L'attribuzione dei testi gramsciani: metodi e modelli matematici

D. Benedetto, E. Caglioti, M. Degli Esposti

25 gennaio 2007

1 Presentazione

In questo contributo descriviamo il metodo di attribuzione per gli articoli gramsciani che abbiamo sviluppato insieme a M. Lana[1]. Le tecniche utilizzate non sono solo il risultato di sperimentazioni, ma si basano su alcune idee importanti della matematica moderna. Presentiamo dunque una breve descrizione del punto di vista matematico, approccio che riteniamo utile per uscire dall'ambito sostanzialmente empirico di questo tipo di ricerche.

2 I testi come sequenze di simboli

La storia dei tentativi di utilizzare idee matematiche nell'analisi dei testi non è recentissima. Non appena i matematici e i fisici hanno iniziato ad interessarsi sistematicamente di sequenze di simboli hanno naturalmente tentato di utilizzare le loro idee anche per lo studio di sequenze generate da fenomeni biologici ed umani (testi, serie temporali associate a indici di borsa, sequenze di DNA).

Le idee che vogliamo esporre sono state formalizzate nei primi decenni del 1900 (nella "teoria dei processi stocastici discreti"). Le illustriamo ispirandoci all'articolo del 1948[2] con cui C.E. Shannon rende quantitativo (e dunque misurabile) il concetto di informazione contenuta in un testo. Secondo questo approccio, un testo va pensato come una sequenza di simboli scelti in un alfabeto, e un autore come ad una "sorgente" di testi.

Assumere che il testo è "solo" una sequenza di simboli vuol dire che non si prendono in considerazione né il contenuto del testo né gli aspetti grammaticali: le lettere dell'alfabeto, i segni di interpunzione, la spaziatura tra parole sono solo simboli astratti, senza gerarchia. Inoltre la parola come elemento del testo non ha maggiore significato rispetto ad altri aggregati di simboli, e il suo ruolo come unità di ordine superiore rispetto al singolo carattere viene preso dall'n–gramma. È utile fare qualche esempio:

– per monogramma (1–gramma) si intende un qualunque simbolo dell'alfabeto;

– per bigramma (2–gramma) si intende qualunque sequenza di di due simboli, ad esempio "`il`" ma anche "`l'`" e anche "`a `" (cioè la "`a`" seguita da uno spazio);

– per trigramma (3–gramma) si intende qualunque sequenza di tre simboli, ad esempio "`del`", ma anche "`e l`";

– con n–gramma si intende una sequenza qualunque di n simboli; ad esempio "`l prolet`" è un 8–gramma.

In questa teoria non solo il testo è pensato come una sequenza astratta di simboli, ma si assume anche che esso venga generato, simbolo per simbolo, da una sorgente. La natura della sorgente non è oggetto di analisi, essa è solo un modello astratto per tutti gli enti che possono generare testi. La sorgente emette i suoi messaggi (testi) scegliendo con regole probabilistiche quale simbolo emettere di volta in volta. Le sorgenti si differenziano tra di loro per le diverse regole probabilistiche con cui generano messaggi.

Con questa teoria in mente, un matematico è portato a immaginare l'autore come un generatore astratto di simboli, e i suoi testi disponibili come "esempi casualmente generati". Dunque se una qualche struttura matematico/probabilistica esiste per l'autore come sorgente (o anche per il singolo testo), essa determina quantitativamente tutti gli oggetti misurabili nel testo; attraverso le misure di tali quantità si deve dunque poter risalire alle caratteristiche della sorgente/autore.

Naturalmente lo schema sorgente/messaggio è troppo rigido ed astratto per essere una ragionevole interpretazione del rapporto autore/testo. In particolare, nei modelli matematici per le sorgenti le regole

[1] Per numerosi e importanti scambi di idee sull'argomento vogliamo ringraziare Sandro Graffi, del Dipartimento di Matematica dell'Università di Bologna, Vittorio Loreto e Andrea Baronchelli, del Dipartimento di Fisica dell'Università di Roma *La Sapienza*); inoltre ringraziamo Chiara Basile, Chiara Farinelli e Andrea Tolomelli (dottorandi in Matematica a Bologna) per il loro contributo teorico e sperimentale a questa ricerca

[2] *A Mathematical Theory of Communication* The Bell System Technical Journal 27, 1948, p. 623.

per la generazione dei simboli sono esplicitabili, mentre è a dir poco dubbio che esse esistano per un autore che scrive un testo. D'altra parte questo approccio dà indicazioni utili, come vedremo nel paragrafo successivo.

