La Matematica nella Società e nella Cultura

Rivista dell'Unione Matematica Italiana

Chiara Basile, Dario Benedetto, Emanuele Caglioti, Mirko Degli Esposti

L'attribuzione dei testi gramsciani: metodi e modelli matematici

La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana, Serie 1, Vol. **3** (2010), n.2, p. 235–269.

Unione Matematica Italiana

<http://www.bdim.eu/item?id=RIUMI_2010_1_3_2_235_0>

La Matematica nella Società e nella Cultura. Rivista dell'Unione Matematica Italiana, Unione Matematica Italiana, 2010.

La Matematica nella Società e nella Cultura
Rivista dell'Unione Matematica Italiana
Serie I, Vol. III, Agosto 2010, 235-269

L'attribuzione dei testi gramsciani: metodi e modelli matematici

Chiara Basile - Dario Benedetto - Emanuele Caglioti
Mirko Degli Esposti

1. – Introduzione

In questo articolo descriviamo, con un certo dettaglio, gli aspetti teorici e matematici di un metodo di attribuzione d'autore, sviluppato in collaborazione con il linguista Maurizio Lana, per la nuova *Edizione Nazionale degli scritti di Antonio Gramsci*, curata dalla *Fondazione Istituto Gramsci* ([1]). Per questa nuova edizione nazionale, il comitato scientifico sta considerando, infatti, anche un grande numero di articoli di giornali anonimi, alcuni dei quali presumibilmente gramsciani. Al fine di riconoscere gli articoli realmente scritti da Gramsci, il comitato ha deciso di affiancare ai tradizionali metodi d'attribuzione, tipicamente qualitativi e filologici, metodi di attribuzione quantitativa specificatamente sviluppati e adattati a questo scopo (una analisi matematica più dettagliata può essere trovata nel lavoro di ricerca [1], che ha ampiamente ispirato questa esposizione divulgativa).

Innanzitutto, perché e come l'*attribuzione di autore* dovrebbe essere l'oggetto di uno studio matematico?

L'idea di applicare tecniche quantitative (non sempre matematicamente fondate) al problema di riconoscere l'autore di un testo anonimo o apocrifo non è certo nuova, ma risale almeno alla fine del XIX secolo, quando in due lavori, il primo del matematico A. De

([1]) http://www.fondazionegramsci.org/

Morgan [6] e il secondo del geofisico T.C. Mendenhall [17], viene proposto di calcolare le distribuzioni delle lunghezze delle parole in testi scritti e di confrontarle, con il fine di determinarne l'autore. La storia della *stilometria* (cioè degli studi delle delle proprietà quantitative dei testi, che presumibilmente ne determinano lo *stile* e ne caratterizzano l'autore) è piuttosto lunga, e attraverso i decenni, ricercatori provenienti dalle più diverse aree della scienza si sono interessati all'argomento (rimandiamo il lettore ai lavori [8, 10] per una descrizione più accurata delle diverse tecniche sviluppate durante questo periodo).

Uno degli aspetti più interessanti che caratterizzano l'evoluzione dei metodi della stilometria, è lo spostamento dell'attenzione da indicatori basati sulle *parole* (elementi naturali da studiare, in quanto componenti fondamentali della lingua), a metodi che non prendono in considerazione nessuna struttura sintattica del testo. Questo punto di vista, in cui il testo viene considerato semplicemente come una sequenza di simboli, è piuttosto naturale da un punto di vista matematico, e fu infatti adottato da A.A. Markov [15, 16] e da C.E. Shannon [20]: in entrambi i casi, le componenti fondamentali del testo sono semplicemente aggregazioni di caratteri, mentre la statistica di sequenze di n caratteri consecutivi (i cosiddetti *n-grammi*) emerge naturalmente come oggetto fondamentale dell'analisi. Sebbene gli approcci basati sulle parole o altre unità sintattiche siano ancora frequentemente usati, nell'ultimo decennio diversi lavori si basano sullo studio degli n-grammi: per esempio R. Clement and D. Sharp, propongono nel 2003 un metodo basato sulle frequenze degli n-grammi [5], mentre nel 2001 D.V. Khmelev e F.J. Tweedie [14] pubblicano dei risultati ottenuti considerando il testo scritto come una catena markoviana del primo ordine. In particolare, calcolano (empiricamente) la matrice di transizione tra coppie di caratteri a partire da un testo di riferimento di un dato autore, e poi usano questa matrice per determinare la probabilità che un dato testo anonimo sia stato effettivamente scritto (*generato*) dall'autore.

A partire da queste e altre importanti idee della matematica moderna, descriviamo in dettaglio il metodo di attribuzione per gli articoli gramsciani che abbiamo sviluppato.

2. – I testi come sequenze di simboli

Come già ricordato nell'introduzione, 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 [20] del 1948 con cui C.E. Shannon rende quantitativo (e dunque misurabile) il concetto di informazione contenuta in un testo (per una più esposizione più estesa si puó vedere il libro di Pierce [18]). 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 sia "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, o un segno di interpunzione o di spaziatura;

– per bigramma (2-gramma) si intende qualunque sequenza di due simboli, per 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, per esempio "`del`", ma anche "`e l`";

