INDIVIDUARE SCRITTI GRAMSCIANI ANONIMI IN UN *CORPUS* GIORNALISTICO. IL RUOLO DEI METODI QUANTITATIVI

Maurizio Lana

1. *Presentazione*. Nei suoi primi di attività Gramsci collaborò con numerosi giornali torinesi, sui quali non era inconsueto che gli articoli fossero pubblicati anonimi. All'interno del *corpus* costituito dagli articoli giornalistici anonimi pubblicati negli anni 1913-1926 su tali giornali («Il Grido del Popolo», «Avanti!», «La Città Futura») si trova quindi certamente un certo numero di scritti gramsciani.
Per questo, nel quadro delle attività editoriali per l'Edizione nazionale delle opere di Gramsci, il presidente della Fondazione Istituto Gramsci di Roma cercò nel 2005 uno o piú esperti in attribuzione di testi allo scopo di individuare gli scritti gramsciani all'interno di quel *corpus* di scritti anonimi, con l'intento di offrire ai curatori delle varie annate degli scritti di Gramsci nuovo materiale da valutare.
Nelle pagine che seguono viene delineato il piú ampio contesto dell'attribuzione di testi con metodi quantitativi, nel quale si colloca il lavoro di attribuzione sui testi gramsciani giunti anonimi realizzato da chi scrive insieme con i colleghi Basile, Benedetto, Caglioti, Degli Esposti. Il resoconto del loro contributo a questa ricerca è stato pubblicato su riviste scientifiche del settore[1].
La prima parte di questo contributo delineerà sinteticamente aspetti metodologici e storici e problematicità dell'attribuzione di testi con metodi quantitativi; la seconda esporrà con maggiore dettaglio le questioni connesse con l'attribuzione dei testi gramsciani pubblicati sui quotidiani. Della storia dell'attribuzione di testi con metodi quantitativi sono oggi disponibili alcune visioni d'insieme (Holmes[2], Craig[3], Love[4]) ma non esiste una ricostruzione condivisa,

[1] C. Basile, D. Benedetto, E. Caglioti, M. Degli Esposti, *An example of mathematical authorship attribution*, in «Journal of Mathematical Physics», 2008, 49, pp. 1-20; C. Basile, D. Benedetto, E. Caglioti, M. Degli Esposti, *L'attribuzione dei testi gramsciani: metodi e modelli matematici*, in «La Matematica nella Società e nella Cultura», 2010, 3, pp. 235-269.
[2] D.I. Holmes, *The Evolution of Stylometry in Humanities Scholarship*, in «Literary and Linguistic Computing», XIII, 1998, 3, pp. 111-117.
[3] H. Craig, *Analysis and Authorship Studies*, in «Companion to Digital Humanities», ed. by S. Schreibman, R. Siemens, J. Unsworth, London, Blackwell, 2005, pp. 273-280.
[4] H. Love, *Attributing Authorship. An Introduction*, Cambridge, Cambridge University Press, 2002.

almeno a grandi linee, di ciò che è accaduto. I cenni di ricostruzione storica che presenteremo nelle pagine successive mostrano che l'analisi quantitativa di tipo matematico (da non confondere con *statistico*) negli studi di attribuzione, adottata per lo studio degli scritti giornalistici di Gramsci giunti anonimi, costituisce lo sviluppo attuale di *una* specifica linea di ricerca presente da tempo in questo campo (descrivere anche solo per cenni le altre linee di ricerca andrebbe al di là delle finalità di questo articolo). Ciò che è specifico del gruppo di ricerca è il fatto che cerchiamo di «far interagire metodi automatici con approcci filologici» (Degli Esposti).

2. *Tra le due culture*. Individuare gli scritti gramsciani all'interno di questo *corpus* di scritti anonimi ricorrendo agli strumenti tradizionali della filologia e della critica storica è reso difficile dal concorso di una serie di fatti: non si possiedono minute o originali autografi degli articoli; gli articoli talora subivano interventi della censura che li mutilava; in quegli stessi anni collaboravano ai medesimi giornali e sui medesimi temi altri giornalisti, pensatori, politici quali Serrati, Tasca, Togliatti, Bordiga, Bianchi, i cui articoli ugualmente venivano talora pubblicati senza firma; i temi e la collocazione ideologica degli articoli sono molto simili nei vari autori; gli articoli, data la loro natura, sono scritti talora brevi.
La scelta di lavorare con metodi matematici sul problema dell'attribuzione dei testi è innovativa e come tale porta con sé una dose di rischio, difficile da quantificare ma sicuramente presente; ma come accade in ogni ricerca, se non si assumono – in modo attento, misurato – dei rischi non si fanno passi avanti. Il passo avanti dal punto di vista dello studio di Gramsci e del suo pensiero è nel fatto che grazie al concorso del lavoro di linguisti e matematici e storici (senza nascondere che all'acutezza critica dello storico spetta l'ultima parola) sarà possibile pubblicare un certo numero di scritti di Gramsci fino ad ora non riconosciuti come suoi e ciò potrà ampliare le prospettive degli studiosi; mentre nell'ambito dell'attribuzione di testi con metodi quantitativi questa ricerca ha fatto nascere un gruppo di lavoro italiano, fatto innovativo anche questo, e ha portato a mettere a punto tecniche di analisi nuove ed efficaci.
Il contributo dato dall'attribuzione di testi con metodi quantitativi allo studio dei testi gramsciani si colloca in una prospettiva simile a quella dell'apporto della filologia allo studio dei testi: preparare il testo per l'edizione concorrendo con altre ad una migliore conoscenza e comprensione del contenuto dello scritto e favorendone una corretta interpretazione. È chiaro che ricercare e individuare nuovi testi gramsciani fino ad ora sconosciuti è un'attività non neutrale: attribuire un testo a un autore significa definire un contesto per il testo e quindi renderne possibile lo studio e la comprensione.
L'attribuzione di testi per mezzo di analisi quantitative ha una storia breve ma interessante. Per quanto è oggi noto ha un precedente nel 1400 con Lorenzo

Valla[5], nasce nel 1851 con Wincenty Lutosławski (cfr. il paragrafo 4 qui sotto) e interseca la statistica, la filologia, la matematica, la critica letteraria, l'informatica, la cladistica (disciplina che ricostruisce le parentele genetiche tra organismi viventi), la fisica e altri ambiti disciplinari ancora. Chi la pratica opera in concreto, anche senza programmi formalmente definiti, l'incontro delle due culture, come le chiamò Ch.P. Snow[6] (ma prima di lui già nel 1882 Matthew Arnold[7]) che oggi forse potrebbero essere dette scienze formali e scienze empiriche, o discipline scientifiche e discipline umanistiche. Se oggi non esiste piú un pregiudizio di aridità e meccanicità degli umanisti verso gli scienziati e le loro ricerche, e di inconsistenza e infondatezza da parte degli scienziati verso gli umanisti e le loro ricerche, rimane vero che le scienze empiriche spesso si trovano a dover giustificare la loro minore precisione e rigore di fronte alle scienze formali e che la ricerca procede sostanzialmente per compartimenti stagni anche perché i territori di frontiera in cui gli uni e gli altri possono fare ricerca insieme non sono facili da individuare e quindi non sono frequentati. Non si può fare a meno di notare che quando si parla di umanisti e scienziati, implicitamente ma chiaramente si afferma che gli umanisti non sono scienziati, il che potrebbe essere ovvio se fosse in gioco solo un'indicazione di ambito disciplinare, mentre si adombra l'affermazione che nell'area umanistica non si fa scienza, che il discorso delle discipline umanistiche non è scientifico, cioè non è rigoroso, non è metodologicamente fondato (e si noti per converso come le discipline umanistiche vengano talvolta chiamate *scienze* umane, come a superare il divario almeno dal punto di vista terminologico). Non è questione di sudditanza psicologica, o disciplinare, o di incapacità a riconoscere le specificità irriducibili del proprio dominio di studi; semmai si tratta di dare fondamenti analitici rigorosi all'attività di studio dei testi – centrale nell'area umanistica – che costituisce la premessa per l'interpretazione dei testi stessi.
L'attribuzione di testi è uno di questi territori di frontiera in cui direttamente entrano in contatto i due ambiti disciplinari, territorio aperto perché non esistono metodi affermati globalmente, metodologie consolidate e provate al

[5] L. Valla, *La falsa Donazione di Costantino, Discorso di Lorenzo Valla sulla Donazione di Costantino da falsari spacciata per vera e con menzogna sostenuta per vera*, a cura di G. Pepe, Firenze, Ponte alle Grazie, 1992 (poi Milano, Tea, 1994).