2.1 La statistica degli n–grammi

Se è vero che i testi non sono generati da sorgenti che seguono regole probabilistiche, è però vero che con tali regole si possono dare delle "approssimazioni" dei testi. Utilizzeremo, come esempio, 100 articoli di giornale gramsciani e non gramsciani del periodo 1914–1919. I testi sono scritti in un alfabeto di 84 simboli: le lettere della lingua italiana (minuscole e maiuscole, accentate e non), con qualche lettera degli alfabeti stranieri; i più comuni simboli di interpunzione; lo spazio separatore.

Una approssimazione di "ordine 0" si ottiene semplicemente estraendo simboli a caso, tutti con la stessa probabilità. Naturalmente i testi che si ottengono sono ben lontani dal somigliare ad un testo italiano, come si vede dal seguente esempio:

```
mZmJMux,1UrsN.u l3HEpf7.hy-!7WForèÈ;1tSàgMfÈFXsa7WX9FXfùrOO
```

L'approssimazione al "primo ordine" si ottiene estraendo i simboli con probabilità uguali alle frequenze relative con cui si trovano in un corpus di riferimento. Un esempio di testo così ottenuto è:

```
illfmbaoaocnn e aai,sfrmrta eeoiddmaoo'iVar legeq arnoh everl dl slB lanl
```

Con l'approssimazione di "secondo ordine" si introduce una differenza significativa: il nuovo carattere si ottiene scegliendolo in funzione del precedente. Ad esempio per scegliere il carattere che segue una "`c`", si misurano nel corpus le frequenze dei bigrammi che iniziano per "`c`", e si dividono per la frequenza di "`c`"; i numeri così ottenuti sono le *frequenze condizionate*: ad esempio, nel nostro corpus, a "`c`" segue "`a`" con frequenza 9%, segue "`e`" con frequenza 13%, segue "`h`" con frequenza 21%. Il carattere che segue "`c`" viene dunque scelto con probabilità uguali alle frequenze condizionate; analogamente viene fatto per tutti gli altri caratteri. Un testo generato con queste regole è ad esempio:

```
Loncueresono astantà chedali co le prora Lafra Seoccoro do li, fi dunqu No o ch
```

Analogamente un modello del terzo ordine sarà ottenuto misurando le frequenze di un carattere in funzione dei due precedenti. Ad esempio a "`ch`" non può seguire il carattere "`a`" (probabilità 0), mentre con probabilità 0.74 segue il carattere "`e`" e con probabilità 0.16 il carattere "`i`".
Qui di seguito riportiamo infine un esempio di testo generato con un modello del decimo ordine.

```
La pietra fondamentale nel contegno delle due alleate, quando si è convertito, è sempre da creare
```

Con l'ordine della approssimazione aumentano le caratteristiche dei testi originali conservate nei modelli: nell'approssimazione del primo ordine la divisione in parole somiglia a quella della lingua italiana; in quella del secondo ordine le sillabe sono sostanzialmente corrette, e sono credibili l'inizio e la fine delle parole; l'approssimazione di ordine dieci riproduce le singole parole e rispetta le regole grammaticali.

Si può supporre, e molti lo hanno infatti supposto[3] che le differenze "stilistiche" tra autori debbano tradursi in differenze numeriche per le frequenze degli n–grammi. Dunque, misurando le frequenze degli n–grammi di un testo sconosciuto e confrontandole con le frequenze degli n–grammi "tipiche" di un autore, si può effettuare l'attribuzione scegliendo l'autore per cui la differenza tra le frequenze sia minima. In particolare, sfruttando le informazioni che si ottengono contando quante volte ciascun n–gramma compare in un dato testo è possibile costruire un indicatore di (dis)similarità che esprime numericamente la *distanza* tra due testi arbitrari. Confrontando le distanze tra un testo di un autore sconosciuto e testi di autori noti, è possibile costruire delle procedure che attribuiscono il testo ad uno degli autori (ad esempio si può scegliere l'autore del testo più vicino).