– con n-gramma si intende una sequenza qualunque di n simboli; per 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 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 vederemo 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, i 100 articoli di giornale gramsciani e non gramsciani elencati nelle successive tabelle 2 e 3. 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,lUrsN.u l3HEpf7.hy-!7WForsÈ;1tSgMfÈFXsa7WX9FXfvrOO
```

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. Per 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*: per 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 per esempio:

Loncueresono astant 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. Per 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 ([2]):

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

([2]) Può essere di un qualche interesse la procedura per costruire questi esempi: Shannon apriva a caso una pagina di un libro e sceglieva il nuovo carattere come quello che seguiva la prima occorrenza dell'n-gramma già scritto; qui è stata usata una versione elettronica di questa procedura.

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, 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. Discuteremo nei successivi paragrafi come queste idee si traducano in una effettiva procedura di attribuzione. Nel prossimo paragrafo invece descriveremo un'importante evoluzione di queste idee, da cui è stata creata la teoria dell'informazione.

2.2 – *La misura del contenuto di informazione*

La teoria dell'informazione nasce nel 1948 con il già citato articolo [20] di Claude E. Shannon *A Mathematical Theory of Communication*, che pone e risolve il problema di definire, appunto, la quantità di informazione contenuta in un "messaggio", per 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 ([3]).

Un esempio: la sequenza di DNA "AGCTTTTCATTCTGACTGCA" è composta di 20 caratteri e un file di testo che la contiene è grande 20 byte. Si potrebbe pensare che questa sequenza contenga dunque $20 \times 8 = 160$ bit di informazione. In realtà, poiché per scrivere sequenze di DNA è sufficiente un alfabeto costituito dalle 4 lettere "A C G T" e poiché per codificare 4 simboli sono sufficienti 2 bit, la sequenza data contiene (al più) $20 \times 2 = 40$ bit di informazione. La codifica influenza quindi la quantità di informazione impiegata per un scrivere un messaggio: mentre per un testo in italiano sono utilizzati 8 bit per carattere, per un 'testo genetico' sono sufficienti 2 bit per carattere. 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 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 stabilisce che la quantità di informazione contenuta in un messaggio è il minimo numero di bit necessari per codificarlo, e definisce l'*entropia* come il minimo numero di bit per carattere. Esistono programmi che cercano di codificare un messaggio impiegando il minor numero possibile di bit: sono i programmi di compressione dati (per esempio winzip sui sistemi Windows, gzip e bzip2 sui sistemi Unix – per una descrizione generale dei compressori si veda per esempio [22]). Il rapporto di compressione (ottenuto confrontando la dimensione del testo compresso con la dimensione del

([3]) La prima codifica, ASCII, era a soli 7 bit; i diversi gruppi di lingue occidentali usano varianti della codifica iso8859 (cfr. http://en.wikipedia.org/wiki/ISO_8859). Unicode è lo standard internazionale per codifiche universali, quelle cioè che permettono di rappresentare i caratteri di qualunque alfabeto umano (http://www.unicode.org). Infine UTF-8 è tra tali codifiche lo standard di fatto (cfr. http://en.wikipedia.org/wiki/UTF-8).

testo originario) permette di stimare l'entropia di un testo (si veda anche [23]). A titolo di esempio nella Tabella 1 riportiamo i rapporti di compressione di alcuni testi della letteratura italiana, ottenuti utilizzando winzip.

TABELLA 1: rapporti di compressione in bit per carattere di alcuni testi della letteratura italiana.

autore	opera	rapporto di compressione
Dante	Commedia De Vulgari Eloquentia Convivio	3.2 3.0 2.7
Boccaccio	Decamerone	2.8
Petrarca	Canzoniere	3.1

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: il concetto di *entropia relativa*. Per illustrarlo è utile descrivere in maggior dettaglio il funzionamento di alcuni metodi (algoritmi) di compressione. I primi ad apparire (Shannon-Fano, Huffman) funzionano utilizzando una conoscenza a priori della statistica dei caratteri nel testo e codificando un solo carattere (o pochi caratteri) alla volta. Il codice assegnato ad un carattere è tanto più corto quanto più il carattere è frequente. Come esempio si può considerare il codice Morse che, pur non essendo un codice di compressione, è stato pensato con un'esigenza analoga a quella dei compressori: rendere veloce la trasmissione di messaggi in lingua inglese. Il codice Morse utilizza 5 caratteri: linea, punto e in-

tervallo breve per codificare le lettere, intervallo medio per separare parole, intervallo lungo per separare frasi. 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. L'entropia è il minimo numero di bit per carattere che servono a codificare una sequenza, dunque 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, 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. Inoltre l'entropia relativa è una quantità che può essere computata efficacemente utilizzando gli algoritmi di compressione o algoritmi basati su idee simili. Già nel 1993 Ziv e Merhav in [25] avevano proposto l'uso dell'entropia relativa per affrontare problemi di categorizzazione, suggerendo anche algorimi concreti per misurarla. Su problemi più specifici questi metodi sono stati proposti ed utilizzati sia nel campo dell'analisi delle sequenze biologiche che nel campo dell'attribuzione dell'autore; qui citiamo, senza pretese di completezza, i lavori di Juola [9], Teahan [21], Khmelev [11], e Benedetto, Caglioti, Loreto [2].

2.3 – *I metodi matematici per l'attribuzione*

Dalla descrizione dei testi come sequenze di n-grammi e dall'entropia come misura del contenuto di informazione, si possono costruire

procedure per l'attribuzione di un testo di autore incognito:

– per gli n-grammi: si misurano le loro frequenze nel testo e si confrontano con quelle dei testi dei diversi autori;
– per l'entropia: si misura l'entropia relativa del testo rispetto ai testi dei diversi autori.

Esistono numerose altre procedure, ma un metodo matematico di attribuzione sarà comunque caratterizzato, in estrema sintesi, da due aspetti:

a) scelta degli "oggetti" di cui si ritiene significativo il conteggio, cioè degli oggetti che si suppone possano essere usati con frequenze sensibilmente differenti dai diversi autori;
b) scelta del modo di tradurre in attribuzioni le misure delle quantità descritte nel punto a).

Frequenza degli n-grammi ed entropia relativa sono due possibili scelte per il punto a). Altre scelte sono per esempio quelle di Khmelev e Tweedie [14] e quelle di Clement e Sharp [5] sopra citate. Per le scelte del punto b), sono prevalentemente usati metodi probabilistico/statistici e metodi metrici o di similarità. I metodi probabilistici e statistici partono dal presupposto che le caratteristiche di un testo (quelle scelte nel punto a), non siano univocamente associabili ai singoli autori, ma compaiano con frequenze differenti per i diversi autori. Dunque, per ogni autore si può determinare la probabilità che in un suo testo si osservino queste caratteristiche, e che tali probabilità varino al variare dell'autore. Esistono tecniche matematiche consolidate (per esempio la formula di Bayes e più in generale i test statistici) che studiano il problema inverso, cioè permettono di calcolare la probabilità che un testo in cui si osservano certe caratteristiche sia di un dato autore.

