UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

Dipartimento di Studi Linguistici e Letterari

Corso di Laurea Magistrale in Linguistica

Classe LM-39

Tesi di Laurea

I metodi quantitativi e metodi qualitativi di analisi dei testi e l'attribuzione d'autore L'attribuibilità ad Antonio Gramsci di una serie di articoli non firmati

Relatore
Prof. Michele Cortelazzo

Laureanda
Sara Maurelli
n° matr.1131527 / LMLIN

Anno Accademico 2017 / 2018

INDICE

INTRODUZIONE

Attribuire un testo anonimo non è un'impresa semplice.

Il problema dell'attribuzione d'autore è stato sempre molto sentito nell'ambito filologico e letterario; occupandosi di capire chi si nasconde dietro ad un testo anonimo e quindi occupandosi anche di lingua, i metodi di attribuzione d'autore sono stati per molto tempo soltanto metodi di tipo qualitativo.

Dalla seconda metà del 1800 inizia a svilupparsi però, all'interno degli studi sull'attribuzione d'autore, un'attenzione non soltanto al dato qualitativo ma anche al dato quantitativo.

Questa attenzione al dato quantitativo nel tempo è diventata sempre più presente, tanto che all'attribuzione d'autore hanno iniziato ad interessarsi non soltanto studiosi di ambito letterario ma anche matematici, statistici ed informatici.

I metodi d'attribuzione quantitativa se all'inizio venivano reputati come poco affidabili e visti con sospetto, soprattutto nel campo letterario, nel tempo sono stati migliorati e ritenuti talmente tanto affidabili da poter essere usati come prove in tribunale (Joula, 2006, p. 311).

Alcuni di questi metodi mettono in crisi particolarmente chi si occupa di stile letterario o di linguaggio perché non soltanto si basano solo su dati quantitativi ma non si basano più nemmeno su dati linguistici, cioè sulla fonologia, sulla morfologia, sul lessico, sulla sintassi o sulla semantica.

Alcuni di questi metodi, come ad esempio gli *n*-grammi, scompongono infatti il testo in una serie di sequenze di simboli.

Il lavoro di Basile e altri (Basile, Benedetto, Degli Esposti, & Caglioti, 2010) su un corpus di testi gramsciani e non gramsciani da attribuire a Gramsci si basa proprio su questi metodi.

Basile e altri sono stati contatti dall'Istituto Fondazione Gramsci perché la fondazione era alla ricerca di un metodo d'attribuzione che potesse aiutare l'analisi qualitativa nell' attribuire a Gramsci alcuni articoli di giornale anonimi.

Basile e altri usando metodi di tipo matematico che non si basano direttamente su dati linguistici, cioè l'entropia relativa e gli *n*-grammi, hanno avuto buoni risultati e hanno attribuito correttamente la maggior parte dei testi.

Nel nostro elaborato abbiamo voluto capire se anche un linguista può ottenere gli stessi risultati ottenuti da Basile e altri utilizzando soltanto un'analisi qualitativa sullo stesso corpus; abbiamo inoltre tentato di capire cosa questi metodi matematici riescano a vedere

che ipoteticamente sfugge all'essere umano e abbiamo tentato di evidenziare dal punto di vista qualitativo quali siano le caratteristiche di alcuni testi gramsciani che invece il metodo di Basile e altri non ha attribuito a Gramsci.

Un obbiettivo più ampio del nostro elaborato è invece quello di fare una panoramica sui metodi quantitativi di analisi dei testi e sulla loro applicazione nell'attribuzione d'autore, ma soprattutto il nostro obbiettivo è quello di comparare metodi qualitativi e metodi quantitativi, in modo da poter vedere quanto possano essere utili o meno per chi si occupa di analisi testuale e di stilistica.

L'elaborato è diviso in tre capitoli: il primo capitolo è introduttivo ai metodi di attribuzione d'autore quantitativi, un secondo capitolo illustra tre casi di attribuzioni d'autore basati su metodi quantitativi, mentre il terzo capitolo espone il lavoro di Basile e altri su un corpus di testi gramsciani e non gramsciani assieme all' analisi di tipo qualitativo da noi condotta sullo stesso corpus.

Il primo capitolo riflette sulle differenze fra analisi qualitativa e quantitativa, mostra i punti di forza e i punti di debolezza di ogni tipo di analisi, indicando i casi in cui è più conveniente soffermarsi su analisi di tipo qualitativo o di tipo quantitativo e in quale occasione poter unire entrambe.

Successivamente viene introdotta la definizione di stilometria e ripercorsa la storia della stilometria dalle sue origini ad oggi.

Vengono definiti gli obbiettivi dell'attribuzione d'autore e quali siano i tratti utilizzati nell'attribuzione d'autore (tratti lessicali, tratti grafici, tratti sintattici, tratti semantici, tratti specifici per applicazione).

Vengono infine illustrati i differenti approcci usati dai metodi quantitativi per l'attribuzione d'autore (approcci profile based, istance based ed approcci ibridi) e la comparazione tra i differenti approcci.

Al secondo capitolo riportiamo tre casi di attribuzione d'autore basati su metodi quantitativi, il caso Robert Galbraith studiato da Patrick Joula (Joula, 2013), un lavoro di Jan Rybicki su corpora di traduzioni (Rybicki, 2012) e uno studio sull'attribuzione d'autore per le tragedie e le commedie di Molière e Corneille fatto da Labbè (Labbè & Labbè, 2001).

Il terzo capitolo vuole essere un confronto fra l'analisi di tipo quantitativo e quella di tipo qualitativo.

Viene riportata l'analisi quantitativa fatta da Basile e altri sul corpus di testi gramsciani e non gramsciani, viene esplicitata la metodologia del lavoro di Basile e altri soffermandosi in

particolare sul concetto di *n*-gramma e di entropia relativa ed infine vengono illustrati i risultati della loro analisi.

Una seconda parte viene dedicata alla nostra analisi qualitativa sullo stile di scrittura di Gramsci e di tutti gli altri autori che compongono il corpus utilizzato da Basile e altri.

I risultati qualitativi dell'analisi vengono poi confrontati fra Gramsci e gli altri autori per vedere da quali tratti stilistici siano accomunati e per vedere per quali tratti stilistici invece si differenzino.

Con i risultati della nostra analisi qualitativa abbiamo provato a vedere se un'attribuzione con i soli dati qualitativi sia possibile o meno e quanto sia efficiente rispetto all'attribuzione quantitativa.

Abbiamo infine provato a vedere se i testi non riconosciuti dai metodi quantitativi potessero avere delle caratteristiche comuni; infine abbiamo dato un nostro giudizio sulla costruzione del corpus.

Nel capitolo conclusivo esponiamo il nostro pensiero sull'utilizzo dei metodi qualitativi e quantitativi nell' attribuzione d'autore ed indichiamo ulteriori domande di ricerca secondo noi possibili oltre alle future applicazioni di questi metodi.

CAPITOLO I

I METODI QUANTITATIVI DI ANALISI DEI TESTI E L'ATTRIBUZIONE D'AUTORE

Immaginiamo di essere chiamati in qualità di consulenti linguistici ad un processo riguardante un plagio di autore, ammettiamo ad esempio di aver eseguito un'analisi stilistica sul testo in esame e di aver analizzato lo stile degli autori coinvolti nel processo.
Esponiamo i nostri risultati davanti alla giuria quando l'avvocato della controparte ribatte, sostenendo che la nostra analisi è una analisi di tipo soggettivo, condizionata dal nostro pensiero sullo stile dell'opera in questione e quindi facilmente confutabile. Come possiamo allora avvalorare le nostre tesi sull'analisi, sottolineando che i risultati a cui siamo giunti sono di fatto oggettivi?
Una delle possibili strade in questo caso è l'uso di metodi quantitativi, sia statistici che matematici.
Certo, chi lavora all'interno degli studi umanistici potrebbe prendere le distanze dall'uso dei metodi quantitativi, pensando ad esempio che nella lingua letteraria ed in quella italiana in particolare è proprio lo scarto tra la lingua comune e la lingua ricercata dall'autore a creare l'individualità stilistica; questo dato sembra infatti avvalorare coloro che ritengono l'analisi qualitativa l'unica strada possibile da seguire nell'analisi del testo. (Cortelazzo, 2012).
Se riflettiamo però:

> quando noi facciamo analisi stilistiche di un autore, o di un testo, facciamo quasi sempre, sia pure implicitamente, analisi statistiche (quando diciamo, per esempio, che una tal parola o un tal costrutto hanno una ricorrenza più o meno rilevante nell' opera dell'autore esaminato, che quella parola o quel costrutto sono tipici di quell'autore – il che significa che risultano specifici di quell'autore rispetto ad un corpus di riferimento, quale può essere la tradizione del periodo, del genere e così via). Si può dire, che un bravo studioso di critica stilistica è, inconsciamente e magari controvoglia, un bravo statistico, anche se non adotta i criteri quantitativi della statistica.

(Cortelazzo, 2012, p. 88-89).

L'analisi nel caso esemplificativo del tribunale è di tipo confermativo, dà la certezza allo studioso che i risultati a cui è giunto non sono condizionati dal suo pensiero sull'opera ma

sono per l'appunto confermati da dati empirici. L'ottica confermativa dell'analisi quantitativa infatti accerta che non si sovrastimino fenomeni esistenti: alcuni fenomeni possono essere molto visibili agli occhi dello studioso ma poi di fatto risultare poco significativi se confrontati con modelli di riferimento. Indubbiamente un'analisi di tipo quantitativo è utile anche come strumento di sintesi, soprattutto quando rielabora i risultati ottenuti in uno schema di tipo grafico (Cortelazzo, 2013).

I metodi quantitativi inoltre possono essere usati non soltanto in ottica confermativa ma anche in ottica esplorativa: durante l'analisi possono sfuggire allo studioso dettagli che purtroppo si perdono all'interno di un testo troppo ampio, indizi troppo dilazionati per poter essere scovati da una mente umana e che non possono essere in alcun modo evidenziati se non riordinando il testo attraverso metodi statistici o matematici (Cortelazzo, 2013).

L'ottica esplorativa permette allo studioso di scoprire nuove piste di ricerca, di formulare nuove ipotesi che altrimenti con l'utilizzo del solo metodo qualitativo difficilmente avrebbe potuto considerare.

I metodi quantitativi hanno come ogni metodo dei punti di forza e dei punti deboli; per questo ha senso avvalersi di questi metodi solo in determinate situazioni.

Un punto di forza dei metodi quantitativi è quello di poter analizzare corpora[1] di grandi dimensioni altrimenti impossibili da gestire per una mente umana, infatti ha senso utilizzare questi metodi quando si può sacrificare la ricchezza di un testo per ordinarlo in grafici, tabelle e qualsivoglia tipo di rappresentazione.

Come abbiamo già detto questi metodi danno inoltre ulteriore prova delle intuizioni avute con l'analisi qualitativa, poiché si basano su dati sistematici permettono infatti generalizzazioni e classificazioni più sicure.

I metodi quantitativi infine consentono la visualizzazione dei rapporti tra testi, tra unità testuali e tra unità testuali e testi; (Cortelazzo, 2013) questa applicazione, come vedremo, si rende particolarmente utile non solo nell'analisi stilistica ma anche in problemi di attribuzione d'autore o in qualsiasi analisi in cui si voglia analizzare i rapporti fra testi.

Naturalmente questi metodi sono anche portatori di alcuni svantaggi, uno di questi è che a seconda del metodo scelto viene analizzato solo un livello linguistico, ad esempio i metodi di analisi statistica lessicale si basano soltanto sul lessico e non tengono conto di altri livelli

[1] Per corpus usiamo la definizione di Tuzzi: «Il materiale testuale oggetto delle analisi prende il nome di corpus e si configura come una collezione di testi. Il corpus raccoglie testi coerenti con gli scopi perseguiti dalla ricerca e questa coerenza è valutabile solo discrezionalmente. Nello studio dell'intera opera di un autore i testi costituenti il corpus possono essere, per esempio, le singole opere inedite e/o inedite di cui si conosce l'esistenza; nello studio di un romanzo i singoli capitoli; nell'analisi dei risultati di un'indagine con intervista a domande aperte le trascrizioni dei colloqui […] nell'analisi di annate di stampa i quotidiani (o i settimanali o mensili ecc.) pubblicati […] ecc.» (Tuzzi, 2003, p. 29).

linguistici come la sintassi o della retorica, allo stesso modo un metodo basato su n-grammi sintattici terrà conto soltanto della sintassi e così via.

Un esempio di svantaggio dato ad esempio dagli strumenti di analisi statistica è quello di riuscire a rappresentare soltanto quello che è l'aspetto formale, ovvero il significante ma non il significato, poiché questi strumenti non riconosco sinonimi ne distinguono tra omografi.

Si può ovviare al problema attraverso la lemmatizzazione, cioè il raggruppamento delle forme in lemmi; questa è però spesso una operazione difficoltosa perché non completamente automatizzabile e pone anche problemi teorico pratici non sempre risolvibili (la distinzione tra participi presenti e passati con funzione verbale e quelli con funzione aggettivale, oppure il trattamento delle forme composte di un verbo).

Soltanto con corpora poco ampi la fatica della lemmatizzazione ha davvero senso in termini di costo per il ricercatore; è vero anche però che il corpus lemmatizzato ripaga con risultati più raffinati in quanto solo con la lemmatizzazione si può superare la variazione dovuta agli accordi con il contesto (Cortelazzo, 2013).

Con questo non stiamo naturalmente dicendo che l'analisi qualitativa sia uno strumento da dimenticare; nell'analisi dei testi lo strumento principale di ricerca rimane il metodo qualitativo; Holmes spiega infatti come il metodo quantitativo non voglia sostituirsi alle analisi di letterati e storici, semplicemente i metodi quantitativi vogliono ampliare le prospettive di ricerca degli umanisti (Holmes, 1998, p. 111).

Vogliamo semplicemente evidenziare infatti come questi metodi che si avvalgono della statistica e della matematica possano venire in aiuto al linguista (ma non solo, anche ad esempio al sociologo, al politologo, al filologo) e di come vengano usati e considerati all'interno di quella branca ibrida del sapere che è l'informatica umanistica, meglio conosciuta con il termine anglosassone Digital Humanities.

Sarà quindi compito dello studioso scegliere a seconda del contesto della sua ricerca se adottare un'analisi di tipo qualitativo o quantitativo e in caso si scelga la strada quantitativa capire quale tipo di analisi utilizzare, se ad esempio un'analisi di tipo statistico o un'analisi di tipo matematico, se usare la distanza intertestuale o gli n-grammi.

Naturalmente si richiede una certa competenza interdisciplinare o quanto meno il potersi avvalere delle competenze di altri studiosi come quelle di statistici, matematici e informatici. Cavalli infatti sottolinea questo problema:

> di fatto si sono consolidate delle specializzazioni per cui chi ha acquisito una competenza nell'uso di metodi qualitativi spesso non sa come si fa una ricerca quantitativa e viceversa.

(Cavalli, 2001, p. 140).

A noi sembra che soltanto l'interdisciplinarietà e la collaborazione tra studiosi del campo umanistico e del campo scientifico possano superare questo problema.

1.1 La stilometria

1.1.1 Definizione

La stilometria è l'analisi statistica dello stile letterario, questa analisi parte dal presupposto che ogni autore di testi ha delle caratteristiche inconsce nel suo stile, caratteristiche quindi che non sono per lui del tutto manipolabili; queste caratteristiche sarebbero inoltre quantificabili ed attraverso il loro studio sarebbe così possibile ad esempio distinguere tra un autore e un altro. (Holmes, 1998).

La principale applicazione della stilometria si trova, come abbiamo già detto, negli studi di attribuzione d'autore, cioè ogniqualvolta si voglia far luce sull'autore di un testo che non ha paternità o la cui paternità è dubbia.

Un'altra applicazione possibile invece è quella in casi di datazione cronologica problematica, in pratica quando si voglia ordinare cronologicamente le opere di un autore la cui datazione è dubbia; l'uso di una analisi stilometrica dovrebbe infatti rilevare la modificazione dello stile di un autore durante la sua produzione e quindi facilitare la datazione di una sua opera.

Secondo Laan lo scopo principale della stilometria rimane però la descrizione dello stile di un autore, mentre stabilire l'autore di un testo o una sua datazione sono a suo parere obbiettivi sempre secondari. In sintesi, secondo la studiosa, per dare validità a analisi di attribuzioni d'autore o cronologiche bisogna sempre fare precedentemente delle analisi descrittive dello stile dell'autore.

Laan mostra infatti come la variazione nello stile non è sempre dettata da una mano d'autore differente o da una diversa datazione cronologica.

Si potrebbe obiettare quindi che la stilometria arriva a due asserzioni molto differenti, da una parte dichiara infatti che lo stile di un autore rimane lo stesso per tutta la sua vita e possa quindi essere analizzato attraverso le "impronte" della sua scrittura, dall'altra che le

caratteristiche inconsce di un autore possano cambiare durante la sua vita e quindi rendere possibile una datazione fra i suoi scritti.
Come spiega però Laan:

> It is, of course, possible that the contradiction between these two hypotheses is indeed only apparent and that they are not incompatible.
>
> This would be the case, for instance, if the unconscious aspect of an author's style consists of two parts, one that stays the same throughout and one that changes.
>
> Thus, both claims would be true and universally valid.
>
> Another possibility seems to be that some authors change in the unconscious features of their style and that others stay the same throughout, in which case each claim is only true in a limited sense.
>
> (Laan, 1995, p. 272).

Lo stile può variare infatti per differenza di genere testuale o di contenuto e può sovrapporsi in situazioni di imitazione o altre forme di intertestualità.
Ecco perché nell'uso della stilometria è bene lavorare con un corpus formato da testi dello stesso genere letterario nell'attribuzione d'autore, mentre nell'attribuzione cronologica è consigliabile lavorare con testi cronologicamente vicini.

1.1.2 Storia della stilometria

L'accezione di stilometria come analisi statistica dello stile di un autore trova le prime attestazioni solo nella seconda metà dell'Ottocento, anche se la paternità della disciplina è dubbia.
Alcuni studiosi, come ad esempio Holmes, sostengono che il padre della disciplina sia Augustus de Morgan, un esperto di logica inglese che nel 1851 avrebbe suggerito in una lettera ad un ecclesiastico per un problema di attribuzione d'autore su un Gospel di osservare se un testo occupasse più parole di un altro, indicando come a suo parere un giorno con questo metodo si sarebbero potuti identificare testi spuri. (Morgan, 1851/1882).
Anche se questa idea potrebbe essere plausibile, dato che autori con un largo vocabolario usano solitamente vocaboli più lunghi, oggi sappiamo che la lunghezza media della parola

non è stabile all' interno di un solo autore e non può essere usata come discriminante tre due autori. (Joula, 2006).

L'idea di Morgan venne ripresa e migliorata nel 1887 dal fisico americano Thomas Mendenhall, altro possibile candidato a fondatore della disciplina.

Egli era convinto di poter creare da un testo uno "spettro di parole" o meglio una "curva caratteristica", una rappresentazione grafica della disposizione di parole in base alla lunghezza e alla frequenza con cui le parole occorrono nel testo, così come nella fisica era possibile, attraverso l'uso di uno spettroscopio, analizzare e scomporre un fascio di luce non omogeneo (Mendenhall, 1887).

Nonostante il suo metodo ad oggi sia ritenuto talmente poco affidabile tanto da consigliare ad ogni serio studioso di scartarlo (Joula, 2006), in realtà Mendenhall con le sue analisi di testi letterari ha scoperto molte affinità tra Shakespeare e Marlowe, quelle stesse affinità che oggi vengono indagate con tecnologie più moderne (Holmes, 1998).

La coniazione del termine "stilometria" si deve invece a Wincenty Lutoslawsky, un filosofo polacco.

In una sua lezione tenuta alla Oxford Philological Society il 21 maggio 1897 egli afferma di avere usato, nelle sue analisi stilistiche e di datazione sui dialoghi platonici[2], un suo metodo per misurare statisticamente affinità tra i testi che chiama "stilometria". (Lutoslawski (A), 1897).

Lutoslawski spiega che la stilometria studia lo stile di porzioni di testo, così come la paleografia si occupa delle peculiarità esterne dei manoscritti.

Lutoslawsy divide poi la stilometria dalla Sprachstatistick ideata dagli studiosi tedeschi perché differisce da questa per nei seguenti punti:

- Solo campioni di testo equivalenti sono comparabili per il numero di particolarità che contengono, mentre prima della stilometria ogni testo veniva analizzato come intero, senza alcuna considerazione per la sua lunghezza. La misura ideale del campione di testo è il numero di parole.

- Grandi numeri di peculiarità stilistiche sono richiesti per arrivare a inferenze corrette. Lutoslawski sottolinea come altri autori nelle loro analisi abbiano considerato poche

[2]Le analisi sui dialoghi platonici sono raccolte nel libro "The Origin and Growth of Plato's Logic With an Account of Plato's Style and of the Chronology of His Writings", pubblicato nel 1987 da Longmans.

peculiarità di stile o abbiano dedotto le loro inferenze addirittura da singole occorrenze o peculiarità, mentre le sue analisi sono basate invece su cinquecento peculiarità che rappresentano cinquantotto mila osservazioni fatte da vari studiosi.

- La differenza di importanza delle peculiarità deve essere considerata: secondo Lutoslawsky bisogna infatti dividere le peculiarità in quattro categorie: accidentali, ripetute, importanti e molto importanti; solo così si può arrivare a una più esatta determinazione delle affinità stilistiche. Per peculiarità accidentale Lutoswlasky intende una parola o locuzione che occorre soltanto in un testo; se questa è comune ai due testi forma il più debole collegamento stilistico tra questi e viene contata come un'unità di misura per l'affinità tra testi. Una peculiarità ripetuta tra due testi o tra un testo e un gruppo di testi corrisponde a due unità di misura, una peculiarità frequente corrisponde a tre unità, mente una molto frequente a quattro. In questo modo ogni testo ha un certo numero di unità di affinità con altri testi o con gruppi di testi, identificando così una maggiore o minore affinità di stile tra questi.

Dallo studio dei dialoghi platonici Lutoslawsky formula quella che chiama *Legge di affinità stilistica*, ovvero:

> Considerando due campioni di testi dello stesso autore e della stessa misura, è più vicino cronologicamente a un terzo campione quello che condivide con questo il maggior numero di unità di affinità.

(Lutoslawski (A), 1897, p. 284).

Da questa legge Lutoslawsky ne fa seguire altre due:

- Il numero sufficiente per determinare il carattere stilistico di un campione di testo deve essere più grande di quanto sia stato usato finora nella storia delle analisi stilistiche (cinquecento peculiarità ad esempio sono sufficienti per un testo di venti pagine).

- La differenza minima del numero di unità di affinità indispensabile per inferenze cronologiche è stimata a una differenza del 10% tra due opere, anche se in alcuni casi si dimostra insufficiente.

Alla fine della sua dissertazione dei risultati nell'attribuzione cronologica dei dialoghi platonici lo studioso sottolinea come sia proprio grazie al metodo stilometrico che i suoi risultati hanno valore oggettivo, dato che sono basati su una rimarchevole quantità di osservazioni sullo stile di Platone.

Nonostante la stilometria di Lutoslawsky sia molto rozza rispetto alle tecniche oggi disponibili, possiamo vedere come egli avesse già allora osservato che uno strumento di analisi e confronto dello stile letterario dovesse basarsi non tanto sulla lunghezza delle parole ma su quelle che noi oggi chiamiamo occorrenze, su porzioni di testo di quantità identica e sul maggior numero possibile di dati per poter ottenere risultati il più possibile oggettivi.

Dopo questi lavori ci fu un salto di oltre vent'anni prima che altri studiosi si dedicassero nuovamente alla stilometria; lo statistico britannico George Udny Yule e il linguista americano George Zipf elaborarono nuovi metodi all'interno della disciplina. (Holmes, 1998).

Zipf riprendendo una scoperta dello stenografo francese Eustop del 1916 e a sua volta, Benoit Mandelbrot, matematico polacco studioso dei frattali, riformulò ulteriormente questa legge, giungendo, per deduzione puramente matematica, a una nuova formula (Tuzzi, 2003, p. 115).

Questa nuova formula, chiamata legge di Zipf (Zipf, 1932) stabilisce un legame matematico tra la frequenza di una word type (forma grafica distinta) e il rango da questo assunto nel vocabolario ordinato per frequenza decrescente. Le parole del vocabolario si distribuiscono nei testi in modo tale che la frequenza f_r e il rango r siano inversamente proporzionali, secondo una costante di proporzionalità c (Tuzzi, 2003, p. 125).[3]

In pratica all'aumentare del rango la frequenza di un vocabolo diminuisce e viceversa.

Nel 1938 Yule suggerì di usare come parametro per l'attribuzione d'autore la lunghezza di frase; provò questa sua ipotesi in uno studio di attribuzione del *De imitatione Christi* (Yule, 1938), concludendo che i metodi statistici basati sulla lunghezza della frase non fossero del tutto degli indicatori validi; nonostante ciò il suo studio fu fondamentale per la disciplina.