Su queste basi abbiamo sviluppato alcuni metodi che abbiamo testato sul campione di 100 testi gramsciani e non gramsciani già citato. Abbiamo utilizzato n–grammi con n=8, perché per questo valore di n otteniamo i risultati più accurati. Elaborando i valori numerici di similarità tra i testi abbiamo calcolato per ogni articolo da esaminare un cosidetto *indice di gramscianità* , che è un numero compreso tra -1 e 1: valori vicini a 1 (ovvero -1) indicano un testo fortemente gramsciano (ovvero non gramsciano), mentre valori dell'indice prossimi allo 0 denotano una situazione di forte indecidibilità .

In figura 1 è riportato il valore di per ognuno dei 100 testi del corpus di riferimento (testi gramsciani in rosso, testi non gramsciani in blu): si può facilmente osservare come l'attribuzione di alcuni testi, ad

[3]vedi, tra gli altri, V. Keselj, F. Peng, N. Cercone, C. Thomas *N–gram–based Author Profiles for Authorship Attribution* Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03, Dalhousie University, Halifax, Nova Scotia, Canada, August 2003, pagg. 255–264.

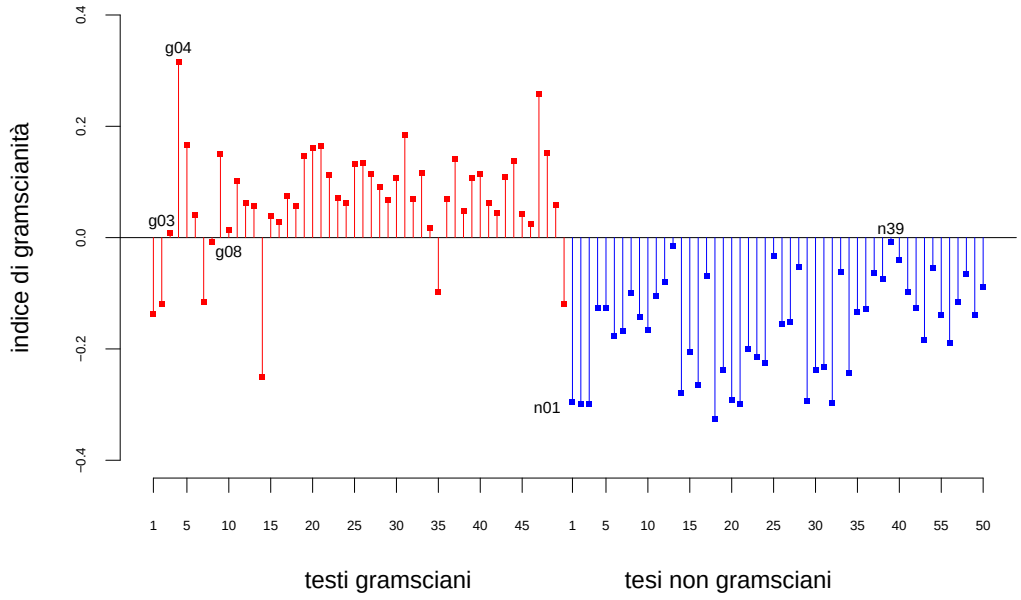

Figura 1: Attribuzioni dei testi con misura dell'affidabilità dell'attribuzione

esempio g03[4], g08[5], n39[6], sia molto meno certa rispetto a quella di g04[7] o di n01[8]. In definitiva, comunque, questo metodo attribuisce correttamente a Gramsci 44 testi gramsciani su 50, e non gli attribuisce nessuno dei 50 testi non gramsciani.

2.2 La misura del contenuto di informazione

La teoria dell'informazione nasce nel 1948 con il già citato articolo di Claude E. Shannon che pone e risolve il problema di definire, appunto, la quantità di informazione contenuta in un messaggio, ad esempio un testo o più in generale una qualunque sequenza di simboli.

L'unità di misura dell'informazione è il *bit* (dall'inglese "binary unit"); misura un bit l'informazione che sceglie uno dei due elementi di un'alternativa: acceso o spento, aperto o chiuso, giusto o sbagliato, vero o falso, 0 o 1 (che sono appunto i due simboli utilizzati dal sistema di numerazione binaria). Con un bit a disposizione si possono fare solo due affermazioni distinte; con due bit si possono invece dire quattro "parole" (in rappresentazione numerica binaria possiamo generare quattro parole: 00, 01, 10, 11); con tre bit se ne possono dire otto, e così via. Alla quantità di informazione corrispondente ad otto bit è stato dato il nome di *byte*; con un byte si possono generare 256 parole differenti. Con 256 possibilità si può codificare un alfabeto delle lingue occidentali. Infatti le lettere - incluse maiuscole, lettere accentate, segni di interpunzione, simboli speciali - non sono più di 256. Ad ogni lettera è dunque assegnata una sequenza di otto bit che la rappresenta, mediante "codici" universalmente accettati.