Un differente approccio consiste nel sintetizzare in un unico numero la differenza/dissimilarità che si osserva misurando le quantità scelte nel punto a). Tale numero sarà una misura della vicinanza tra testi o tra testo e autore; in genere, sarà un numero tanto più piccolo quanto più piccola sarà la differenza misurata, cioè quanto più vicini/simili saranno i testi. Un matematico a questo punto preferirà che questa vicinanza sia in realtà una vera e propria "distanza" (o "metrica"), cioè un

concetto matematico preciso, ottenuto astraendo le caratteristiche della usuale distanza tra punti dello spazio ([4]). Il vantaggio rispetto ad una generica misura di "vicinanza" è duplice: la "distanza" è un concetto matematicamente solido, non ambiguo, inoltre permette di utilizzare altri strumenti matematici che proprio dalla nozione di distanza sono stati sviluppati (la descrizione metrica permette tre le altre cose la costruzione di "alberi filogenetici", in cui l'attribuzione corrisponde all'appartenenza ai vari rami dell'albero ([5])).

3. – I testi: messa a punto dei metodi

Il problema dell'attribuzione dei testi gramsciani ha caratteristiche specifiche rispetto ad altri problemi di attribuzione. Una prima caratteristica è che si deve 'solo' decidere se un testo è o non è di Gramsci, e non è necessario attribuire al giusto autore ogni testo; questa peculiarità rappresenta un elemento di semplificazione rispetto ad un problema di attribuzione generico con molteplici autori possibili. È invece un elemento di complicazione la seconda caratteristica peculiare del problema che stiamo affrontando. I testi incogniti sono 'omogenei' sotto vari punti di vista: stessa fonte (articoli di giornale), stesso periodo, argomenti analoghi, autori distinti che però presumibilmente condividono in ampia misura il linguaggio e molte idee di base sugli argomenti dei quali scrivono. Come vedremo, queste caratteristiche verranno tenute in considerazione per la determinazione del metodo di attribuzione.

Abbiamo dunque sperimentato varie tecniche in una fase preliminare di messa a punto, in cui sono stati utilizzati 100 testi, 50 di Gramsci e 50 di altri autori, elencati nelle tabelle 2 e 3. A questa fase preliminare

([4]) In particolare la distanza è caratterizzata dall'obbedire alla "disuguaglianza triangolare" che in sostanza afferma: se andando da A a B si fa una deviazione per C, il percorso si allunga. La "vicinanza" tra testi è in genere una nozione oscura proprio perché non soddisfa questa proprietà piuttosto naturale.

([5]) In un altro ambito, il passaggio dalle distanze agli alberi è stato usato da Cavalli-Sforza [4] nelle sue ricerche di grande rilevanza sulla vicinanza genetica tra le popolazioni umane.

è seguito un test cieco, con lo scopo di verificare l'adeguatezza dei metodi sviluppati.

TABELLA 2: testi gramsciani.

pubblicati su "Il Grido del popolo"

1. Neutralità attiva e operante, "Il Grido del popolo", 31 ottobre 1914.
2. Dopo il congresso socialista spagnuolo, "Il Grido del popolo", 13 novembre 1915.
3. La luce che si è spenta, "Il Grido del popolo", 20 novembre 1915.
4. [L'idea nazionale], "Il Grido del popolo", 27 novembre 1915.
5. La Festuca, "Il Grido del popolo", 11 dicembre 1915.
6. Il Sillabo ed Hegel, "Il Grido del popolo", 15 gennaio 1916.
7. Pietro Gavosto, "Il Grido del popolo", 22 gennaio 1916.
8. Socialismo e cultura, "Il Grido del popolo", 29 gennaio 1916.
9. Armenia, "Il Grido del popolo", 11 marzo 1916.
10. Il Mezzogiorno e la guerra, "Il Grido del popolo", 1° aprile 1916.
11. La paura del "Dumping", "Il Grido del popolo", 13 maggio 1916.
12. Il Dumping germanico, "Il Grido del popolo", 20 maggio 1916.
13. L'eroe, "Il Grido del popolo", 17 giugno 1916.
14. Beneficenza, "Il Grido del popolo", 12 agosto 1916.
15. Contro il feudalesimo economico, "Il Grido del popolo", 12 agosto 1916.
16. Mônssù Bôtegari, "Il Grido del popolo", 13 gennaio 1917.
17. Carattere, "Il Grido del popolo", 3 marzo 1917.
18. Note sulla rivoluzione russa, "Il Grido del popolo", 29 aprile 1917.
19. Il perfido straniero, "Il Grido del popolo", 9 giugno 1917.
20. La scuola di Stenterello, "Il Grido del popolo", 15 giugno 1917.
21. Abbruciamenti, "Il Grido del popolo", 21 luglio 1917.
22. I massimalisti russi, "Il Grido del popolo", 28 luglio 1917.
23. L'orologiaio, "Il Grido del popolo", 18 agosto 1917.
24. Letture, "Il Grido del popolo", 24 novembre 1917.
25. Intransigenza-tolleranza, intolleranza-transigenza, "Il Grido del popolo", 8 dicembre 1917
26. La rivoluzione contro il "Capitale", "Il Grido del popolo", 5 gennaio 1918 [già in "Avanti!", 24 dicembre 1917].
27. La critica critica, "Il Grido del popolo", 12 gennaio 1918.
28. La Lega delle nazioni, "Il Grido del popolo", 19 gennaio 1918.
29. Achille Loria, "Il Grido del popolo", 19 gennaio 1918.
30. La funzione sociale del Partito socialista nazionalista, "Il Grido del popolo", 26 gennaio 1918.
31. La famiglia, "Il Grido del popolo", 9 febbraio 1918; Anche in "Avanguardia", 3 marzo 1918.
32. Il nostro Marx, "Il Grido del popolo", 4 maggio 1918.
33. Libero pensiero e pensiero libero, "Il Grido del popolo", 15 giugno 1918.
34. L'utopia russa, "Il Grido del popolo", 27 luglio 1918 [anche in "Avanti!" del 25 luglio 1918 col titolo Utopia].

pubblicati su "Avanti!"

35. Morgari in Russia, "Avanti!", 20 aprile 1917.
36. Il canto delle sirene, "Avanti!", 10 ottobre 1917.
37. Contro un pregiudizio, "Avanti!", 24 gennaio 1918.
38. Il sindacalismo integrale, "Avanti!", 31 marzo 1918.
39. Il cieco Tiresia, "Avanti!", 18 aprile 1918.
40. La tua eredità, "Avanti!", 1° maggio 1918.
41. I contadini e lo Stato, "Avanti!", 6 giugno 1918.
42. L'irresponsabilità sociale, "Avanti!", 7 agosto 1918.
43. I liberali italiani, "Avanti!", 12 settembre 1918.
44. Uomini, idee, giornali e quattrini, "Avanti!", 23 ottobre 1918.
45. I cattolici italiani, "Avanti!", 22 dicembre 1918.

pubblicati su "La città futura"

46. Tre principi, tre ordini, "La città futura", 11 febbraio 1917.
47. Indifferenti, "La città futura", 11 febbraio 1917.
48. Analfabetismo, "La città futura", 11 febbraio 1917.
49. Margini, "La città futura", 11 febbraio 1917.
50. [La città futura], "La città futura", 11 febbraio 1917.

TABELLA 3: testi non gramsciani in ordine alfabetico per autore.

1. Giuseppe Bianchi, Il mio atto di fede, "Il Grido del popolo", 1° maggio 1915.
2. Giuseppe Bianchi, Ai lombrichi dell'Azione socialista, "Il Grido del popolo", 12 giugno 1915.
3. Giuseppe Bianchi, Di male in peggio, "Il Grido del popolo", 19 giugno 1915.
4. Amadeo Bordiga, La rivoluzione russa. I, "L'Avanguardia", 21 ottobre 1917.
5. Amadeo Bordiga, La rivoluzione russa.II, "L'Avanguardia", 4 novembre 1917.
6. Amadeo Bordiga, La rivoluzione russa. III, "L'Avanguardia", 11 novembre 1917.
7. Amadeo Bordiga, La rivoluzione russa. IV, "L'Avanguardia", 2 dicembre 1917.
8. Amadeo Bordiga, Le direttive marxiste della nuova internazionale, "L'Avanguardia", 26 maggio 1918.