[6] Sir Charles Percy Snow tenne a Cambridge nel 1959 una conferenza intitolata *The two cultures and the scientific revolution*, pubblicata in «New Statesman», 6 ottobre 1956, ma poi diffusasi grazie alla ripubblicazione su «Encounter», maggio 1959, e in «Spectator», 9 marzo 1962.

[7] M. Arnold aveva tenuto una Rede Lecture a Cambridge, pronunciandosi a sua volta sull'argomento delle due culture in una conferenza intitolata *Literature and Science* (M. Arnold, *Lectures and Essays in Criticism*, ed. by R.H. Super and T.M. Hoctor, Ann Arbor, The University of Michigan, 1973).

di là di ogni dubbio. Su tematiche che richiamano da vicino quelle di questa ricerca osserva McCarty:

> We nevertheless observe a striking, non-trivial resemblance between humanities computing and experimental science: both are data-centred, equipment-orientated activities that centrally involve modelling and tend to be collaborative. This resemblance is significant not because humanities computing requires the honourific title of a «science», rather because in establishing the field we need to ask where its kinship lies and what intellectual assistance we may derive from them[8].

La molla che mette in movimento verso l'attribuzione con metodi quantitativi è la percezione, nello studioso di scienze umane, che l'attribuzione di un testo dubbio richiede dapprima di individuare e successivamente di elencare le caratteristiche distintive del testo, per poterlo confrontare con altri (un testo dubbio di Senofonte sarà confrontato con altri di Senofonte, prima di tutto); ma l'elenco, il *catalogo* delle caratteristiche diventa rapidamente *conteggio* di quelle caratteristiche, che sono caratteristiche di stile. Nel momento in cui dal catalogo si passa al conteggio si entra in una zona dove confinano e si intersecano matematica, statistica, informatica. A questo punto lo studioso di scienze umane può eventualmente iniziare – con ottimismo e ingenuità di principiante – a studiare matematica, informatica, statistica, cladistica, per capire che cosa esattamente esse possano dare al suo ambito di studio e di ricerca e per utilizzare in proprio strumenti di analisi specifici di quegli ambiti, scoprendo rapidamente la difficoltà di padroneggiare molteplici discipline ad alto livello; oppure studiosi di scienze umane e studiosi di scienze fisiche, matematiche o naturali costituiscono un gruppo di ricerca con un interesse comune, come è avvenuto per l'attribuzione degli scritti giornalistici anonimi di Gramsci.

3. *Stilometria o attribuzione con metodi quantitativi?*. Attribuzione di testi con metodi quantitativi (*quantitative authorship attribution*) è una perifrasi appropriata a descrivere gli studi di attribuzione che si basano sul conteggio di elementi interni ai testi. Ricorrono e si incontrano intorno a questa idea molte espressioni e termini variamente interconnessi: stilometria, *statistique stylistique*, analisi stilometrica, *computational stylistic*, *authorship attribution*, *non-traditional authorship attribution research, automatic inference of authorship, statistical authorship attribution, stylistic fingerprint*. La proliferazione delle denominazioni è conseguenza della storia relativamente breve di questo tipo di studi ma è anche indice, occorre riconoscerlo, di identità incerta.
Talora si parla anche semplicemente di *stilometria*, cioè di «misurazione dello stile», fondata sulla convinzione che sia possibile misurare, quantificare, le ca-

[8] W. McCarty, *Humanities computing*, Preliminary draft entry for «The Encyclopedia of Library and Information Science», New York, Dekker, 2003, http://www.cch.kcl.ac.uk/legacy/staff/wlm/essays/encyc/.

ratteristiche stilistiche di un testo; e l'espressione *stylistic fingerprint*, «impronta digitale stilistica», contiene l'idea che lo stile di un testo sia come un'impronta digitale che ne rivela l'autore a chi sappia rilevarla, che rivela inconfondibilmente sia l'autore che si maschera (testi pseudepigrafi) sia l'autore che si nasconde (testi anonimi); non a caso ricorrono in questa prospettiva anche termini come *detection*, *prove*, *fraud*, senza dimenticare tentativi di uso della stilometria in ambito giudiziario.

Is it truly the case that any two authors can always be distinguished on the basis of their style, so that stylometry can provide unique stylistic fingerprints for any author, given sufficient data? Despite the long history of authorship attribution, almost all stylometric studies have been carried out on the assumption that stylometric fingerprinting is possible[9].

La misurazione che costituisce lo scopo della stilometria corrisponde per opposizione ai procedimenti tradizionali dello studio dei testi, saldamente fondati sulla convinzione che la comprensione del significato di un testo abbia sempre aspetti non formalizzabili in procedure e che quindi richieda in ultima analisi le capacità interpretative di un essere umano, di una mente umana. La stilometria procede quantificando le caratteristiche del testo, scelte solitamente tra le caratteristiche linguistiche ancorabili alle parole (frequenze di parole di specifici tipi, per esempio: congiunzioni, preposizioni; oppure altri tipi di misure quali lunghezza media delle parole, rapporto nomi-aggettivi, o nomi-verbi, o quello nomi-pronomi, ampiezza del lessico in rapporto alla ripetitività dell'uso dei termini [*type-token ratio*], frequenza delle sequenze di parole di lunghezza da 2 a un numero prefissato di caratteri, ecc.): misurando e contando gli elementi caratteristici dello stile si mira a scoprire le caratteristiche distintive di uno specifico autore. Alla base c'è infatti l'idea che ogni autore e ogni testo abbiano uno stile caratteristico, e che quindi testi che hanno caratteristiche stilistiche identiche siano del medesimo autore, e testi di stile differente siano di autori diversi (il modo in cui questo criterio si attua nella realtà è ovviamente ben piú complesso di questa formulazione un po' riduttiva). Il termine stile è quindi usato non nell'accezione piú frequente di *qualità estetica distintiva*, o di norma espressivo-compositiva («manuale di stile»), ma nell'accezione di *caratteristica espressiva individuale*, con un significativo trapasso dall'indicare qualcosa di inafferrabile benché riconoscibile, all'indicare una realtà che si presume indagabile e misurabile benché rispondere alle domande: *che cos'è lo stile? in che cosa consiste lo stile?* non sia per nulla semplice (ai problemi di tipo teorico si aggiunge, per la stilometria, che non sono utilizzabili gli elementi di

[9] H. Baayen *et al.*, *An experiment in authorship attribution*, in *JADT 2002: 6es Journées internationales d'Analyse statistique des Données Textuelles*: http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2002/PDF-2002/baayen_vanhalteren_neijt_tweedie.pdf.

stile non formalizzabili, non descrivibili come specifici oggetti individuabili con procedure definite e non ambigue). La stilometria assume che ogni autore abbia aspetti di stile consci e inconsci (il lessico connesso con il contenuto dello scritto è un carattere stilistico che deriva prevalentemente da scelte conscie; mentre la scelta, per esempio di usare una delle due forme *tra* e *fra* è un carattere stilistico che deriva da scelte inconscie):

At [the heart of stylometry] lies the assumption that authors have an unconscious aspect to their style, an aspect which cannot consciously be manipulated but which possesses features which are quantifiable and which may be distinctive. Bailey lists the general properties which quantifiable features of a text should possess: «They should be salient, structural, frequent and easily quantifiable, and relatively immune from conscious control». By measuring and counting these features, stylometrists hope to uncover the «characteristics» of an author[10].