[3] $f_r r^a$=c dove a è una misura di ricchezza lessicale del corpus che prende il nome di ricchezza del vocabolario. Mentre per Eustop il prodotto tra il rango e la frequenza di una parola deve essere costante al divergere dell'ampiezza del testo, per Zipf deve invece essere costante il prodotto tra la frequenza di una parola e il rango elevato a un certo parametro di potenza (costante nel testo, ma variabile da persona a persona). (Tuzzi, 2003, p. 115)

Nel 1944 Yule ideò una misurazione della ricchezza lessicale[4] che non dipendesse però dalla lunghezza del testo, creando quella che viene chiamata "costante caratteristica K di Yule"[5] (Yule, 1944).

Questa costante si basa sulle dimensioni delle classi di frequenza, Yule dimostrò come l'occorrenza di una data parola sia basata sulla probabilità e possa essere modellata da una distribuzione di Poisson[6]

Studi successivi dimostrarono che la costante K da sola non è una prova affidabile nella attribuzione d'autore. (Holmes, 1998).

Una misura simile alla costante caratteristica K è il D di Simpson[7] del 1949 (Simpson, 1949), come K anche D misura il tasso in cui le parole vengono ripetute nei testi e può essere usato come una valutazione dell'inverso della ricchezza lessicale, ovvero come un valore crescente al decrescere della ricchezza lessicale. (Tuzzi, 2003, p. 129)

Williams nel 1940 aveva invece scoperto che tracciando le distribuzioni di frequenza dei logaritmi del numero di parole per frase si otteneva un'approssimazione a una distribuzione normale per ogni autore (Williams, 1940), questa scoperta venne riutilizzata poi da Wake nei suoi studi sugli autori greci (Wake, 1957).

Cox e Brandwood nel 1959 usarono la stilometria in uno studio di datazione cronologica delle opere di Platone, studiarono la distribuzione delle ultime cinque sillabe di ogni frase, classificando ogni sillaba come lunga o corta.

Con questo metodo i due studiosi trovarono una marcata distinzione tra la distribuzione delle sillabe nella *Repubblica* e nelle *Leggi*; cercarono di ordinare poi le restanti opere di Platone attraverso un ordine di decrescente affinità con la *Repubblica* (Brandwood & Cox, 1959).

Lo studio che però definitivamente sancì al mondo il potenziale della stilometria nell'attribuzione d'autore fu uno studio degli anni '60 di due statistici americani, Mosteller e Wallace (Holmes, 1998).

I due autori cercarono di comprendere, attraverso l'uso di metodi statistici, chi potesse essere l'autore dei *Federalist Papers*, una raccolta di articoli pubblicata anonima tra il 1787-88 che

[4] Il rapporto fra numero di parole diverse (word type) e il numero di parole totali (word token) spesso viene usato come indice di ricchezza lessicale, ovvero $\frac{V(N)}{N}$, detto anche type token ratio.

[5] $K = \frac{\sum_m V_m(N) m^2 - N}{N^2} \times 10.000$, dove N è la lunghezza del corpus in word token (forme grafiche totali), $V_m(N)$ è il numero di word type (forme grafiche distinte) presenti nel corpus, *m* rappresenta la classe di frequenza

[6] Una distribuzione di probabilità discreta (definita su un insieme discreto S) che esprime le probabilità per il numero di eventi che si verificano successivamente e indipendentemente in un dato intervallo di tempo, sapendo che mediamente se ne verifica un numero λ.

[7] $D = \sum_m V_m(N) \frac{m(m-1)}{N(N-1)} \times 10.000 = \frac{\sum_m V_m(N) m^2 - N}{N^2} \times 10.000$

aveva lo scopo di persuadere i cittadini dello stato di New York a firmare la costituzione (Mosteller & Wallace, 1964).

È noto che il testo è stato scritto da Alexander Hamilton, John Jay e James Madison; tutti dichiararono infatti di aver contribuito alla stesura, il problema fu che dodici degli ottantacinque articoli vennero attribuiti sia a Madison che ad Hamilton.

Basandosi sul lavoro che Ellegård aveva compiuto sull'analisi d'attribuzione d'autore delle *Letters of Junius* usando le frequenze di occorrenza delle parole funzione, Mosteller e Wallace utilizzarono parole funzione come preposizioni, congiunzioni e articoli come discriminanti.

Ad esempio, la parola *upon* ha una media di 3.24 per 1.000 parole negli articoli attribuiti per certi ad Hamilton, ma solo di 0.23 negli articoli di Madison; basandosi su queste analisi i due statistici usarono probabilità numeriche per esprimere gradi di giudizio di verità su ipotesi come "Hamilton ha scritto il saggio numero 52" utilizzando poi il teorema di Bayes[8] per affinare le probabilità sui risultati della loro analisi.

Mosteller e Wallace attribuirono così i dodici articoli contesi a Madison, attribuzione che trovò pieno consenso fra gli storici; nonostante i risultati avessero trovato conferma anche fra gli studiosi dell'area umanistica Mosteller e Wallace sottolinearono come l'analisi fosse stata però particolarmente ardua in quanto gli articoli presentavano non solo similarità per il loro contesto politico ma anche per lo stile di Madison e Hamilton.

Questo studio segna una svolta all'interno della stilometria in quanto i due statistici riuscirono ad arrivare ai risultati soltanto sulla base di probabilità dedotte statisticamente e sul teorema di Bayes.

Da questo studio in poi i *Federalist Papers* diventeranno il perfetto test per i nuovi metodi d'attribuzione poiché sono facilmente reperibili, hanno un set di candidati come autori ben definito e sono un esempio perfetto di un testo anonimo scritto da tutti i possibili candidati allo stesso tempo, sullo stesso argomento e scritto per una pubblicazione su uno stesso media e dello stesso genere. (Joula, 2006).

Dopo lo studio sui *Federalist Papers* gli studiosi dell'area umanistica non accettarono di buon grado l'intrusione della statistica negli studi di attribuzione, ribadendo come nessun parametro al di fuori dell'analisi linguistica potesse essere valido nell' analisi dello stile di un autore. (Holmes, 1998).

[8] Detto anche teorema della probabilità delle cause, viene impiegato per trovare la probabilità di una causa che ha scatenato l'evento verificato.

Uno dei maggiori dibattiti fu sul metodo di Morton (Morton, 1978), un metodo che affermava di poter identificare testi di autori scritti in inglese studiando l'occorrenza di una parola in una determinata posizione, osservando come una specifica parola preceda o segua un'altra specifica parola e comparando l'uso di una specifica parola al posto di un'altra.
Questo metodo fu usato da Merriam (Merriam, 1979-1980) nei suoi studi su Shakespeare; il metodo fu però accusato da Smith di essere poco preciso: mentre Morton infatti (Morton, 1986) era convinto di poter distinguere un autore da un altro grazie allo studio della posizione degli hapax, secondo Smith (Smith, 1987) non vi era alcuna evidenza che provasse il posizionamento degli hapax come paramento discriminante nell'attribuzione d'autore.
In quegli anni altri metodi basati sulla statistica, come ad esempio quello di Thisted ed Efron e di Foster (Foster, 1989) (Thisted & Efron, 1987) furono poi in seguito confutati da altri statistici (Valenza, 1990) (Valenza & Elliot, 1996), aumentando così la diffidenza degli studiosi dell'area umanistica nei confronti di questi metodi.
Nonostante molti metodi di quel periodo siano stati confutati, proprio alla fine degli anni Ottanta ne venne ideato uno usato ancora tutt'oggi nelle attribuzioni d'autore, il *Delta* di Burrows (Burrows, 1987).
Burrows analizzò la frequenza delle 150 parole più frequenti in un corpus dei poeti della Restaurazione inglese, per ognuna di queste calcolò poi una distribuzione *z* (una stima della frequenza media di una parola così come una stima della varianza per quella frequenza). Ad ogni testo è stato poi assegnato un punteggio basato su ognuna delle 150 parole, registrando quanto fossero distanti in valore superiore o inferiore dalla norma: un punteggio *z* positivo indica una parola più comune della media, un punteggio *z* negativo indica una parola meno comune della media, un punteggio *z* 0 indica una parola che appare esattamente nella media.
Il Delta è la media della differenza assoluta tra i punteggi zeta per un dato set di parole in un dato gruppo di testi e i punteggi zeta per lo stesso set di parole in un testo target: minore è la misura del Delta fra due testi più questi saranno vicini e quindi simili.
Questo metodo ha avuto successo anche perché i risultati dell'analisi possono essere rappresentati graficamente su un piano cartesiano, mostrando così le relazioni presenti tra il testo target e i testi del corpus.
Burrows ha applicato questo metodo con successo alle opere di Austen, delle sorelle Bronte, di Scott e Byron, mostrando che la distinzione tra un autore e un altro può essere fatta studiando il modo in cui gli scrittori usano le parole funzione più comuni.
Se il Delta fu uno dei più grandi risultati nella stilometria, nei primi anni Novanta venne ideata una tecnica che si rilevò un fallimento per l'attribuzione d'autore, il Cusum.

Il Cusum[9], detto anche tecnica Qsum, abbreviazione per cumulative sum o somma cumulativa, è un metodo statistico visuale usato per osservare similarità tra sequenze di misure.

Morton (Morton, 1991) propose di utilizzare questo metodo per l'attribuzione d'autore basandosi sull'idea che ogni individuo usi nel linguaggio sia scritto che parlato un set di caratteristiche uniche; queste caratteristiche sarebbero quantificabili in quanto componenti delle frasi di un individuo e si baserebbero in particolare sulle parole funzione, su quelle parole che Morton chiama "parole vocale", cioè parole inizianti per vocale, e sull'insieme di parole funzione e di parole vocali.

Secondo Morton, nell'attribuzione d'autore, se in un testo parole di una determinata classe appaiono ad un tasso che è consistente ma che differisce significativamente da quello dei testi del candidato all'attribuzione, allora l'autore del testo preso in analisi non è il candidato all'attribuzione.

Il Cusum richiede che vengano generate due tracce, una per la lunghezza delle frasi e l'altra per il numero di volte che una determinata parola caratteristica si trova in ogni frase, le due tracce vengono poi sovrapposte.

I due valori (la lunghezza della frase e il numero di parole caratteristiche) dovrebbero essere l'un l'altro paralleli nel linguaggio di ogni individuo. Una divergenza dei due valori dimostrerebbe una diversa autorialità, perché indicherebbe una differenza nel tasso d'uso della caratteristica.

Questo test fu molto usato all'interno dei tribunali inglesi e irlandesi come prova ma fu poi confutato da studi compiuti da altri statistici (Canter, 1992) (Hann & Schils, 1993), dimostrando così la sua inaffidabilità.

Nonostante questi insuccessi le ricerche nel campo della stilometria continueranno, aprendo dalla metà degli anni Novanta in poi le porte a metodi che utilizzano reti neurali (Matthews, 1994) (Merriam, 1993) e a metodi matematici basati sull'uso di algoritmi genetici (Holmes & Forsyth, 1995) e di compressione come ad esempio l'entropia, o ancora il metodo degli n-grammi, un metodo che come vedremo nel prossimo paragrafo non si baserà più sulle parole ma scomporrà il testo in una sequenza di simboli.

[9] Prendiamo una sequenza di numeri come {8, 6, 7, 5, 3, 0, 9, 2 . . . } e calcoliamo la media {5}. Calcoliamo poi per la sequenza di numeri la differenza dalla media { 3, 1, 2, 0, −2, −5,4, −3 . . . } da questa sequenza si traccia la somma cumulativa {3, 4, 6, 6, 4, −1, 3, 0 . . . } (Joula, 2006, p. 244).

1.2 L’attribuzione d’autore

1.2.1 Definizione e contesto

Come abbiamo già accennato nel paragrafo precedente una definizione ampia di attribuzione d’autore potrebbe essere quella di metodo che deduce le caratteristiche dell’autore di un testo dalle caratteristiche del testo costruito da quell’autore. (Joula, 2006)
Nello specifico il termine “attribuzione d’autore”, secondo Joula, viene usato nei casi in cui:

- Dato un campione di testo sicuramente attribuito ad un autore compreso in un set di autori, si debba riconoscere a quale degli autori compreso nel set appartiene il campione di testo preso in esame

- Dato un campione di testo ritenuto essere di uno degli autori compresi nel set, si debba riconoscere se eventualmente il campione di testo preso in esame appartenga ad uno degli autori compresi nel set

- Dato un documento, si debba riconoscere chi è l’autore di quel documento

Il termine “profilazione” secondo alcuni autori sarebbe invece usato nei casi in cui da un campione di testo si debbano estrarre le caratteristiche dell’autore di testo, caratteristiche che indubbiamente aiutano nell’attribuzione d’autore, come ad esempio: il testo ha un solo autore o più autori? L’autore del testo è parlante nativo della lingua in cui è stato scritto il testo? L’autore del testo è una donna o un uomo? Ecc. (Joula, 2006, p. 239)
Stamatatos (Stamatatos, 2009, p. 540) indica come obbiettivi dell’attribuzione d’autore:

- La verifica dell’autore (decidere se il testo dato in esame è stato scritto da un determinato autore o no)

- Il riconoscimento del plagio (trovare similarità tra testi)

- La profilazione dell’autore e caratterizzazione (estrarre dal documento informazioni come età, sesso e grado di istruzione dell’autore di un documento)

- Il riconoscimento di incoerenze stilistiche (incoerenze che si trovano, ad esempio, quando un testo è stato scritto da più autori)

Forse il lettore, dopo aver letto il paragrafo 1.1.2, si potrà chiedere come questi metodi quantitativi, così discussi e spesso respinti, possano davvero contribuire agli obbiettivi che l'attribuzione d'autore si pone.
Questi metodi innanzitutto a partire dalla metà degli anni Novanta sono stati decisamente migliorati.
Prima degli anni Novanta le tecniche quantitative precedenti avevano infatti diversi limiti metodologici (Stamatatos, 2009, p. 540):

- I dati testuali erano troppo lunghi (per le tecniche dell'epoca) e non omogenei stilisticamente

- Il numero di candidati possibili fra gli autori del testo preso in esame era troppo esiguo (solitamente due o tre)

- I corpora per lo studio dello stile d'autore non avevano spesso lo stesso argomento

- La valutazione dei metodi d'attribuzione era per lo più intuitiva

- La comparazione tra i differenti metodi risultava difficoltosa per la mancanza di dati di riferimento

Dalla fine degli anni Novanta i metodi di attribuzione d'autore sono stati migliorati proprio perché l'espansione del world wide web e la comparsa di enormi quantità di testi elettronici (e-mail, blog, forum online, ecc.) ha non solo permesso ma anche obbligato a dover catalogare al meglio queste informazioni, infatti:

- Il settore di studio del recupero di informazioni (Information Retrival[10]) ha sviluppato tecniche per rappresentare e classificare grandi quantità di testi.

[10] Un insieme di tecniche che si occupano della ricerca di materiale (solitamente documenti) di natura non strutturata (solitamente testi) che soddisfino una richiesta di informazione all'interno di grandi collezioni (solitamente immagazzinate in calcolatori) (Manning, Raghavan, & Schütze, 2008, p. 1).

- Sono stati creati potenti algoritmi nell'apprendimento automatico (Machine Learning)[11] capaci di gestire dati multidimensionali e frammentari permettendo migliori rappresentazioni.

- Le ricerche sull'elaborazione del linguaggio naturale (Natural Language Processing[12]) hanno creato tecniche di analisi del testo efficienti e offerto nuove forme di misura per la rappresentazione dello stile.

Potremmo quindi pensare oggi, date queste nuove innovazioni nel campo, di poter individuare l'autore sconosciuto di un testo tra miliardi di possibili candidati?
Naturalmente questo non è possibile, le tecniche di attribuzione d'autore non sono infatti "formule magiche" che permettono di trovare un autore tra un set di candidati potenzialmente infinto, bisogna avere un indizio o almeno un'idea su chi possa essere il possibile candidato o i possibili candidati ad autori di un testo.
È quindi estremamente importante la scelta dei testi che formerà il corpus da studiare, una scelta che dev'essere ben ragionata.
Non è detto ad esempio che un testo possa essere soltanto il prodotto di un solo autore: nella creazione di un romanzo, anche nel caso in cui un testo sia stato scritto da un singolo autore e non a quattro mani, possono infatti incidere l'editor, la casa editrice, a volte persino il tipografo.
Inoltre, anche se si avesse la certezza che nessuno oltre l'autore abbia lavorato sul testo in esame, alcuni elementi all'interno del testo potrebbero non essere di pugno dell'autore, come ad esempio nel caso delle citazioni: nonostante le citazioni siano spesso dichiarate dall'autore del testo, queste però possono scombinare il computo statistico nell'attribuzione, sono elementi quindi da tenere in considerazione all' interno dell'analisi (Joula, 2006).
Infine, nel caso in cui si voglia vedere se un testo appartenga realmente all'opera di un determinato autore, bisogna escludere dal corpus di riferimento tutti i testi di dubbia attribuzione per quel determinato autore, in quanto è meglio perdere un testo non certamente attribuito piuttosto che rischiare di alterare l'analisi.
Date queste premesse, nel lettore potrebbe sorgere una domanda: questi metodi funzionano?

[11] Un insieme di metodi sviluppato da diverse discipline che ha come scopo primario di dare ad una macchina l'abilità di imparare senza essere stata specificatamente programmata per farlo.
[12] Scienza che si occupa dell'elaborazione dei linguaggi naturali mediante calcolatori elettronici.

La riposta potrebbe essere: dipende da quale tipo di documenti si dispone e da qual è l'obbiettivo della ricerca.
Come abbiamo già ripetuto prima, la costruzione del corpus è di fondamentale importanza: se di un possibile candidato ad autore disponiamo ad esempio di un saggio filosofico, di un romanzo e di un articolo di giornale sarà praticamente impossibile usare tutti e tre i testi nell'analisi di attribuzione, in quanto un metodo che funzioni contemporaneamente con generi di testo diversi è per gli studiosi del campo estremamente complicato da realizzare.
Allo stesso modo, nonostante ci siano diversi studi di stilometria basati sulla traduzione in diverse lingue (si veda il paragrafo 2.2), non è possibile avere all'interno del corpus dell'analisi testi scritti in lingue diverse.
Le tecniche e i metodi inoltre, anche se valide in sé, potrebbero non esserlo in un determinato ambito di ricerca, ad esempio un ricercatore che voglia studiare le differenze tra uomini e donne nella scrittura potrebbe essere interessato non tanto alle differenze in sé ma alle ragioni dietro alle differenze (Joula, 2006, p. 247), ragioni che queste tecniche e metodi non riescono a spiegare.
Sta quindi allo studioso scegliere il metodo più adatto all'impostazione della sua ricerca e ai risultati che vuole ottenere; il vero problema, a parte la validità o meno delle diverse tecniche e metodi d'attribuzione d'autore, è proprio questo: non esiste un'unica tecnica o metodo valido per l'attribuzione d'autore ma ne esistono moltissimi.
Nel prossimo paragrafo illustreremo i diversi tratti su cui si basano le tecniche usate nell'attribuzione d'autore e i metodi utilizzati nell'attribuzione d'autore, rifacendoci in particolare all'articolo di Stamatatos (Stamatatos, 2009).

1.3 Tratti utilizzati nell'attribuzione d'autore

1.3.1 Tratti lessicali

Un testo potrebbe essere visto come un insieme di unità formate da frasi, ogni frase a sua volta può essere costituita da parole, numeri e segni di punteggiatura.
L'analisi dei tratti lessicali è stata il primo tipo di analisi nell'ambito della stilometria e dell'attribuzione d'autore a cominciare da valutazioni molto generali: le primissime analisi nell'ambito sono state l'analisi della lunghezza delle frasi e la lunghezza delle parole (a tal proposito si veda il paragrafo 1.2.1).

L'utilizzo di tratti lessicali porta il vantaggio di poter utilizzare testi scritti in una qualsiasi lingua naturale (anche se si potrebbe incontrare qualche difficoltà con alcune lingue naturali come ad esempio il cinese) avvalendosi soltanto di un *tokenizer*, uno strumento che segmenta il testo in unità, in questo caso parole.

Nelle attribuzioni d'autore sono molto studiate la ricchezza di vocabolario, misurata attraverso la type-token ratio[13] e il numero di hapax presenti nei testi, tenendo sempre in considerazione però che la ricchezza del vocabolario dipende anche dalla lunghezza dei testi: il vocabolario cresce al crescere della lunghezza del testo ed è infatti una funzione monotona non decrescente della dimensione, con tasso di accrescimento decrescente (Tuzzi, 2003, p. 119).

Uno degli approcci più usati nell'attribuzione d'autore al testo è quello del bag-of-words, approccio che considera il testo come un set di parole in cui ogni parola ha una frequenza di occorrenza.

Se nell'approccio basato sulla classificazione per argomento si utilizzano molto le parole contenuto o parole piene, nell'attribuzione d'autore si studiano soprattutto le parole funzione o le parole vuote (articoli, proposizioni, pronomi, ecc.), in quanto sembra che siano proprio le parole funzione ad essere utilizzate inconsciamente dall'autore, oltre ad essere indipendenti rispetto all'argomento del testo.

Per decidere quali sono le parole funzione da studiare spesso si estraggono le parole funzione più frequenti nel corpus e si sceglie un algoritmo da usare in base all'analisi che si vuole effettuare.

Oltre al tokenizer sono richiesti altri strumenti nell'analisi come strumenti per la lemmatizzazione, strumenti per ridurre le parole a tema morfologico (in inglese stemmers) e strumenti per identificare parole omografe.

Per tenere conto anche del valore contestuale delle parole una misura proposta è stata quella degli *n*-grammi parola (word *n*-grams), cioè *n* parole contigue, dette anche collocazioni di parola.

Alcuni studi (Sanderson e Guenter 2006; Villasenor-Pineda, et al. 2006) hanno però dimostrato che gli *n*-grammi parola non sono sempre più affidabili rispetto allo studio delle parole funzione, in quanto spesso gli *n*-grammi parola sono più legati al contenuto del testo che non allo stile dell'autore.

[13] Il rapporto fra numero di parole diverse (word type) e il numero di parole totali (word token) ovvero $\frac{V(N)}{N}$

Anche l'analisi degli errori, ovvero la ricerca degli errori nel testo, può essere utilizzata quantitativamente nell'attribuzione d'autore: sono stati definiti dei set di errori ortografici (omissioni e inserzioni di lettere) e di formattazione (ad esempio tutte le parole scritte in maiuscolo), si è infine creato un metodo per estrarre da un testo queste informazioni autonomamente (Koppel & Schler, 2003).

1.3.2 Tratti grafici

Un testo può essere visto anche come un insieme di sequenze di caratteri grafici.
Esistono diverse definizioni di misura dei caratteri: numero dei caratteri di tipo alfabetico, numero dei caratteri di tipo digitale, numero di caratteri maiuscoli e minuscoli, frequenza delle lettere, numero dei segni di interpunzione, ecc.
Queste informazioni sono facili da reperire in qualsiasi tipo di corpus basato su lingue naturali e si sono dimostrate utili nella quantificazione dello stile letterario.
Un tipo di misura delle sequenze di caratteri è quella degli *n*-grammi: una sequenza di *n* segni alfanumerici presenti in un testo, dove *n* può essere un qualsiasi valore a discrezione dello studioso.
Gli *n*-grammi non corrispondono a nessun criterio di segmentazione tradizionale del testo in quanto nel numero di *n* non rientrano solo lettere ma anche gli spazi e i segni di interpunzione.
Un modo per esemplificare gli *n*-grammi può essere quello di pensare ad una finestra di lunghezza di *n* segni che scorre sul testo (Lana, 2010, p. 36); ad esempio, se leggiamo la sequenza all'interno del riquadro, un 8-gramma può essere:

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

Altri esempi di 8-grammi sono:

segmenta
ma rende
: un n-g

La segmentazione del testo in *n*-grammi si è dimostrata utile nell'attribuzione d'autore e rende l'analisi molto più semplice in lingue naturali in cui l'utilizzo di un tokenizer non è così semplice (ad esempio il cinese).
Come per le parole, sono gli *n*-grammi più frequenti ad essere analizzati nell'attribuzione d'autore, la procedura di estrazione degli *n*-grammi varia in base alla lingua e non richiede altri strumenti.
Nell'analisi è molto importante la decisione del valore di *n*, un numero alto di *n* cattura meglio le informazioni lessicali, contestuali e tematiche, viceversa valori bassi di *n* (2 o 3) sono più vicini alla sillaba che non alla parola e viene quindi persa l'informazione contestuale.
La scelta di *n* dipende quindi sia dal tipo di analisi che si vuole effettuare sia dal tipo di lingua in cui è stato scritto il corpus.
Un'altra tipologia di analisi mediante caratteri è quella che usa metodi basati su algoritmi di compressione[14]: inizialmente tutti i testi attribuiti con certezza a un determinato autore sono prima uniti in un grande file *Xa*, il file viene poi compresso dall'algoritmo nel file compresso *C(Xa*).
Il testo d'autore sconosciuto *X* viene aggiunto a ogni testo di *Xa* e l'algoritmo di compressione è usato ancora per ogni insieme di *Xa* e di *X*, ovvero *C(Xa + X*).
La differenza di dimensione di bit fra i due file compressi, cioè $d(X, Xa) = C(Xa + X) - C(Xa)$, indica la similarità del testo sconosciuto con ogni candidato autore.