9. Amadeo Bordiga, L'illusione elezionista, "Il Soviet", 9 febbraio 1919.
10. Amadeo Bordiga, Formiamo i "Soviet"?, "Il Soviet", 21 settembre 1919.
11. Attilio Carena, Pasqua di risurrezione, "Il Grido del popolo", 7 aprile 1917.
12. Attilio Carena, Fede e programmi secondo Benedetto Croce, "Il Grido del Popolo", 3 novembre 1917.
13. Attilio Carena, Libera la tua volonta!, "Il Grido del Popolo", 24 agosto 1918.
14. Gino Castagno, I pretesi errori confederali, "Il Grido del Popolo", 24 giugno 1916.
15. C.D., Per l'Ufficio dell'assistenza popolare, "Avanti!", 15 luglio 1917.
16. Alessandro De Giovanni, L'internazionale sarà, "Il Grido del popolo", 28 ottobre 1916.
17. C.F., Dall'ex barriera di Casale, "Avanti!", 9 luglio 1917.

18. Leo Galetto, La pace futura, “Il Grido del Popolo”, 5 giugno 1915.
19. Leo Galetto, La guerra della democrazia, “Il Grido del popolo”, 19 giugno 1915.
20. Leo Galetto, L’avvenire nostro, “Il Grido del popolo”, 3 luglio 1915.
21. Leo Galetto, Impressioni e commenti, “Il Grido del popolo”, 24 luglio 1915.
22. Adolfo Giusti, La fungaia malefica, “Il Grido del popolo”, 23 gennaio 1915.
23. Adolfo Giusti, Ostracismi sindacali, “Il Grido del popolo”, 6 febbraio 1915.
24. Adolfo Giusti, La fungaia malefica [2], “Il Grido del popolo”, 6 febbraio 1915.
25. Adolfo Giusti, I profitti dell’industria laniera, “Avanti!”, 3 settembre 1915.
26. Alfonso Leonetti, Il centenario della nascita di Carlo Pisacane: Pisacane socialista, “Il Grido del Popolo”, 24 agosto 1918.
27. Alfonso Leonetti, I comunisti e le elezioni, “L’Ordine nuovo”, 9 agosto 1919.
28. Ottavio Pastore, L’assemblea dei pescicani, “Avanti!”, 8 aprile 1916.
29. Mario Santarosa, L’eccellenza mentisce, “Il Grido del popolo”, 1 luglio 1916.
30. Giacinto Menotti Serrati, Scampoli. Il Primo Maggio di Maria, “Avanti!”, 1º maggio 1917.
31. Giacinto Menotti Serrati, Discutendo tra relativisti ed intransigenti. Replica, “Avanti!”, 9 maggio 1918.
32. Giacinto Menotti Serrati, Salutatemi la disciplina, “Avanti!”, 3 settembre 1918.
33. Angelo Tasca, Il mito della guerra, “Il Grido del popolo”, 24 ottobre 1914.
34. Angelo Tasca, Triplice alleanza e triplice intesa, “Il Grido del popolo”, 13 marzo 1915.
35. Angelo Tasca, Battute di preludio, “L’Ordine nuovo”, 1º maggio 1919.
36. Angelo Tasca, Il programma massimalista, “L’Ordine nuovo”, 30 agosto 1919.
37. Angelo Tasca, Cultura e socialismo, “L’Ordine nuovo”, 28 giugno-5 luglio 1919.
38. Umberto Terracini, Il protezionismo: decima moderna, “Il Grido del popolo”, 26 maggio 1917.
39. Palmiro Togliatti, Lotta economica e guerra, “Il Grido del popolo”, 20 ottobre 1917.
40. Palmiro Togliatti, Le due Italie, “Il Grido del popolo”, 3 novembre 1917.
41. Palmiro Togliatti, Il mito dell’indipendenza economica, “Il Grido del popolo”, 3 luglio 1918.
42. Palmiro Togliatti, La disfatta di A. Lanzillo, “L’Ordine nuovo”, 1º maggio 1919.
43. Palmiro Togliatti, Parole oneste sulla Russia, “L’Ordine nuovo”, 1º maggio 1919.
44. Palmiro Togliatti, “Guerra e fede” di Giovanni Gentile, “L’Ordine nuovo”, 1º maggio 1919.
45. Palmiro Togliatti, Parassiti della cultura, “L’Ordine nuovo”, 15 maggio 1919.
46. Palmiro Togliatti, “Franche parole alla mia Nazione” di Arturo Farinelli, “L’Ordine nuovo”, 15 maggio 1919.
47. Palmiro Togliatti, Postilla, “L’Ordine nuovo”, 19 luglio 1919.
48. Palmiro Togliatti, La battaglia delle idee (G. Prezzolini, “Dopo Caporetto”), “L’Ordine nuovo”, 25 ottobre 1919.
49. Palmiro Togliatti, Creare una scuola, “L’Ordine nuovo”, 15 novembre 1919.
50. Andrea Viglongo, Il concetto dell’educazione, “Il Grido del popolo”, 16 marzo 1918.

Su questo corpus sono stati sperimentati i vari metodi, con la seguente procedura:

– viene isolato un testo (come se fosse incognito) dagli altri 99 (testi noti);

– il testo incognito viene confrontato con i testi noti per effettuarne l'attribuzione;

– la procedura viene ripetuta per ognuno dei 100 testi.

I risultati sono sintetizzati in quattro numeri, che, usando il linguaggio proprio dei test diagnostici per la medicina, sono

– il numero di *veri positivi*, cioè il numero di testi gramsciani attribuiti a Gramsci;

– il numero di *veri negativi*, cioè il numero di testi non gramsciani correttamente non attribuiti a Gramsci;

– il numero di *falsi positivi*, cioè il numero di testi non gramsciani erroneamente attribuiti a Gramsci;

– il numero di *falsi negativi*, cioè il numero di testi gramsciani erroneamente non attribuiti a Gramsci;

Un metodo efficace e 'certo' darebbe come risultato 50 veri positivi e 50 veri negativi. Naturalmente non si ottengono risultati così buoni. Quello che si tenta di fare è cercare di ottenere il massimo numero di veri positivi ottenendo nel contempo un numero minimo di falsi positivi.

Sostanzialmente sono due le ipotesi operative sopravvissute alla fase di messa a punto: un miglioramento del metodo degli n-grammi di Vlado Kešelj [12] e un metodo di tecniche entropiche ottenuto a partire dal lavoro [2] del 2002.

Questi due metodi sono confluiti nella nostra strategia complessiva di attribuzione, che li utilizza entrambi con lo scopo di diminuire il numero di falsi positivi (in sostanza attribuiamo a Gramsci i soli testi che entrambi i metodi attribuiscono a Gramsci). Nelle pagine seguenti queste sintetiche annotazioni verranno presentate in modo più ampio e approfondito. Qui ci limitiamo a dare due indicazioni metodologiche piuttosto rilevanti.

Preparazione dei testi

I testi sono stati sottoposti ad una lavoro di preparazione che ha uniformato la loro grafia. In particolare è stata mantenuta la distinzione tra lettere minuscole e maiuscole, sono stati tenuti i segni di interpunzione e lo spazio separatore; tutti gli spazi multipli sono stati ridotti a spazi singoli, sono stati eliminate le note di trascrizione come [...] e [?] e grafie particolari (per esempio gli accenti acuti piuttosto che gravi, e simili) sono stati sostituiti con le loro varianti più ricorrenti; infine è stato eliminato il terminatore di linea (il simbolo di a-capo), ritenendo che la divisione in capoversi dipendesse più da esigenze tipografiche che da scelte (consapevoli o meno) degli autori.

Confronto testo/autore e confronto testo/testo

Un approccio ricorrente nell'attribuzione di testi con metodi quantitativi consiste nel confrontare il singolo testo di attribuzione incerta con un "modello" o "profilo" del possibile autore, costruito attraverso i suoi testi noti. Noi abbiamo preferito considerare il singolo testo come oggetto "naturale" al centro dell'analisi e dunque abbiamo preferito confrontare tra loro i singoli testi; interpretando solo successivamente i risultati ottenuti in termini di autore. Abbiamo fatto questa scelta, supportata dai risultati sperimentali, perché convinti che sia rilevante per l'attribuzione la presenza di "aspetti particolari" nella produzione di un autore: il confronto tra singoli testi può mettere in luce il ripetersi di queste caratteristiche rare, mentre un modello o profilo di autore è per definizione "medio" e rischia di nasconderle sullo sfondo.

3.1 – *Il metodo degli n-grammi di Kešelj*