Costituisce una questione complessa e non risolta se gli aspetti di stile inconsci siano stabili nel corso del tempo, oppure siano soggetti ad evoluzione.
Nella stilometria c'è forte attenzione sull'attribuzione dei testi per mezzo delle caratteristiche di stile inconscie, perché – sillogisticamente – *se* le caratteristiche di stile inconscie sono strettamente connesse con l'identità dell'autore, *e se* le caratteristiche inconscie reperite in una serie di testi di un autore si ripresentano in uno o piú testi dubbi, *allora* si può concludere che i testi dubbi sono di quell'autore, con tutta la forza logica derivante dal procedimento (come se l'astratta chiarezza razionale del sillogismo in qualche modo si riverberasse sul caso concreto – concreto e per ciò stesso inevitabilmente impastato di ambiguità e incertezza ineliminabili – dell'attribuzione di un dato testo ad un dato autore).

The search for stylistic markers which are outside the conscious control of the writer has led to a divergence between literary interpretation and stylometry, since, as Horton puts it, «the textual features that stand out to a literary scholar usually reflect a writer's conscious stylistic decisions and are thus open to imitation, deliberate or otherwise» (quoted in Lancashire 1998: 300). In support of these unconscious style markers are the studies that show that much of language production is done by parts of the brain which act in such swift and complex ways that they can be called a true linguistic unconscious (Crane 2001: 18). Lancashire (1997: 177) adds: «This is not to say that we cannot ourselves form a sentence mentally, edit it in memory, and then speak it or write it, just that the process is so arduous, time-consuming, and awkward that we seldom strive to use it»[11].

[10] Holmes, *The Evolution of Stylometry in Humanities Scholarship*, cit., p. 111.

[11] Craig, *Analysis and Authorship Studies*, cit., p. 285. Le citazioni interne al passo provengono nell'ordine da: I. Lancashire, *Paradigms of Authorship*, in «Shakespeare Studies», XXVI, 1998, p. 300; M.T. Crane, *Shakespeare's Brain: Reading with Cognitive Theory*, Princeton,

A fronte di tutte queste questioni, prima delle quali è l'assenza di consenso degli studiosi su quali debbano essere gli indicatori di stile da misurare e se e come cambi lo stile di un autore nel corso del tempo e nel passaggio da un genere ad un altro, da un contenuto ad un altro, sembra preferibile un atteggiamento piú cauto che pur riconoscendo – come anche questa ricerca mostra – che con metodi quantitativi si possono raggiungere risultati interessanti nell'attribuzione di testi, siamo ben lungi dall'aver individuato e chiarito senza margini di dubbio che cosa permette il riconoscimento della paternità di un testo e come si possano confrontare due testi per individuare se abbiano il medesimo autore. Appare sostenibile e scientificamente piú corretto un approccio piú cauto che non pretenda di individuare indicatori e misure assolute dello stile ma che costruisca un metodo complesso capace di classificare i testi in base alla loro somiglianza e differenza (alla loro vicinanza e distanza) con testi di riferimento senza aspirare a, o pretendere di, individuare la misura dello stile di un testo, costantemente ridefinendo la misura della somiglianza tra un nuovo testo sconosciuto e i campioni di riferimento noti. Il metodo utilizzato è stato quindi, in termini generali, quello del costante confronto dei testi anonimi con due gruppi di testi: uno di testi gramsciani e uno di testi di altri autori che scrivevano sui medesimi giornali, in modo da individuare comparativamente le somiglianze e differenze.

Un'altra questione delicata nell'attribuzione di testi è quella dell'interferenza tra caratteristiche di stile autoriali e caratteristiche di stile connesse con il contenuto, descritta in modo efficace da Ross Clement e David Sharp:

> Documents, be they articles, plays, fiction or nonfiction books do not only contain markers indicating authorship, but also include markers indicating the topic of the document. […] both authorship attribution and topic identification can sometimes be seen as classification task, i.e. we assume a document is by one of a number of authors (or is on one of a number of topics), and we wish to examine features of the document to decide to which class (author or topic) it should be assigned. […] In a crude sense, analysis of documents from the viewpoints of authorship or topic can be viewed as measuring different signals in the document. Compare this with a recording of a sung song that encodes two separate (lyrics and melody) signals in a single waveform. Both K[knowledge] M[anagement] and A[uthorship] A[ttribution] are in effect trying to extract one of the two (topic and authorship) signals from a document while ignoring the other. […]
>
> Darwin's frequent use of words related to biology may appear to be an authorship signal when compared with the works of Conan Doyle. While the use of these content markers may often predict the correct author, a work such as Conan Doyle's *The Lost World* including frequent reference to dinosaurs may be incorrectly attributed. There is no reason to believe that any feature of a document will be exclusively associated

Princeton University Press, 2001, p. 18; I. Lancashire, *Empirically Determining the Shakespeare's Idiolect*, in «Skasperare's Studies», XXV, 1997, p. 177.

with either a topic or authorship signal. […] To make things even messier, it is also not certain that there are just two signals present in a document. There may be many, dependent, independent, or partially dependent, signals in a document, supplying evidence for either topic or authorship. For example, character usage (e.g. bigrams) may, or may not be completely independent of sentence length variation[12].

La questione del riconoscimento dello stile dell'autore piuttosto che del contenuto del testo è di grande importanza: se si fa una ricerca di attribuzione con metodi quantitativi su un *corpus* di autori differenti e di testi differenti per contenuto, è difficile escludere che ciò che può passare per riconoscimento degli autori sia invece classificazione di contenuti (ad esempio se si analizza una serie di capitoli di romanzi inglesi e in alcuni di essi ricorrono i nomi Robinson e Venerdí non occorre grande finezza di analisi per individuare che quei capitoli fanno parte di una medesima opera: il riconoscimento dell'argomento si traduce in riconoscimento dell'autore). In altre parole la scelta dei fenomeni oggetto di misurazione concorre in modo rilevante, anche se ancora non ben chiarito, a determinare se l'attribuzione con metodi quantitativi individua le differenze di stile o le differenze di contenuto.

4. *I precursori degli attuali studi di attribuzione con metodi quantitativi*. Il primo scritto organico veramente importante in materia di attribuzione fu quello di Wincenty Lutosławski (1863-1954), polacco, filologo e studioso di filosofia platonica. Egli nel 1897 pubblicò un importante studio sulla cronologia dei dialoghi di Platone, *The origin and growth of Plato Logic*[13], basato su criteri stilometrici[14]. Il termine stilometria appare essere una coniazione originale di Lutosławski, attestato in due conferenze che egli tenne a Oxford alla Philological Society nel 1897 («He thanks the Oxford Philological Society for the attention paid to his first attempt to lecture in English, which happens also to be the first public explanation of the method of stylometry») e a Parigi all'Académie des Inscriptions et Belles-Lettres nel 1898 («Cela nous conduit à une nouvelle science de la stylométrie, puissant instrument auxiliaire de l'histoire de la pensée humaine»)[15].