1.3.3 Tratti sintattici

Un testo può essere letto come una serie di informazioni sintattiche.
Uno degli assunti dell'attribuzione d'autore è proprio quello che nella sintassi di un autore si trovino i tratti inconsci del suo stile.
L'analisi sintattica risulta però complicata in quanto ogni lingua naturale ha una diversa sintassi e sono quindi necessari programmi di Natural Language Processing molto potenti e specifici per lingua, oppure è necessario un POS tagger (part of speach tagger), uno strumento che etichetta ogni word token con informazioni di tipo morfo-sintattico basate sul valore contestuale.

[14] Algoritmi che permettono la riduzione della quantità di bit necessari alla rappresentazione di un'informazione in forma digitale: WinRAR, WinZip e StuffIt sono alcuni esempi di programmi basati su algoritmi di compressione usati da pubblico non specialista.

Molto spesso però, nonostante questi strumenti possano essere anche molto accurati, producono dati non puliti in quanto, data l'enorme difficoltà di un'analisi sintattica automatica, non sono esenti da errori.
Una prima analisi nell'attribuzione d'autore di tipo sintattico è stata quella di Baayen, Van Halteren e Tweedie (Baayen, Van Halteren, & Tweedie, 1996) : gli studiosi hanno usato un corpus formato da testi scritti in inglese annotati sintatticamente, comprendente anche alberi sintattici generati in maniera semiautomatica per ogni frase del corpus.
Dallo studio di questo corpus gli studiosi sono riusciti a trovare la frequenza delle regole sintattiche, dove ognuna di queste regole esprimeva una parte dell'analisi sintattica, ad esempio la regola:

$$A:PP \rightarrow P:PREP+PC:NP$$

indica che un sintagma preposizionale (PP) di tipo avverbiale (A) è costituito da una preposizione (P:PREP) seguita da un sintagma nominale (NP) usato come complemento preposizionale (PC).
Questo tipo di misurazione sembra dare risultati più affidabili rispetto alla ricchezza del vocabolario o a misure di tipo lessicale.
Stamatatos e altri (Statamatos, Fatokakis , & Kokkinakis, 2001) hanno invece utilizzato uno strumento di NLP capace di suddividere un testo in greco moderno non preparato all'analisi sintattica in frasi e sintagmi, ad esempio:

NP[*Another attempt*] VP[*to exploit*] NP[*syntactic information*] VP[*was proposed*] PP[*by Stamatatos, et al. (2000)*]

dove NP sta per sintagma nominale, VP per sintagma verbale e PP per sintagma preposizionale. Questa tecnica è molto meno laboriosa rispetto a quella usata da Bayen e altri, inoltre può essere estratta automaticamente dal testo senza particolari preparazioni.
Stamatatos ha poi utilizzato uno strumento di NLP che analizza il testo in più passaggi: il primo passaggio analizza i casi più semplici mentre l'ultimo passaggio combina i risultati del primo passaggio per produrre risultati più complessi.
Questo tipo di strumento dà un'informazione indirettamente sintattica.
Nonostante anch'esso sia linguaggio specifico ha però il vantaggio di estrarre informazioni sintattiche da testi illimitati.

Hirst e Feiguina (Hirst & Feiguina, 2007) hanno trasformato invece i risultati parziali di un'analisi sintattica in un flusso ordinato di etichette sintattiche, la frase "*a simple example*" ad esempio è formata da:

NX DT JJ NN

cioè da un sintagma nominale (NX) che consiste di un determinante (DT), di un aggettivo (JJ) e di un nome (NN).
Da questo flusso hanno poi estratto una serie di frequenze di bigrammi per rappresentare delle informazioni sintattiche contestuali e hanno visto come queste informazioni siano utili per discriminare l'autore di un testo anche quando il testo esaminato sia molto corto (ad esempio 200 parole).
Koppel e Schler (Koppel & Schler, 2003) si sono invece concentrati sugli errori sintattici che possono essere presenti in un testo, ad esempio frammenti di frasi, frasi agrammaticali, consecutio temporum errata ecc.
Koppel e Schler hanno utilizzato un correttore automatico di tipo commerciale per individuare gli errori e poter quindi discriminare tra i possibili candidati autori.
Sfortunatamente il correttore da loro usato si è rivelato però molto impreciso, come accade spesso per i correttori automatici di uso comune.
Karlgren ed Eriksson infine (Karlgren & Eriksson , 2007) hanno usato caratteristiche sintattiche per individuare sequenze di pattern per poi descrivere l'uso di questi pattern in frasi che si susseguono nel testo; hanno cioè cercato le proprietà di distribuzione delle caratteristiche sintattiche nel testo, una tecnica promettente che può evidenziare le impronte stilistiche dell'autore.

1.3.4 Tratti semantici

Un'analisi di tipo semantico è difficilmente automatizzabile, per questo sono stati fatti pochi studi sull'uso di tratti semantici e pragmatici.
Uno dei lavori più importanti nell'ambito è quello di Argamon (Argamon, Whitelaw, Chase, & Hota , 2007).
Il suo lavoro è stato ispirato dalla teoria della grammatica sistemico funzionale di Halliday.
La grammatica sistemico funzionale si focalizza sul contesto, sulla funzione e sul ruolo sociale del linguaggio: ad esempio nella grammatica sistemico funzionale lo schema

CONGIUNZIONE denota come una data frase si espanda di alcuni concetti rispetto al contesto precedente.
Alcuni tipi di espansione nello schema CONGIUNZIONE possono essere ELABORAZIONE (esemplificazione), ESTENSIONE (aggiunta di nuova informazione), AUMENTO (qualificazione): alcune parole o frasi rientreranno nello schema congiunzione e saranno quindi collegate a dei tipi di espansione, ad esempio la parola "specialmente" verrà utilizzata per una CHIARIFICAZIONE di una ELABORAZIONE di una CONGIUNZIONE e così via.
Per rilevare queste informazioni semantiche Argamon ha utilizzato un lessico formato da parole e frasi prodotto semi-automaticamente e basato su thesaurus online.
Ogni input nel lessico è stato poi associato a una parola o a una frase con un set di limiti sintattici e di proprietà semantiche.
Con queste associazioni Argamon ha poi potuto vedere quante CONGIUNZIONI fossero poi espanse a ELABORAZIONI o quante ELABORAZIONI venissero elaborate a CHIARIFICAZIONI e così via.
L'accuratezza di questa misura non è stata però mai dichiarata; sembra però che i tratti semantici possano aiutare nell'attribuzione d'autore se associati con tratti basati sul lessico.

1.3.5 Tratti specifici per applicazione

I tratti prima descritti, cioè quelli lessicali, grafici, sintattici e semantici, sono indipendenti dalla loro applicazione, dato che con strumenti adeguati possono essere estratti da qualsiasi tipo di testo.
Al di là di questi tratti esistono però delle misure specifiche per alcune tipologie di testo e di formato di testo, dei tratti specifici per applicazione che possono rappresentare al meglio ogni possibile sfumatura di stile: ad esempio nell'analisi di e-mail, messaggi e forum online si sono infatti create delle misure specifiche per quantificare lo stile dell'autore come l'uso dei saluti, i diversi tipi di firma, l'uso dell'indentazione, la lunghezza dei paragrafi, ecc.
Alcune misure per testi in formato HTML possono essere invece la distribuzione dei tag, il colore dei font e la loro misura, ecc.
Queste misure, che sembrano essere minori in quanto relegate a domini specifici, diventano molto importanti ad esempio quando si dispone di testi molto corti che non possono essere sottoposti a tecniche indipendenti dall'applicazione.

Come abbiamo già detto nell'attribuzione d'autore le informazioni specifiche al contesto non vengono molto considerate proprio perché l'analisi non venga alterata da questo.
Quando però tutti i testi dell'analisi parlano di uno stesso argomento e sono di uno stesso genere si possono definire alcune parole che ricorrono frequentemente per un determinato argomento o genere testuale; queste informazioni se selezionate accuratamente possono rivelare alcune scelte autoriali.
Alcuni tipi di tratti specifici per applicazione possono essere validi soltanto per una determinata lingua naturale: Tambouratzis e altri (Tambouratzis, et al., 2004) hanno ad esempio proposto un set di terminazioni verbali tipiche del Katharevousa e della Dimotiki, rispettivamente il registro formale e informale nel greco moderno, in maniera da poter discriminare automaticamente fra i due diversi registri.

1.3.6 Selezione dei tratti e estrazione

Poiché alcuni tipi di tratti, come ad esempio quelli lessicali o grafici, possono di molto aumentare la variabilità del set di tratti, vengono allora usati degli algoritmi di selezione dei tratti per ridurne la variabilità.
Di solito i tratti selezionati da questi metodi sono sempre esaminati individualmente per discriminare gli autori di un dato corpus, anche se a volte, alcuni di questi tratti che sembravano irrilevanti esaminati individualmente, possono essere molto utili se combinati con altre variabili: per questo esistono algoritmi di selezione di tratti che selezionano la migliore combinazione di tratti da usare nell'analisi.
Al di là dell'uso di questi algoritmi, la caratteristica più importante della selezione di un set di tratti è la loro frequenza, in generale più un tratto è frequente e più riesce a catturare caratteristiche dello stile dell'autore.
Autori come Koppel, Adiva e Dagan (Koppel, Akiva, & Dagan, 2006) parlano invece di instabilità di tratti: più un tratto tende a modificarsi, ovvero è instabile, a differenza ad esempio di parole come *e* o *il* che sono molto stabili in quanto non hanno sinonimi, più quel tratto sarà un indicatore delle scelte stilistiche dell'autore.
Un altro approccio per ridurre la variabilità dei tratti è quello dell'estrazione di tratti: in pratica si crea un set di caratteristiche "sintetiche" combinando il set di caratteristiche iniziali.
La principale tecnica di estrazione dei tratti è l'analisi in componenti principali, una tecnica per la semplificazione dei dati che rappresenta i testi in uno spazio bidimensionale.

1.4 Metodi per l'attribuzione d'autore

In ogni problema di attribuzione d'autore abbiamo un set di candidati autori, un set di campioni di testo attribuiti con certezza a tutti i candidati autori (il training corpus) e un set di campioni di testo di autore sconosciuto o dubbio (test corpus), ognuno dei quali dovrebbe essere attribuito a un candidato autore.

Stamatatos distingue tra due principali metodi per l'attribuzione d'autore: *profile based approaches*, da non confondere con i metodi di profilazione d'autore, cioè estrarre informazioni sull'autore come età, sesso, grado di istruzione ecc, e *instance based approaches*, da non confondere con i metodi di *istance based learning*.

I primi, i *profile based approaches*, uniscono tutti i test del training corpus in unico grande file da cui si estrae una rappresentazione cumulativa dello stile dell'autore, mentre i secondi, gli *instance based approaches*, rappresentano ogni testo del training corpus individualmente come un esemplare separato dello stile di un autore.

In generale i *profile based approaches* tentano di individuare lo stile generale di un autore, mentre gli approcci basati su esemplare mostrano uno stile individuale per ogni singolo documento.

Esistono poi anche dei metodi ibridi, metodi che combinano le caratteristiche dei due approcci.

1.4.1 Approcci profile based

Come abbiamo già detto i *profile based approaches* uniscono tutti i test del training corpus in un singolo file di testo (un singolo file di testo per ogni autore del training corpus) in modo da estrarre lo stile di tutti gli autori che compongono il training corpus (in figura 1 Xa e Xb, cioè vettori di tratti di rappresentazione dei testi dell'autore A e dei testi dell'autore B).

Il profilo del testo d'autore sconosciuto (Xu) viene confrontato con i profili di testo creati per ogni autore del training corpus (Xa e Xb), si misura quindi lo stile d'autore più simile al testo d'autore sconosciuto attraverso una misura di distanza, trovando così il candidato più probabile al testo di autore sconosciuto.

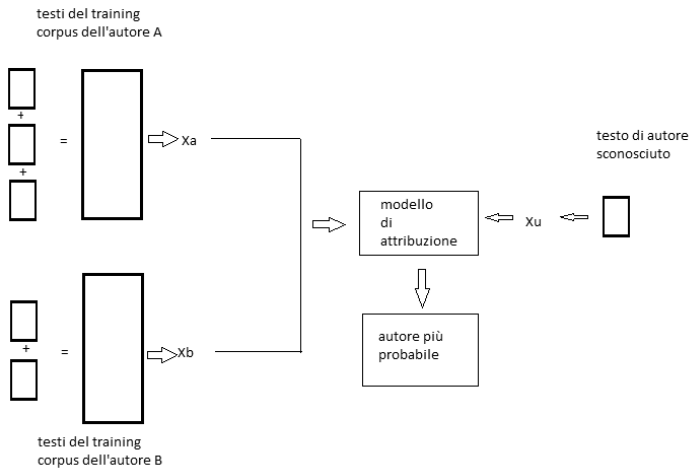


Figura 1 Figura ripresa e tradotta da Stamatatos 2009, p.13

Questo tipo di approccio può essere usato con modelli di tipo probabilistico, cioè metodi che massimizzano la probabilità P(*x*|*a*) di un testo *x* di appartenere al candidato autore *a*.

Un esempio dell'uso di modello di tipo probabilistico può essere visto nel lavoro di Mosteller e Wallace (Mosteller e Wallace 1964, per ulteriori informazioni sul loro lavoro si rimanda al paragrafo 1.1.2).

La metodologia basata su profilo usa molto anche gli algoritmi di compressione (questa metodologia è stata già descritta al paragrafo 1.3.2).

Altro approccio molto usato nei *profile based approaches* è quello degli *n*-grammi comuni (Common n-grams o CNG), descritto da Keselj e altri (Keselj, Peng, Cercone, & Thomas, 2003).

Secondo questo metodo il profilo PR(*x*) di un testo *x* è composto dagli *L* più frequenti *n*-grammi del testo.

Viene poi calcolata una misura di distanza per vedere la similarità tra un testo *x* e un testo *y*, questa distanza computa la dissimilarità di due profili calcolando lo scarto relativo tra i loro *n*-grammi in comune, mentre tutti gli *n*-grammi che non sono in comune contribuiscono con un valore costante alla distanza.

Il metodo CNG ha due importanti parametri che devono essere settati: la misura di profilo *L* e la lunghezza degli *n*-grammi *n*, cioè quante stringhe e di quale lunghezza debbano costituire il profilo.

Un grosso problema nell'attribuzione d'autore si presenta quando la distribuzione degli autori nel training corpus è irregolare, ad esempio quando si hanno molti testi nel training corpus di un candidato autore e pochi di un altro.

Questo problema nel *machine learning* viene chiamato *class imbalance problem*; si è visto come il CNG fallisca in questi casi.

Per questo è stata creata da Frantzeskou e altri (Frantzeskou, Stamatatos, Gritzalis, & Katsikas, 2006) una distanza più semplice chiamata SPI *Simplified Profile Intersection*, una distanza che conta semplicemente il numero di *n*-grammi comuni tra i due profili, in cui l'autore più probabile è quello con il valore SPI più alto.

La SPI dà però un valore in sovrastima quando tutti i candidati autori hanno testi molto corti eccetto uno, per questo Stamatatos (Stamatatos 2007) ha proposto una riformulazione della CNG che funziona meglio sia in questo caso che in altri casi di *imbalance problem*.

1.4.2 Approcci istance based

Negli *instance based approaches* ogni testo del training corpus viene considerato come un'unità che contribuisce in maniera separata al problema di attribuzione, cioè ogni campione di testo di autore conosciuto viene visto come un'istanza del problema di attribuzione.

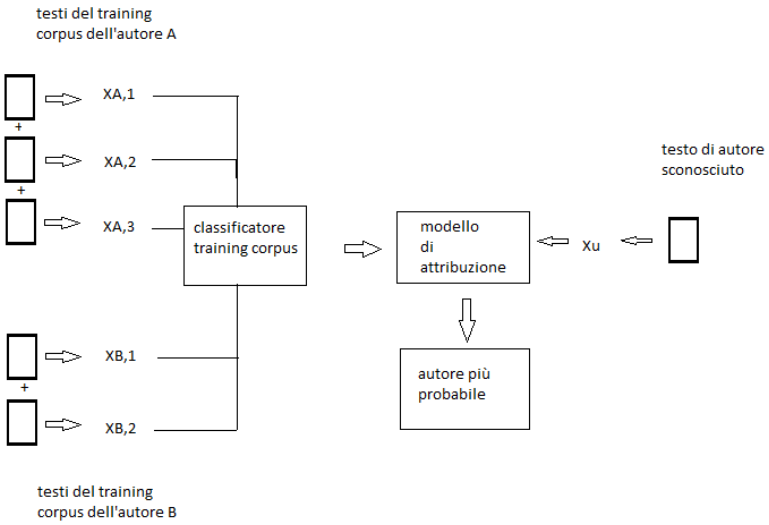


Figura 2 Figura ripresa e tradotta da Stamatatos 2009, p.13

Ogni campione di testo del training corpus è rappresentato da un vettore di attributi (X), in seguito un algoritmo di classificazione viene addestrato usando un set di esemplari di autori conosciuti (il training set), in modo da sviluppare un modello di attribuzione; in seguito questo modello di attribuzione stabilirà l'autore del testo sconosciuto.

Bisogna però sottolineare come questi algoritmi di classificazione richiedano più esemplari di training set per estrarre un modello affidabile, inoltre se abbiamo un solo testo molto lungo

per un candidato autore (ad esempio un intero libro) questo dovrà essere segmentato in più parti, probabilmente di uguale lunghezza.
Quando invece ci sono più campioni di testo di lunghezza variabile per autore, la lunghezza di testo dell'esemplare del training set deve essere normalizzata: infatti i testi di ogni autore del training test vengono segmentati in campioni di uguale lunghezza.
In ogni caso però i campioni di testo dovrebbero essere lunghi abbastanza in modo che i tratti di rappresentazione del testo possano rappresentare adeguatamente lo stile, in letteratura sono state proposte varie misure di lunghezza.
Con l'approccio *istance based* può essere usato il *vector space model*, un modello in cui ogni testo viene visto come un vettore in uno spazio multivariato; diversi algoritmi di apprendimento statistici e matematici sono stati usati per la creazione di un modello di classificazione.
Sono stati usati anche modelli basati su similarità, cioè modelli che calcolano una misura di similarità tra il testo dell'autore sconosciuto e tutti i testi del training, la stima dell'autore più vicino è basata sull'algoritmo *nearest neighbour* (l'algoritmo vicino più vicino); un esempio di modello di similarità è il Delta di Burrows, (descritto al paragrafo 1.1.2).
Esistono anche modelli di *meta learning*[15] usati nell'approccio basato su esemplare, uno di questi modelli è *l'unmasking method* di Koppel e altri (Koppel, Schler, & Bonchek-Dokow, 2007).

1.4.3 Approcci ibridi

Un metodo ibrido, a metà fra *profile based approaches* e *instance based approaches* è stato descritto da Van Halteren (Van Halteren, 2007).
Come negli *instance based approaches*, nel metodo di Van Halteren tutti i campioni di testo del training set sono rappresentati singolarmente, viene però prodotto un singolo profilo per ogni autore, come nei *profile based approaches*.
Si calcola poi una distanza tra il testo sconosciuto e ogni profilo d'autore basata su tre fattori: uno per la differenza fra i valori dei tratti del profilo di testo sconosciuto e del profilo d'autore, uno per l'importanza del tratto del testo sconosciuto e uno per l'importanza del tratto per un particolare autore.

[15] Il meta learning è un settore del machine learning: algoritmi di apprendimento automatici vengono applicati a dei meta dati su esperimenti di machine learning. Lo scopo finale è migliorare la performance degli algoritmi di apprendimento già esistenti.

1.4.4 Comparazione fra approcci

Quando abbiamo soltanto testi molto corti nel training corpus (ad esempio e-mail o messaggi di forum online) conviene utilizzare un approccio del tipo *profile based*, proprio perché l'insieme dei diversi testi, in quanto troppo corti se presi singolarmente, può produrre una rappresentazione più valida.

Conviene usare l'approccio *profile based* anche nel caso in cui ci sia un solo testo molto lungo o pochi testi molto lunghi per un autore, poiché nell'approccio *instance based* il testo dovrebbe per forza essere segmentato.

Gli approcci *instance based* riescono però a gestire dati di grandi dimensioni e poco ordinati grazie all'uso di potenti algoritmi, riescono anche a combinare diversi tratti stilometrici in maniera molto più semplice rispetto agli approcci *profile based*.

Inoltre, alcuni tratti definibili soltanto a livello testuale, come l'uso dei saluti e della firma, non possono essere facilmente utilizzati negli approcci *profile based*, proprio perché questo approccio rappresenta un profilo generale dell'autore rispetto alle caratteristiche specifiche di un testo.

Mentre negli approcci *profile based* il costo per il tempo di training è basso, negli approcci *istance based* è relativamente alto; in entrambi gli approcci invece il costo del tempo di calcolo è basso; soltanto con l'uso di metodi basati sulla compressione il costo del tempo di calcolo aumenta.

Negli approcci i*stance based* lo squilibrio della classe (*class imbalance*) può dipendere soltanto dal numero dei testi e a volte può apparire anche quando lunghi testi nel training test sono stati segmentati in più parti, mentre negli approcci *profile based* lo squilibrio di classe dipende soltanto dalla lunghezza e può apparire nel caso in cui si abbia uno stesso numero di campioni di testo per due autori ma i testi del primo autore sono più corti, mentre quelli del secondo autore più lunghi: la concatenazione dei testi di training per ogni autore produrrà infatti due file che differiranno molto per la lunghezza.

CAPITOLO II

TRE CASI DI ATTRIBUZIONE D'AUTORE CON UTILIZZO DI METODI QUANTITATIVI

In questo capitolo andremo ad illustrare tre studi che hanno utilizzato metodi quantitativi per l'attribuzione d'autore: lo studio di Patrick Joula su *The Cuckoo's Calling* di Robert Galbraith (pseudonimo di J.K. Rowling) (Joula, 2015) (Joula, 2013), lo studio di Jan Rybicki su metodi di attribuzione d'autore applicati alle traduzioni (Rybicki, 2012) e infine lo studio di Cyril Labbé e Dominique Labbé sull'attribuzione d'autore applicata alle opere di Molière e Corneille (Labbè & Labbè, 2001)

2.1 Il caso Robert Galbraith

The Cuckoo's Calling è un poliziesco scritto da Robert Galbraith, pubblicato nel 2013 dalla casa editrice Sphere (in Italia è stato pubblicato nel 2013 da Salani).

All'epoca del lancio del libro si era a conoscenza del fatto che Robert Galbraith fosse uno pseudonimo: si pensava infatti che dietro a Galbraith si nascondesse un ex militare della Royal Military Police e proprio per proteggere la sua identità si sarebbe quindi utilizzato uno pseudonimo.

Nessuno avrebbe mai sospettato che dietro al nome Robert Galbraith si potesse nascondere non solo una donna, ma persino la scrittrice inglese più famosa d'Inghilterra: J. K. Rowling, la creatrice della saga di Harry Potter.

Quando un reporter del Sunday Times di Londra contattò il professor Patrick Joula gli disse che aveva ricevuto una soffiata sul fatto che Robert Galbraith fosse in realtà J.K. Rowling.

Subito dopo la soffiata il reporter aveva trovato alcune prove a sostegno di questa: Galbraith si soffermava infatti in descrizioni di abiti femminili particolarmente dettagliate (fatto insolito per un ex-militare), un'altra prova era quella che Galbraith e la Rowling avevano non solo lo stesso editor ma anche lo stesso agente.

Il reporter sapeva però che queste prove non potevano di certo far confessare la Rowling, serviva qualcosa di più per sostenere una simile tesi e proprio per questo chiese a Joula di condurre un'analisi quantitativa su *The Cuckoo's Calling*.

2.1.1 L'analisi su *The Cuckoo's Calling*

Joula ha eseguito questa analisi con un software da lui sviluppato chiamato JGAAP (Java Graphical Authorship Attribution Program), un software che con analisi di tipo matematico riesce a trovare un grado di similarità fra due autori analizzando un grosso numero di caratteristiche, troppe per essere analizzate da un essere umano.
Per condurre questa analisi Joula ha avuto bisogno non soltanto di una copia elettronica di *The Cuckoo's Calling,* ma anche delle copie elettroniche di *The Casual Vacancy* (giallo scritto dalla Rowling col suo vero nome) e copie elettroniche di gialli di scrittrici britanniche contemporanee alla Rowling: *The St. Zita Society* di Ruth Rendell, *The Private Patient* di P.D. James e *The Wire in the Blood* di Val McDermid.
I romanzi delle autrici inglesi servivano infatti da "distrattori" durante l'analisi: è stato visto infatti come l'attribuzione d'autore risponda meglio a problemi di classe chiusa, cioè problemi che rispondano alla domanda: "Quale fra questi autori è con maggiore probabilità l'autore di questo testo?" (Joula, 2015, p. 106).
I testi sono stati poi ripuliti: sono state eliminate le prefazioni e i sommari, infine sono state fatte alcune operazioni di normalizzazione dei testi (in questo caso eliminazione delle virgolette, semplificazione della punteggiatura, riduzione degli spazi bianchi).
The Cuckoo's Calling è stato poi diviso in sezioni di 1000 righe l'una; ognuna di queste sezioni è stata comparata singolarmente con un modello di base per ognuna delle quattro autrici.
Sono stati eseguiti quattro diversi tipi di analisi per differenti tipi di caratteristiche linguistiche, una delle caratteristiche analizzate è stata la distribuzione della lunghezza delle parole.
Usando una formula di distanza matematica (la formula di distanza normalizzata del coseno) Joula ha elaborato un indice di similarità, che va da 0.0 (situazione di autore identico) a numeri progressivamente più alti (maggiore dissimilarità).