L'uso degli n-grammi nell'ambito dell'attribuzione d'autore è, come già detto, un'idea abbastanza recente. Dopo un primo esperimento presentato nel 1976 da William R. Bennett [3] che utilizzava le frequenze dei 2-grammi (bigrammi), Vlado Kešelj insieme ai suoi collaboratori pubblicò nel 2003 un articolo [12] nel quale le frequenze degli n-grammi vengono utilizzate per definire una distanza di similarità tra testi. Descriveremo ora con maggiore dettaglio l'approccio e la formula

di Kešelj, poiché rappresentano il punto di partenza per lo sviluppo del nostro metodo.

Kešelj innanzitutto definisce per ogni autore un profilo. Il profilo dell'autore A si costruisce in questo modo: fissato un valore di n, di solito compreso tra 4 e 8, si calcolano le frequenze dei possibili n-grammi usando tutti i testi a disposizione appartenenti all'autore A. Successivamente questi n-grammi vengono disposti in ordine decrescente di frequenza e vengono presi in considerazione solo i primi L, dove L è un ulteriore parametro da scegliere.

In presenza di più autori Kešelj costruisce in questo modo una sequenza di profili da usare poi nella fase di attribuzione. La medesima operazione (estrazione delle frequenze degli n-grammi e successivo ordinamento) viene poi eseguita sul testo incognito X da attribuire.

Indichiamo con il simbolo ω un arbitrario n-gramma, con $f_X(\omega)$ la frequenza con cui ω appare nel testo X, e con $f_A(\omega)$ la frequenza con cui ω appare nei testi dell'autore A. Con queste notazioni un testo X può essere confrontato con un profilo A mediante la seguente formula, che definisce una misura della vicinanza tra il testo X e l'autore A:

$$d_n(X,A) = \sum_{\omega} \frac{(f_A(\omega) - f_X(\omega))^2}{(f_A(\omega) + f_X(\omega))^2},$$

dove la sommatoria è eseguita su tutti gli n-grammi ω che compaiono almeno una volta in uno dei due testi ([6]). Il testo X viene quindi attribuito all'autore A per cui la distanza è minima. Kešelj afferma di essere giunto a questa formula ispirato dall'articolo di Bennett [3], nel quale l'autore utilizzava come indicatore di (dis)similarità la distanza definita semplicemente come la somma dei quadrati delle differenze tra le frequenze in A e in X:

$$d_n(X,A) = \sum_{\omega} (f_A(\omega) - f_X(\omega))^2.$$

Si noti che nella formula di Kešelj riportata in precedenza, e a differenza della formula di Bennet, ogni termine della somma è pesato

([6]) In presenza del parametro L, la somma è ristretta ai primi L in ordine di frequenza decrescente.

con l'inverso del quadrato della somma delle frequenze di quel particolare n-gramma in A e in X, in modo da dare un peso maggiore nella sommatoria al contributo delle "parole rare", ovvero agli n-grammi con frequenze più basse. In questo modo, ad esempio, una differenza di 0,1 per un n-gramma con frequenze 0,9 e 0,8 nei due profili avrà un peso minore rispetto alla stessa differenza per un n-gramma con frequenze 0,2 e 0,1. È anche utile osservare che $d_n(X,A)$ non è in effetti una vera distanza dal punto di vista matematico (infatti per esempio non soddisfa la diseguaglianza triangolare). Ma come si usa fare solitamente, continueremo a riferirci a questa funzione chiamandola distanza.

Kešelj e collaboratori verificarono l'efficacia del loro metodo su diversi corpora di testi: opere letterarie di 8 autori di lingua inglese di diverse epoche; articoli di giornale di 10 autori diversi, scritti in greco moderno; alcuni romanzi sulle arti marziali di 8 scrittori cinesi moderni. Come è possibile verificare in dettaglio nel lavoro sopra citato, i risultati finali sono piuttosto soddisfacenti e comunque uguagliano o superano in quasi tutti i casi (con la sola eccezione del corpus di opere cinesi) quelli raggiunti con i metodi precedentemente sperimentati sia da Kešelj sia da altri ricercatori sugli stessi insiemi di testi. È bene però osservare, come in effetti fa lo stesso Kešelj, che la dipendenza da uno o due parametri (n ed eventualmente L) dà luogo ad un problema metodologico: come scegliere il valore ottimale di n ed L per un problema di attribuzione reale, di cui cioè non si conosca la soluzione? In una brevissima pubblicazione [13] Kešelj propose a questo scopo l'introduzione di un *voto pesato*: per ogni coppia (n, L), con n compreso fra 3 e 8 e L fra 1000 e 5000, e per ogni testo incognito X, se d_1 è la distanza di X dall'autore A a lui più vicino e d_2 è la distanza di X dal secondo autore più vicino, il voto di A per quei valori dei parametri (n, L) sarà $1 - d_1/d_2$. Tale voto è maggiore quanto maggiore è la differenza tra le distanze d_1 e d_2 dei primi due vicini da X; il voto favorisce cioè le situazioni in cui il primo autore in classifica è fortemente distaccato dal secondo. Ora l'attribuzione per X è decisa prendendo il massimo tra i voti calcolati al variare di n ed L, così da rendere il risultato finale indipendente dai valori dei due parametri.

3.2 – *Il metodo degli n-grammi modificato*

Per gli esperimenti di riconoscimento di Gramsci abbiamo utilizzato e successivamente modificato le idee di Kešelj, sviluppando un metodo differente e più adatto al problema specifico.

Innanzitutto, come ricordato, anziché unire in un solo profilo tutti i testi a disposizione per i vari autori, la distanza viene calcolata per tutte le coppie di testi disponibili. Peraltro, in questo caso, una costruzione dei profili degli autori, secondo il metodo di Kešelj, si scontrerebbe con le caratteristiche del corpus dei 100 articoli in esame, in cui la suddivisione dei testi tra i vari autori è fortemente non omogenea (cfr. Tabella 4), così che l'unificazione di tutti i testi di un autore in un unico profilo avrebbe

TABELLA 4: lunghezza in caratteri totale e media degli articoli degli autori usati nella fase preliminare.

Autore	Numero degli articoli	Lunghezza totale degli articoli	Lunghezza media degli articoli
Antonio Gramsci	50	326910	6538,2
Palmiro Togliatti	11	91339	8303,5
Amedeo Bordiga	7	47895	6842,1
Angelo Tasca	5	48688	9737,6
Leo Galetto	4	18625	4656,3
Adolfo Giusti	4	14349	3587,3
Giuseppe Bianchi	3	12933	4311,0
Attilio Carena	3	23559	7853,0
Giacinto Menotti Serrati	3	12860	4286,7
Alfonso Leonetti	2	16515	8257,5
Gino Castagno	1	8146	8146,0
C. D.	1	5612	5612,0
Alessandro De Giovanni	1	6700	6700,0
C. F.	1	2659	2659,0
Ottavio Pastore	1	4177	4177,0
Mario Santarosa	1	5053	5053,0
Umberto Terracini	1	9435	9435,0
Andrea Viglongo	1	7451	7451,0

dato luogo a profili con significato statistico molto diverso: si noti per esempio la grande sproporzione tra la lunghezza totale dei testi a disposizione per Gramsci e per Viglongo.

Inoltre, vista la brevità degli articoli e la scelta di confrontarli singolarmente, il parametro L diviene superfluo ed è necessario considerare tutti i possibili n-grammi con frequenza non nulla. In un testo singolo e per n grande, infatti, come risulta dall'esempio in Tabella 5, la maggior parte degli n-grammi si presenta una sola volta cosicché considerare solo gli L più frequenti equivarrebbe a scegliere gli n-grammi in modo arbitrario.

Infine, per eliminare la grande dipendenza della formula di Kešelj dalla variabilità delle lunghezze dei testi in esame, la distanza viene divisa per la media del numero di n-grammi presenti nei due testi. I risultati dell'attribuzione con i primi vicini dei 100 testi sono riportati in figura 1. Lungo l'asse orizzontale è riportata la lunghezza

TABELLA 5: occorrenze e percentuali degli n-grammi che compaiono una sola volta nel testo gramsci 27.

n	n-grammi totali	n-grammi che compaiono una sola volta	percentuale di n-grammi che compaiono una sola volta