Sono possibili codifiche differenti, quale sia la "migliore" dipende dal contesto. Ad esempio le sequenze di DNA sono sequenze di sole 4 lettere: "A C G T": per codificare 4 simboli sono sufficienti 2 bit per carattere, invece degli 8 che si usano per lingue naturali. Si possono immaginare codifiche più fantasiose: per la sequenza

TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT

(50 T di seguito), il contenuto informativo è di 400 bit se essa è codificata nell'alfabeto italiano, di 100 se codificata nell'alfabeto del DNA, di pochi byte in un qualunque linguaggio di programmazione ricorrendo

[4] Gramsci, La luce che si è spenta, "Il grido del Popolo" 20 novembre 1915
[5] Gramsci, Socialismo e cultura, "Il Grido del Popolo", 29 gennaio 1916
[6] Togliatti, Lotta economica e guerra, "Il Grido del Popolo", 20 ottobre 1917
[7] Gramsci, L'idea nazionale, "Il Grido del Popolo", 27 novembre 1915
[8] Bianchi, Il mio atto di fede, "Il Grido del Popolo", 1 maggio 1915

ad un'istruzione che in linguaggio umano equivalga a "scrivi 50 T". Ma allora quant'è grande l'informazione della sequenza?

Nel suo lavoro del 1948 Shannon stabilì che la quantità di informazione contenuta in un messaggio è il minimo numero di bit necessari per codificarlo, e definisce l'*entropia* come il rapporto tra quel numero di bit ed il numero di caratteri. Esistono programmi che cercano di codificare un messaggio impiegando il minor numero possibile di bit: sono i programmi di compressione dati (ad esempio winzip sui sistemi Windows, gzip e bzip2 sui sistemi Linux). Dividendo la dimensione in bit del testo compresso per il numero di caratteri del testo originario si ottiene una stima della sua entropia. A titolo di esempio nella Tabella 1 riportiamo i valori calcolati utilizzando winzip per alcuni testi della letteratura italiana.

autore	opera	bit/carattere
Dante	Commedia	3.2
	De Vulgari Eloquentia	3.0
	Convivio	2.7
Boccaccio	Decamerone	2.8
Petrarca	Canzoniere	3.1

Tabella 1: Rapporti di compressione in bit per carattere di alcuni testi della letteratura italiana

La teoria di Shannon ha una formulazione rigorosa e coerente solo per oggetti matematici ben definiti, però ai matematici viene naturale utilizzare le sue idee anche nel campo dell'analisi dei testi: si può infatti ipotizzare che misurando il rapporto di compressione per i testi di un dato autore si stia misurando una quantità intrinseca della sorgente/autore. Shannon stesso, con un esperimento, stimò che la quantità di informazione media della sorgente "lingua inglese" è compresa tra 0.6 e 1.3 bit per carattere. Nonostante le caratteristiche entropiche degli scritti di un autore siano interessanti, esse non sono particolarmente utili per il problema dell'attribuzione, come si vede dalla Tabella 1.

Sviluppando le idee di Shannon si può però ottenere uno strumento più efficace per il problema dell'attribuzione: la misura dell'*entropia relativa* tra testi, anch'essa ottenuta con i programmi di compressione. Per illustrarne il significato si può considerare il codice Morse, che pur non essendo un codice di compressione, è stato pensato con un'esigenza analoga: rendere veloce la trasmissione di messaggi in lingua inglese. Il codice Morse utilizza 2 caratteri: linea e punto. Le lettere più probabili in lingua inglese vengono appunto codificate con una sequenza più corta, cioè più velocemente trasmissibile: la lettera e viene codificata con . mentre la lettera z viene codificata con –... La frequenza delle lettere fu studiata a priori: Morse si recò in tipografia per ottenerla. Poiché l'entropia è il minimo numero di bit per carattere che servono a codificare una sequenza, se si codifica una sequenza in modo non ottimale si impiegano più bit del necessario; l'entropia relativa tra due sequenze è proprio il numero di bit per carattere che si aggiungono codificandone una nel codice ottimale per l'altra. L'esempio del codice Morse aiuta a capire il concetto. Supponiamo che il codice Morse sia ottimale per la lingua inglese: se lo si utilizza per codificare un messaggio in italiano si ottiene un testo più lungo di quello che si sarebbe ottenuto se si fosse utilizzato un codice Morse ottimale per la lingua italiana. La differenza di lunghezza (per carattere) è una misura dell'entropia relativa tra l'inglese e l'italiano.