Usando questa formula in riferimento alla distribuzione della lunghezza delle parole per le undici sezioni di *The Cuckoo's Calling* sei sono risultate più vicine alla Rowling, cinque alla James.
Un'altra analisi è stata eseguita sulle cento parole più comuni (molte di queste saranno parole funzione, cioè in questo caso *the*, *of,* ecc.)
Con questa analisi, usando una formula di similarità, quattro sezioni sono state attribuite alla Rowling, quattro alla McDermid, le altre tre invece sono state divise tra James e Rendell.
Sono stati eseguiti in seguito due test: il primo basato sugli *n*-grammi con valore settato a quattro (per la definizione di *n*-gramma si veda il paragrafo 1.3.2), il secondo invece sugli *n*-grammi parola (anche per questa definizione si veda il paragrafo 1.3.1).
L'analisi con i *4*-grammi ha mostrato una preferenza per la McDermid, attribuendole otto sezioni, tre sono state attribuite invece alla Rowling.
L'analisi con gli *n*-grammi parola ha invece mostrato una netta preferenza per la Rowling; con questo metodo le sono state attribuite nove sezioni, due invece sono state attribuite alla McDermid.

2.1.2 I risultati finali

Cosa possono dire questi risultati?
Fra tutte le candidate le uniche possibili sarebbero la McDermid e la Rowling, in quanto nessun test ha puntato verso la Rendell e solo un test (secondo Joula poco affidabile) ha selezionato la James.
I valori della distribuzione della lunghezza delle parole della McDermid sono però molto distanti da quelli di Galbraith; l'unica certezza è che la Rowling appare in tutti i risultati, venendo scelta sempre come autrice più probabile o come seconda possibilità.
Joula spiega come questo risultato non sia naturalmente una prova schiacciante del fatto che la Rowling abbia scritto *The Cuckoo's Calling*, da questa analisi risulta soltanto come la candidata più probabile.
Sicuramente l'analisi ha dato però un risultato molto significativo: quando il reporter del Sunday Times con questi dati alla mano chiese direttamente all'agente della Rowling se fosse stata lei a scrivere *The Cuckoo's Calling*, il giorno dopo J.K.Rowling ammise di essere l'autrice del romanzo.

Joula porta questa analisi come modello per l'attribuzione d' autore perché tutti i testi di confronto hanno un autore certo, i testi di confronto sono stati reperiti già digitalizzati quindi con minori errori rispetto a testi da digitalizzare, i testi "distrattori" sono stati scelti con criteri adeguati, è stato proposto uno specifico candidato ad autore e l'attribuzione risponde ad una domanda semplice e netta (è questo l'autore del testo? L'unica risposta possibile è sì o no).
È interessante notare come Joula sottolinei che questo metodo di attribuzione d'autore e i metodi quantitativi di attribuzione d'autore in generale siano ben lontani dall'essere una minaccia per la privacy delle persone: la sua analisi è entrata in gioco solo dopo la soffiata ricevuta al Sunday Times e quindi non ha minato la privacy della Rowling perché il dubbio era presente già a priori, i suoi metodi l'hanno soltanto confermato.

2.2 La mano invisibile del traduttore: metodi quantitativi e traduzioni

Rybicki fa notare al lettore come la figura del traduttore sia spesso bistrattata: sconosciuti al pubblico non specialista, spesso mal pagati, i traduttori in realtà compiono un nobile servizio alla letteratura e alla società in generale.
Il traduttore traduce opere letterarie (e non solo) dalle lingue in cui sono state originariamente scritte alla lingua di interesse, rendendo possibile leggere, per un pubblico vastissimo, scritti che non sarebbero mai potuti essere letti nella lingua originale.
La domanda che pervade i lavori qui presentati di Rybicki è la seguente: è possibile con i metodi quantitativi scoprire l'impronta stilistica del traduttore? Oppure è lo stile dello scritto originale ad essere maggiormente visto da questi metodi?
Per trovare risposte a queste domande Rybicki ha confrontato diversi corpora costituiti da corpus formati da originali e traduzioni.

2.2.1 L'analisi su corpora di traduzioni

Rybicki ha utilizzato un'analisi multivariata basata sui punteggi *z* della formula del Delta di Burrows (descritta al paragrafo 1.1.2).

I punteggi *z* sono stati elaborati in un'analisi *cluster* [16](detta anche analisi dei gruppi) per produrre dei diagrammi ad albero per un dato set di parametri come il maggior numero di parole più frequenti, l'eliminazione dei pronomi e il *rate culling*.

Il *rate culling* si esprime in percentuale; questa percentuale specifica il numero di testi nel corpus in cui una data parola deve apparire per poter essere presa in esame: una *rate culling* del 100% limita l'analisi a tutte le parole che appaiono almeno una volta in ogni testo del corpus, una *rate culling* del 50% limita l'analisi a parole che appaiono almeno nella metà dei testi del corpus, una *rate culling* dello 0% analizza tutte le parole presenti nei testi del corpus, senza nessuna esclusione.

Questi risultati, prodotti per una grande varietà di combinazioni, sono stati inseriti in una procedura di *bootstrap:*[17] in pratica una serie di diagrammi ad albero individuali dell'analisi *cluster* formano un albero *bootstrap* finale.

L'albero *bootstrap* finale è stato fatto poiché è stato dimostrato da diversi studi che l'analisi *cluster* da sola non è affidabile.

Tutta la procedura è stata eseguita utilizzando il programma statistico R: i testi elettronici sono stati processati per creare una lista di tutte le parole utilizzate nei testi con le frequenze di ogni singolo testo, per poi creare una matrice di input iniziale di parole (righe) per testi singoli (colonne); ogni cella conteneva quindi una data frequenza di parole in un dato testo. Le frequenze sono state poi normalizzate, sono state selezionate per l'analisi delle parole con determinate frequenze, è stata poi eseguita la cancellazione dei pronomi e il *rate culling*.

Si sono comparati i risultati per ogni test individuale, sono stati eseguiti i calcoli del Delta per ogni set di parametri, sono state ottenute le similarità/dissimilarità di Delta attraverso il clustering e infine sono stati prodotti gli alberi finali con la procedura di *bootstrap* (per la procedura completa si veda Eder, Maciej e Rybicki, 2011).

La validità di questo metodo è stata provata con un corpus composto da sessantacinque romanzi inglesi di diciannove autori diversi, sono stati utilizzati duecentocinquanta diagrammi di analisi *cluster* creati con diversi parametri per formare il diagramma bootstrap ad albero mostrato in figura 3.

[16] Il Clustering in statistica è un insieme di tecniche di analisi multivariata dei dati adibita alla selezione e al raggruppamento di elementi omogenei in un insieme di dati.

[17] Tecnica statistica di ricampionamento con reimissione per approssimare la distribuzione campionaria di una statistica.

Come si può vedere dalla figura 3, le opere di uno stesso autore sono state correttamente raggruppate ognuna in un "ramo" unico per autore, inoltre le opere sono raggruppate fra di loro nella figura anche in termini di vicinanza di corrente letteraria.

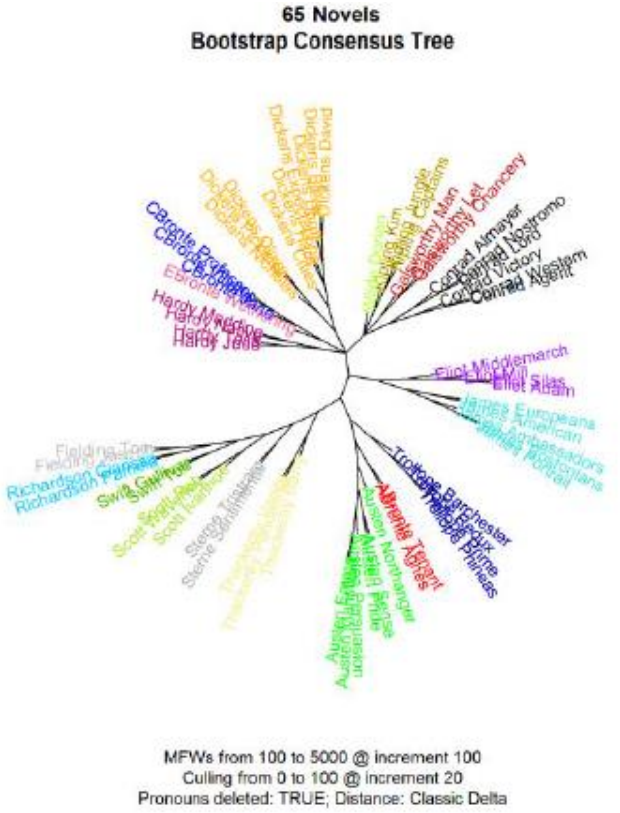


Figura 3 Figura presa da Rybicki, 2012

2.2.2 I risultati dell'analisi

Rybicki oltre ad essere un esperto nel campo della digital humanities è soprattutto un prolifico traduttore dall'inglese al polacco e viceversa; gli è sembrato quindi naturale provare a testare il metodo con le sue traduzioni.

Rybicki ha infatti analizzato un corpus di nove traduzioni da lui eseguite dall'inglese al polacco comprendente due autori tradotti più volte: il canadese Douglas Coupland con *Generation X, Polaroids from the Dead, The Gum Thief* e il britannico John le Carrè con *A Perfect Spy, Absolute Friends, The Missions Song, Tinker Tailor Soldier Spy, A Most Wanted Man, A Small Town in Germany*.

Come si può vedere dalla figura 4 il risultato del grafico è quello di un unico traduttore che traduce due diversi autori: i romanzi infatti si accoppiano a seconda del loro autore.

Figura 4 Figura presa da Rybicki 2012

Se si espande il corpus a lavori di altri autori e traduttori la divisione per autore originale e non per traduttore diventa ancora più evidente.

In figura 5 vediamo il risultato dell'analisi su un corpus costituito da traduzioni in polacco fatte da diversi traduttori di sessantacinque romanzi di undici autori inglesi, francesi ed italiani.

In questo caso è interessante osservare come i lavori di uno stesso autore si raggruppino insieme nonostante siano stati tradotti da traduttori diversi, sebbene però a volte si possano osservare, all'interno di uno stesso ramo, raggruppamenti per uno stesso traduttore (ad esempio nel caso delle traduzioni della Austen).

È possibile vedere anche come i raggruppamenti separati di autori tradotti da uno stesso traduttore siano però vicini (ad esempio nel caso delle traduzioni di Coupland e le Carrè fatte da Rybicki) e si può notare inoltre come le tre traduzioni dei volumi singoli di una stessa saga vengano raggruppati per volume invece che per traduttore (in questo caso i volumi delle traduzioni di Tolkien in polacco).

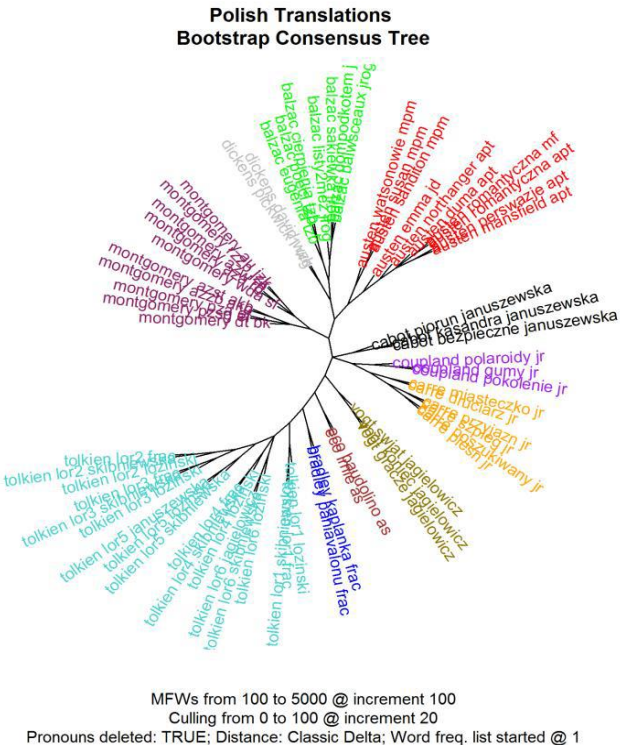


Figura 5 Figura presa da Rybicki 2012

La figura 6 mostra il risultato su un corpus di traduzioni dal polacco all'inglese delle opere di Sienkiewicz.

Anche in questo caso troviamo una trilogia tradotta da due traduttori differenti, ma in questo caso il raggruppamento per volume della trilogia è limitato soltanto alle traduzioni di Curtin e Binon, mentre il lavoro di Kunziack è separato da questi.

Rybicki spiega come questo risultato non sia affatto inatteso: la traduzione della trilogia di Kunziack segue un modello di traduzione molto moderna e esplicativa, tanto che è stata definita da alcuni critici un adattamento più che una vera e propria traduzione (Segel, 1991). La traduzione di Kunziack ha infatti una lunghezza dei token espansa del 150-170% (mentre il valore standard dell'espansione nella lunghezza dei token nelle traduzioni dal polacco all'inglese è del 120-130%), proprio perché Kunziack alla traduzione ha unito delle aggiunte di suo pugno per spiegare alcuni passaggi del romanzo e ha fatto inoltre delle pesanti modifiche al testo originale.

Le traduzioni di romanzi singoli invece si raggruppano, come abbiamo visto negli altri casi, per autore e non per traduttore; il grafico inoltre riflette bene anche la divisione fra i romanzi storici di Sienkiewicz (nella metà in basso della figura) e quelli invece ambientati nel tempo contemporaneo all'autore, il XIX secolo (nella metà in alto della figura).

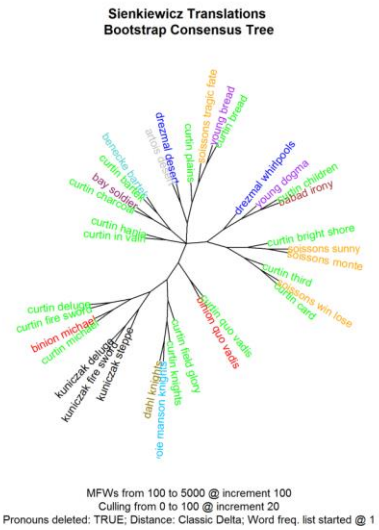


Figura 6 Figura ripresa da Rybicki, 2012

Altra interessante analisi è quella mostrata in figura 7 ed eseguita su un corpus costituito da traduzioni dall'inglese al Polacco di settanta romanzi thriller di Clancy, Coben, Grisham, Koontz e Ludlum, comprendendo anche le collaborazioni di Ludlum. Il Delta non solo riesce a raggruppare i lavori di ogni autore, ma sembra riuscire anche a distinguere le collaborazioni di Ludlum dalle sue opere individuali, nonostante spesso il traduttore sia in comune.

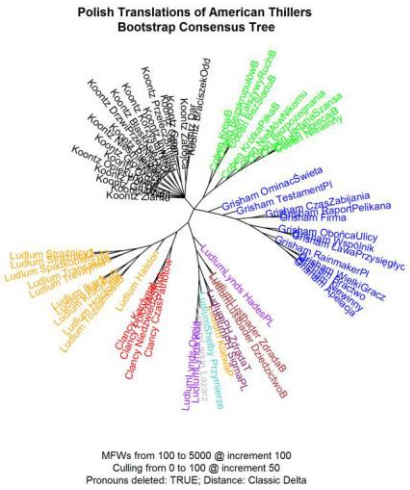


Figura 7 Figura ripresa da Rybicki, 2012

In figura 8 è possibile vedere un'analisi su di traduzioni di romanzi dall'inglese al francese; si può vedere come sia ancora ben visibile il raggruppamento per autore e non per traduttore. Le poche eccezioni al raggruppamento per autore sono facilmente spiegabili: le due traduzioni della Austen che non compaiono nel raggruppamento della Austen sono delle

prime traduzioni compendiate e la prefazione in una di esse suggerisce lo stesso editore per entrambe, quindi con molta probabilità anche lo stesso traduttore/compendiatore.
Molto interessante notare come il raggruppamento fra Charlotte e Emily Brontë rifletta molto bene la vicinanza fra le due sorelle, riprendendo quindi i risultati della figura 3.

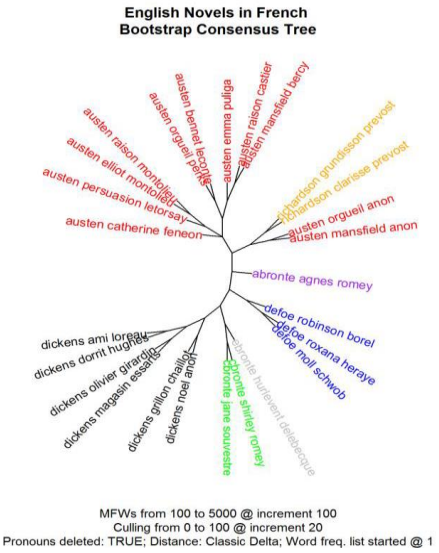


Figura 8 Figura ripresa da Rybicki, 2012

Se a questo corpus vengono aggiunti due romanzi in francese scritti da due dei traduttori non si ha una situazione di uniformità, come si può vedere in figura 9: mentre *Manon Lescaux* di Abbé Prévost si avvicina nella figura alle sue traduzioni di Richardson, *Caroline de Lichtfield* della Baronessa di Montolieu invece non si colloca vicino alle sue traduzioni di Jane Austen.

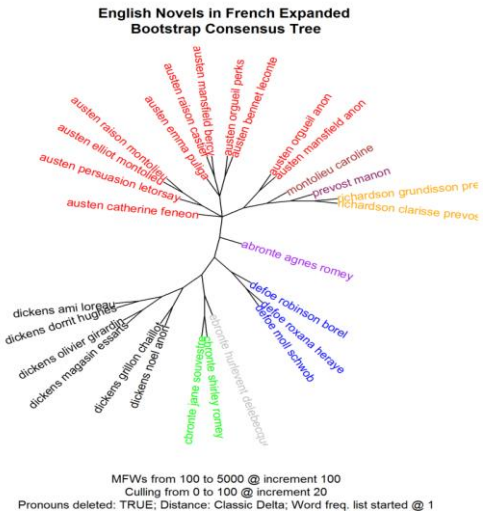


Figura 9 Figura ripresa da Rybicki, 2012

2.2.3 Conclusioni finali

Dall'insieme di tutti questi risultati la risposta alla domanda iniziale di Rybicki sembra essere proprio che con il Delta non si riesca a vedere la mano del traduttore ma soltanto quella dell'autore originale della traduzione.

Il traduttore sembra quindi relegato all'invisibilità anche dalla statistica, poiché viene riconosciuto dal metodo soltanto quando fa qualcosa di estremamente sbagliato o molto controverso dal punto di vista della traduzione classica, come ad esempio eliminare parti di un romanzo o inserire le proprie considerazioni personali all'interno di questo.

Naturalmente tutti coloro che ritengono che il lavoro del traduttore debba essere relegato all'invisibilità, poiché il traduttore deve soltanto far parlare il testo dell'autore nella sua lingua non originale, saranno lusingati da questi risultati del Delta.

Il Delta usa però le 5000 parole più frequenti, una parte molto piccola delle 50000 parole che un adulto usa (Miller, 1996): molto probabilmente non riesce ad evidenziare in maniera totalmente precisa la libertà della scelta di parola o la sua manipolazione (Burrows, 1987).

L'uso di una determinata parola, anche se è l'uso più inconscio possibile di una parola funzione e non di una parola contenuto, da sola non può rappresentare lo stile di un autore o almeno non unicamente questo.

Questi risultati infatti non mostrano che il traduttore non lascia mai una traccia: molti lavori nell'ambito degli studi sulla traduzione si concentrano su quelle che vengono chiamate le "impronte" del traduttore o le "tendenze deformanti" del traduttore e non sono di certo errati.

Quello che invece si può affermare guardando a questi risultati è che il Delta sembra non riuscire a catturare queste "impronte" o "tendenze deformanti" per le ragioni già dette prima.

Secondo Rybicki infatti Delta rappresenta maggiormente le scelte consapevoli di contenuto, infatti raggruppa i romanzi del corpus più per autore e volume che per traduttore; nelle traduzioni questa rappresentazione viene probabilmente amplificata in quanto due traduzioni diverse di uno stesso testo nella stessa lingua hanno molto più in comune che qualsiasi due testi letterari scritti nella stessa lingua.

Tutto ciò non significa naturalmente che i metodi quantitativi non potranno mai trovare "la mano" dell'autore, altre misure sempre ideate da Burrows come Zeta e Iota, che rispettivamente calcolano le frequenze di parola medie e basse, potrebbero riuscire a evidenziare le caratteristiche stilistiche dei testi di due o più traduttori.

2.3 Uno studio d'attribuzione d'autore su Molière e Corneille

Già da tempo si è insinuato il dubbio negli studiosi di letteratura francese che alcune opere di Molière siano in realtà state scritte da Corneille.
In una edizione delle opere di Molière, più esattamente nel frontespizio di *Psyche* del 1671, l'editore avverte infatti che l'opera è stata scritta per due terzi da Corneille, ma che in precedenza era stata rappresentata col nome di Molière.
In molti, come il poeta P. Louys e gli scrittori belgi Wouters e Ville de Goyer (Wouters & Ville De Goyet, 1990), hanno sostento che Molière non è l'autore di tutte le sue opere.
Nello studio qui presentato Labbè ha creato una formula, da lui definita distanza intertestuale, con cui ha analizzato un corpus composto da opere dei due celebri drammaturghi francesi per fare luce sulla questione.

2.3.1 La distanza intertestuale

La distanza formulata da Labbè si basa sulle frequenze di tutti i type di un testo e dà una distanza di quanto un testo sia simile o dissimile da un altro testo.
La distanza che Labbè voleva formulare doveva essere:

- non sensibile alla diversa lunghezza dei testi comparati

- applicabile a più testi e possibilmente a tutti i testi scritti in una stessa lingua

- con valori che partono da 0 (stesso vocabolario e stessa frequenza di ogni type nei due testi confrontati) a 1 (nessun type in comune); senza salti o effetti soglia attorno ad alcuni valori

- simmetrica, nel senso che dati due testi A e B, la distanza tra il testo A e il testo B e la distanza tra il testo B e il testo A deve essere la stessa, cioè $\delta(A,B) = \delta(B,A)$.

- il più transitiva possibile: quando vengono uniti due testi, la distanza di questa unione rispetto a un altro testo deve essere la stessa di quella precedente all'unione: se la distanza dei testi singoli è $\delta(A,B) < \delta(A,C) < \delta(B,C)$ allora $\delta(A,B) < \delta\{A,(B \cup C)\}$

- il più solida possibile (un cambiamento marginale in uno dei due testi deve riflettere anche un cambiamento marginale nella loro distanza).

Erano già state create precedentemente due formule di distanza da Muller e Brunet, dove la distanza assoluta tra due testi A e B veniva data dall'unione dei due testi meno la loro intersezione, cioè dalla somma delle differenze fra le frequenze assolute di ogni type nel testo.
Queste formule[18] però avevano diversi problemi:

- le due formule sono equivalenti soltanto nel caso in cui le lunghezze dei testi presi in esame siano eguali, mentre se nessun type è in comune le due formule danno un risultato 1 indipendentemente dalla lunghezza del testo.

- il minimo teorico 0 può essere raggiunto soltanto nel caso di testi di uguale lunghezza.

- in entrambe le formule l'intersezione dei due testi viene contata due volte, perciò viene data maggiore importanza ai type in comune piuttosto che al vocabolario specifico di ogni testo.

Per ovviare a questo problema Labbè in pratica riduce la dimensione del testo più grande alla dimensione del testo più piccolo[19].

18

1) $\delta(A,B) = \frac{\sum_{VA}|FiA - \text{FiB}| + \sum_{VB}|FiB - FiA|}{NA+NB}$

2) $\delta(A,B) = \frac{1}{2}\left(\frac{\sum_{VA}|FiA - \text{FiB}| + \sum_{Vb}|FiB - FiA|}{NA+NB}\right)$

dove *VA* è il numero di type nel testo A, *VB* il numero di type nel testo B, *FiA* la frequenza del type di tipo *i* nel testo A, *FiB* la frequenza del type di tipo *i* nel testo B, *NA* il numero dei token presenti nel testo A e *NB* il numero dei token presenti nel testo B.