[12] R. Clement, D. Sharp, *Ngram and Bayesian Classification of Documents for Topic and Authorship*, in «Literary and Linguistic Computing», XVIII, 2003, 4, pp. 423-447, p. 426.

[13] W. Lutosławski, *The Origin and Growth of Plato's Logic*, London-New York-Bombay, Longmans, Green & Co., 1897 (*reprint* Hildesheim, Georg Olms, 1983).

[14] Secondo Leonard Brandwood (*The Chronology of Plato's Dialogues*, Cambridge, Cambridge University Press, 1990) almeno altri tredici studiosi prima del 1897 avevano pubblicato studi di tipo stilometrico sulla cronologia dei dialoghi di Platone.

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

Tutto il lavoro di Lutosławski si fondava una serie di presupposti chiari ed espressamente formulati, che permettevano l'attuazione pratica dello studio stilometrico:

– si presuppone che nei testi di ogni autore esista uno stile individuale e caratteristico, e che esso sia indipendente dal contenuto;
– è possibile risolvere la questione dell'identità dell'autore sulla base delle proprietà stilistiche dei suoi testi, proprietà che sono da considerare come caratteristiche esterne;
– la stilometria è analoga alla grafologia, e ha la medesima efficacia potenziale;
– le caratteristiche di stile rilevanti sono numerose ma limitate;
– in base al numero di caratteristiche di stile condivise è possibile misurare e quantificare la somiglianza tra testi;
– lo stile di un individuo muta nel corso del tempo;
– si devono confrontare campioni testuali di eguali dimensioni;
– e altri ancora.

La ricerca di Lutosławski sulla cronologia dei dialoghi platonici fu condotta su 58.000 passi di testo nei quali venne misurata la presenza di 500 fenomeni linguistici, quali ad esempio:

– risposte denotanti assenso soggettivo meno di 1 volta su 60 risposte;
– aggettivi di grado superlativo in risposte affermative con frequenze superiori alla metà degli aggettivi di grado positivo, ma non prevalenti sui positivi;
– proposizioni interrogative con *ara* costituenti tra il 15 e il 24% di tutte le interrogative;
– preposizione *perí* collocata dopo la parola a cui si riferisce costituente piú del 20% di tutte le occorrenze di *perí*[16].

È importante notare che il marcatore stilistico veniva individuato ad un livello sintattico alto, complesso, tale per cui nemmeno oggi, pur con la disponibilità dei computer, sarebbe agevole immaginare un'individuazione automatica delle occorrenze.
Thomas Corwin Mendenhall, un fisico, fu attratto dalla tecnica di analisi dei testi basata sulla distribuzione di frequenza delle parole a causa della somiglianza con l'analisi spettroscopica, che negli ultimi decennio del XIX secolo era in primo piano nell'attenzione degli scienziati. Scrisse infatti su «Science» nel 1887:

It is proposed to analyse a composition by forming what may be called a «word spectrum» or «characteristic curve» which shall be a graphic representation of the arrangement of words according to their length and the relative frequency of their occurrence[17].

[16] Citato in A. Kenny, *The Computation of Style*, Oxford, Pergamon Press, 1982, p. 3.
[17] T.C. Mendenhall, *The characteristic curves of composition*, in «Science», XI, marzo 1887, ns-9, vol. 214, pp. 237-249; il passo citato è a p. 241.

Nell'articolo, intitolato *The characteristic curves of composition*[18], venivano mostrate le distribuzioni di frequenza delle parole di varie opere e scrittori in lingua inglese e in altre lingue. Successivamente pubblicò nel 1901 su «The Popular Science Monthly» un articolo intitolato *A mechanical solution of a literary problem*[19], in cui venivano studiati e confrontati gli scritti di Shakespeare, Marlowe e Bacone in quanto già prima di allora c'era una questione aperta sull'identità dell'autore delle opere trasmesse dalla tradizione sotto il nome di Shakespeare. Lo scopo di Mendenhall era verificare se lo stile delle opere di Shakespeare fosse unico e particolare, o se fosse simile (o identico!) a quello delle opere di Marlowe e Bacone, nella convinzione che la distribuzione di frequenza della lunghezza delle parole fosse l'indicatore di stile appropriato.

5. *Il XX secolo: Ellegård, Mosteller e Wallace*. Le *Lettere di Junius* sono una serie di 69 lettere scritte fra il 1769 e il 1772 in Inghilterra, e pubblicate dal «Public Advertiser», in cui l'autore si propone uno scopo formativo verso i lettori che vuole rendere attenti ai pericoli del potere arbitrario. La loro attribuzione fu fin dagli anni della pubblicazione variamente discussa.
Nel 1962 Alvar Ellegård, studioso svedese di storia della letteratura, studiò il problema esponendo i risultati in due volumi intitolati *Who was Junius?* e *A statistical method for determining authorship: The Junius Letters 1769-1772*[20], in cui incrociava un uso attento delle tradizionali prove storiche e biografiche con un metodo statistico di sua concezione (studio delle frequenze di 500 parole ed espressioni che nelle lettere di Junius sono o molto piú frequenti o molto meno frequenti che negli scritti dei contemporanei; ad esse si aggiungevano circa 50 termini che Junius sceglieva all'interno di coppie o terne di termini che risultavano approssimativamente sinonimi, come *on* e *upon*; *kind* e *sort*; e cosí via). Sul piano metodologico lo studio di Ellegård si caratterizzava per il fatto di lavorare su un *corpus* di testi anonimi senza alcuna indicazione di uno o piú possibili autori. Ellegård costruí dunque il *corpus* dei contemporanei che secondo lui potevano essere autori delle *Lettere di Junius*. Ma questo modo di procedere rimane aperto al dubbio che il vero autore possa essere un altro e non sia stato inserito nel gruppo dei contemporanei di cui analizzare i testi, fintantoché non si presentino altre prove o forti indizi indipendenti dai testi. Per questo Ellegård studiò anche le evidenze esterne, storiche e biografiche, che riguardavano gli autori dei testi studiati.

[18] *Ibidem*.
[19] T.C. Mendenhall, *A mechanical solution of a literary problem*, in «The Popular Science Monthly», vol. LX, 1901, n. 7, pp. 97-105.
[20] A. Ellegård, *Who was Junius?*, Stockholm, Almqvist & Wiksell, 1962, p. 159; Id., *A statistical method for determining authorship: The Junius Letters 1769-1772*, Göteborg, 1962.

Negli anni 1787-88 Alexander Hamilton, John Jay e James Madison scrissero 85 «federalist papers» per persuadere i cittadini di New York a ratificare la Costituzione degli Stati Uniti. Furono tutti pubblicati anonimi sotto lo pseudonimo «Publius». Si sa che Hamilton scrisse 51 *papers*, Madison 14, Jay 5, che 3 furono scritti da Hamilton e Madison; la paternità dei rimanenti 12 *papers* è disputata tra Hamilton e Madison.
Nel 1964 Frederick Mosteller and David Wallace pubblicarono il loro studio sull'attribuzione, *Inference and Disputed Authorship: The Federalist*[21]. Anche Mosteller e Wallace come già Ellegård cercarono parole identificative, che avessero un uso significativamente differente fra Hamilton e Madison. Dopo averne esaminate molte, Mosteller e Wallace si fermarono su 28 che avevano la piú alta efficacia discriminante: *also*, *upon*, *by*, *of*, *on*, *there*, *this*, *to*, e altre, tra le parole grammaticali; e *innovation*, *language*, *probability*, e altre tra le parole piene. Confrontando l'uso di queste parole Mosteller e Wallace conclusero che i 12 *papers* dubbi erano da attribuire a Madison.
I metodi di Mosteller e Wallace risultarono efficaci benché applicati ad un numero di indicatori stilistici molto piú ristretto rispetto a Ellegård (28 contro i 500). Sul piano metodologico la situazione con cui Mosteller e Wallace si confrontarono era piú semplice di quella delle *Lettere di Junius*: i 12 *papers* da analizzare erano con certezza o di uno o di un altro autore, per entrambi dei quali era disponibile un ampio *corpus* testuale molto vicino per molteplici caratteristiche di stile ai 12 *papers*. Pertanto il *corpus* dei testi dei candidati all'attribuzione non dovette essere costruito con un complesso lavoro storico indiziario ma era già dato.