L'entropia relativa è uno strumento molto potente per quantificare la differenza tra sequenze[9], e dunque tra autori: è ragionevole aspettarsi che l'entropia relativa tra due testi di Manzoni sia più bassa di quella tra un testo di Pirandello e uno di Manzoni. In questa direzione di ricerca citiamo, senza pretesa di completezza Joula[10], Teahan[11], Khmelev[12], Benedetto, Caglioti, Loreto[13]. In particolare, in quest'ultimo lavoro viene suggerita una via molto pratica per stimare l'entropia relativa tra due testi, comprimendo opportune concatenazione dei file.

[9] J. Ziv, N. Merhav *A measure of relative entropy between individual sequences with application to universal classification* IEEE Transactions of Information Theory, 39 (4), 1993, pagg. 1270-1279.

[10] P. Juola *Cross-entropy and linguistic typology* Proceeding of New Methods in Language Processing 3, Sidney, 1998.

[11] W.J. Teahan *Text classification and segmentation using minimum cross -entropy* Proceedings of the International Conference on Content-based Multimedia Information Access (RIAO 2000), pages 943-961. C.I.D.-C.A.S.I.S, Paris, 2000.

[12] D.V. Khmelev in O.V. Kukushkina, A.A. Polikarpov, D.V. Khmelev *Using literal and grammatical statistics for authorship attribution* Problemy Peredachi Informatsii, 37 (2), 2000, pagg. 96–108, translated in English in Problems of Information Transmission 37 (2001) 172-184.

[13] D. Benedetto, E. Caglioti, V. Loreto *Language Trees and Zipping* Phys. Rev. Lett. 88, n. 4, 048702-1, 048702-4 (2002)

3 La procedura complessiva ed il test cieco

La strategia complessiva di attribuzione si basa dunque sui due metodi descritti: 8–grammi ed entropia relativa. In pratica calcoliamo un indice di gramscianità con entrambi i metodi e attribuiamo a Gramsci i soli testi che tutti e due i metodi attribuiscono a Gramsci.

Abbiamo applicato questa procedura per attribuire 40 ulteriori testi, consegnati anonimi dalla Commissione Nazionale. In Tabella 2 ne riportiamo l'elenco, con la loro indicazione bibliografica, che ovviamente non ci era nota al momento del test cieco.

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. 0 .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], 11 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, Il 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 efuturi, "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.

Tabella 2: Autori e titoli dei 40 testi utilizzati per il test cieco

L'applicazione del metodo a questi 40 testi dà il risultato mostrato in figura 2 (i punti-testo sono accompagnati dal numero che li identifica nell'elenco).
L'asse orizzontale rappresenta l'indice di gramscianità fornito dal metodo degli n-grammi: a valori positivi corrisponde l'attribuzione a Gramsci, a valori negativi la non attribuzione. I punti più a destra sono quelli di attribuzione più certa a Gramsci, quelli più a sinistra sono i testi che con maggior certezza il metodo degli n-grammi non attribuisce a Gramsci. Sull'asse verticale è riportato il valore dell'analogo indice fornito dal metodo dell'entropia relativa; in tal caso procedendo dal basso verso l'alto si va da testi meno attribuibili a Gramsci a testi più attribuibili a Gramsci.

Nel quadrante in alto a destra ci sono dunque i testi che entrambi i metodi attribuiscono a Gramsci. Tra di essi non vi è nessun punto blu, dunque non c'è nessuna falsa attribuzione a Gramsci, vale a dire che il 100% dei testi non gramsciani viene correttamente riconosciuto come tale. In totale i testi correttamente attribuiti a Gramsci sono 18 su 20, pari al 90%.

Nel quadrante in alto a sinistra ci sono i testi attribuiti a Gramsci dal metodo dell'entropia relativa, ma non attribuiti a Gramsci dal metodo degli n-grammi. Nel quadrante in basso a destra non c'è nessun testo: sarebbero quelli attribuiti a Gramsci dal metodo degli n-grammi ma non dal metodo entropico. Infine nel quadrante in basso a sinistra ci sono i testi non attribuiti a Gramsci da entrambi i metodi.