[19] Abbiamo due testi A e B, dove A è più corto di B.
Possiamo ridurre la frequenza dei type di B in ragione della dimensione di A con:

$F^*iB : FiB = NA : NB$

Otteniamo una stima della frequenza che ogni type di B assumerebbe se B avesse la stessa dimensione di A:

$F^*iB = FiB\frac{NA}{NB}$

Questa nuova formulazione non risolve però la doppia conta dell'intersezione, inoltre, se i due testi sono molto differenti, tutti i type del testo più lungo non potranno essere usati in quello più corto.
Per questo è stato proposto di considerare l'intersezione di testi soltanto una volta e limitare i calcoli a tutti i type del testo più corto e di considerare nel testo più lungo soltanto i type la cui frequenza è tale da aspettarne almeno uno nel testo più corto.
Labbè osserva come la precisione quantitativa della distanza sia leggermente ridotta dall'arrotondamento; la distanza infatti non dovrebbe essere utilizzata con testi troppo corti e non dovrebbe essere utilizzata una scala di misure troppo ampia.
Prima di calcolare la distanza i testi devono però essere normalizzati e tutti i token devono essere lemmatizzati; se si calcolasse la distanza senza aver fatto prima la normalizzazione la distanza dividerebbe ad esempio, in un corpus di testi in prosa e poesia, tutti i testi in prosa da quelli in poesia anche se non vi sono grandi differenze di contenuto.
La distanza intertestuale è stata testata su svariati corpora (da romanzi a discorsi dei primi ministri francesi e canadesi, fino ad arrivare a trascrizioni di interviste) e da questi studi è stata tratta una scala empirica per la distanza (questa scala empirica non è stata confermata però da altri studi successivi che utilizzano la formula di Labbè, per ulteriori informazioni si veda Cortelazzo, Tuzzi e Nadalutti 2013):

- per i testi di uno stesso autore le distanze sono sempre minori rispetto a quelle fra due autori differenti e contemporanei (quando parlano dello stesso argomento)

- distanze minori di 0.20 solitamente non esistono fra due autori differenti (su testi dello stesso tipo con simili argomenti). Nel caso di un autore sconosciuto, l'attribuzione con questi valori è praticamente certa. Se si è sicuri che nessuno dei due autori possa essere l'altro, uno dei due autori si è molto ispirato all'altro.

Si può quindi modificare la formula (1) della nota 19 sostituendo *FiB* con *F*iB*, ottenendo così una stima della distanza intertestuale:

$$\delta(A,B) = \frac{\sum_{i \in V_{A \cup B}} |FiA - F^{*}\mathrm{iB}|}{2NA}$$

Il valore zero viene raggiunto quando il testo più piccolo è un modello di quello più grande, in questo caso tutti i type di A sono presenti in B con la frequenza *FiA= F*iB* e il numeratore della formula sarà zero.
Il valore sarà invece 1 quando A e B non condividono nessun type.

- un valore tra 0.20 e 0.25 indica il caso in cui i testi sono molto simili. Se questi valori appaiono fra le opere di un solo autore significa che vi è stato un cambio nel tema o nel genere. Se uno dei due autori è sconosciuto l'attribuzione è possibile ma non certa e deve essere confermata da altre prove stilistiche.

- valori sopra 0.25 indicano che gli autori sono differenti o che genere e argomento delle opere sono troppo distanti per compararli.

2.3.2 L'analisi su Molière e Corneille

Una prima analisi della distanza intertestuale è stata eseguita su alcune delle più famose opere di Molière (i valori sono riportati in figura 10).
Si può notare come ci sia una forte similarità fra tutte le opere, sebbene i loro argomenti siano molto differenti.
I valori più bassi sono fra *Tartuffe* e *le Misanthrope* (.167), due opere scritte in alessandrini e che usano un linguaggio abbastanza lineare.
I valori più alti sono invece fra (.239) fra *le Misanthrope* e *le Bourgeois gentilhomme o le Malade imaginaire.*
La prima opera è in versi mentre le altre due sono in prosa e contengono molti neologismi formati dal turco e dal latino.

	Ecole des femmes	Tartuffe	Dom Juan	Le Misanthrope	L'Avare	Bourgeois gentilh.	Femmes savantes	Malade imaginaire
Ecole des femmes	0	.183	.205	0.194	0.200	.231	.198	.223
Le Tartuffe		0	.199	.167	.199	.230	.170	.219
Dom Juan			0	.204	.170	.207	.219	.205
Le Misanthrope				0	.210	.239	.173	.239
L'Avare					0	.194	.214	.187
Bourgeois gentilh.						0	.234	.196
Femmes savantes							0	.226
Malade imaginaire								0

Figura 10 Distanza fra le opere di Molière più conosciute, ripresa da Labbè, 2001.

Distanze maggiori di .20 separano *l'Ecole des femmes*, *Tartuffe*, *le Misanthrope* e *les Femmes savantes*, tutti scritti in versi, e *Dom Juan*, *l'Avare*, *le Bourgeois gentilhomme* e *le Malade imaginaire*, scritti tutti in prosa. Tenendo conto di queste differenze di struttura testuale è chiaro che tutti questi lavori appartengono allo stesso autore.

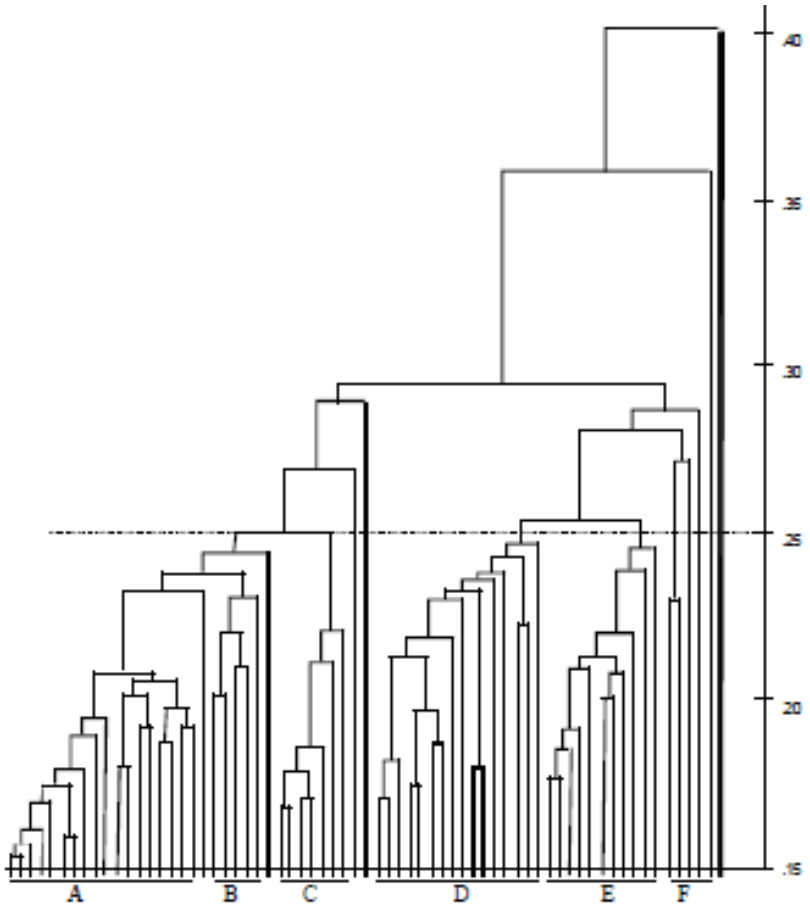
La tesi dell'unico autore è confermata anche da opere come *Tartuffe* e *Dom Juan*, la prima scritta in versi e la seconda in prosa: nonostante *Dom Juan* abbia un linguaggio formato da patois, un fattore che aumenta quindi la distanza, rimangono molto vicine (.199)
È stata trovata poi una media di distanza fra tutte le opere di Molière, la media generale è di .249.
Le sue commedie più famose in particolare sono quelle più vicine alla media generale, mentre alcune opere si distanziano di più, come ad esempio le prime commedie che Molière ha rappresentato prima di vivere a Parigi (la *Jalousie du barbouillé, le Médecin volant*), alcune recite scritte per particolari occasioni (*la Critique de l'Ecole des femmes* e *l'Impromptu de Versailles*) e altre opere come *les Précieuses ridicules* e *Dom Garcia*.
A parte queste eccezioni si può affermare con buona probabilità che tutte le opere sono di un singolo autore.
Un'ultima parte dell'analisi mostra come la parte di *Psychè* scritta da Molière sia molto lontana dai valori delle sue opere (.305) e soprattutto più lontana rispetto alla parte scritta da Corneille, anche se comunque distante dalla media dalle opere di Corneille.
Da questi risultati si può affermare che *Psychè* si pone in una posizione atipica sia nelle opere di Molière che di Corneille.
Un'altra analisi è stata invece fatta su un corpus formato da 64 opere di Molière e Corneille, comprendente 917000 tokens; le opere rientrano nel periodo storico 1630-1673.
È stata fatta prima un'analisi cluster su una matrice di distanza, le due opere più vicine sono state unite e la distanza di questo nuovo set rispetto a tutte le altre opere è stata ricalcolata per tutti gli altri raggruppamenti; questi risultati sono poi stati riassunti in un dendrogramma.
Il dendrogramma mostra accorpamenti per le tragedie mature di Corneille (in figura 11 il gruppo A), per le sue prime tragedie che lo hanno reso famoso (B) e per le sue commedie (C).
Il dendrogramma per Molière separa le commedie in versi (D) da quelle in prosa (E, F); infine alcuni raggruppamenti corrispondono a quelli delineati dai critici, come ad esempio la vicinanza fra opere di Molière come *les Femmes savantes*, *l'Ecole des Mar*is e *l'Ecole des femmes*.
Sono state trovate però alcune anomalie (le linee in grassetto in figura 11):

- Un'opera di Molière, il *Dom Garcie*, si trova fra quelle di Corneille. Presumibilmente quest'opera non è di Molière: infatti è molto vicina alle opere che Corneille ha scritto nello stesso periodo.

- Le due *Psychè* sono entrambe fra i lavori di Corneille.

- Due commedie di Corneille (*Menteur* e *Suite du Menteur*) sono state posizionate in mezzo alle commedie di Molière. Questa classificazione risulta molto bizzarra, in quanto le due commedie sono datate 1642-1643, mentre la prima opera di Molière è datata 1656.

Proprio l'attribuzione di *Menteur* e *Suite du Menteur* a Molière rende incerti i risultati: questa attribuzione non si spiega bene, è molto improbabile che Molière sia l'autore di queste opere. Una possibile spiegazione potrebbe essere quella che le due opere di Corneille siano state usate come modello da Molière.
Per chiarire meglio i risultati è stata eseguita un'analisi ad albero (figura 12).
In questa analisi testi o gruppi di testo non sono classificati uno per uno ma, per ogni uno, la migliore rappresentazione dei suoi vicini è stata comparata con tutte le altre.
Potremmo leggere questo grafico come una serie di testi (foglie) con una serie di collegamenti fra questi (rami e tronchi), dove i collegamenti che vi sono fra i vari testi misurano la loro prossimità.
Tutte le opere di Corneille (numeri da 1 a 33) appaiono raggruppate in basso a sinistra, con due divisioni all'interno: tragedie e tragi-commedie.



From left to right :
A. Corneille :
Tite et Bérénice
Pulchérie
Suréna
Agésilas
Othon
Sertorius
Sophonisbe
Attila
Nicomède
Don Sanche
Polyeucte
Théodore
Héraclius
Pertharite
Andromède
Toison d'Or
Rodogune
Oedipe

Dom Garcie
B : Corneille
Cinna
Pompée
Le Cid
Horace
Médée
Psyché Corneille
C Corneille comédies
Galerie du Palais
La Suivante
Mélite
La Veuve
La Place Royale
L'illusion comique
Clitandre
Comédie des Tuileries
Psyché Molière

D. Molière (verse)
Le Tartuffe
Le Misanthrope
Femmes savantes
L'étourdi
Dépit amoureux
L'école des maris
L'école des femmes
Amphytrion
Sganarelle
Le Menteur 1
Le Menteur 2
Mélicerte
Les fâcheux
Princesse d'Elide
E. Molière (prose)
Amants magnifiques
Le sicilien
Georges Dandin

L'avare
Dom Juan
Fourberies Scapin
Médecin malgré lui
M. Pourceaugnac
Malade imaginaire
Bourgeois gentil.
L'amour médecin
Mariage forcé
Ctesse d'Escarb.
F. Molière :
Critique de l'école
L'impromptu
Précieuses ridicules
Médecin volant
La jalousie
Psyché Quinault

Figura 11 Cluster Analysis delle opere di Corneille e Molière. Figura presa da Labbè, 2010.

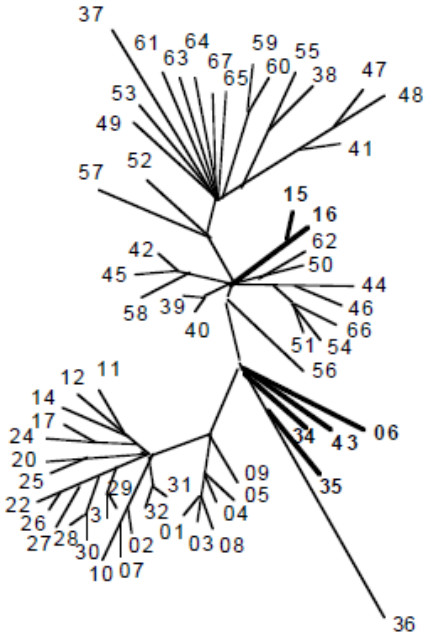


Figura 12 Classificazione ad albero delle opere di Molière e Corneille. Figura presa da Labbè 2010

Le opere di Molière (da 36 a 67) appaiono invece nella parte alta del grafico, chiaramente divise in due gruppi: opere scritte in prosa (in alto) e opere scritte in versi (nel mezzo).

In basso a destra sono invece presentate delle anomalie: un pezzo di *Psyché* molto corto attribuito a Quinault (36), le due parti di *Psychè* attribuite a Corneille (34) e Molière (35); *Dom Garcie* (43) e *Comédie des tuileries* (06), un lavoro commissionato a Corneille dal cardinale Richelieu 37 anni prima di *Psyché*.

Interessante notare invece come *Menteur* e *Suite du Menteur* (15-16) siano state nuovamente attribuite a Molière, in quanto le due opere si trovano al centro del suo raggruppamento.

Questo non vuol dire che Molière sia necessariamente l'autore delle due opere, ma indica sicuramente una vicinanza nello stile dei due autori; ulteriori verifiche devono essere eseguite per poter avere una conferma certa.

Molière potrebbe aver infatti preso l'idea di una satira contemporanea da Corneille; oppure, dato che Molière era solito rappresentare opere di Corneille tanto da saperne a memoria i versi, sarebbe stato influenzato da questi nella scrittura.

Molte ipotesi sono plausibili, sicuramente bisognerà indagare che cosa accomuni lo stile di Corneille a quello di Molière e allo stesso tempo cosa li differenzi.

Questo articolo di Labbè, a nostro parere, mostra come i metodi quantitativi possano, in ottica esplorativa, aprire nel campo degli studi letterari interessanti ipotesi di ricerca, altrimenti difficilmente esplorabili.

CAPITOLO III

ANALISI QUANTITATIVA E QUALITATIVA SU UN CORPUS DI TESTI GRAMSCIANI E NON GRAMSCIANI

In questo capitolo presenteremo l'analisi su un corpus di articoli di giornale gramsciani e non gramsciani basata su metodi e modelli matematici di Basile, Benedetto, Caglioti e Degli Esposti (Basile, Benedetto, Degli Esposti, & Caglioti , 2010; Lana, 2010; Lana, 2011) e la confronteremo con i risultati della nostra analisi qualitativa.
Scopo della nostra analisi è infatti capire che cosa accomuna e che cosa differenzia lo stile di scrittura gramsciano da quello degli altri collaboratori dei giornali in cui scriveva Gramsci, cercando quindi di capire ipoteticamente quali siano i tratti che il modello matematico, come vedremo in seguito, riesce a differenziare con tanta precisione.

3.1 L'analisi quantitativa: lo studio di Basile, Benedetto, Degli Esposti e Caglioti

Le redazioni dei giornali per cui Gramsci lavorava nel periodo 1913-1926 pubblicavano spesso articoli non firmati; non si può determinare a priori se questi articoli possano essere attribuiti a lui o a collaboratori di giornali in cui egli scriveva.
La *Fondazione Istituto Gramsci*, in vista di una nuova *Edizione Nazionale degli scritti di Antonio Gramsci*, proprio per fare luce sui numerosi editoriali anonimi, ha deciso di utilizzare, oltre ai metodi d'attribuzione qualitativi e filologici, metodi di attribuzione quantitativa basati su modelli matematici.
Per questi motivi sono stati interpellati i fisici matematici Basile, Benedetto, Degli Esposti e Caglioti che, affiancati dal linguista Maurizio Lana, hanno utilizzato metodi matematici per attribuire articoli gramsciani e non.
Sono stati scelti metodi a base matematica e non a base statistica in base alle caratteristiche degli articoli anonimi attribuibili a Gramsci, questi testi sono infatti (Lana, 2010, p. 33):

- per lo più brevi (raramente superano il migliaio di parole).

- non si differenziano tra loro per il contenuto e quindi non si differenziano tra loro per il lessico.

Proprio per queste caratteristiche si è scelto di non utilizzare un modello a base statistica ma a base matematica: l'analisi statistica infatti riesce meglio su testi molto lunghi o testi che si differenzino per contenuto e quindi ovviamente anche per lessico.
La statistica riduce infatti una grande mole di dati ad una sintesi, semplificando quindi il dato iniziale; inoltre dati molto sparsi in statistica non sono reputati validi per poter trarne delle conclusioni (Lana, 2010, p. 34).
In questo caso invece i dati sono utilizzati per raggiungere delle conclusioni anche quando sono sparsi, poiché ogni dato, anche quello sparso, è rilevante per l'attribuzione.
L'idea di base è che i testi siano quantificabili con modelli matematici (gli stessi che si utilizzano per l'analisi del DNA o l'analisi del suono, ad esempio) che vedano i testi come sequenze di simboli.
Il modello matematico ipotizza che l'autore di un testo utilizzi delle sequenze di simboli secondo regole probabilistiche che gli sono proprie e che lo differenziano da altri autori: studiando queste sequenze si possono ricostruire le regole probabilistiche che le hanno generate ed è possibile quindi distinguere tra differenti autori.

3.1.1 Gli n-grammi e l'entropia relativa

Lo studio sui testi gramsciani anonimi utilizza due metodi di analisi: gli n-grammi e l'entropia relativa.
Degli n-grammi abbiamo già parlato al paragrafo 1.3.2.
In questo caso le frequenze degli n-grammi di un testo sconosciuto vengono confrontate con le frequenze di n-grammi di un testo noto: se c'è una somiglianza nelle frequenze di n-grammi esiste una somiglianza del testo sconosciuto con il testo noto.
Le sequenze di n-grammi riproducono infatti con buona precisione le regole di produzione di un testo, riportiamo un esempio molto esplicativo che fa Lana, parlando di riprodurre un testo di un autore per sequenze successive di diversi ordini:

Si potrebbe iniziare con un'approssimazione *di ordine 0* scegliendo a caso, con pari probabilità, all'interno dell'insieme dei simboli disponibili, uno dopo l'altro, i simboli da utilizzare e si potrebbero ottenere testi di questo tipo:

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

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

illfmbaoaocnn e aai,sfrmrta eeoiddmaoo

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

Loncueresono astantà chedali co le prora Lafra Seoccoro

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

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

Si potrebbe procedere vincolando la scelta di un simbolo alla sequenza dei 3 che lo precedono, e così via. (Lana, 2010, p. 35).

Altro metodo utilizzato in questa analisi è invece quello dell'entropia informativa relativa. L'entropia è il numero di bit per carattere necessari per codificare un messaggio.

Quando comprimiamo un file attraverso gli algoritmi di compressione (ne abbiamo già parlato al paragrafo 1.3.2) possiamo avere una stima dell'entropia: maggiore sarà la sua compressione, minore sarà la sua entropia.

Riprendiamo un esempio fatto da Basile e altri per spiegarci meglio (Basile , Benedetto, Degli Esposti, & Caglioti , 2010, p. 243): il codice Morse, pur non essendo un algoritmo di compressione, ha la stessa esigenza degli algoritmi di compressione: rendere veloce la trasmissione di messaggi.

Nel codice Morse infatti le lettere più probabili vengono codificate con una sequenza di simboli più corta; se l'entropia è il minimo numero di bit per carattere che servono per codificare una sequenza, quando si codifica una sequenza in maniera non ottimale si impiegheranno più bit del necessario; l'entropia relativa tra due sequenze è il numero di bit per carattere che si ottengono codificando una sequenza in un codice ottimale per un'altra.

Ammettiamo che il codice Morse sia ottimale per la lingua inglese: se lo usiamo per codificare un messaggio in italiano avremmo un testo più lungo rispetto a quello che si sarebbe ottenuto utilizzando un codice Morse ottimale per la lingua italiana.

La differenza di lunghezza (per carattere) è una misura dell'entropia relativa fra inglese ed italiano; allo stesso modo l'entropia relativa fra due testi di Manzoni sarà più bassa rispetto a quella fra un testo di Pirandello e uno di Manzoni.

Se si confronta l'entropia fra due testi si avrà quindi una misura della loro entropia relativa e poichè l'entropia relativa quantifica la differenza fra sequenze, si potrà distinguere fra differenti autori.

3.1.2 I metodi matematici e l'analisi

L'analisi fatta da Basile e altri è stata divisa in due test: un primo test dove agli studiosi sono stati dati 100 articoli, 50 attribuiti sicuramente a Gramsci e 50 attribuiti ad altri autori che collaboravano negli stessi giornali in cui lavorava Gramsci e negli stessi anni[20]; questa fase è servita per verificare e mettere a punto per il metodo.

Una secondo test è stato eseguito su 40 articoli divisi tra articoli sicuramente attribuibili a Gramsci e articoli attribuiti a giornalisti del primo test, consegnati però agli studiosi in maniera anonima: questa fase è stata l'effettiva prova della bontà del metodo.

[20] Tutti gli articoli usati nell'analisi di Basile, Benedetto, Degli Esposti e Caglioti saranno elencati in una sezione a parte della bibliografia.

Nel primo test ognuno dei 100 testi veniva isolato (come se fosse un testo incognito) dagli altri 99, il testo incognito veniva poi confrontato con i testi noti per effettuare l'attribuzione; la procedura veniva poi ripetuta per ognuno dei 100 testi.

Tutti i testi sono stati sottoposti ad un lavoro di normalizzazione: è stata mantenuta la distinzione fra lettere maiuscole e minuscole, sono stati tenuti i segni di interpunzione e lo spazio separatore.

Gli spazi multipli sono stati ridotti a spazi singoli, sono state eleminate le note di trascrizione come [...]e [?] oltre alle grafie particolari (accenti acuti e gravi).

Si è eliminato anche il terminatore di linea poiché si è ritenuto che la divisione in capoversi dipendesse più da scelte tipografiche che da scelte autoriali.

Invece di confrontare ogni testo di attribuzione incerta con "un modello" o "profilo" del possibile autore, si è preferito confrontare ogni singolo testo con tutti i testi del corpus, interpretando solo successivamente i risultati.

Gli studiosi hanno scelto questa strada perché convinti di mettere in luce meglio caratteristiche rare, indispensabili in questo specifico caso di attribuzione.

Per l'analisi è stato utilizzato il metodo degli *n*-grammi di Keselj (Kešelj, Fuchun , Cercone, & Thomas, 2003) modificato però a causa della brevità dei testi e per il fatto che all'interno del corpus dei testi non gramsciani ci fossero molti autori differenti e in numero diseguale fra loro.

Il valore *n* è stato settato ad 8 [21] e si è elaborato un punteggio di gramscianità che va da valori di -1 a 1 (più si è vicini a 1 più un testo è considerato gramsciano e viceversa).

[21] Sono state analizzate sequenze di 8 simboli, dove per simbolo non si intende solo una lettera o un segno di interpunzione, ma anche lo spazio. Esempi di 8-grammi sono:

un n-g
segmenta
ma rende

(Lana, 2010, p. 36)

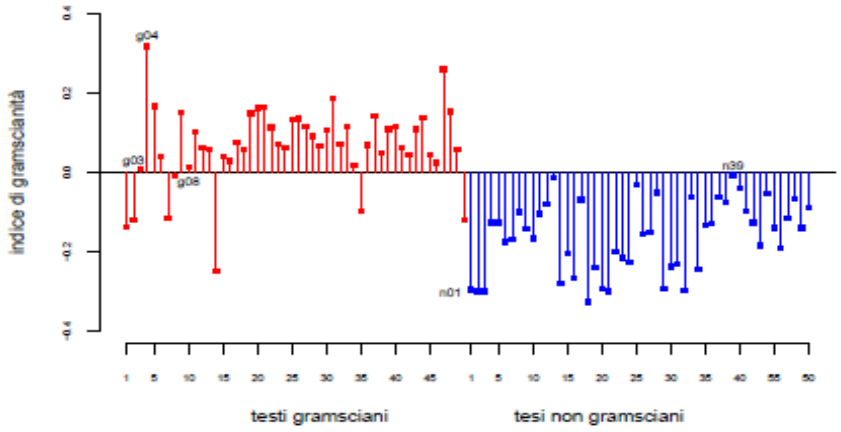


Figura 13 Figura ripresa da Basile e altri. (2010). Attribuzioni dei testi con misura dell'attribuibilità

La figura 13 mostra i risultati dell'analisi descritta (testi gramsciani in rosso, testi non gramsciani in blu).