6. *Un punto di svolta: Gerard Ledger, «Re-counting Plato»*. I metodi e gli studi precedentemente menzionati si basavano su conteggi di parole o di combinazioni di parole. Nel 1989 G.R. Ledger pubblicò uno studio su autenticità e cronologia delle opere di Platone (*Re-counting Plato: A Computer Analysis of Plato's Style*[22]) che si inserisce nella già menzionata tradizione di studi stilometrici sulle opere platoniche, ma che si caratterizza per un approccio radicalmente innovativo. L'oggetto dell'analisi non appartiene piú ai territori della linguistica ma segna il passaggio verso territori *altri*, non riconducibili agevolmente alle denominazioni disciplinari note e correnti. Nelle parole di Ledger stesso:

[21] F. Mosteller, D.L. Wallace, *Inference and Disputed Authorship: The Federalist*, Addison-Wesley, Reading, Mass. 1984. Il loro lavoro si trova anche citato con il titolo della seconda edizione: *Applied Bayesian and Classical inference: The Case of the Federalist Papers,* New York, Springer-Verlag, 1984.
[22] G.R. Ledger, *Re-counting Plato: A Computer Analysis of Plato's Style*, Oxford, Clarendon Press, 1989, p. 2.

If we accept that language is a regulated system (few would deny it), it is inconceivable that statistics could have nothing to say about it. Stylometry must have a basis in reality because of the orderly nature of the material with which it is dealing, not because authors might differ in their particular usage of selected words, or are idiosyncratic in other ways, although it is very probable that this is the case. What matters more is that there is inherently, in the very structure of language, an orderly matrix, a network of predictability. […]
I have also departed from the traditional approach of stylometry by ignoring entirely meanings and grammatical functions, measuring instead the frequencies of words according to their orthographic content.

Ledger sceglie quindi (ma si potrebbe anche dire *costruisce*) tre tipi di oggetti di analisi:

1. parole che contengono una specifica lettera;
2. parole che terminano con α, ε, η, ι, ν, ο, ς, υ, ω;
3. parole con α, δ, ε, η, ι, ο, τ, υ, ω in penultima posizione[23].

L'approccio di Ledger è innovativo e ha una ragione interna: se il linguaggio è un sistema regolato la statistica ha titolo per occuparsene; e se la statistica se ne occupa si possono cercare nuovi tipi di strutture del linguaggio, che non necessariamente sono comprensibili, o si manifestano in modo descrivibile in termini verbali, a livello semantico-grammaticale-sintattico.
Ledger costituisce quindi un punto di svolta di grande importanza per gli sviluppi successivi in quanto la sua ricerca è la prima a lasciare il livello semantico, lessicale, linguistico, grammaticale, sintattico del testo, per contare «oggetti» le cui quantità costituiscano elementi stilistici discriminanti. Tali quantità vengono poi misurate e analizzate; ma la scelta del tipo di «oggetti» è piú significativa del tipo di conteggi ed analisi. Il limite del lavoro di Ledger consiste nel fatto che occorrerebbe provare su casi noti che queste caratteristiche di stile siano parole che contengono una specifica lettera valide e significative per l'attribuzione e la cronologia di un testo. Il problema della verifica dei metodi è un'altra questione aperta, a cui abbiamo tentato di dare risposta in questa ricerca.

7. *Una struttura matematica latente dei testi*. La tendenza che si manifestava per la prima volta in evidenza nel lavoro di Ledger avrebbe via via acquistato importanza anche in relazione al fatto che i metodi stilometrici tradizionali mostrano anche altri limiti quando vengono adottati in un contesto matematico/informatico, come notava Peng nel 2003:

There are several problems with [stylometric] approach however. First, techniques used for style marker extraction are almost always language dependent, and in fact differ dramatically from language to language. […] Second, feature selection is not a trivial

[23] Ivi, p. 6.

process, and usually involves setting thresholds to eliminate uninformative features. These decisions can be extremely subtil, because although rare features contribute less signal than common features, they can still have an important cumulative effect. Third, current authorship attribution systems invariably perform their analysis at the word level. However, although word level analysis seems to be intuitive, it ignores the fact that morphological features can also play an important role, and moreover that many Asian languages such as Chinese and Japanese do not have word boundaries explicitly identified in text[24].

È utile quindi segnalare che tra la fine del secolo scorso e l'inizio del secolo attuale accanto alle procedure di analisi qualitative si afferma la consapevolezza che esistono nei testi strutture matematiche che si possono descrivere solo in termini quantitativi e nascono metodi di analisi che si basano su questa consapevolezza.
A questa concezione si arriva con un percorso non ovvio ma chiaro: di fronte ad una serie di problemi nello studio di un testo, per esempio l'attribuzione, si può ricorrere ai consueti (per le scienze umane) strumenti di tipo semantico, linguistico, storico: si cerca di ampliare e approfondire le conoscenze sul contenuto del testo, sulla lingua in cui è scritto, sulla storia della sua composizione e trasmissione. Può succedere però che non si facciano progressi significativi e si può allora decidere di ricorrere allo studio di «altre» caratteristiche del testo. Qui si affrontano due passaggi importanti: in primo luogo ci si rende conto che il modello di studio del testo basato sul «contenuto» non risponde alle domande sul testo che lo studioso sta facendo; in secondo luogo si ipotizza, con un atteggiamento esplorativo, che si debbano cercare le risposte costruendo un differente modello in cui contenuto, lingua, storia del testo vengono messi in secondo piano. Per far questo, il primo nodo concettuale è quello del passaggio dagli oggetti ai numeri, dall'individuazione degli oggetti al loro conteggio, dell'assegnare numeri (reali) agli oggetti; o in termini piú formalizzati, di trasformare un sistema di relazioni qualitative (il testo) in un sistema di relazioni quantitative (l'insieme dei dati che contiene le informazioni sugli oggetti dell'analisi) grazie ad una o piú operazioni di *classificazione* del testo. La scelta degli oggetti da studiare è, per ogni testo, veramente molto ampia: si potrebbe ipoteticamente pensare alle sequenze di caratteri che iniziano con una «a» e che hanno una «z» a distanza di quattro caratteri dalla «a», per contare e vedere come si distribuiscono nel testo da studiare e in altri vicini. Solo per indicare che le scelte possibili, se non si è piú vincolati alle parole, sono amplissime. Da un sistema qualitativo si passa cosí ad un sistema

[24] F. Peng, D. Schuurmans, V. Keselj, S. Wang, *Language Independent Authorship Attribution using Character Level Language Models*, in *Proceedings 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, 2003, pp. 267-274.

quantitativo (Doležel[25]). Se i due sistemi corrispondono perfettamente (se sono isomorfi) la situazione risulta per certi versi insoddisfacente perché significa che il passaggio dal sistema qualitativo a quello quantitativo non ha messo in luce nulla di rilevante, la costruzione di un nuovo sistema non ha fatto apparire nulla di nuovo. Ma se delle discrepanze tra i due sistemi appaiono, se si percepiscono differenze, allora si è su una buona strada (la differenza è, e genera, informazione), perché diventa necessaria una riorganizzazione delle conoscenze allo scopo di capire come si rapportano i dati dei due sistemi e che cosa significano le differenze tra l'uno e l'altro. Alla base di questa riorganizzazione della conoscenza, che è nuova conoscenza, sta il riconoscimento che i dati quantitativi fungono da indicatori della presenza nel testo di proprietà qualitative che non appaiono in evidenza:

> Sometimes when a structure is not objectively definable, there exists a series of objectively definable indicators which when taken collectively are almost co-extensive with the given structure[26].