1	62	9	15%
2	416	107	26%
3	1576	689	44%
4	2948	1805	61%
5	3960	2948	74%
6	4611	3806	83%
7	5030	4405	88%
8	5297	4806	91%
9	5480	5086	93%
10	5611	5294	94%
11	5707	5453	96%
12	5777	5565	96%
13	5837	5660	97%
14	5888	5741	98%
15	5931	5807	98%

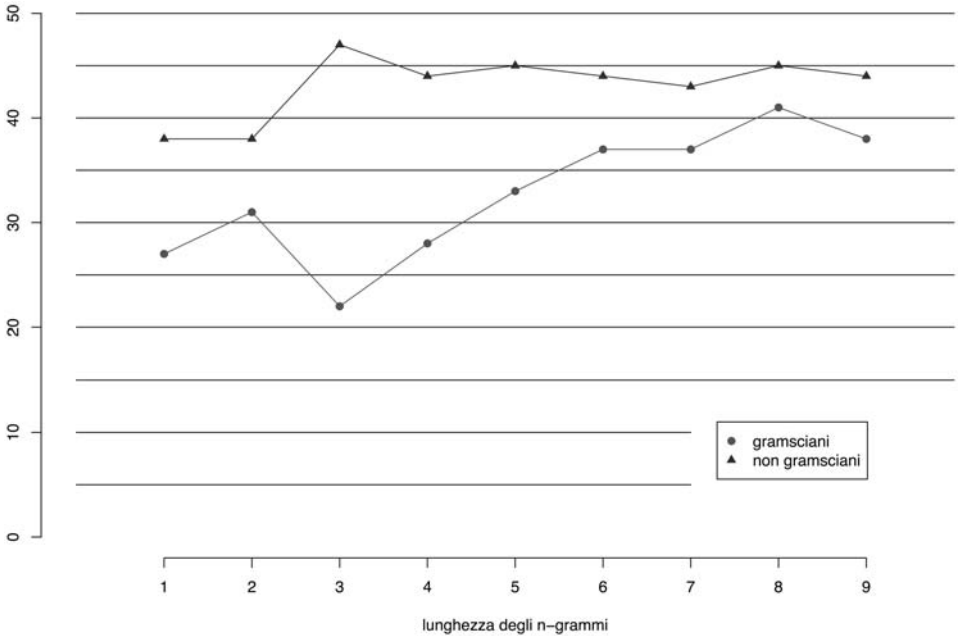

Fig. 1. – Metodo degli n-grammi modificato, numero di testi gramsciani e non gramsciani correttamente attribuiti al variare della lunghezza degli n-grammi.

degli n-grammi, per n da 1 a 9; in corrispondenza di ogni valore di n sono riportati in verticale due valori: il cerchio rosso rappresenta il numero dei testi di Gramsci che sono correttamente attribuiti dal metodo a Gramsci (i veri positivi) e il triangolo blu il numero di testi non gramsciani riconosciuti come tali (i veri negativi).

Poiché lo scopo principale della ricerca è quello di separare i testi gramsciani da quelli di altri autori, si può osservare direttamente dalla figura 1 che non esiste un valore di n che massimizzi contemporaneamente entrambe le quantità: il numero massimo di testi di Gramsci a lui correttamente attribuiti si ottiene infatti per $n = 8$, mentre il numero minimo di falsi positivi (testi non gramsciani erroneamente attribuiti a Gramsci), si ha per valori più piccoli del parametro, $n = 3, 4, 5$ ([7]).

([7]) I falsi positivi si ricavano dalla tabella, ricordando che i testi non gramsciani da attribuire sono 50, e quindi il numero falsi positivi è pari a 50 diminuito del numero di veri negativi.

Per i successivi esperimenti è stato scelto n = 8: per tale valore, infatti, si ottengono i migliori risultati di attribuzione (41 testi su 50) pur non perdendo troppo in *precisione* (solo 5 falsi positivi).

I risultati sono stati ottenuti prendendo in considerazione unicamente il primo vicino di ogni testo. Questa scelta trascura il fatto che l'insieme dei testi di riferimento conta ben 100 distinti articoli rispetto ai quali confrontare un testo dato. Da qui alcune interrogativi:

– che cosa ci possiamo aspettare riguardo alla distanza di un articolo di Gramsci dagli altri 49 testi gramsciani?

– questi ultimi saranno "mediamente più vicini" al testo in esame rispetto ai 50 testi di altri autori?

– si può utilmente tenere in conto anche le distanze di tutti gli altri testi di riferimento, oltre a quello più vicino?

Per tenere conto di tali considerazioni si definisce, per un testo X, un *indice di gramscianità* $g(X)$ nel seguente modo: vengono ordinati in una lista, dal piú vicino ad X al piú lontano, tutti i testi di

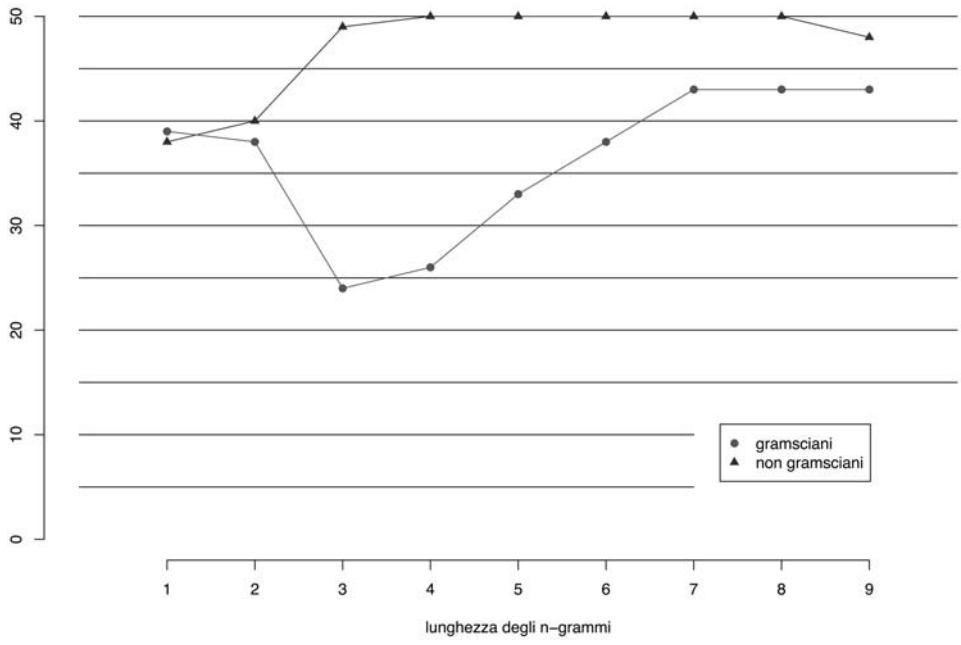

Fig. 2. – Metodo degli n-grammi modificato, attribuzioni corrette mediante gli indici di gramscianità e non gramscianità al variare della lunghezza degli n-grammi.

riferimento; al j-esimo testo di Gramsci della lista viene assegnato il punteggio $k(j)/j$, dove $k(j)$ è la sua posizione nella lista (in altre parole al sua posizione in classifica); l'indice di gramscianità $g(X)$ è la somma dei punteggi dei 49 testi di Gramsci che compaiono nella classifica. Si definisce anche *l'indice di non gramscianità* $ng(X)$ di X, attraverso la somma di analoghi punteggi dei primi 49 testi non gramsciani.

L'indice di gramscianità $g(X)$ sarà tanto più piccolo quanto più il testo incognito risulterà vicino al gruppo dei testi gramsciani ($ng(X)$ ha l'analoga proprietà per i testi non gramsciani). Il testo verrà quindi attribuito a Gramsci se l'indice di gramscianità $g(X)$ è inferiore all'indice di non gramscianità $ng(X)$.

La figura 2 riassume, con le stesse modalità usate in figura 1, i risultati dell'esperimento sul corpus dei 100 testi ottenuti utilizzando il metodo degli indici con n-grammi di lunghezza da 1 a 9.

Si osservi come neanche in questo caso vi sia un valore di n che massimizza il numero di testi gramsciani e non gramsciani correttamente attribuiti. I risultati però suggeriscono $n = 7$ o $n = 8$ come valori del parametro per costruire il metodo di attribuzione, infatti in tal caso si hanno gli ottimi valori per i testi di Gramsci riconosciuti (43 testi su 50) e l'assenza di falsi positivi.

Utilizzare questi indici ha anche un altro vantaggio: la loro differenza dà in modo naturale una misura dell'affidabilità dell'attribuzione. Più precisamente, dato un articolo da attribuire X, se $g(X)$ e $ng(X)$ sono i due indici di gramscianità e non gramscianità appena definiti, il numero

$$v(X) = \frac{ng(X) - g(X)}{ng(X) + g(X)}$$

sarà sempre compreso tra -1 e 1: valori di v 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 3 è riportato il valore di v per ognuno dei 100 testi del corpus di riferimento (i testi gramsciani sono rappresentati da un cerchio rosso, quelli non gramsciani da un triangolo blu; i testi sono ordinati seguendo la numerazione delle tabelle 2 e 3): si può facilmente

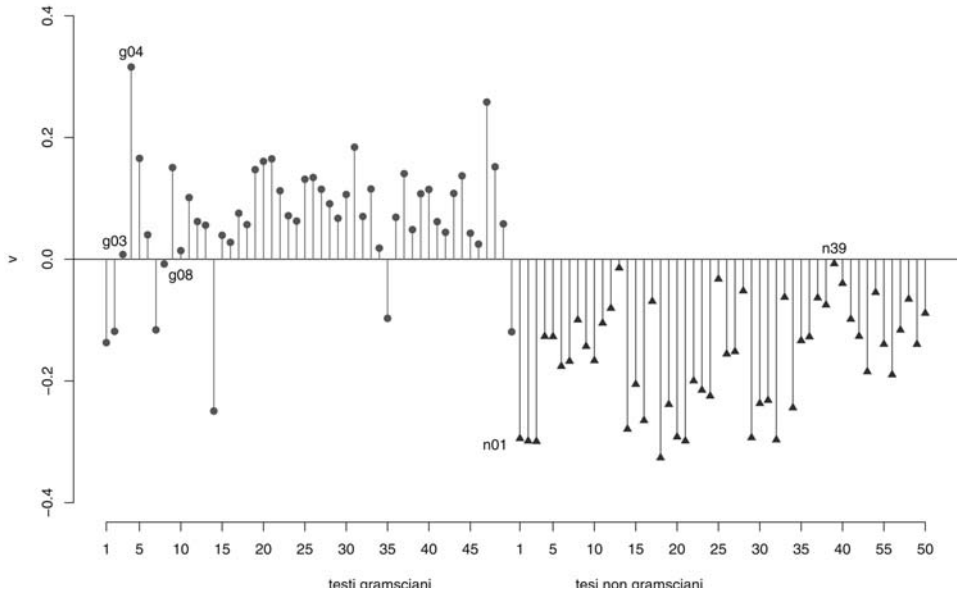

Fig. 3. – Metodo degli n-grammi modificato con n = 8, attribuzioni e indici di gramscianità e di non gramscianità.

osservare come l'attribuzione di alcuni testi, per esempio *gram_03*, *gram_08*, *nongram_39*, sia molto meno certa rispetto a quella di *gram_04* o di *nongram_01*.

L'efficacia del metodo si può sintetizzare in un unico valore calcolando la cosidetta *misura F* (o *F-score*) definita come la media armonica tra "precisione" e "richiamo":

$$\text{misura F} = 2 \times \text{precisione} \times \text{richiamo} / (\text{precisione} + \text{richiamo})$$

dove, se VP è il numero di veri positivi, VN il numero dei veri negativi, e FP e FN sono rispettivamente il numero di falsi positivi e dei falsi negativi

$$\text{precisione} = VP/(VP+FP), \quad \text{richiamo} = VP/(VP+FN).$$

Considerando che la misura F varia tra 0,5 (nel caso in cui ogni testo sia attribuito in modo casuale a Gramsci o a "non Gramsci") e 1 (quando FP=FN=0), l'efficacia del nostro metodo degli 8-grammi con il voto sui testi di riferimento è molto buona avendo una misura F di 0,92.

3.3 – *Il metodo dell'entropia relativa*

Come già descritto, la misura dell'entropia di un singolo testo si può ottenere semplicemente comprimendo il testo e misurandone il rapporto di compressione, inoltre l'entropia relativa può essere un indicatore di vicinanza di testi. Alcuni algoritmi di compressione, ed in particolare quello che ci apprestiamo a descrivere, permettono anche di ottenere stime dell'entropia relativa tra due testi, e dunque di misurarne la vicinanza.

Nel 1977 Ziv e Lempel in [24] introducono l'algoritmo LZ77, che è alla base dei programmi di compressione Zip/Winzip/gzip. È un grande salto nella storia degli algoritmi di compressione: i messaggi non vengono più codificati un carattere alla volta e neanche un numero fissato di caratteri alla volta. L'algoritmo LZ77 comprime una stringa sequenzialmente, partendo dall'inizio secondo questo algoritmo:

– se un carattere non è stato ancora incontrato, lo riscrive come è;
– se un carattere è già stato incontrato, cerca la sottostringa più lunga, già incontrata, che eguaglia la sottostringa che inizia con il carattere appena letto; scrive la lunghezza della sottostringa trovata e il numero di caratteri che la separano dall'altra.

La stringa è dunque complessivamente codificata in una sequenza di caratteri e coppie di numeri. Conviene fare un esempio: consideriamo la seguente frase, in cui lo spazio è stato per comodità sostituito con il trattino basso "_":