Come possiamo vedere l'attribuzione per alcuni testi (g08, g03, n39) è molto incerta rispetto a testi come g04 o a n01.

In ogni caso questo metodo ha attribuito a Gramsci 44 testi su 50, non creando nessun falso positivo (nessun testo di altri autori è stato attribuito a Gramsci).

Successivamente è stata eseguita un'analisi con l'entropia relativa per gli stessi 100 tesi.

Questa analisi inizialmente però non è stata altrettanto buona, in parte perché l'entropia è molto sensibile alla dimensione dei testi (nel corpus ci sono testi molto corti che creano quindi uno scompenso durante l'analisi).

Si è quindi deciso di riunire tutti i testi in unico file e poi successivamente di tagliarlo in tante porzioni uguali. Questi nuovi file sono diventati il corpus di riferimento e i risultati sono di molto migliorati.

Come per gli *n*-grammi anche in questo caso si aggiunge una votazione, limitando però la somma dei punteggi ai primi tre testi gramsciani e ai primi tre testi non gramsciani.

Riassumendo quindi l'attribuzione si basa sugli 8 grammi con voto esteso a tutti i test di riferimento ed entropia relativa con testi riassemblati e voto per i primi tre classificati.

3.1.3 I risultati dell'analisi

Il risultato del primo test comprendente i 100 articoli è riassunto in figura 14.

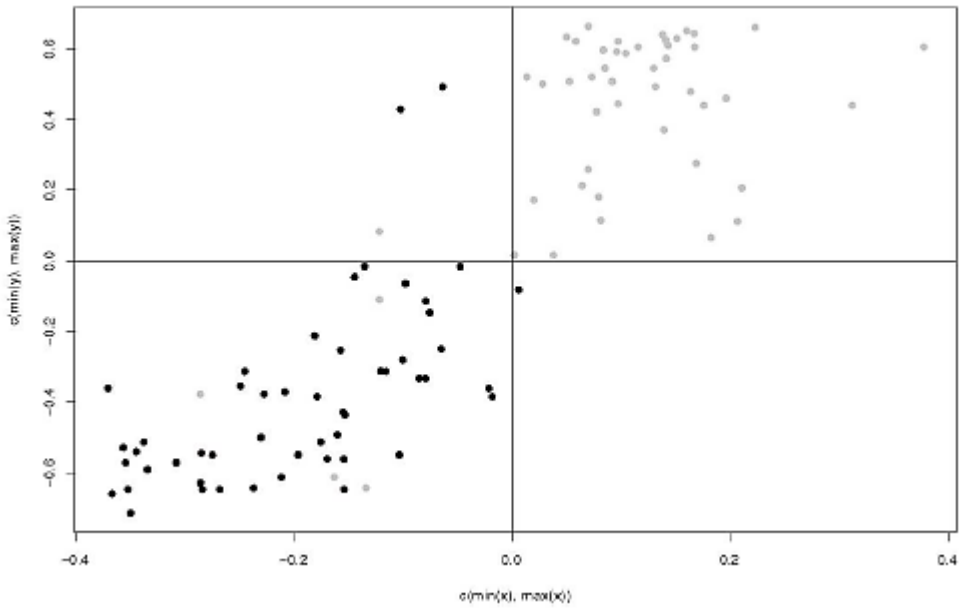


Figura 14 Figura ripresa da Lana, (2010). Attribuzione agli scritti anonimi al termine della fase di test in chiaro

L'asse orizzontale rappresenta l'indice di gramscianità fornito col metodo degli *n*-grammi: i punti più a destra sono quelli di attribuzione più certa a Gramsci, quelli più a sinistra sono invece i testi che il metodo con maggiore probabilità non attribuisce a Gramsci.
Sull'asse verticale è invece riportato il valore dell'indice dato dall'entropia relativa; in basso abbiamo i testi non attribuibili a Gramsci con maggiore probabilità, in alto i testi attribuibili a Gramsci con maggiore probabilità.
Nel quadrante in alto a destra ci sono i testi che entrambi i metodi attribuiscono a Gramsci (punti grigi), tra questi non c'è nessun falso positivo.
Nel quadrante in alto a sinistra ci sono i testi attribuiti a Gramsci dall'entropia relativa ma non dal metodo degli n-grammi: sono i testi Gramsci numero 7, 8, 35.
Nel quadrante in basso a destra non c'è nessun testo: sarebbero i testi attribuiti a Gramsci dal metodo n-grammi ma non da quello dell'entropia relativa.
Nel quadrante in basso a sinistra ci sono testi non attribuiti a Gramsci da entrambi i metodi, ovvero i testi Gramsci 1, 2, 14, 50.
La stessa analisi è stata eseguita sui 40 testi consegnati in forma anonima, i risultati dell'analisi sono visibili in figura 15.

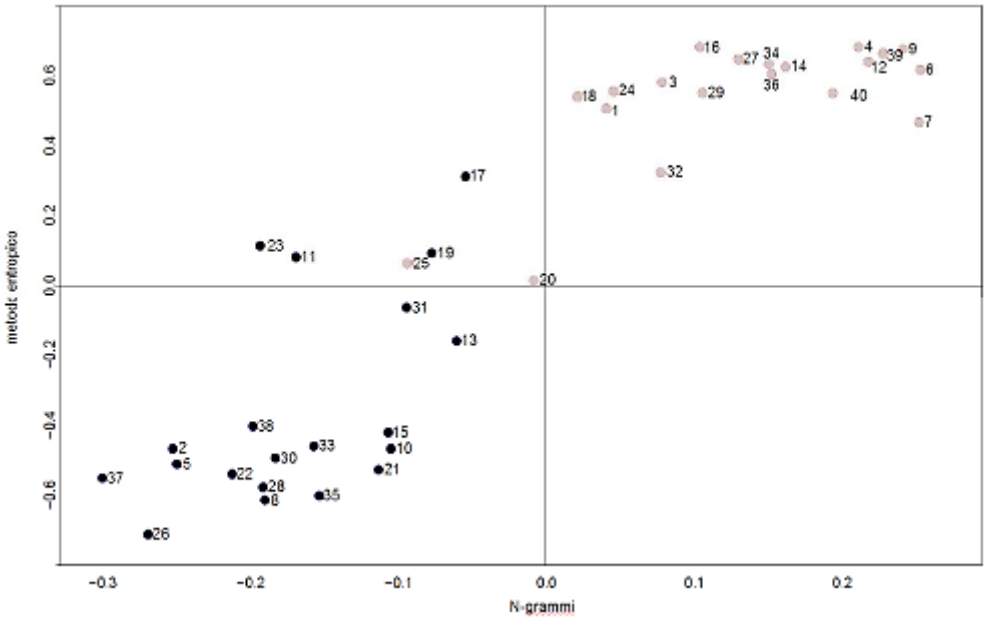


Figura 15 Figura ripresa da Lana, 2010. Esito del test cieco

La figura 15 si legge allo stesso modo della figura 14: in ascissa abbiamo l'indice di gramscianità fornito dagli n-grammi, mentre in ordinata il valore dato dell'entropia relativa. A Gramsci vengono attribuiti 18 testi su 20, una percentuale di affidabilità del 90%, senza falsi positivi.

I due testi gramsciani non riconosciuti dagli *n*-grammi in questo caso sono il numero 20 e il numero 25.

Questi risultati mostrano che l'analisi non è perfetta ma arriva a delle percentuali di successo accettabili (l'86% di successo nella prima analisi, il 90% di successo nella seconda) e soprattutto nell'analisi non c'è stato nessun falso positivo, cioè l'analisi non ha attribuito nessun test non gramsciano a Gramsci.

Alla vista di questi risultati ci siamo posti alcune domande:

Cosa riescono a vedere questi metodi che l'occhio umano non riesce a vedere?

Ma soprattutto un'analisi qualitativa riuscirebbe a discriminare con analoga efficacia fra un testo di Gramsci e uno non di Gramsci?

3.2 L’analisi qualitativa

Per questi motivi abbiamo pensato che fosse interessante comparare i risultati dello studio di Basile e altri con un’analisi sui testi di tipo qualitativo.
La nostra analisi si è principalmente concentrata sui testi del primo test di Basile e altri; questi testi, come abbiamo già detto, sono stati divisi fra 50 testi sicuramente attribuiti a Gramsci e 50 testi attribuiti sicuramente ad altri giornalisti che collaboravano negli stessi anni e nelle stesse testate dove aveva lavorato Gramsci.
Stileremo un unico profilo linguistico per Gramsci e un profilo diverso per ognuno dei giornalisti dei 50 testi non attribuibili a lui.

3.2.1 Profilo linguistico di Antonio Gramsci

I cinquanta testi gramsciani sono articoli pubblicati dal 1914 al 1918 in tre diverse testate: *Il Grido del Popolo, Avanti*! e *La città futura*.
Dal punto di vista retorico all’interno dei cinquanta articoli possiamo distinguere tra due diverse procedure di generazione del discorso: il discorso politico polemico e il discorso politico didattico (Desideri, 1984, p. 42-44).
Il discorso politico polemico consiste nel portare il destinatario della comunicazione alle argomentazioni espresse dall’emittente, un tentativo da parte dell’emittente di far immedesimare il destinatario con lui.
Questa immedesimazione è creata attraverso tecniche di *embrayage* o *innesto*, avvicinamento *attananziale*.
Sono tipiche di questa forma l’uso di pronomi personali (*io*, *noi*), degli aggettivi possessivi (*mio*, *nostro*), l’uso di deittici come avverbi di spazio e di tempo (*qui*, *oggi*).
Dal punto di vista sintattico viene invece adottata una sintassi breve, che predilige frasi coordinate scandite da figure retoriche di ripetizione. Per finire, il lessico è colorato ed enfatico (Dell'Anna, 2010, p. 20).
Esempio di discorso politico-polemico all’ interno del corpus gramsciano potrebbe essere questo estratto dal testo numero 4:

> Sentite, sentite... A Roma le autorità ecclesiastiche vogliono che i suffragi celebrati per i morti in guerra siano estesi a tutti i caduti, di qualsiasi nazionalità essi siano. Il nazionalista salta su, come

> il babau dalla scatoletta di cartone, e protesta. La morte o la vita per la sua anima di princisbecco, non hanno altra risonanza che quella stridula del caricatore del fucile: il suo spirito legnoso di machiavellino da strapazzo deturpa tutte le bellezze morali come il bruco lascia le tracce della sua digestione sul verdore dei giardini.

Il discorso politico-didattico invece non presuppone un confronto tra destinatario e emittente sulle tesi di quest'ultimo: l'emittente infatti rimane sullo sfondo, mentre il destinatario si identifica di più con il contenuto dei messaggi dell'emittente più che con l'emittente.
Il contenuto dei discorsi sarà quindi formato da una serie di sequenze oggettivamente vere, come l'uso di sequenze di tipo storico, descrittivo e scientifico.
Le tecniche testuali, in questo caso, saranno quelle del *debryage* o del *disinnesto, allontanamento attananziale*, col risultato di produrre allontanamento e distacco fra emittente e destinatario.
Questa idea di distacco si esplicita con l'uso di verbi in forma impersonale, con una sintassi più complessa, formata da subordinate, nell'assenza di deittici di spazio o tempo, nell'uso di un lessico tecnico (Dell'Anna, 2010, p. 22).
Un esempio di discorso politico didattico all'interno del corpus gramsciano può essere visto in questo estratto dal testo numero 30:

> Si è notato ancora come i nazionalisti italiani, fra i quali non abbonda lo spirito inventivo, hanno spesso ricalcato i loro amici-nemici di Francia anche in particolari iniziative, che avevano una loro ragione d'essere oltralpe, ma erano completamente disambientate in Italia. Ma si è troppo trascurato di porsi il quesito del perché il Partito nazionalista abbia finito con l'affermarsi vittoriosamente, del come si sia venuto intimamente modificando, e sia divenuto sociale, cioè sia venuto acquistando concretezza politica, per il fatto che una parte delle sue ideologie è stata fatta propria da determinati ceti economici della borghesia, che nel Partito nazionalista hanno visto il loro partito, che negli scrittori nazionalisti hanno visto i loro scrittori, i teorici dei loro interessi, dei loro bisogni e delle loro aspirazioni.

Dell'Anna (2010, p. 23) afferma però come il discorso politico polemico e il discorso politico didattico non siano completamente distinti, né rappresentino delle alternative l'uno all'altro: in Gramsci, infatti, anche nei discorsi che potrebbero venir categorizzati come politico-didattici, troviamo un uso insistito di figure di ripetizione e l'uso dei deittici di tempo (come ad esempio *ora*) proprio perché Gramsci sta sempre cercando di convincere delle sue tesi un ipotetico lettore.

Troviamo spesso anche l'uso di pronomi personali, in Gramsci più frequentemente *noi* (68 occorrenze) rispetto ad *io* (17 occorrenze).
Lo stile del discorso politico polemico appare invece molto evidente quando Gramsci attacca il nemico, visto sia come una ideologia contrapposta alla sua sia come un esponente politico o culturale di una ideologia a lui antitetica (per questo ultimo caso si vedano ad esempio i testi 14, 16 e 29).
In questo tipo di discorsi politico polemici spesso viene esposto un fatto di cronaca in cui appare l'avversario e questo viene svilito anche per mezzo di metafore e similitudini molto colorite.
I discorsi politici didattici possono essere invece dei commenti ragionati a fatti accaduti (in questo caso con un uso naturalmente preponderante del passato) o delle esposizioni di tesi: queste tesi possono essere introdotte e poi confutate, portando il lettore al ragionamento di Gramsci o, al contrario, viene introdotta una tesi di Gramsci per poi contrapporla a quelle di altri e, attraverso una serie di esempi, ne viene saggiata la sua bontà. L'argomentazione si serve spesso di una serie di domande retoriche a cui Gramsci risponde (l'uso di domande retoriche è in realtà presente anche nei discorsi politico polemici).
Riportiamo qui un esempio dal testo 34:

> Perché gli individui, nella loro maggioranza, compiono solo determinati atti? Perché essi non hanno altro fine sociale che la conservazione della propria integrità fisiologica e morale[...].

Dal punto di vista strutturale invece tutta la costruzione del discorso gramsciano ruota attorno a strutture parallele: al suo interno infatti vi si trovano molte forme di correctio, come ad esempio in Gramsci 46:

> Dello stato che *non deve* essere lasciato in balìa delle forze libere spontanee degli uomini, *ma deve* in ogni cosa, in ogni atto imprimere il suggello di una volontà di un programma stabilito, preordinato dalla ragione.

o ancora in Gramsci 12:

> Evidentemente *non è l'oro* che forma la ricchezza e dà il benessere, *ma l'armonico equilibrio degli scambi* fra i vari paesi che provvede a ciascuno al minor prezzo ciò che ciascuno per le varie condizioni in cui lavora può produrre a migliori condizioni.

Presente è anche la comparazione, soprattutto nel costrutto *tanto…quanto*, ad esempio in Gramsci 11:

> Dagli articoli dei cosidetti competenti, alle notizie di cronaca, è uno stillicidio continuo, diuturno di elementi di dubbio, di odio, di sproposito, *tanto più* pericoloso *quanto più* il giornale conservatore milanese ha sempre mantenuto nella polemica politica un atteggiamento moderato, se non addirittura germanofilo, certo non sfavorevole alla Germania.

e ancora in Gramsci 23:

> [...]*quanto più* queste oscillazioni diventano irregolari e capricciose, *tanto più* si dice che i tempi sono calamitosi.

Le figure retoriche maggiormente presenti all' interno del testo sono figure retoriche di ripetizione: queste servono non soltanto a costruire parallelismi ma anche a dare ritmo all'interno del testo e sostenere quindi l'attenzione del lettore.
Possono essere ripresi con anafore nomi, come ad esempio in Gramsci 43:

> *Lo scrupolo* nasceva in lui per il "Corriere", le cui molte centinaia di migliaia di copie avrebbero veramente servito a indirizzare l'opinione borghese; *lo scrupolo* non esisteva per le poche migliaia di lettori dell'"Unità", già tutti d'accordo.

Possono essere ripresi con anafore verbi, sempre in Gramsci 43 (in questo caso preceduti da negazione):

> Essi *non hanno* mai combattuto per le loro idee; si sono posti al servizio del parassitismo capitalista; *non hanno* neppure tentato di iniziare l'opera educativa tra le masse [...]

Possono essere riprese con anafore congiunzioni, sempre in Gramsci 43, ad esempio con valore causale:

> Oggi il negromante si spaventa, *perché* vede di non poter più dominare le forze demoniache che ha scatenato, *perché* ogni giorno che passa complica il problema.

Riprese in catene anaforiche si trovano anche preposizioni che introducono sintagmi preposizionali, spesso come complementi di specificazione, come ad esempio in Gramsci 09:

> Vuol dire rimanere isolati, chiusi nel proprio dolore, senza possibilità *di* aiuti, *di* conforto.

Sebbene di solito vi sia una ripresa anaforica binaria, questa ripresa più raramente può diventare ternaria, in particolare in alcuni costrutti; troviamo infatti la ripresa ternaria con la ripresa del verbo *essere*, molto spesso alla terza persona singolare del presente, in figure di definizione, in cui una parola o un concetto vengono appunto spiegati al lettore.
Un esempio di questo costrutto può essere visto in Gramsci 46:

> Liberismo *è* la formula che comprende tutta una storia di lotte, di movimenti rivoluzionari per la conquista di singole libertà. *È* la forma mentis venutasi creando attraverso questi movimenti. *È* la convinzione venutasi formando nel sempre maggior numero di cittadini che vennero attraverso queste lotte a partecipare all'attività pubblica[...]

O ancora in Gramsci 18:

> *È* la liberazione degli spiriti, *è* l'instaurazione di una nuova coscienza morale che queste piccole notizie ci rivelano. *È* l'avvento di un ordine nuovo, che coincide con tutto ciò che i nostri maestri ci avevano insegnato.

La struttura ternaria viene utilizzata anche con riprese anaforiche di preposizioni, come ad esempio in Gramsci 41:

> Egli dispone, meno che l'industriale e il commerciante, dei mezzi *per* controllare l'operato dei commissari di requisizione, *per* protestare contro gli arbitrii, *per* difendersi dalle spogliazioni brutali.

Altra figura di ripetizione presente all'interno dei testi gramsciani, anche se in misura minore rispetto all'anafora, è l'epifora, ad esempio in Gramsci 49:

Amico mio, ci ripetiamo sconsolatamente, il tuo era *l'uovo di Colombo*. Ebbene, non mi importa di essere lo scopritore *dell'uovo di Colombo*.

Sempre in Gramsci 49:

[...]senza che lo abbiano appagato nello stesso tempo tutti gli altri individui della sua *classe*. E perciò l'egoismo proletario crea immediatamente la solidarietà di *classe*.

Anafora ed epifora possono combinarsi per dar luogo a simploche, come nel caso di Gramsci 48:

[...]*perché per la vita* comunale e famigliare basta il *dialetto, perché la vita* di relazione si esaurisce tutta quanta nella conversazione in *dialetto.*

Altra figura di ripetizione presente nei testi è quella dell'anadiplosi, ad esempio in Gramsci 17:

II suo cuore non è che la coscienza del suo essere classe, la coscienza dei suoi fini, la coscienza *del suo avvenire. Dell'avvenire,* che è solamente suo, [...]

Oppure ancora in Gramsci 20:

Una lettera che riceviamo ci fornisce dati preziosi per stabilire il genuino carattere del *prof. Arnaldo Monti. Il prof. Arnaldo Monti* è stato per quattro anni pedagogo in casa Giolitti.

Più rara l'epanadiplosi, come ad esempio in Gramsci 39:

[...]: *poesia,* niente altro che *poesia.*

Altre figure di ripetizione presenti nei testi di Gramsci sono le figure etimologiche come «ha preveduto il prevedibile» in Gramsci 26 e i polittoti come «giorno per giorno» in Gramsci 46, oppure «continuiamo e continueremo» in Gramsci 31.

Nei testi gramsciani il ritmo binario viene richiamato anche attraverso l'uso insistito di dittologie sinonimiche, come ad esempio «particolareggiata e documentata» Gramsci 06, «mentalità chiusa e gretta» Gramsci 18, «di saggezza e di sapienza» Gramsci 20, «di errare e spropositare» Gramsci 25 «complesso e imbrogliato» Gramsci 34.

In generale queste strutture binarie di coordinazione per mezzo di *e* fra aggettivi e nomi sono molto insistite all'interno dei testi anche senza che queste siano sinonimi l'uno dell'altro, come ad esempio «economiche e sociali» Gramsci 37, «l'uomo e la realtà» Gramsci 27, o anche in relazione di antitesi come «giganti e pigmei» Gramsci 29, «incitamento all'azione e al pensiero» Gramsci 50.

Dal punto di vista del lessico si possono distinguere quattro grandi campi: uno di questi campi sarà incentrato naturalmente sul lessico dell'area semantica della politica (di stampo socialista), uno sul lessico dell'area semantica dell'economia (come vedremo i due campi verranno spesso associati fra di loro), uno su un lessico dell'area semantica di tipo filosofico-storico e l'ultimo si rifarà invece al mondo della biologia.

Per il lessico della politica ad esempio troviamo: *propaganda sovversiva, sommovimento sociale, bolsceviki, classe dirigente, capitalismo, partito politico, Internazionale, socialismo, lotta di classe, classe borghese, marxisti, proletariato russo, guerra, socialisti, Carlo Marx, ambasciatore rosso*.

Troviamo inoltre parole legate a un lessico di tipo economico come: *scienza economica, tariffe protettive, cambiale scaduta, protezionismo agrario, guerra economica, problema economico, dumping, economista, industria, commercio, produttore, consumatore, mercati, feudalismo economico, compratore, venditore*.

Interessante vedere come i lessici dei campi semantici della politica e dell'economia si uniscano, (proprio come ci aspetterebbe da testi che si rifanno sulle teorie del materialismo storico marxiano) spesso per mezzo di congiunzione, come ad esempio in: *rivoluzionari economici, caos economico e caos politico, dottrina politica e dottrina economica, unità politica e economica, la nazione e la produzione*.

Legati a lessico di tipo filosofico-storico troviamo invece: *storia, coscienze individuali, filosofia, vita interiore, cultura, utopia, dialetto, ideali, uomo, storico, idee, spirito, necessità storica, lingua italiana, lingua internazionale, tessuto storico*.

Per il lessico del mondo naturale: *gregge proletario italiano, virgulto rigonfio di succhi vitali, sterminati banchi coralliferi, vaccina condannata al macello, povera festuca, microbi della tubercolosi e della sifilide, greppaioli, ranocchie del piccolo corpo sonoro, le cellule*

e i tessuti, aquile latine, uovo di pidocchio, ghianda, rigida e gonfia carogna, carnaio enorme di belve, enorme vespaio, gli uccelli dell'aria, enorme polipaio umano, siepi.

Il lessico della biologia viene usato soprattutto nei discorsi di tipo polemico per attaccare il nemico attraverso l'uso di metafore e similitudini; nei discorsi di tipo didattico per creare esemplificazioni (nel testo di Gramsci numero 36 viene ad esempio riportata un'intera favola di Fedro).

Altro lessico usato in maniera minore con similitudini e metafore è quello mitologico e letterario, ad esempio: *sirena, basilisco, lillipuziani, gli ulissi, Stenterello italico*.

Sono evidenti casi di suffissazione diminutiva, spesso usati con intenzione dispregiativa, come: *affaruccio, volumetto, mezzucci, giornaletto, staterelli, rotellina, studentucolo, avvocatuzzo*.

Molto presente anche la suffissazione in *-ismo,* ad esempio in: *parassitismo, internazionalismo, confusionarismo, confusionismo, dilettantismo, acrobatismo, bizantinismo*.

Presenti anche alcuni composti in cui l'unione dei due costituenti è segnalata graficamente dal tratto grafico -, come ad esempio: *idee-limiti, idee-forze, amici-nemici, parole-fatti, nazione-ipotesi, non-spirito, organismo-forza*.

Sono presenti inoltre molti derivati con prefisso negativo *in-* come ad esempio: *indicibile, insolute, ingiustizie, intransigente, inespugnabile, ininterrotta, infelice, inattaccabile, invisibile, incomprensibile, infedeli, indiscutibili, inconsapevolmente, intollerante, incapacità, insufficiente*.

L'aggettivazione, come abbiamo visto già nelle dittologie sinonimiche, è molto presente e serve a portare enfasi e forza retorica al discorso, ad esempio: *doveri eccezionali, colpe imperdonabili, avversione innata, battaglia strenua, aspre lotte, inesorabile processo, verità sacrosanta, opera assidua*.

Proprio per dare al discorso maggiore enfasi sono presenti aggettivi al grado superlativo come: *tristissimo stato, vivissima simpatia, dileziosissimo romanzo, difficoltà gravissime, balordissimo senso comune, pubblicazioni recentissime, valore altissimo, dramma altissimo, vilissimo materasso*.