Negli ultimi anni questa tendenza al riconoscimento dell'esistenza di strutture matematiche nei testi[27] ha trovato espressione in varie ricerche, tra le quali è interessante segnalare l'acquisizione di tipo metodologico che emerge dal lavoro di Clement e Sharp (citato sopra, al par. 3) secondo i quali l'attribuzione di testi con metodi quantitativi non sempre è in grado di distinguere le caratteristiche stilistiche autoriali da altre caratteristiche quali l'argomento del testo, o da caratteristiche spurie derivanti dall'accostamento dei testi scelti per la ricerca. Ciò costituisce una sorta di aporia per problemi di attribuzione del mondo reale in cui si studiano testi incogniti per i quali non sono disponibili prove o controprove sperimentali: al di là della questione della falsificabilità dei risultati (altri studiosi, con altri metodi, operano sui medesimi materiali testuali e giungono a conclusioni differenti), si pone con evidenza la questio-

[25] L. Doležel, *A note on quantification in text theory*, in *Text Processing. Text Analysis and Generation. Text Typology and Attribution. Proceedings of Nobel Symposium 51*, ed. by S. Allén, Stockholm, Almqvist & Wiksell International, 1982, pp. 539-552, pp. 540-542.

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

[27] D. Khmelev, F. Tweedie, *Using Markov chains for identification of writers*, in «Literary and Linguistic Computing», XVI, 2001, 4, pp. 299-307; R.H. Baayen, H. van Halteren, F. Tweedie, *Outside the cave of shadows. Using syntactic annotation to enhance authorship attribution*, ivi, XI, 1996, pp. 121-131; D. Benedetto, E. Caglioti, V. Loreto, *Language Trees and Zipping*, in «Phys. Rev. Lett.», LXXXVIII, 2002, n. 4, 048702-1, 048702-4; Clement, Sharp, *Ngram and Bayesian Classification of Documents for Topic and Authorship*, cit.; K. Luyckx, W. Daelemans, *The effect of author set size and data size in authorship attribution*, in «Literary and Linguistic Computing», XXV, 2010, 1, pp. 35-55; G.B. Schaalje, P.J. Fields, M. Roper, G.L. Snow, *Extended nearest shrunken centroid classification: A new method for open-set authorship attribution of texts of varying sizes*, ivi, XXVI, 2011, 1, pp. 71-88.

ne della verificabilità, in quanto le attribuzioni dei testi non possono essere verificate sulla base di conoscenze fattuali. Negli esperimenti controllati su casi di studio costruiti ad hoc, invece, si ricorre a corpora testuali di autore noto e quindi si può individuare subito se le attribuzioni sono corrette e se l'argomento, diverso da autore ad autore, abbia avuto un qualche peso nella classificazione dei testi.

Un'altra testimonianza significativa della progressiva crescita di importanza degli approcci matematici all'attribuzione dei testi è costituita dalla «ad-hoc authorship attribution competition» (aaac) bandita nel 2003 da Patrick Juola, matematico della Duquesne University che da tempo si interessava di problemi di attribuzione[28]. La gara aveva un intento metodologico: tutti i partecipanti si confrontano sull'attribuzione di un medesimo insieme di testi composto di vari sottoinsiemi e quindi si verifica *in termini oggettivi, con controllo pubblico,* se esistono metodi migliori di altri.

La gara si svolse su 13 set di campioni testuali (ancora oggi a disposizione di chiunque sia interessato, all'URL http://www.mathcs.duq.edu/~juola/authorship_materials2.html). La descrizione dei campioni testuali ovviamente non era nota ai partecipanti. Le lingue presenti nei campioni erano inglese elisabettiano, medievale e contemporaneo, angloamericano del XIX secolo, francese, antico slavo, latino, olandese. In ciascuno dei set di testi si trovano vari campioni appartenenti a due o piú autori; per ogni autore ci sono due o piú campioni che si devono attribuire correttamente (non nel senso di indicare *chi* ne sia l'autore, ma nel senso di individuare quelli del medesimo autore). In alcuni set è presente un campione testuale di argomento simile ma che non appartiene a nessuno degli autori di cui sono forniti gli altri due o piú campioni; un campione spurio, quindi, che permette di verificare la finezza operativa del metodo di analisi che non deve classificare «a forza bruta».

I migliori risultati (con percentuali di attribuzioni corrette superiori all'80%; tutti i partecipanti avevano sostanzialmente fallito nell'analisi di un set testuale costituito da lettere) interessanti per questa prospettiva di ricerca furono quelli di Vlado Keselj (Dalhousie University, Halifax, Nuova Scozia, Canada) e di Patrick Juola stesso. Entrambi avevano adottato un metodo di classificazione dei testi basato sulle frequenze degli *n*-grammi, un metodo che scompone il testo in sequenze di *n* caratteri, delle quali si misurano e si confrontano le frequenze, cioè su un metodo che studiava caratteristiche quantitative del testo non riconducibili

[28] P. Juola, *What Can We Do with Small Corpora? Document Categorization via Crossentropy*, in *Proceedings of an Interdisciplinary Workshop on Similarity and Categorization*, Edinburgh, Department of Artificial Intelligence, University of Edinburgh, 1997; P. Juola, H. Baayen, *A Controlled Corpus Experiment in Authorship Attribution by Crossentropy*, in *Proceedings of ACH/ALLC 2003*, Athens, Georgia; ma si veda anche, negli anni sucessivi alla gara: P. Juola, J. Sofko, P. Brennan, *A Prototype for Authorship Attribution Studies*, in «Literary and Linguistic Computing», XXI, 2006, pp. 169-178.

(o solo difficilmente e vagamente riconducibili) alla descrizione di caratteristiche qualitative. I metodi basati sulla tradizionale stilometria o su analisi del testo di tipo qualitativo scomparvero.
Alla fine dell'800 Lutosławski, come si è visto, riteneva che i marcatori stilistici si collocassero in strutture grammaticali e sintattiche complesse («risposte che denotano assenso soggettivo meno di 1 volta su 60», e cosí via; cfr. par. 4). All'inizio del 1900 Mendenhall per i suoi lavori di stilometria costruiva le curve di frequenza della lunghezza delle parole e in generale fin verso la fine del 1900 oggetto delle misurazioni stilometriche erano le parole e le unità piú ampie composte di parole (sintagmi, segmenti, frasi). «In anticipo sui tempi», lo studio di Ledger, *Recounting Plato,* del 1989, indicò la tendenza degli anni di fine secolo, per cui vengono studiate caratteristiche del testo non descrivibili in termini qualitativi o in termini semantici. Infine, nel 2003, in occasione dell'aaac tra i metodi di analisi complessivamente piú efficaci ci furono, come si è visto, quelli che operano sui testi al livello degli *n*-grammi. Il testo «diventa» un'unica lunghissima stringa letta a blocchi di *n* caratteri. Anche in questa situazione – se gli *n*-grammi sono brevi – il testo può essere frammentato per l'analisi ad un livello a cui nessun approccio linguistico tradizionale, per quanto analitico, arriverebbe: non strutture sintattiche, non sintagmi, non parole, non fonemi. E che cosa *significhi* un dato *n*-gramma di un dato testo è difficile dire.
Da Lutosławski all'aaac si manifesta quindi una tendenza per cui dall'analisi di caratteristiche sintattiche complesse, individuabili solo da parte di una mente umana molto competente, si passa all'analisi di caratteristiche sempre piú elementari fino ad arrivare per esempio agli *n*-grammi, sequenze di caratteri individuabili in modo rapido e automatico da parte di un computer.