```
lascia_l'ascia_all'uscio
```

Iniziando da sinistra, non si trovano caratteri ripetuti fino alla seconda "a", dunque il compressore scrive "lasci"; la seconda "a" è un carattere che è stato già incontrato, ma la coppia "a_" no, dunque la seconda "a" viene codificata come (1,4): infatti il singolo carattere che stiamo leggendo è lo stesso che si trova andando indietro di 4. Segue "_" che è un nuovo carattere, e, dopo, la seconda "l" che viene codificata come (1,7); segue l'apostrofo, che è ancora un nuovo carattere, e poi l'intera sequenza "ascia_", che viene codificata con la coppia di numeri (6,8), infatti la sequenza di 6 caratteri che inizia con "a" è identica

alla sequenza di pari lunghezza che inizia 8 caratteri prima. Il risultato finale è il seguente:

testo:	lasci	a	_	l	′	ascia_	a	l	l′	u	sci	o
codifica:	lasci	(1,4)	_	(1,7)	′	(6,8)	(1,2)	(1,9)	(2,10)	u	(10,3)	o

La seconda sequenza (codifica) ha esattamente lo stesso contenuto della prima (testo): durante la decodifica, il decompressore interpreta la coppia di numeri (x,y) come l'istruzione "torna indietro di y caratteri e da lì in avanti copiane x". Quanto più lunghe saranno le stringhe trovate, tanto più piccola sarà la dimensione del file compresso. In altre parole LZ77 utilizza le ridondanze (ripetizioni) interne al testo per scriverne una versione più corta.

Il funzionamento di LZ77 ha suggerito a Benedetto, Caglioti, Loreto [2] un metodo per stimare l'entropia relativa, e dunque la "vicinanza" tra testi: supponiamo di comprimere il testo A+X, cioè il testo che si ottiene mettendo in sequenza il testo A e il testo X. L'algoritmo di compressione, che è sequenziale, codificherà prima tutti caratteri di A e poi inizierà a codificare i caratteri di X, cercando le stringhe nella parte già letta, cioè dentro il testo A. Tanto più i due testi saranno simili tanto più lunghe saranno le stringhe di X trovate in A, e quindi più efficacemente sarà compresso il file complessivo. Il compressore infatti può in questo caso utilizzare non solo la ridondanza all'interno dei singoli testi, ma anche la ridondanza tra i due testi, migliorando il rapporto di compressione. La differenza di lunghezza tra la versione compressa del testo A+X e del testo A, divisa per la lunghezza di X, è una misura dell'entropia relativa del testo X rispetto ad A. Tale numero è tanto più piccolo quante più parti di X vengono trovate in A o, in modo più suggestivo, quanto più facile è esprimere il contenuto del testo X conoscendo il contenuto di A (per un'analisi dettagliata di cosa succede quando si comprime un file seguito da un altro si veda [19]).

Il metodo descritto è effettivamente implementabile attraverso l'uso di winzip/gzip, e dà risultati ragionevoli. Va però osservato che nei compressori che fanno uso delle idee di LZ77, alla fase di codifica ne segue un'altra in cui opportuni algoritmi ricodificano le coppie di numeri per risparmiare ulteriori bit. Abbiamo dunque realizzato un

programma in cui questa ricodifica viene ottimizzata per migliorare le capacità di attribuzione.

I risultati sui 100 testi utilizzati nella fase preliminare di messa a punto dell'attribuzione gramsciana non sono sufficientemente buoni: 32 veri positivi e 14 falsi positivi (misura F = 0.67). Il fatto è che il metodo entropico è molto sensibile alla dimensione dei file dei testi di confronto. In generale tutti i metodi che confrontano singoli testi tendono a scegliere tra quelli di dimensione maggiore il testo più vicino al testo incognito. Testi grandi, infatti, sono relativamente più ricchi di statistica e informazione, e quindi hanno maggiore possibilità di avere caratteristiche in comune con i testi incogniti. D'altra parte, mentre per gli n-grammi i testi che appaiono per primi nelle attribuzioni hanno una dimensione che è circa 1,5 volte quella media dei 100 testi, per il metodo entropico questo rapporto è superiore a 2.

Si procede dunque utilizzando come testi di confronto dei testi "riassemblati": il corpus di riferimento, per esempio dei testi gram-

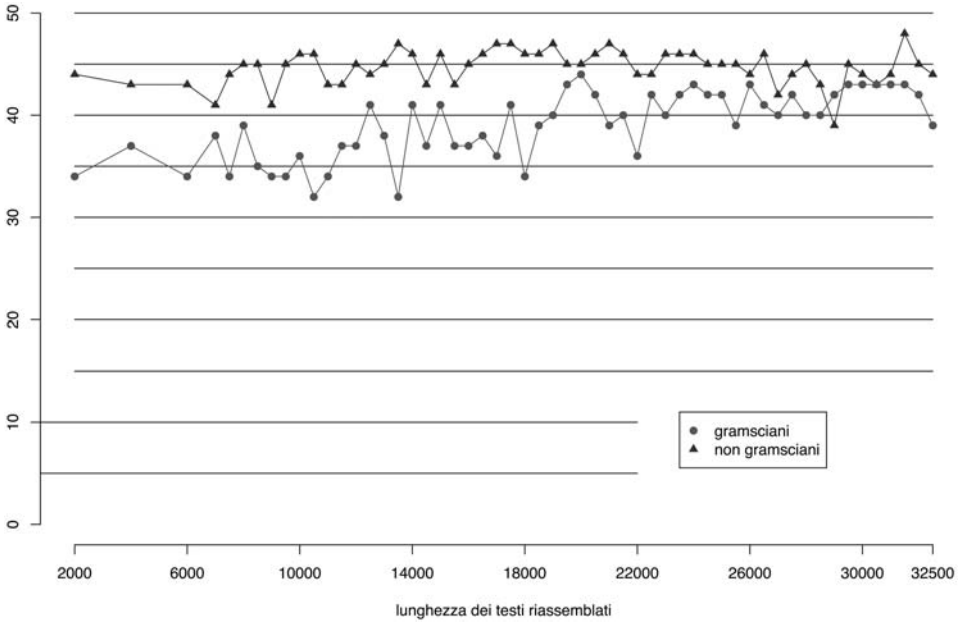

Fig. 4. – Metodo entropico, attribuzioni corrette al variare della lunghezza dei testi riassemblati.

sciani, è stato prima unificato in un solo grande file e poi successivamente tagliato in tante porzioni uguali (perdendo in questo modo la divisione originaria in articoli). Questi nuovi file di eguale dimensione sono così divenuti il corpus di riferimento. I risultati sono migliorati sensibilmente; li riportiamo in figura 4, al variare della lunghezza scelta per i testi di confronto.

Ai risultati di questi metodi si aggiunge la votazione descritta per gli n-grammi, ma limitando in questo caso la somma dei punteggi ai primi tre testi gramsciani e ai primi tre testi non gramsciani classificati: il miglior risultato si ottiene per 29 500 byte di lunghezza, ed è di 46 veri positivi e un falso positivo (misura $F = 0{,}95$).

4. – La procedura complessiva ed il test cieco

La strategia complessiva di attribuzione si basa dunque sui due metodi descritti:

8-grammi con voto esteso a tutti i testi di riferimento;

entropia relativa con testi riassemblati di 29 500 caratteri e voto per i primi tre classificati.

I due metodi funzionano in base a principi completamente diversi. D'altra parte si può temere che nei fatti diano le stesse indicazioni, senza nulla aggiungere alla accuratezza del metodo complessivo. Ci siamo dunque accertati, con metodi opportuni, dell'indipendenza statistica delle classifiche di vicinanza tra i testi date dai due metodi.

Complessivamente, infine, abbiamo attribuito a Gramsci i soli testi che tutti e due i metodi attribuiscono a Gramsci. Inoltre, entrambi i metodi danno un valore numerico all'attribuzione, dunque è possibile (e molto utile) una rappresentazione grafica dei risultati, che riportiamo in figura 5.

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

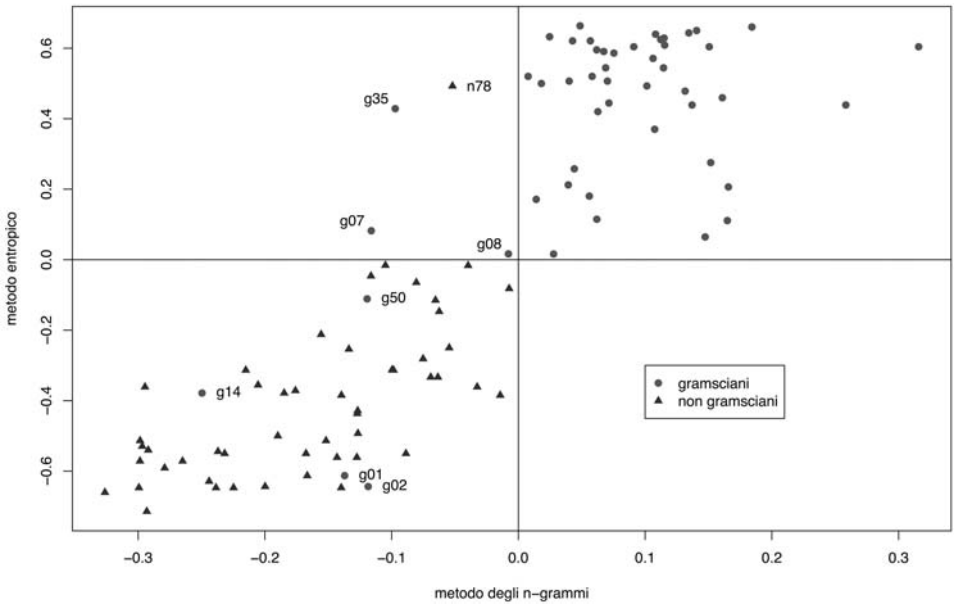

Fig. 5. – Attribuzioni dei 100 testi per la messa a punto, al termine della fase preliminare.

indice fornito dal metodo dell'entropia relativa; in tal caso procedendo dal basso verso l'altro si va da testi meno attribuibili a Gramsci a testi più attribuibili a Gramsci.

Nel quadrante in altro a destra ci sono dunque i testi che entrambi i metodi attribuiscono a Gramsci. Tra di essi non vi è nessun triangolo blu, dunque non c'è nessun falso positivo (nessuna falsa attribuzione a Gramsci). In totale i testi correttamente attribuiti a Gramsci sono 43, pari all'86%. Nel quadrante in altro a sinistra ci sono i testi attribuiti a Gramsci dal metodo dell'entropia relativa, ma non attribuiti a Gramsci dal metodo degli n-grammi: sono i testi di Gramsci n. 7, 8, 35 (quello molto vicino all'origine è il numero 8).

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. Tra di essi ci sono i testi di Gramsci n. 1, 2, 14, 50.

Abbiamo applicato questa procedura per attribuire 40 ulteriori testi, consegnati anonimi dalla Commissione Nazionale. In Tabella 6 ne

riportiamo l'elenco, con la loro indicazione bibliografica, che ovviamente non ci era nota al momento del test cieco.

TABELLA 6: autori e titoli dei 40 testi utilizzati per il 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. 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.
l6. 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!", 27 luglio 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.

L'applicazione del metodo a questi 40 testi dà il risultato mostrato in figura 6 (i punti-testo sono accompagnati dal numero che li identifica nell'elenco), costruita con lo stesso procedimento usato per la figura 5 con i 100 testi di messa a punto: 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.

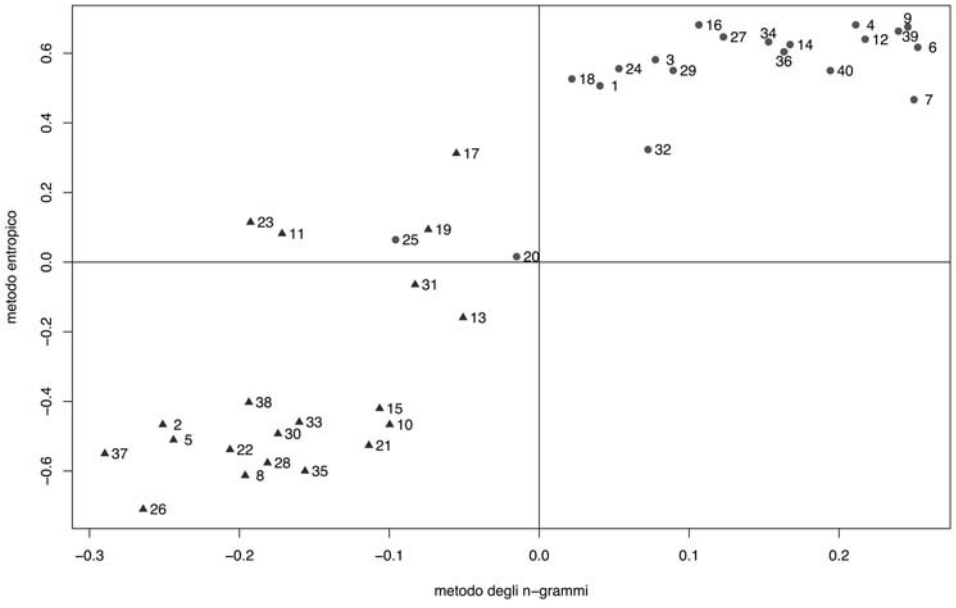

Fig. 6. – Attribuzioni dei 40 testi per il test cieco.