Come abbiamo già detto prima per l'uso dei tempi e dei modi verbali si deve distinguere a seconda del tipo di discorso: nel discorso polemico troveremo un uso più insistito dell'indicativo al presente, con qualche incursione del futuro; è presente anche l'imperativo, in quest'ultimo caso infatti Gramsci si rivolge direttamente ai suoi lettori o ai suoi avversari. Nel discorso didattico troveremo invece verbi alla forma impersonale ed un uso più insistito

della forma passiva e del passato (molto spesso infatti Gramsci in questo tipo di discorsi commenta fatti già avvenuti).
Anche dal punto di vista sintattico bisogna sempre distinguere tra discorso polemico e discorso didattico: nel discorso polemico tenderà a prevalere appunto la paratassi sull'ipotassi, mentre nel discorso politico didattico il contrario.
A nostro parere, in generale Gramsci preferisce un periodare secco e basato più sulla paratassi che non sulla ipotassi, con una prevalenza della giustapposizione piuttosto che della coordinazione.
O meglio: la coordinazione è molto presente fra nomi e aggettivi (nel caso ad esempio delle dittologie sinonimiche) ma non tra periodi.
Molto usato è infatti l'asindeto: numerose infatti all'interno dei tesi sono le enumerazioni, ad esempio in Gramsci 11:

> Essi venderanno *all'Italia, alla Francia, all'Inghilterra, agli Stati Uniti, al Giappone, alla Russia* i loro prodotti al disotto del costo, costringendo quindi gli industriali di tutti questi paesi a chiudere bottega.

Segnaliamo anche che sono presenti, sebbene in maniera rara all'interno dei testi (soltanto in 3 testi), elenchi puntati e numerati, che hanno il compito di sintetizzare e evidenziare diversi argomenti o sviluppi di una tesi.
Dal punto di vista della coordinazione va segnalato come, all' interno dei testi di Gramsci, a volte i periodi possano iniziare con *e* maiuscola o con *ma* maiuscolo, dove il *ma* maiuscolo ha spesso valore pragmatico di dare maggiore importanza a quello che lo segue; rarissimo invece l'uso di *o* maiuscolo ad inizio periodo (presente in 5 soli testi all'interno del subcorpus gramsciano).
Numeroso è all'interno dei testi l'uso di parentetiche: queste in Gramsci vengono inserite tra due parentesi tonde, a volte accompagnate soltanto da una parentetica introdotta da lineetta. Solitamente le parentetiche sono poste tra soggetto e verbo della frase in cui sono inserite, o fra verbo e complementi; un esempio può essere visto in Gramsci 34:

> La quantità (struttura economica) vi diventa qualità poiché diventa strumento di azione [...].

Le parentetiche sono maggiormente presenti nei discorsi di tipo politico didattico, proprio per la loro natura di aggiunta e di ulteriore specificazione e spiegazione all' interno del testo. È presente all' interno del testo l'anastrofe: questa si presenta principalmente con un soggetto posposto alla fine del periodo, in modo da isolarlo e quindi metterlo in risalto, ad esempio in Gramsci 13:

> Preferite una volgare parola, voi.

Oppure in Gramsci 44:

> [...]per i quali invece è ragione essenziale di vita la libertà degli scambi.

3.2.2 Profilo linguistico di Giuseppe Bianchi

Nel corpus dei testi non gramsciani sono presenti solamente tre testi di Giuseppe Bianchi: il testo 1, il testo 2 e il testo 3.
Il primo testo è una presentazione ai lettori (e soprattutto ai nemici politici) in veste di direttore del Grido del popolo, il secondo una risposta ad una accusa che gli è stata mossa da avversari politici; entrambi i testi rientrano quindi nel discorso di tipo politico polemico.
Troviamo infatti l'uso di deittici (*ora, adesso, oggi*), e l'uso di pronomi personali, in questo caso un suo insistito del pronome *io*, a differenza di Gramsci dove viene più spesso utilizzato *noi* rispetto ad *io (*nei 3 testi di Bianchi *io* ha 9 occorrenze su 2007 word-token, mentre *noi* ha 3 occorrenze*)*.
È presente la congiunzione *ma*, soprattutto ad inizio frase, presente è infatti la correctio; è presente l'anafora, ma in maniera scarsa.
Sono presenti molte dittologie sinonimiche e strutture binarie, come ad esempio: *tenace e energica, affinato e approfondito, vili e subdoli, personalismo e opportunismo*.
Troviamo un'aggettivazione enfatica (*erotismo parossistico, aspre prove, tragica ora, ignobili responsabili, attacco pazzesco e idiota*), con uso di superlativi (*schifosissima, purissimo, amicissimo*).
Sono presenti derivati con suffissazione in *-ismo* come: *ministeriabilismo, erotismo, misterialismo, personalismo, opportunismo*.

Presenti anche composti con i due costituenti uniti da lineetta, come: *idea-forza*.
Il lessico è attinente all'area semantica della politica (*organismi politici e sindacali*, *socialismo*, *sciopero generale*, ecc.); segnaliamo l'uso, nel testo 1 in particolare, della parola *compagni* (in Gramsci presente con solo 11 occorrenze in tutto il corpus, nei testi di Bianchi 5 occorrenze).
Troviamo un lessico appartenente al campo semantico biologico naturale, usato naturalmente per deridere l'avversario (*lombrichi dell'azione socialista*, *schifosissima carogna*, *motriglia graveolente*).
La sintassi è paratattica e basta più sulla giustapposizione che sulla coordinazione; sono presenti incidentali inserite fra lineette.
Segnaliamo inoltre la presenza dell'iterazione *ah* e dell'uso di frasi esclamative, oltre che dei puntini di sospensione (rari in Gramsci, le esclamative sono presenti in soli due testi all'interno del subcorpus gramsciano, mentre i punti di sospensione in soltanto quattro testi).
Il testo 3 è invece un resoconto dei sindacati e dei partiti socialisti in Germania, è infatti un testo con struttura più didattico-argomentativa che polemica.
Anche qui troviamo uno scarso uso di figure di ripetizione, oltre alla presenza di dittologie e strutture binarie.
C'è un uso enfatico degli aggettivi, accompagnato però a un lessico che si attiene strettamente all'area politica.
Anche in questo caso la sintassi resta sempre molto paratattica, ma essendo un discorso politico didattico sono presenti più subordinate rispetto ai testi precedenti.
Sono presenti parentetiche, in questo caso aperte oltre dalle lineette anche dalle parentesi tonde; presenti anche i puntini di sospensione.
Nel testo sono inoltre presenti interrogative retoriche.

3.2.3 Profilo linguistico di Amadeo Bordiga

I testi di Amadeo Bordiga sono i testi numero 4, 5, 6, 7, 8, 9, 10.
Dal testo 4 al testo 7 abbiamo dei resoconti della rivoluzione Russa, i rimanenti testi sono invece commenti su teorie marxiste e socialiste; questi testi quindi appartengono di più al profilo didattico-argomentativo che non a quello polemico, anche se non mancano alcune caratteristiche di quest'ultimo.
All'interno dei testi è presente la correctio.

Sono presenti figure di ripetizione, come l'anafora, ad esempio nel testo 4, associata a climax:

> *Non si poteva* pensare, *non si poteva* parlare, *non si poteva* stampare, *non si poteva* associarsi.

Abbiamo come in Gramsci anche qui la ripresa anaforica di preposizioni che introducono sintagmi preposizionali, sempre al testo 4:

> [...]che lo Stato militarmente più moderno è quello in cui le risorse *dell'*industria, *del* commercio, *dell'*amministrazione, *della* finanza, sono maggiori[...]

Presente inoltre l'anadiplosi, ad esempio nel testo 9:

> [...]fatalmente la maggioranza delle assemblee legislative sarebbe stata costituita da *rappresentanti socialisti*. I quali *rappresentanti*[...]

Anche all'interno dei testi di Bordiga sono presenti dittologie sinonimiche e strutture binarie coordinate dalla congiunzione *e*, ad esempio: *dispotico e tirannico*, *acuta e prolungata*, *economico e sociale*, *assurda e impossibile*, *semplici e grandiose*, *la civiltà e il progresso*, *fatti sociali e politici*.
Allo stesso modo sono però presenti all'interno dei testi strutture ternarie con asindeto, ad esempio nel testo 6:

> [...]al livello delle borghesie più progredite dei paesi *produttori, esportatori, coloniali* dell'Ovest.

O ancora nel testo 9:

> *A regolare, organizzare, disciplinare* i nuovi rapporti sociali fondati non più sul diritto di proprietà privata ma sulla associazione dei lavoratori.

L'asindeto è presente, oltre che nelle figure ternarie, anche in una serie di enumerazioni, ad esempio nel testo 5:

> [...]: l'assolutismo sostenuto dall'alta burocrazia, dalla casta militare, dal clero, dalla nobiltà terriera;

Il lessico sarà naturalmente concentrato sull'area politica, in questo caso però legato alla Russia come ad esempio*: borghesia russa, Czar, rivoluzione russa, Santa Russia, soviet.*
Troviamo legato strettamente all'area politica invece: *capitalismo, proletariato moderno, sindacalisti e anarchici, borghesia russa, élites industriali*.
Altro campo semantico usato è quello biologico, in particolare il lessico afferente al corpo umano e alle sue malattie: *daltonismo bellico, ossatura ancora salda, pieni polmoni, ossigeno vivificatore, reumatismo, dente caduto, polpa rivoluzionaria, vecchia crosta*.
Questo lessico viene usato con metafore e similitudini all'interno del testo in maniera a volte polemica, a volte enfatica; è presente addirittura una personificazione della rivoluzione proletaria nel testo 9:

> Povera rivoluzione proletaria in berretto da notte e pantofole con un tantino di reumatismo e qualche dente caduto!

Anche in questo caso abbiamo un'aggettivazione enfatica, con anche l'uso del grado superlativo (*numerosissimi intellettuali e studenti, larghissime concezioni, affascinate miraggio, assoluta intransigenza*).
La sintassi è tendenzialmente paratattica ma, essendo testi più didattici più che polemici, sono presenti diverse subordinate.
In particolare, in Bordiga, a volte il soggetto è posposto per metterlo in risalto, ad esempio nel testo 5:

> [...]perchè sotto lo scettro dello czar gemevano in una identica oppressione, non certo preferibile a quella austriaca od ottomana, cento diverse nazionalità.

Sono presenti diverse parentetiche, aperte sia da lineette che da parentesi tonde nel testo 4, nel testo 5 e nel testo 6 ci sono soltanto parentetiche aperte da lineette, configurazione mai presente in Gramsci.
Sono presenti diverse interrogative retoriche e anche alcune esclamative.

Tipico della scrittura di Bordiga è l'uso parecchio insistito dei due punti, come ad esempio nel testo 8:

> [...]dottrina: interpretazione marxista della storia e della società; programma: conquista violenta del potere ed esercizio di esso per attuare la socializzazione dei mezzi di produzione; metodo: azione politica intransigente di classe con disciplina collettiva.

O ancora al testo 5:

> Ed allora: lo czarismo si era messo d'accordo con i tedeschi e tramava la pace, la rivoluzione si è fatta per intensificare la guerra a lato dell'Intesa.

Presente spesso all'interno dei testi anche l'uso dei puntini di sospensione (raro in Gramsci).

3.2.4 Profilo linguistico di Attilio Carena

Nel corpus dei testi non gramsciani i testi di Attilio Carena sono il numero 11, 12 e 13.
Il testo 11 è intitolato Pasqua di resurrezione, ed è un'allegoria sulla Rivoluzione russa vista come una rinascita e quindi una Pasqua di resurrezione. Non avendo particolari intenti polemici di fatto però le caratteristiche del testo lo fanno rientrare in questa tipologia.
Gli altri due testi sono invece di tipo didattico-argomentativo.
Possiamo notare come i testi di Carena siano estremamente intrisi di figure di ripetizione, in particolare il testo 11.
Troviamo al suo interno infatti non solo anafora, epifora, e anadiplosi ma anche simploche ad esempio al testo 12:

> *La religione o filosofia* non è questa o quella particolare *religione o filosofia*; [...]

O ancora troviamo epizeusi al testo 11:

> *Troppi, troppi* interessi sono in giuoco per non farlo e il popolo *troppo, troppo* è buono per sottrarsi alla malia sentimentale di chi oggi è tutto popolo e per il popolo.

Troviamo inoltre numerosi polittoti: *hanno tratto e traggono* (testo 11); *che variano col variare* (testo 12).
Presenti anche in questi testi sia la correctio (soprattutto nei testi numero 12 e 13 con forme come *non è… ma è*), sia il parallelismo, ad esempio con il costrutto *tanto…quanto*.
Sono presenti dittologie sinonimiche e strutture binarie che convivono con strutture ternarie asindetiche come al testo 12:

> I programmi sono contingenti e mutevoli, perciò devono essere *parziali, particolari, specializzati*;

Il lessico del testo 12 si differenzia molto da quello degli altri testi del corpus poiché è legato alla religione e in particolare della Pasqua, troviamo infatti: *idolo colossale, occhio per occhio, osanna, palme, pace sublime, Pasqua di vera resurrezione, resurrezione, osanna*.
Nei testi 13 e 14 troviamo invece un lessico di tipo filosofico-storico: *atomismo sociale, concetti astratti, divenire storico, falsa fede, falsa religiosità, infime categorie, divenire storico, spirito, volontà umana*.
In tutti e tre i testi è presente un'aggettivazione enfatica.
Dal punto di vista sintattico la scrittura di Carena si distingue per periodi brevissimi e giustapposti e per l'uso insistito di domande retoriche.
Molto presenti sono anche le parentetiche, racchiuse graficamente nelle parentesi tonde.

3.2.5 Profilo linguistico di Leo Galetto

I testi di Leo Galetto sono i testi 18, 19, 20, 21.
Mentre i primi tre testi hanno più una struttura didattico argomentativa, il quarto testo ha invece una struttura più polemica.
Comuni a tutti i testi sono forme e strutture basate sulla correctio; in tutti i testi sono presenti dittologie sinonimiche e strutture binarie.
È presente l'anafora, ma in maniera scarsa; ad essere ripreso anaforicamente è spesso il pronome *noi*.
In tutti i testi è presente un'aggettivazione enfatica, in particolare nel testo 21, con superlativi e diminutivi con significato spregiativo.

Dal punto di vista del lessico, troviamo un lessico di stampo politico-socialista in cui compare frequentemente la parola *pace* (10 occorrenze in Gramsci, 26 occorrenze solo nei testi di Galetto).
Nel testo 21 troviamo moltissime interrogative retoriche con altrettante esclamative.
Nell'ultimo testo, poichè è quello più polemico, il periodare diventa più incalzante e giustapposto, dove però il collegamento fra i diversi periodi viene mantenuto con *e* iniziale.
Sono presenti diverse parentetiche, aperte da lineette, mentre spesso il soggetto viene posposto in modo che venga evidenziato, ad esempio al testo 18:

> [...] si apriranno gloriose, se nobilmente tracciate, le vie del divenire socialista.

3.2.6 Profilo linguistico di Adolfo Giusti

I testi di Adolfo Giusti sono i testi 22, 23, 24, 25.
Sono testi brevi, dove 22, 23 e 24 sono polemiche contro la Società di Mutuo Soccorso fra Macchinsti e Fuochisti, mentre il testo 25 parla di uno sciopero dei lavoratori della lana.
All'interno dei testi sono presenti dittologie sinonimiche e figure binarie; qualche ripresa anaforica è presente ma anche questa in maniera scarna e non in tutti i testi.
Dal punto di vista del lessico troviamo termini dell'area biologica come *fungaia malefica* o *tisici*.
Giusti si differenzia però per l'utilizzo del dialetto, troviamo infatti nel testo 24:

> Ma el tacon è riuscito al Mortara peio del buso.

Giusti si differenzia anche dal punto di vista grafico per l'uso del maiuscoletto, ad esempio al testo 23:

> L'Associazione Generale degli Operai conta ben UNDICIMILA soci[...]

O ancora al testo 24:

[...]che non avendo potuto funzionare la lega della categoria Macchinisti e fuochisti PER MANCANZA DI ADERENTI, furono costretti a cambiar rotta [...]

La sintassi dei testi di Giusti è più spesso ipotattica che paratattica, in particolare per l'uso di lunghissime relative o completive introdotte da *che*.
Sono inoltre presenti interrogative retoriche, esclamative e molte parentetiche aperte dalle lineette.

3.2.7 Profilo linguistico di Alfonso Leonetti

I testi di Alfonso Leonetti sono i testi 26 e 27.
Il testo 26 parla delle teorie e le tesi di Pisacane, mentre il testo 27 parla dei pro e contro dell'astensionismo durante le elezioni.
Sono due testi di carattere didattico, basati principalmente sull'argomentazione.
Dal punto di vista retorico possiamo trovare anche qui la correctio, l'uso del costrutto *tanto quanto* e delle definizioni introdotte da *è*.
Presente inoltre anche l'uso della doppia negazione, ad esempio nel testo 26:

Ancora oggi noi *non possiamo non dire*: che peccato che egli sia morto così giovane!

Vi sono figure di ripetizione (soprattutto nel testo 27) come l'anafora e l'anadiplosi.
Anche in Leonetti è inoltre presente la ripresa anaforica di preposizioni che introducono sintagmi preposizionali.
Sono presenti dittologie sinonimiche e strutture binarie insieme a strutture ternarie collegate asindeticamente.
Il lessico è quello tipico della sfera politica socialista, segnaliamo l'uso del sintagma *partito comunista*, mai presente nei 50 testi gramsciani.
Troviamo il partito visto metaforicamente come una nave al testo 27:

Non è dunque vero che occorre oggi cambiar rotta, ma solo intensificare la pressione delle caldaie ed accelerare la marcia per giungere in porto [...]

Anche qui troviamo un'aggettivazione enfatica; dal punto di vista sintattico troviamo invece diverse interrogative retoriche.
C'è una prevalenza della paratassi sull'ipotassi, dove il soggetto può venire posposto.

3.2.8 Profilo linguistico di Giacomo Menotti Serrati

I testi di Giacomo Menotti nel corpus sono il testo 30, il testo 31 e il testo 32.
Sono tutti testi di tipo polemico: il 31 e il 32 sono in polemica contro dei compagni di partito, il testo 30 più che un testo polemico è un testo parodico su Maria Ryger e si differenzia molto poiché è strutturato come un piccolo racconto in cui vengono descritte in maniera ironica le azioni e i pensieri della donna.
In tutti i testi sono presenti forme di parallelismo, ad esempio *così…come,* e correctio nella forma *non solo…ma anche, non così…ma come*.
Sono presenti più strutture binarie che non dittologie sinonimiche, sono presenti anche strutture ternarie, collegate sia con asindeto che con polisindeto.
Per le figure di ripetizione troviamo al testo 30 un forte uso dell'anafora, in questo caso formando alliterazione:

> *Si* presero, *si* ripresero, *s*'avvinghiarono, squassarono l'un l'altro

L'anafora è meno presente nei testi 31 e 32, dove però nel testo 31 abbiamo però la ripresa anaforica di preposizioni che introducono sintagmi preposizionali.
Da segnalare nel testo 30 l'uso della perifrasi *il primo giorno del mese dei fiori* per indicare il primo maggio e la metafora *violetta senza profumo* per indicare un primo maggio senza lotta di classe.
Il testo 30 inoltre si chiude con una frase in dialetto romanesco:

> Te possino ammazzatte, te possimo.

In entrambi i testi abbiamo sempre un'aggettivazione enfatica con presenta di superlativi e nel testo 30 di diminutivi come *cappellino, leccatina* che vengono usati per dare un'aria leziosa alla protagonista del testo.

La sintassi è paratattica più che ipotattica; nei testi troviamo molte domande retoriche e molte esclamative.
Menotti spesso incomincia i periodi con *che* iniziale, utilizzando molto i puntini di sospensione.
Sono presenti molte parentetiche aperte da lineette.

3.2.9 Profilo linguistico di Angelo Tasca

I testi di Angelo Tasca sono i testi 33, 34, 35, 36, 37.
Tranne il testo 33, che è polemico contro l'interventismo di Mussolini, tutti gli altri testi sono del tipo didattico-argomentativo.
Come al solito è presente sia la correctio che il parallelismo; presente è anche la definizione, introdotta da *è* ripreso anaforicamente.
Sono presenti dittologie sinonimiche e strutture binarie, accompagnate da strutture ternarie collegate asindeticamente.
Sono presenti molte enumerazioni, anch'esse collegate asindeticamente, ad esempio al testo 36:

> [...] ma la vita intima dei vari partiti, le loro tendenze effettive, gli stati d'animo, le disposizioni, la preparazione loro restavano un mistero.

Per le figure retoriche molto frequente è l'anafora, anche con ripresa di preposizioni che introducono sintagmi preposizionali.
Altra figura retorica frequente è l'anadiplosi; sono presenti inoltre molti polittoti come *ha portato e la porterà, la potevamo e la possiamo*, *giorno per giorno*.
Il lessico è quello politico-socialista dove spesso è presente e ripresa anaforicamente è la parola *guerra*.
Al testo 37 troviamo però un lessico legato alla cultura e alla scuola come: *mucchi di libri*, *congerie di carta stampata*, *cultura*, *laurea*, *coscienza di sé*, *filosofia*, *scuola*, *metodo pedagogico*, *cultura socialista*.
Troviamo poi *diapason* inteso come propagazione non di suono ma di idea rivoluzionaria:

> Possiamo dire che la tensione rivoluzionaria ha toccato allora *il suo diapason.*

Troviamo inoltre similitudini, sempre al testo 37:

> [...] una specie di torta di cui toccano i grossi quarti ai pochi, le briciole ai più

Anche qui troviamo una aggettivazione enfatica, con superlativi e diminutivi in senso spregiativo.
La sintassi è tendenzialmente ipotattica, sono presenti interrogative: non soltanto interrogative retoriche ma anche delle vere e proprie domande rivolte al lettore.
Presente l'uso del vocativo (al testo 33: *o Mussolini*; *o occhi tuoi*; *o amico Mussolini*) e l'uso delle parentetiche, racchiuse più spesso fra parentesi tonde che non fra lineette.
Spesso il soggetto della frase è posposto per essere messo in risalto.

3.2.10 Profilo linguistico di Palmiro Togliatti

Togliatti ha il maggior numero di testi all'interno del corpus non gramsciano: suoi sono i testi numero 39, 40, 41, 42, 43, 44, 45, 46, 47, 48 e 49.
Sono tutti testi di carattere didattico-argomentativo in cui però Togliatti inserisce sempre una vena polemica e, come i testi gramsciani, sono costruiti o portando delle tesi avversarie e smontandole o portando esempi a favore di una propria tesi.
Anche in questo caso i testi sono basati sul parallelismo e sulla correctio, troviamo infatti strutture come: *non più soltanto di… ma di*, *non è… ma è*.
Presente è anche la definitio, dove il verbo essere alla terza persona presente singolare è spesso ripreso anaforicamente, ad esempio al testo al testo 47:

> Orbene, la pratica del socialismo *è* la migliore scuola per la diffusione di questa concezione, *è* dimostrazione continua delle verità del pensiero moderno, ed *è* una dimostrazione che non procede per via discorsiva, ma si serve di quei veri e concreti sillogismi che sono i fatti.

Oltre al presente, si può avere lo stesso schema con il passato remoto, ad esempio al testo 42:

[...]: *fu* la passione esacerbata di milioni di uomini, l'amarezza, il risentimento degli individui ripresi dal turbine delle istintive passioni bestiali dormienti sotto la vernice di civiltà, *fu* la libertà compressa, la personalità negata; e *fu* pure lo sfrenarsi delle brame di ogni egoismo, nella speranza del bottino, nella visione del regno del benessere dischiuso dalla rapina e dalla distruzione.

Le dittologie sinonimiche e le strutture binarie che convivono con strutture ternarie; in questo caso però le strutture ternarie possono essere sia collegate asindeticamente, ad esempio in 46:

Curioso, benevolo, attento, si era rivolto l'animo loro ad ascoltare voci diverse [...]

O possono essere coordinate con congiunzione, solitamente tra ultimo e penultimo termine, ad esempio al testo 44:

[...] noi viviamo, lavoriamo e ci tormentiamo.

Come si sarà già notato molto spesso all'interno di queste strutture ternarie è possibile trovare climax, molto spesso ascendente, ad esempio nel testo 46:

Ecco: *studio, serietà, disciplina*, ecc.

Sono presenti anche enumerazioni, spesso asindetiche ma anche polisindetiche come:

ecco la Gironda e i Giacobini, il Terrote e la Vandea, Robespierre e Carlotta Corday, e l'animo delle folle e la psicologia dei tribuni.

Moltissime sono le figure di ripetizione all'interno del testo, non soltanto anafora (anche con ripresa anaforica di preposizioni), ma anche epifora, ad esempio al testo 39:

che essa non è una politica *di bottegai*, ma è la politica di un governo che si trova sotto l'influenza di alcune classi *di bottegai*.

Troviamo inoltre anadiplosi e polittoto, ad esempio: *furono e sono, fu ed è, di sconfitta in sconfitta, si sono fatte e si fanno*.

Dal punto di vista del lessico troviamo tre diverse aree.

La prima è la tipica area politico-socialista dove ad esempio troviamo: *borghesia liberale italiana, classi borghesi, lotta di classe, lotta sociale, operaio, popolo italiano, sindacalismo, socialismo, classe dirigente*.

La seconda è l'area economica dove troviamo: *organismi economici, campo economico, categoria di beni, scambi, commercio internazionale, mercato nazionale, prodotti agricoli, agricoltura patriarcale, merci, industria*.