I due testi gramsciani non riconosciuti sono il n. 20 (un testo breve, di poche righe, quindi oggettivamente piuttosto difficile da attribuire) e il n. 25 che non presenta caratteristiche singolari.

4 Conclusioni e prospettive

I risultati del test cieco sono estremamente positivi, soprattutto in considerazione delle caratteristiche peculiari del problema. Infatti l'attribuzione con metodi quantitativi di testi di un corpus è spesso inscindibile da una classificazione per contenuto. Inversamente, le nostre procedure hanno dovuto operare (e così sarà in futuro) proprio su testi **molto simili** dal punto di vista degli argomenti trattati e quindi dei termini utilizzati.

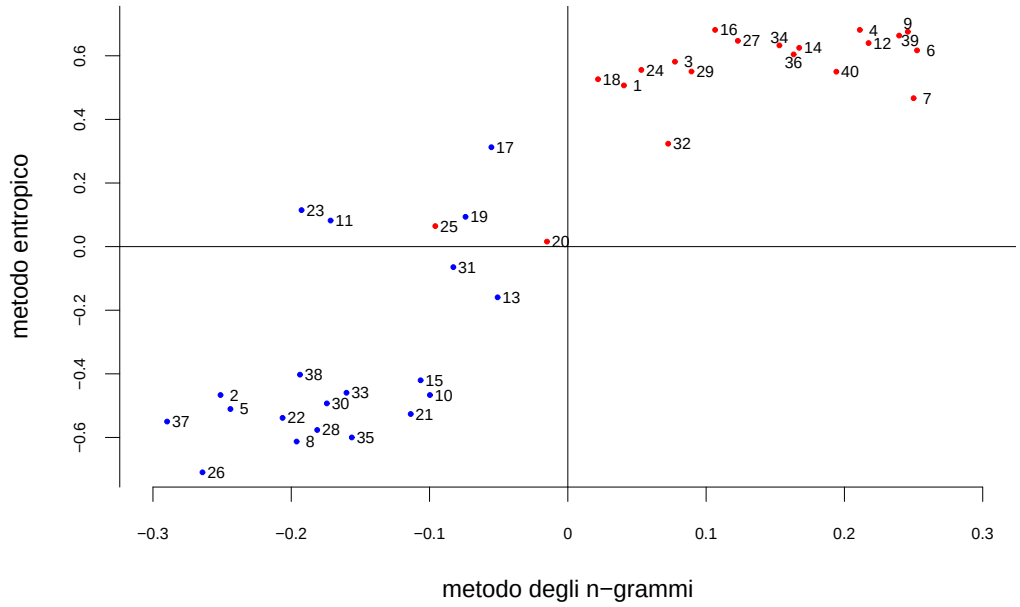

Figura 2: Attribuzioni dei 40 testi per il test cieco

Comunque è necessario non trascurare alcune questioni di metodo che si pongono nella pratica dell'attribuzione: superata la fase del test cieco non si possono più avere controprove sperimentali, inoltre i metodi dovranno ragionevolmente essere ricalibrati per le diverse annate di articoli. D'altra parte siamo rassicurati dall'aver seguito un protocollo rigoroso che ha operato in modo efficace. Le attribuzioni poi non saranno recepite come oracoli indiscutibili, ma i testi attribuiti verranno valutati dai curatori che decideranno se introdurli nell'edizione.

Dal punto di vista matematico, ci sono varie osservazioni da fare, in particolare sul metodo degli n–grammi. Considerare gli n-grammi come elementi costitutivi del testo vuol dire guardarne di traverso il livello delle strutture grammaticali e sintattiche: in questo modo però si tengono forse in conto aspetti diversi e tra loro correlati: le frequenze di caratteri, delle parole corte, dei segni di interpunzione, delle lettere iniziali di una frase; per n-grammi sufficientemente lunghi, anche le frequenze di coppie o terne di parole.

Le analisi preliminari indicano interessanti direzioni di ricerca: ad esempio il valore n=8 sembra collocare il metodo degli n-grammi nella zona di confine tra metodi filologici che osservano l'apparire in un testo di singole caratteristiche e metodi statistici che misurano solo caratteristiche che si presentano in numero abbastanza grande. È una regione intermedia, matematicamente inesplorata, ed è analoga a quella che incontrano i matematici e i fisici che collaborano con gli scienziati della vita. Riteniamo che nello studio di queste "regioni intermedie" la matematica e le altre discipline potranno fare, insieme, significativi passi avanti.