cui si è parlato sopra, effettuarono tutte le attribuzioni correttamente; nell'attribuzione dei 20 testi degli scrittori inglesi Margery Allingham e Michael Innes su cui avevano lavorato Baayen *et al*. in un importante lavoro del 1996 [20] ottennero un risultato di 19 attribuzioni corrette su 20.

Khmelev e Tweedie conclusero osservando che

> *The data used for the Markov chain can perhaps be described as linguistically microscopic-the unit is too small for meaningful conclusions to be reached regarding characteristics of the texts by the individual authors. Comparison of transition matrices may allow the re-*

[17] Non si vuole con ciò dimenticare che sono oggi numerose le ricerche che utilizzano tecniche statistiche multivariate (l'analisi dei componenti principali, per esempio) per estrarre e mostrare l'informazione contenuta nelle matrici di dati. Ma il versante delle ricerche che riconoscono l'esistenza di strutture matematiche nei testi pare più interessante e lo studio di attribuzione di testi con metodi quantitativi effettuato per individuare gli scritti gramsciani giornalistici giunti anonimi si colloca in questa linea.

[18] D. V. Khmelev, *Disputed authorship resolution through using relative entropy for Markov chains of letters in human language texts*, in "Journal of Quantitative Linguistics" 7 (2000), pp. 115-26, accessibile all'URL: http://www.philol.msu.ru/~lex/ khmelev/published/jql/khmelev.html.

[19] D. Khmelev, F. Tweedie, *Using Markov chains for identification of writers*, in "Literary and Linguistic Computing", 16/4 (2001), pp. 299-307.

[20] R. H. Baayen, H. van Halteren, F. J. Tweedie, *Outside the cave of shadows. Using syntactic annotation to enhance authorship attribution*, in "Literary and Linguistic Computing", 11 (1996), pp. 121-31.

> *searcher to comment that Hamilton uses 'p' followed by 'a' more than Madison, for example, but this does not add to the stylistic interpretation of the texts.* [21]

La chiarezza con cui definiscono la questione del rapporto tra analisi del testo con strumenti matematici e analisi con strumenti operanti al livello semantico mostra bene anche a livello concettuale la svolta degli ultimi anni verso metodi intrinsecamente matematici.

Nel 2002 D. Benedetto, E. Caglioti, V. Loreto e altri, pubblicarono l'articolo già ricordato sulla misurazione dell'entropia informativa, *Language Trees and Zipping* [22].

Nel 2003 Clement e Sharp pubblicarono un articolo intitolato *Ngram and Bayesian Classification of Documents for Topic and Authorship* [23] in cui mostrarono di ottenere attribuzioni di testi molto soddisfacenti utilizzando *n*-grammi.

Nel 2004 si svolse una gara internazionale di attribuzione di testi, la *ad-hoc authorship attribution competition*, promossa da P. Juola che vide chiari vincitori i metodi di attribuzione basati su n-grammi [24]. Juola stesso due anni dopo rilasciò un *toolkit* per attribuzione di testi denominato JGAAP (Java Graphical Authorship Attribution Program; http://www.mathcs.duq.edu/~ryan/jgaap-4_1.zip) che permette di confrontare differenti approcci alla classificazione e attribuzione di testi.
Nel 2008 Basile, Benedetto, Caglioti, Degli Esposti pubblicarono *An example of mathematical authorship attribution* [25] esponendo i risultati delle attribuzioni sulle prime annate di scritti anonimi possibilmente gramsciani.

---

[21] D. Khmelev, F. Tweedie, *op. cit.*, p. 304.

[22] D. Benedetto, E. Caglioti, V. Loreto, *op. cit.*

[23] R. Clement, D. Sharp, *op. cit.*, pp. 423-447.

[24] P. Juola, *Ad-hoc authorship attribution competition*, Paper presented at the Proc. 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004) in Göteborg, Sweden.

[25] C. Basile, D. Benedetto, E. Caglioti, M. Degli Esposti, *An example of mathematical authorship attribution*, in "Journal of Mathematical Physics" 49/12 (2008).