Come si può osservare vengono correttamente attribuiti 18 testi gramsciani su 20, pari al 90%, senza falsi positivi. 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).

5. – Conclusioni e prospettive

I risultati del test cieco sono stati estremamente positivi, soprattutto in considerazione delle caratteristiche peculiari del problema: l'attribuzione 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.

Inoltre, nella fase di messa a punto si rischiava di commettere un insidioso errore metodologico: eccedere nell'ottimizzazione dei metodi per ottenere un perfetto sistema per l'attribuzione dei testi iniziali, ma privo della flessibilità necessaria per attribuire testi simili ma differenti. Il test cieco ha anche avuto la funzione di indicare che la procedura opera altrettanto bene anche su un diverso campione.

Negli ultimi mesi, il lavoro è proseguito ed è entrato in una fase di analisi sistematica, dove centinaia di articoli, probabilmente gramsciani, sono stati analizzati ed i risultati passati agli studiosi della *Fondazione Istituto Gramsci* per considerazioni filologiche molto più accurate e assolutamente indispensabili e necessarie per una attribuzione più solida e conclusiva. Pur non essendo questi risultati ancora pubblicati, riscontriamo (senza una vera validazione scientifica) un'ottima concordanza tra le nostre *attribuzioni* e le opinioni degli esperti di scritti gramsciani, a conferma, almeno dal nostro punto di vista, dell'utilità e delle potenzialità che scaturiscono dall'interazione della matematica con discipline estremamente diverse.

Comunque è necessario non trascurare altre 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 devono ragionevolmente essere ricalibrati per le diverse an-

nate di articoli. D'altra parte siamo rassicurati dall'aver seguito un protocollo rigoroso che ha operato fin dall'inizio in modo efficace. Le attribuzioni poi non sono recepite come oracoli indiscutibili, ma i testi attribuiti vengono valutati dai curatori che decideranno se introdurli nell'edizione.

Dal punto di vista matematico, ci sono varie osservazioni da fare. Per quel che riguarda il metodo dell'entropia relativa, sviluppi interessanti potrebbero venire da alcuni nuovi algoritmi per misurare l'entropia. In particolare, quello suggerito recentemente da Grassberger [7] permette di selezionare le sequenze di caratteri importanti per un testo. Inoltre, anche il metodo di entropia si basa sugli n-grammi, semplicemente non ne viene fissata a priori la lunghezza. Questo fatto suggerisce di sperimentare limitazioni per la lunghezza delle stringhe codificate dagli algoritmi di compressione.

Il metodo degli n-grammi pone questioni di altra natura. 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 inziali di una frase; per n-grammi sufficientemente lunghi, anche le frequenze di coppie o terne di parole.

D'altra parte gli esperimenti indicano un valore ottimale ($n = 8$) decisamente grande rispetto alla dimensione dei testi. Per chiarire questo punto si può fare il confronto tra bigrammi ed 8-grammi. I bigrammi presenti nei testi sono praticamente gli stessi, a parte qualche rara eccezione. Inoltre ogni bigramma compare più volte nei singoli testi, e dunque la misura della loro frequenza è statisticamente significativa. Al contrario, nel confronto tra due testi (per esempio *gram_26* e *gram_27*), l'87% degli 8-grammi compare una sola volta e il 95% compare in uno solo dei due testi. In altre parole: "l'alfabeto di 8-grammi" in cui sono scritti i due testi ha solo una piccola porzione in comune, e la frequenza del singolo 8-gramma non è in genere statisticamente significativa. Inoltre, nel complesso dei testi molti 8-grammi compaiono più volte, ma sempre abbastanza al di sotto della soglia di significatività statistica.

Questa analisi preliminare indica interessanti direzioni di ricerca: il valore $n = 8$ sembra collocare il metodo degli n-grammi, in questo caso, 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 ([8]). Riteniamo che nello studio di queste "regioni intermedie" la matematica e le altre discipline potranno fare, insieme, significativi passi avanti.

RIFERIMENTI BIBLIOGRAFICI

[1] C. Basile - D. Benedetto - E. Caglioti - M. Degli Esposti, *An example of mathematical authorship attribution*, Journal of Mathematical Physics, **49**, 1-20 (2008).

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

[3] W. R. Bennett, *Scientific and engineering problem-solving with the computer*, Prentice-Hall, Inc. Englewood Cliffs, New Jersey (1976).

[4] L. L. Cavalli-Sforza - P. Menozzi - A. Piazza, *Storia e geografia dei geni umani*, Milano, Adelphi 2000.

[5] R. Clement - D. Sharp, *Ngram and Bayesian Classification of Documents for Topic and Authorship*, Lit. Ling. Comp. **18**, n. 4 423 (2003).

[6] A. de Morgan, in *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters*, (Longman's Green and Co., London, 1851/1882).

[7] P. Grassberger, *Data compression and entropy estimates by non-sequential recursive pair substitution*, ArXiv:physics/0207023

[8] J. W. Grieve, *Quantitative Authorship Attribution: a History and an Evaluation of Techniques*. http://hdl.handle.net/1892/2055, Lit. Ling. Comp. **22**, 251 (2007).

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

([8]) Per esempio una proteina è una sequenza di numerosi amminoacidi, che non sono così pochi da poter essere studiati con le leggi della meccanica, né sono così tanti da poter essere studiati con le leggi della termodinamica.

[10] P. Juola, *Authorship Attribution*, Foundations and Trends in Information Retrieval, vol. **1**, no. 3, 233-334 (2006).
[11] D. V. Khmelev - 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.
[12] 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.
[13] V. Keselj - N. Cercone, *CNG Method with Weighted Voting* Ad-hoc Authorship Attribution Competition (AAAC), June 2004. Part of ALLC/ACH 2004 conference.
[14] D. V. Khmelev - F. J. Tweedie, *Using Markov Chains for Identification of Writers*, Lit. Ling. Comp. **16**, 3: 299-307 (2001).
[15] A. A. Markov, *Primer statisticheskogo issledovanija nad tekstom 'Evgenija Onegina' illjustrirujuschij svjaz' ispytanij v tsep. (An example of statistical study on the text of 'Eugene Onegin' illustrating the linking of events to a chain.)*, Izvestija Imp. Akademii nauk **VI**, 153-162 (1913).
[16] A. A. Markov, *Ob odnom primeneni statisticheskogo metoda. (On some application of statistical method)*, Izvestija Imp. Akademii nauk serija **VI**, 4: 239-42 (1916).
[17] T. C. Mendenhall, *The characteristic curves of composition*, Science, vol. **IX**, 237-249 (1887).
[18] J. R. Pierce, *La Teoria dell'Informazione*, Milano, Mondadori, 1963.
[19] A. Puglisi - D. Benedetto - E. Caglioti - V. Loreto - A. Vulpiani, *Data compression and learning in time sequences analysis*, Phys. D 180, no. 1-2, 92-107 (2003).
[20] C. E. Shannon, *A Mathematical Theory of Communication*, The Bell System Technical Journal **27**, 1948, p. 623.
[21] 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.
[22] I. H. Witten - A. Moffat - T. C. Bell, *Managing Gigabytes*, second edition, Morgan Kaufmann Publishers, 1999.
[23] A. D. Wyner, *Typical sequences and all that: Entropy, Pattern Matching and Data Compression*, 1994 Shannon Lecture, IEEE Information Theory Society Newsletter, July 1995.
[24] J. Ziv - A. Lempel, *A universal algorithm for sequential data compression*, IEEE Transactions on Information Theory, IT-23 no. 3, pagg. 337-343 (1977).
[25] 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.

Chiara Basile, Mirko Degli Esposti
Dipartimento di Matematica, Università di Bologna
e-mail: basile@dm.unibo.it desposti@dm.unibo.it

Dario Benedetto, Emanuele Caglioti
Dipartimento di Matematica, Università di Roma "La Sapienza"
e-mail: benedetto@mat.uniroma1.it caglioti@mat.uniroma1.it