8. *L'attribuzione degli scritti giornalistici di Gramsci*. Lo studio per l'individuazione degli scritti gramsciani tra gli articoli pubblicati anonimi si colloca nella linea evolutiva verso metodi automatici di individuazione e rilevamento degli oggetti di analisi appena descritta e si caratterizza per il fatto di costituire un problema del mondo reale, non un caso di studio di laboratorio: a parte rari casi non esistono infatti criteri oggettivi che possano provare la correttezza o l'erroneità delle attribuzioni. Il gruppo di ricerca decise perciò con la Fondazione di procedere, preliminarmente all'inizio del lavoro di attribuzione, attraverso due fasi di test: un test preliminare in chiaro su 50 testi gramsciani e 50 non gramsciani (Bianchi, Bordiga, Carena, De Giovanni, Galetto, Giusti, Leonetti, Pastore, Santarosa, Serrati, Tasca, Togliatti, Viglongo) e un test cieco su 40 testi anonimi gramsciani e non gramsciani le cui attribuzioni, note e certe agli studiosi, non fossero note (né comunicate) al gruppo di ricerca. Il primo destinato a mettere a punto il metodo, il secondo destinato a mostrare alla Fondazione quale livello di qualità si fosse in grado di ottenere nelle attribuzioni. La ricerca sarebbe passata dalla fase sperimentale a quella produttiva

solo se il test cieco fosse stato superato con successo: si doveva arrivare ad un'alta percentuale di attribuzioni corrette con una percentuale minima, meglio se nulla, di falsi positivi.
Le tecniche per lo studio dei testi messe a punto durante la fase preliminare furono due: un'evoluzione ad opera di Degli Esposti del metodo degli *n*-grammi che era stato usato da Keselj nella gara di attribuzione del 2003[29], e il metodo dell'entropia informativa relativa di Benedetto, Caglioti e Loreto[30]. Prima una breve descrizione della misurazione delle distanze per mezzo degli *n*-grammi e poi una per l'entropia informativa.
Dato il testo: *questo è un testo*, la sua scomposizione in *n*-grammi – per esempio – di lunghezza 4 (4-grammi) darà luogo a queste metaparole:

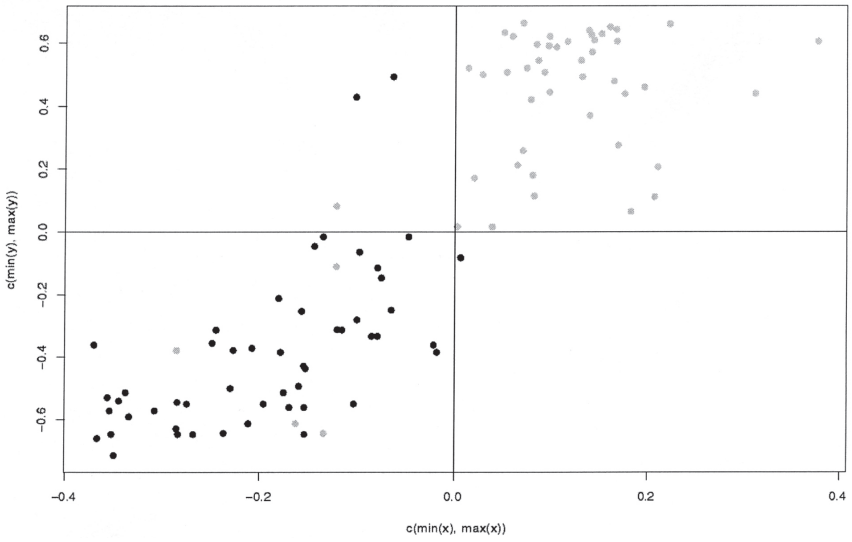
ques	uest	esto	sto<spazio>
to<spazio>è	o<spazio>è<spazio>	<spazio>è<spazio>u	è<spazio>un
<spazio>un<spazio>	un<spazio>t	n<spazio>te	<spazio>tes
test	esto		

che definiscono un metadizionario del testo in questione. Esso, confrontato a due a due con quelli di altri testi permetterà di individuare somiglianze date dalla presenza e frequenza delle medesime metaparole in testi differenti. Come si vede bene, gli *n*-grammi danno luogo a metaparole che non hanno corrispondenza con nessuna entità nota in linguistica; peraltro *n*-grammi lunghi (per esempio 8-grammi, come furono quelli scelti per i testi Gramsci) catturano parti abbastanze ampie della frase (nell'esempio precedente gli 8-grammi sarebbero questo<spazio>è, uesto<spazio>è<spazio>, esto<spazio>è<spazio>u, sto<spazio>è<spazio>un, to<spazio>è<spazio>un<spazio>, o<spazio>è<spazio >un<spazio>t, eccetera) tali da far trasparire frammenti di sequenze linguisticamente e semanticamente significative.
L'entropia informativa (descritta per la prima volta nel 1948 da Shannon[31]) è – in termini discorsivi – la quantità di informazione contenuta in una sequenza di segni alfanumerici, ed è minore in presenza di regolarità espressive («canta, canta, canta!») mentre è maggiore in presenza di irregolarità, non ripetitività, imprevedibilità del messaggio («canta, salta, dormi!»). L'entropia è fenomeno che si manifesta chiaramente, e si può misurare solo con strumenti matematici.

[29] V. Kešelj, N. Cercone, *CNG Method with Weighted Voting*, in P. Joula, *Ad-hoc Authorship Attribution Competition. In Proceedings 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*, Göteborg, 2004.
[30] Benedetto, Caglioti, Loreto, *Language Trees and Zipping*, cit.
[31] C.E. Shannon. *A Mathematical Theory of Communication*, in «The Bell System Technical Journal», XXVII, 1948, pp. 379-423, 623-656.

In termini pratici, utilizzando un algoritmo di compressione (quali ad esempio quelli dei programmi WinZip o WinRAR) un testo molto ripetitivo si può comprimere maggiormente di un testo meno ripetitivo. Se prendiamo un articolo A, anonimo, e lo comprimiamo, generando un dizionario di compressione (che permetterà la decompressine corretta del documento); e poi prendiamo un articolo T di Togliatti, e lo comprimiamo utilizzando il dizionario di compressione del testo A, la compressione sarà massima nel caso che il testo A e il testo T siano identici; sarà nulla nel caso che i due testi non condividano neppure un bigramma cioè siano totalmente e assolutamente differenti. Generalizzando, in questo procedimento la percentuale di compressione del secondo testo è una misura della sua somiglianza con il primo testo. I due metodi confluirono nella strategia complessiva di attribuzione, che utilizzò i risultati di entrambi allo scopo di diminuire il numero di falsi positivi: in sostanza attribuire a Gramsci i soli testi che entrambi i metodi riconoscono come gramsciani. Alla fine della fase preliminare si era in grado di attribuire correttamente 43 testi su 50 (pari in percentuale all'86%). In Figura 1 il quadrante in alto a destra contiene i testi attribuiti a Gramsci (rappresentati dai punti grigi). I punti grigi negli altri quadranti rappresentano le attribuzioni non riuscite. Nel quadrante in alto a destra non ci sono testi non gramsciani (punti neri), cioè il sistema non genera falsi positivi (testi attribuiti a Gramsci ma che non sono gramsciani).