Il caso Gary-Ajar

Un caso molto particolare e interessante è quello dello scrittore R. Gary che dopo aver vinto il premio Goncourt nel 1956 con *Les racines du ciel*, non soddisfatto di essere apprezzato in quanto vincitore del Goncourt invece che per le sue specifiche qualità, per provare di essere capace di conquistare il favore dei lettori e della critica partendo da zero, pubblicò alcuni romanzi sotto lo pseudonimo E. Ajar, ottenendo una seconda volto il premio Goncourt nel 1975 con *La vie devant soi* [26]. L'intera vicenda (anche più complessa di quanto qui accennato) fu rivelata da uno scritto postumo di Gary pubblicato con il titolo *Vie et mort d'Émile Ajar*.

L'elemento interessante è che lo scrittore scelse di mutare le caratteristiche della scrittura di Gary, e di inventare il modo di scrivere di Ajar che fu apprezzato dai commissari del (secondo) premio Goncourt. Sorge quindi la domanda: i commissari avrebbero potuto scoprire l'inganno se avessero avuto a disposizione strumenti di analisi testuale appropriati? Domanda che in realtà ha due aspetti: da una parte implica la ricerca della risposta alla questione se lo stile evolva nel tempo oppure no, questione connessa con l'altra se lo stile possa essere controllato coscientemente (e questa a sua volta è collegata a quella se lo stile sia totalmente cosciente o se vi siano componenti che non sono controllabili al di là dell'intenzione); dall'altra la disponibilità di strumenti di analisi non avrebbe aiutato i commissari perché l'attribuzione ha bisogno di termini di confronto per operare (prima di tutto bisogna sospettare dell'attribuzione del testo che si ha davanti; e se si sospetta si deve scegliere con quali testi di possibili autori confrontarlo).

V. Tirvengadum che ha studiato dal punto di vista dell'attribuzione le due fasi della scrittura di Gary/Ajar [27] conclude così:

> *The statistical tests done in this paper point to the same conclusion:* La Vie devant soi *is significantly different from the other Gary novels, as well as the other novels in the test. They also suggest that*

[26] Il regolamento del premio espressamente esclude l'assegnazione a un precedente vincitore!

[27] V. Tirvengadum, *Linguistic fingerprints and literary fraud*, in "CH Working Papers" (1996); http://journals.sfu.ca/chwp/index.php/chwp/article/viewArticle/A.9/69.

> *high frequency words and pairs of synonyms, which are considered by many experts on style to constitute the unconscious elements of an author's style, can indeed be consciously manipulated by the author. The notion that function words (and synonyms) constitute a genetic fingerprint of an author's style is, therefore, disputed by the case of Romain Gary / Émile Ajar.*

Cioè, stando al suo studio, un autore è in grado di modificare coscientemente il suo modo di scrivere al punto da modificare le coppie sinonimiche e le frequenze delle parole più usate. Salvo ulteriori verifiche, il suo lavoro smonta i presupposti sui quali si erano basati i lavori di Mosteller e Wallace e di Ellegård, i due capisaldi dell'attribuzione di testi basata sulle parole. Sarebbe interessante analizzare con strumenti matematici gli stessi testi di Gary-Ajar studiati da Tirvengadum per vedere se invece lavorando a un livello differente da quello delle parole come unità elementari di analisi si possa individuare qualcosa di differente.

PER CONCLUDERE

Scrive Juola che nel corso del tempo sono stati utilizzati più di 1000 metodi differenti per l'attribuzione di testi (un tempo molto breve e un numero di metodi molto grande):

> *The unfortunate situation is that a scholar with a question of authorship is now faced with a bewildering array of possible methods to use, most of which have been proven to be* better than chance *(as though that was meaningful) but with little guidance as to how much better and under what circumstances maximum accuracy can be achieved. With this situation (and more than 1000 techniques to choose from), what is necessary is a common framework and comparative evaluation to give guidance among candidate techniques.* [28]

[28] P. Juola, *JGAAP: A System for Comparative Evaluation of Authorship Attribution*, in "Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science" 1/1 (2009); http://jdhcs.uchicago.edu/.

Un quadro di riferimento comune e una valutazione comparativa delle tecniche mancano, e sono necessari. La gara internazionale *aaac* aveva svolto alcuni anni fa un ruolo importante in questo senso, ma non si è ripetuta. Occorre riprendere l'idea e darle attuazione.