Figura 1. *Le attribuzioni dei cento testi, al termine della fase preliminare*



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

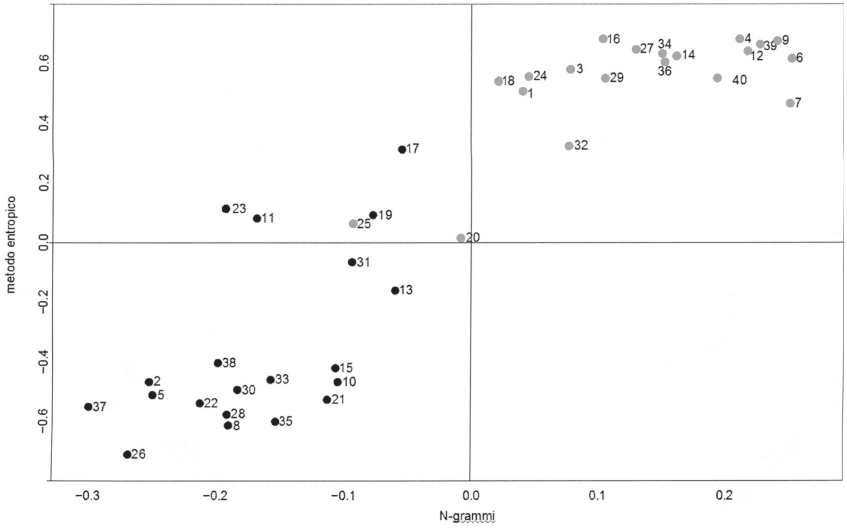
completo, noto solo dopo la presentazione dei risultati alla Fondazione, si trova qui sotto in Tabella 1).

Tabella 1. *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.
9. Gramsci, *Il Cottolengo e i clericali*, «Avanti!», 30 aprile 1917.
10. A.T. [Angelo Tasca], *Sempre piú 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 piú 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!», 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.

I risultati del test cieco, presentati a Roma alla Fondazione Istituto Gramsci il 7 luglio 2006 sono mostrati in Figura 2 (i punti-testo sono accompagnati dal numero che li identifica nell'elenco), costruita con lo stesso procedimento del grafico precedente: nel quadrante in altro a destra ci sono i testi che entrambi i metodi (*n*-grammi e entropia informativa relativa) attribuiscono a Gramsci.

Figura 2. *Attribuzioni del test cieco*



Come si può osservare vennero correttamente individuati e attribuiti 18 testi gramsciani su 20, pari al 90%, senza falsi positivi. I due testi gramsciani non riconosciuti sono il n. 20 e il n. 25 dell'elenco in Tabella 1. La prova era quindi superata con successo. Ne seguí un'altra, imprevista e non annunciata, quando si fece l'analisi di attribuzione per un *corpus* di articoli pubblicati su l'«Avanti!» nel 1916, mai attribuiti a Gramsci ma pubblicati in una rubrica («Sotto la Mole») curata per lo piú da Gramsci: nessuno degli articoli venne riconosciuto come gramsciano, concordemente con il giudizio degli studiosi.

Per i testi effettivamente incogniti, che costituiscono il nucleo della ricerca, si fa quindi affidamento su un metodo di lavoro che è risultato corretto, che è stato verificato, e per il quale è stato definito un protocollo operativo. Le proposte di attribuzione per i testi anonimi vengono presentate agli studiosi curatori dei vari volumi dell'Edizione che valutano uno per uno i testi attribuiti con la procedura quantitativa e decidono anche sulla base di altri criteri di analisi e valutazione se introdurli nell'edizione oppure no. Questo aspetto, che potrebbe apparire in certo modo una diminuzione del significato della ricerca, indica che l'attribuzione con metodi quantitativi è uscita dal laboratorio, è entrata

nel mondo reale, si confronta con problemi reali ed è utilizzata come un nuovo strumento filologico.

Conclusioni. Quali prospettive si aprono? Che la misurazione delle distanze basata sugli *n*-grammi di lunghezza 8 e sull'entropia informativa relativa riesca a identificare gli scritti di Gramsci è assodato; perché *n*-grammi e entropia ci riescano non è del tutto chiaro («computational models, however finely perfected, are better understood as temporary states in a process of coming to know rather than fixed structures of knowledge»[32]): non accontentarsi di magiche scatole nere[33] è un aspetto fondamentale della ricerca scientifica. Ma è utile ricordare quanto scrive Doležel nell'articolo già ricordato al paragrafo 7: per i dati derivanti dal conteggio di «oggetti» nei testi è possibile anche un'interpretazione puramente quantitativa, in cui i dati non sono trattati come indicatori di fenomeni qualitativi ma hanno significato in se stessi perché nel testo esistono strutture matematiche che si possono descrivere solo in termini quantitativi. La presunta scatola nera dunque non è nera, ma semplicemente appartiene ad un altro dominio di conoscenza e funziona secondo regole difficili da capire, che la fanno apparire una scatola nera.
Lo studioso di testi e di storia potrebbe a questo punto ritenere che questa ricerca abbia in qualche modo ristretto il territorio di sua competenza, eroso il suo spazio vitale. Non è necessariamente cosí, perché proprio la volontà di capire perché *questi* metodi diano risultati interessanti su *questi* testi chiama in gioco, alla fin fine, le competenze degli storici e dei linguisti. I quali sono cosí portati ad osservare il loro dominio di studi da un'altra angolazione, o con altri occhiali, il che di fatto significa un allargamento del dominio di studi se nuovi punti di vista, nuove prospettive si aprono. Il primo, interessante punto di vista che si apre è questo: quali sono gli *n*-grammi che piú contribuiscono a definire le attribuzioni? (dal momento che gli *n*-grammi utilizzati per l'attribuzione gramsciana sono di lunghezza 8, dentro sequenze di 8 caratteri trovano posto la parte finale di un aggettivo e la parte inziale del sostantivo che lo segue: un 8-gramma come «prima in» e uno come «rima int» rimandano immediatamente alla Prima Internazionale, ed è rapida la verifica se questa ipotesi corrisponda a ciò che effettivamente si trova nei testi). In questo modo si possono individuare dei nessi di parole che permettono allo studioso di ricostruire le componenti di pensiero che con la loro espressione a livello lessicale caratterizzano un determinato gruppo di scritti gramsciani. Questo lavoro di analisi e di studio non può che essere svolto da un essere umano, uno studioso,

[32] W. McCarthy, *Modeling: A Study in Words and Meaning*, in *A Companion to Digital Humanities*, ed. by S. Schreibman, R. Siemens, J. Unsworth, London, Blackwell, 2005, p. 257.
[33] Su questo tema si veda H. Craig, *Analysis and Authorship Studies*, ivi, p. 285.

il computer non spodesta lo studioso: perché siamo ben lungi dalla conoscenza del mondo e del pensiero necessaria perché un computer possa riconoscere che «ima inte» se rimanda alla Prima Internazionale è interessante, mentre non lo è se rimanda ad una «rima intenzionale» o ad una «massima interna».
Analoghe prospettive si aprono se, lasciati gli *n*-grammi, prendiamo in esame la misura dell'entropia informativa relativa, che ha al cuore la costruzione di un dizionario di compressione, un dizionario di sequenze di caratteri che si ripetono e che possono quindi essere efficacemente rimpiazzate da altre piú brevi. Analizzare e studiare le sequenze di caratteri (le metaparole) presenti in quel dizionario permette di individuare quelle che ricorrono meno frequentemente e che quindi risultano piú caratteristiche del testo (quindi del suo stile *e* del suo contenuto). Anche in questa situazione appare chiaro che allo studioso si aprono nuovi territori di indagine: il dizionario di compressione, come un insieme di tessere di un mosaico, rivela le informazioni che contiene solo a chi sappia intravedere il disegno in cui esse possono occupare un posto significativo.