La terza area è quella umanistico filosofica, troviamo infatti: *concezione della vita, coscienza, filosofo, insegnamento, libri, progresso dell'uomo, scuola, spirito umano, spirito dogmatico, verità, volontà*.

Troviamo inoltre diverse metafore e similitudini; interessante notare come in alcune di queste si possa trovare un lessico legato alla letteratura, come ad esempio: *Atta Troll, letto di Procuste, storiella del padre Adamo, Sturm und Drang culturale*.

In senso polemico troviamo, anche se raro, troviamo lessico legato alla sfera naturale-biologica come ad esempio: *parassiti della cultura, larvato, frasario nuvoloso*.

Troviamo anche l'uso di un modo di dire al testo 41:

> campa cavallo che l'erba cresce.

L'aggettivazione è sempre presente e sono presenti sia superlativi che diminutivi.

Sono presenti domande retoriche ed esclamative, accompagnate spesso da interiezioni come *oh* e *uh*.

La sintassi tende più all'ipotassi che non alla paratassi, ma è formata da periodi piuttosto brevi che rendono scorrevole il testo al lettore.

Sono presenti parentetiche, più spesso all'interno di parentesi tonde che non lineette, dove Togliatti dà informazioni aggiuntive quali ad esempio i riferimenti dei testi che lui cita.

Segnaliamo inoltre l'uso di forme di abbreviazione, spesso alla fine di elencazioni, come: *ecc. ecc, e così via*, oltre all'uso dei puntini di sospensione.

Il soggetto può essere posposto, ad esempio in 42:

> È sicuro il Lanzillo [...]

Segnaliamo inoltre la presenza nel testo 42 della formula *gli è che* ad inizio del periodo.

3.2.11 Testi singoli di singoli autori

All'interno del corpus sono presenti molti testi singoli di singoli autori: sono i testi 14, 15, 16, 17, 28, 29, 38 e 50.

Il testo 14 è il testo di Gino Castagno.
Questo testo è un resoconto di un'azione politica commentata però in maniera polemica; anche in questo caso avremo una commistione fra discorso didattico-argomentativo e polemico.
Il testo si basa su forme di correctio (*non per… ma per*); sono presenti dittologie sinonimiche e strutture binarie.
Sono poco presenti le figure di ripetizione.
Anche qui troviamo un'aggettivazione enfatica, il lessico è quello tipo attinente alla sfera politica del socialismo; una menzione può essere fatta per l'uso di alcuni termini francesi come *union sacrè* e *récivement* e per l'uso della parola *compagni* (poco utilizzata in Gramsci).
Dal punto di vista della sintassi troviamo invece un periodare più ipotattico rispetto agli altri testi, dove vi possono essere premesse lunghissime, come:

> Dati i precedenti della Confederazione francese sui quali è inutile qui soffermarsi; dati i conosciuti vincoli massonici di tutta quella brava gente dell'intesa operaia francese, inglese, italiana; dato lo sfruttamento che del convegno di Londra si andava facendo, ai fini guerraioli, della stampa borghese; dato il voluto ristretto invito alle sole organizzazioni dell'intesa, con esclusione anche di quelle dei paesi neutrali; dato infine il carattere eminentemente politico (praticamente poi inutile, perché sulle proposte della Confederazione americana è già in corso un referendum fra tutte le Organizzazioni) del Convegno di Londra, non era naturale chi aveva aderito entusiasticamente al programma di Zimmerwald e si era impegnato a lavorare per esso in tutte le Organizzazioni, vedesse un pericolo grave nel Convegno stesso e ne temesse le impressioni sulle masse operaie e sulle Organizzazioni da esso escluse, sì da rendere poi più difficile, se non impossibile, il lavoro propostosi?

Nel testo, oltre a domande retoriche, sono frequentissime e ravvicinate le parentetiche, tanto da trovare parentetiche all'interno di parentetiche, inoltre sono molto più frequenti le parentetiche aperte da lineette:

> La cosa è diversa, Rigola e Quaglino sono andati a Parigi - commettendo, per il momento scelto (si teneva appunto a Parigi la kermesse interparlamentare dell'intesa) una inopportunità politica - a discutere con i rappresentanti della Confederazione francese del lavoro, di problemi tecnici riflettenti l'emigrazione, i rapporti tra le due Confederazioni e le pressioni sui rispettivi governi per prevenire i pericoli dell'emigrazione stessa.

Il testo numero 15 è firmato C.D.
È un testo a favore della creazione di un ufficio di assistenza popolare, non ha quindi una vena particolarmente polemica.
Anche in questo testo è molto presente la correctio, in forme del tipo *non dico questo ma dico quest'altro, non intendo questo ma intendo quest'altro*.
Sono presenti dittologie sinonimiche e strutture binarie.
Delle figure di ripetizione è presente l'anafora, in misura scarsa.
Nel lessico troviamo molte sigle e nomi di associazioni come: *Camera del lavoro*, *Segretariato del Popolo*, *Umanitaria di Milano*, *Segretariato dell'Emigrazione*, *Ufficio d'assistenza*, *Ufficio di collocamento*.
La sintassi è paratattica e lineare, con un'unica parentetica introdotta da lineetta.

Il testo numero 16 è il testo di Alessandro de Giovanni.
In questo testo De Giovanni polemizza con un compagno di partito sul futuro dell'Internazionale.
Anche in questo caso è presente correctio, introdotta in particolare da *ma* in posizione inziale.
Sono presenti dittologie sinonimiche e figure binarie, sono presenti inoltre figure di ripetizione come l'anafora (in particolare abbiamo anche qui la ripresa anaforica di preposizioni che introducono sintagmi preposizionali) e epizeusi come:

Ah! *Verranno, verranno*, compagno Luzzari, verranno anche per noi i bei giorni sereni [...]

O ancora polittoti come: *l'hanno tradita e l'hanno tradito*, *rafforzata e rafforzare*, *non possono e non potranno*.
Interessante notare come all'interno del lessico entrino termini che si rifanno alla navigazione come: *triste bufera*, *l'ultima voce del cannone*, *la Penelope della leggenda*, *spaventosa tempesta*, *tali marinari*, *tali nocchieri*, *possenti flutti della lotta di classe*.
Alla fine del testo è infatti presente una lunga metafora dove l'Internazionale diventa una barca guidata dai compagni marinari; moltissime sono anche le similitudini all'interno del testo.
Evidenziamo anche qui l'uso della parola *compagni*.
Vi sono molti aggettivi usati in maniera enfatica, ne troviamo alcuni oltre al superlativo anche al grado diminutivo con valore di spregio, come ad esempio*: giornalucoli pseudoletterari*.
Dal punto di vista sintattico la paratassi prevale sull'ipotassi e sono presenti numerose domande retoriche; a volte il soggetto della frase viene posposto per essere messo in evidenza.
Da segnalare la chiusura del testo con il motto:

> In alto i cuori, compagni, l'Internazionale sarà.

Il testo numero 17 è firmato C.F
Questo testo è un testo molto corto, ed è un invito all'apertura di un nuovo circolo socialista.
Si trovano dittologie e strutture binarie, qualche anafora, in particolare con la ripresa di preposizioni che introducono sintagmi preposizionali.
Anche in questo caso vi è una forte aggettivazione enfatica; inoltre si trovano una serie di sigle di organizzazioni e di toponimi.
All'interno del testo è presente una lunga catena asindetica di persone a cui l'invito è rivolto, anticipando così i complementi rispetto al verbo.

Il testo di Ottavio Pastore, il testo 28, è un testo altamente polemico che descrive un'assemblea di industriali "pescicani".
Sono presenti figure di ripetizione come l'anafora, in particolare con la ripresa anaforica di preposizioni che introducono sintagmi preposizionali.
Presente anche l'epizeusi:

Pesciolini, pesciolini, ubbidite se no…i pescicani approvano;

Nel testo sono presenti più strutture binarie che non dittologie sinonimiche, l'aggettivazione è anche qui enfatica, con presenza di superlativi.
Come metafora troviamo la figura dei pescicani e quella dei pesciolini e altro lessico marittimo come *burrasche superate, oceano della vita*. Troviamo però anche: *grano, farina, pane, zuccherieri, vermouth, panettieri*.
Nel testo si trovano domande retoriche, esclamative e la presenza di interiezioni come: *uff! ah, ahimè*; ma soprattutto Pastore si distingue dagli altri per l'uso insistito dei tre puntini di sospensione all'interno del testo.

Il testo 29 di Mario Santarosa, è un testo polemico contro il ministro Bonomi.
Nel testo sono presenti molte dittologie sinonimiche e strutture binarie che si accompagno a strutture ternarie, a volte asindetiche e a volte no, come ad esempio:

Ed il profondo, il pratico, l'eccellentissimo ministro Bonomi[...]

O ancora:

[...]i dileggi e le infamie e le persecuzioni[...]

Per le figure retoriche molto presenti sono l'anafora, in particolare la ripresa anaforica di preposizioni che introducono sintagmi preposizionali e l'anadiplosi; sono presenti inoltre diversi polittoti.
Il lessico è quello politico-socialista, l'aggettivazione è anche qui molto enfatica, con presenza di superlativi, tanto da produrre quasi allitterazioni, come nel caso di *pingue pianura padana.*
Dal punto di vista sintattico troviamo diverse domande retoriche e molte parentetiche aperte da lineette che spezzano periodi già molto lunghi e ipotattici, dove a volte il soggetto viene posposto.

Il testo 38 è il testo di Umberto Terracini.
Questo testo è un testo polemico contro il protezionismo nazionalista e i suoi dazi, al suo interno è presente infatti un lungo elenco di beni di prima necessità e il loro prezzo in lire.
Anche in questo caso sono presenti forme di correctio, come ad esempio *non già*, *ma bensì*.
Sono presenti inoltre dittologie sinonimiche e strutture binarie.
Troviamo anche qui figure retoriche di ripetizione come anafora, anche con ripresa di proposizione che introduce sintagmi preposizionali e anadiplosi.
Presente anche il polittoto, ad esempio *è pagata e sarà sempre pagata*.
Il lessico è in questo caso basato sulla sfera economica, troviamo infatti: *altissimi tassi doganali, consumatori, bilancio di una famiglia, tributi medioevali, imposte proporzionali e progressive, bilancio pubblico*.
Troviamo l'uso di *grucce* per intendere aiuti:

> [...]nonostante le grucce donate dallo stato alle nostre industrie[...]

E anche l'uso metaforico con antinomia di *grandi fiumi e piccoli ruscelli*.
Anche in questo caso troviamo un'aggettivazione enfatica con presenza di superlativi.
La sintassi è più ipotattica che paratattica.

Il testo numero 50 è il testo di Antonio Viglongo.
È un testo di tipo argomentativo didattico in cui Viglongo recensisce per i lettori le teorie pedagogiche di tale Lombardo-Radice.
Anche qui è presente la correctio, sono presenti poche strutture binarie.
Sono presenti figure di ripetizione come l'anafora.
Il lessico è un lessico di tipo filosofico, incentrato sulla pedagogia, a cui alla fine del testo si associa il lessico socialista, ad esempio troviamo: *educatore ed educando, concetto di educazione, compenetrazione di anime, nostro spirito, differenza di grado di umanità, noi socialisti, la nostra propaganda, apostolato d'educazione*.
La sintassi è paratattica ma la lettura viene resa poco scorrevole da innumerevoli citazioni e parentetiche a commento del testo.
Presente è all'interno del testo un elenco numerato.
Segnaliamo inoltre l'uso di abbreviazioni come *p. es*. al posto di "per esempio".

3.3 Risultati dell'analisi qualitativa

Vogliamo riassumere quello che abbiamo trovato nell'analisi dei 100 testi gramsciani e non gramsciani.
Come si è potuto vedere dall'analisi i testi gramsciani hanno molti elementi in comune con i testi non gramsciani, in particolare:

- La presenza di discorsi sia didattico-argomentativi che polemici
- La presenza di parallelismo, correctio e definizione
- La presenza di figure binarie, tra cui rientrano le dittologie sinonimiche, assieme a strutture ternarie
- La presenza di figure retoriche di ripetizione
- La presenza di un lessico appartenente alle stesse aree semantiche e quindi molto simili
- La presenza di metafore e similitudini
- La presenza di molti aggettivi
- La presenza di interrogative retoriche
- La presenza di esclamative
- La presenza di parentetiche

Naturalmente esistono però anche delle differenze fra lo stile gramsciano e quello di altri autori come:

- La quantità di figure retoriche di ripetizione

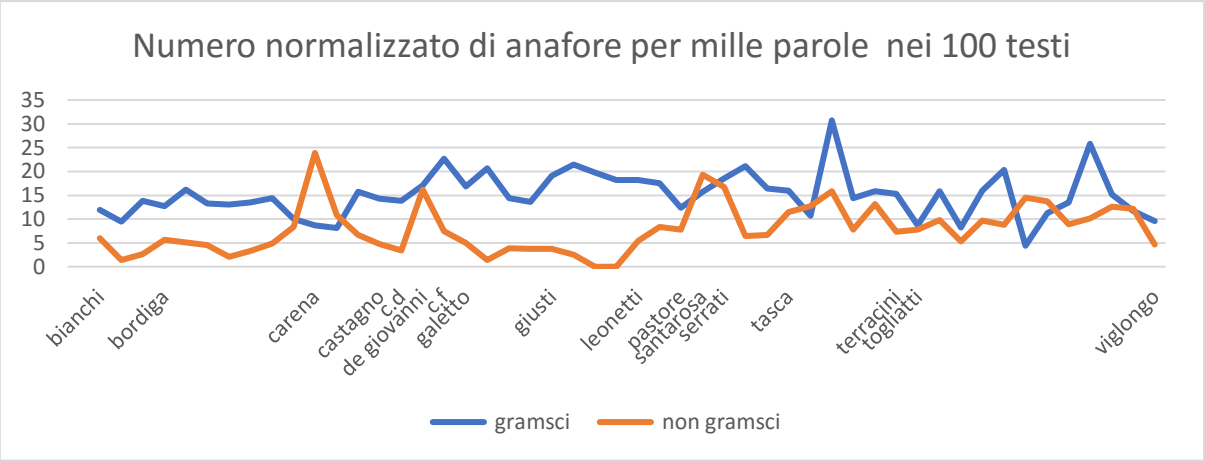


Figura 16 Numero normalizzato di anafore nei testi gramsciani e non gramsciani

In figura 16 abbiamo costruito un grafico contenente il numero di anafore; abbiamo scelto di considerare l'anafora in quanto è la figura retorica più utilizzata da Gramsci e dagli altri autori.

Per ovviare alla diversa lunghezza dei testi del corpus si è scelto di normalizzare il numero di anafore per mille parole per ognuno dei 50 testi gramsciani (linea blu) e dei 50 testi non gramsciani (linea rossa); i testi sono ordinati come appaiono in bibliografia (il primo punto della serie Gramsci rappresenta il numero normalizzato di anafore nel testo 1 di Gramsci, il punto uno della serie non Gramsci rappresenta il numero di anafore normalizzato nel testo 1 degli autori non gramsciani, ovvero il primo testo di Bianchi).

Nonostante questa normalizzazione sia un po' grezza mostra chiaramente come Gramsci usi molto di più l'anafora degli altri autori e come però alcuni autori, come ad esempio Tasca o Togliatti, siano più vicini di altri all'uso che ne fa Gramsci.

- L'assenza o la presenza di determinate parole o sintagmi e la loro quantità

In Gramsci ad esempio non è mai presente il sintagma *partito comunista*; nel subcorpus non gramsciano non è mai presente *Stenterello*. Alcune parole sono meno usate da Gramsci rispetto agli altri autori (*pace* ha un tasso dello 0.9 con 50 occorrenze nel sub-corpus non gramsciano, mentre in quello gramsciano ha un tasso dello 0.1 con 10 occorrenze).

- La quantità di esclamative

In Gramsci le esclamative sono rare (8 in tutti i testi gramsciani, di cui però 7 in un solo testo, il numero 14), mentre altri autori usano più spesso le esclamative.

- Il tipo di segno grafico contenente le parentetiche (parentesi tonde o lineette) e la quantità delle parentetiche.

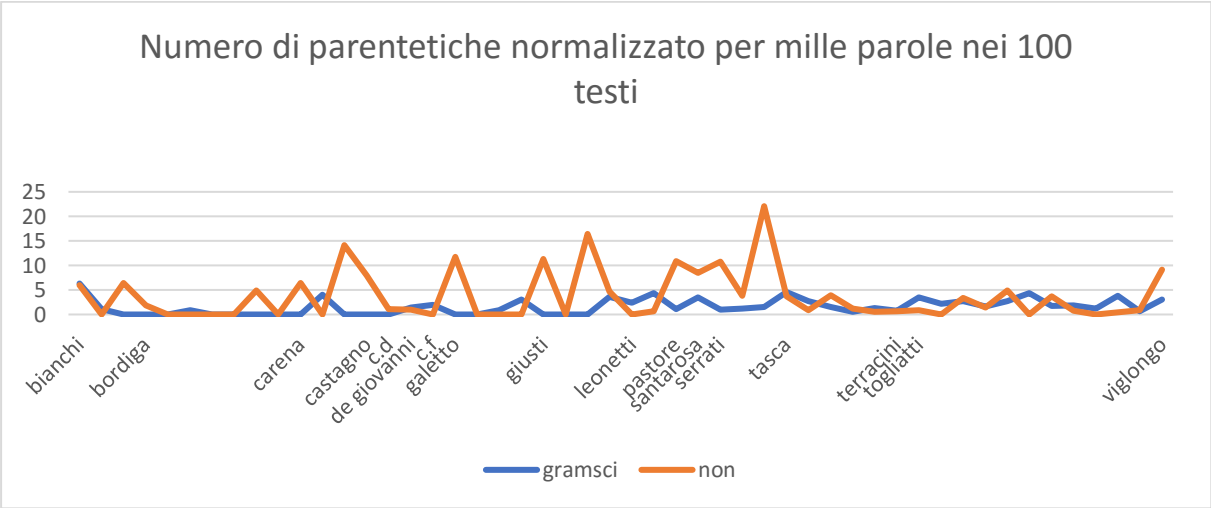


Figura 17 Numero normalizzato di parentetiche all'interno dei 100 testi gramsciani e non gramsciani

Come possiamo vedere in figura 17 abbiamo costruito un grafico contenente il numero normalizzato di parentetiche, ovvero il numero di parentetiche per mille parole in ogni testo. Anche questa normalizzazione è un po' grezza ma permette di vedere nel complesso come alcuni autori (ad esempio Giusti e Serrati) usino molte più parentetiche rispetto a Gramsci. Nell'uso delle parentetiche c'è inoltre una differenza dovuta al simbolo grafico che introduce la parentetica: mentre in Gramsci nel caso di più parentetiche sono presenti più parentetiche aperte dalla parentesi tonda e mai più di una parentetica introdotta da lineetta, in altri autori è invece l'opposto, con un maggior numero di parentetiche aperte esclusivamente da lineetta o in maniera maggiore rispetto a quelle aperte da parentesi tonda.

- La presenza o assenza di interiezioni
- La presenza o l'assenza di caratteri maiuscoli
- La presenza del dialetto

Come si può vedere la maggior parte delle differenze fra gli autori riguardano più l'aspetto quantitativo che l'aspetto qualitativo; alcuni di questi aspetti inoltre non sono prettamente linguistici (ad esempio il tipo di segno grafico che apre la parentetica).
Certo, come abbiamo visto un'analisi qualitativa non è impossibile, ma a noi pare che la nostra analisi abbia messo in risalto proprio questo: la differenza fra i vari autori, proprio perché lo stile retorico e il lessico di questi è molto omogeneo, non sta tanto nella differenza del lessico o delle strutture retoriche, (differenza che in alcuni casi è comunque presente) quanto nella quantità con cui queste strutture si presentano fra di loro.
Per la sintassi andrebbe fatto un discorso a parte, in quanto ci sono sicuramente delle differenze dal punto di vista qualitativo.
Poiché però la sintassi di Gramsci cambia parecchio a seconda che il testo sia tipo didattico-argomentativo o polemico, non è secondo noi un elemento altamente discriminante.
Certamente se ci troviamo di fronte a periodi estremamente lunghi, dove la struttura sintattica presente in Gramsci viene di molto alterata, saremo meno propensi per la non attribuzione del testo a Gramsci, ma appunto deve essere un cambio estremamente evidente rispetto alla sintassi gramsciana e questo cambio non appare né spesso né in tutti gli autori.
Dobbiamo ricordare infatti che non stiamo discriminando fra un autore e un altro autore ma fra un autore e altri diciassette autori; quanto è probabile che una delle differenze di tipo qualitativo che abbiamo trovato appaia in un testo?
Ad esempio, alcune differenze qualitative da noi rilevate sono l'assenza o la presenza del dialetto e l'assenza o la presenza delle interiezioni all'interno dei testi, ma quanto è probabile che un elemento di questo tipo appaia all'interno di un testo da attribuire?
È davvero un fattore discriminante importante in questo caso di attribuzione?
Ma soprattutto, dai risultati di questa analisi qualitativa, possiamo riuscire ad attribuire testi gramsciani a Gramsci e soprattutto non attribuire nessun testo a Gramsci che non sia di Gramsci?

3.4 Confronto fra analisi qualitativa e quantitativa

Esaminando i testi non gramsciani del corpus usato per il test cieco possiamo trovare alcuni autori (ad esempio Leo Galetto o Giuseppe Bianchi) che usano poco figure retoriche di ripetizione; testi di questo tipo sono facili da escludere dall'insieme dei testi gramsciani.

Un testo gramsciano si differenzia da quelli degli altri autori non soltanto per il forte uso delle figure retoriche di ripetizione, ma anche per una sintassi mai troppo contorta o dai periodi eccessivamente lunghi, per la quantità delle esclamative, delle parentetiche e del tipo di segno grafico che introduce le parentetiche.

Il lessico può essere d'aiuto in qualche, sia pur raro, caso: se troviamo all'interno di un testo *Stenterello,* parola usata nel corpus solo da Gramsci, abbiamo una forte probabilità di trovarci davanti a un testo gramsciano.

Il problema appare quando in un testo la quantità di figure retoriche presenti e altre strutture elencate sono molto simili a quelle gramsciane.

Prendiamo ad esempio il 11, un testo di tipo argomentativo-didattico: si parla della condotta del Papa e dei cattolici italiani durante la guerra.

Anche questo testo è basato sulle forme del parallelismo e della correctio con costruzioni come ad esempio *non…ma* e strutture binarie come: *anguillesca ed ondeggiante, moderata e liberale, borghese e conservatore, ricchi e potenti*.

Troviamo però anche strutture ternarie come: *organismo giovane, audace, pieno di vita*.

Troviamo inoltre molte figure di ripetizione, anche con ripresa anaforica di preposizioni che introducono sintagmi preposizionali.

Per il lessico, oltre a quello dell'area politico-socialista, abbiamo un lessico riguardante l'area religiosa, ad esempio: *Chiesa, papa, partiti politici cattolici, quotidiani clericali, religione cristiana*.

Se da una parte è vero che nei testi da noi analizzati Gramsci ha sempre parlato poco di questi argomenti, non possiamo però escludere del tutto che non si sia mai interessato di questi (si veda il testo 45 *I cattolici italiani*).

La sintassi invece, nonostante si tratti di un testo di tipo argomentativo, tende più alla paratassi che all'ipotassi, è molto simile quindi a quella gramsciana; troviamo però un'esclamativa, presente nei testi gramsciani ma in maniera rara.

Anche il testo 33 ha molte somiglianze con una scrittura di tipo gramsciano: un misto fra struttura di tipo didattico-argomentativo e polemica contro chi infanga la rivoluzione russa.

Anche in questo caso il testo è basato sul parallelismo e sulla correctio, ed è presente una serie di definizioni con ripresa anaforica del verbo *essere*:

> Ai primi albori rivoluzionari Kerenski *era* - per le borghesie nazionaliste - il pazzo che correva dietro alla visione esistente, in realtà, solo nel suo cervello malato*: era* l'uomo che non vedeva

la sopravveniente rovina del proprio paese; *era* l'intruso nel Governo prepotentemente cacciato a quel posto dal consesso di irresponsabili nominati a far parte del Soviet.

Sono infatti molto presenti figure di ripetizione, come anafore, anadiplosi e polittoti, anche con ripresa anaforica di preposizioni che introducono sintagmi preposizionali.

Il lessico è quello tipico socialista con la presenza però di alcuni termini della sfera biologica e in particolare della terminologia medica come: *diagnosticare*, *malattia*, *paralisi*, *scienza medica*.

Anche qui abbiamo un'aggettivazione enfatica, con uso di superlativi.

La sintassi è molto simile a quella gramsciana, inoltre come in Gramsci troviamo sia delle parentetiche aperte a volte dalla lineetta, a volte dalla parentesi tonda, l'unica cosa che differisce da Gramsci è la distribuzione dei diversi simboli grafici che introducono la parentetica all'interno del testo: nei testi gramsciani le parentetiche aperte da lineetta non sono mai più di una e vi sono invece diverse parentetiche aperte dalle parentesi tonde, in altri autori è esattamente il contrario.

Nel testo 33 sono molte più le parentetiche aperte dalle lineette che non dalle parentesi tonde.

Riassumendo sia per il testo 11 che per il testo 33 abbiamo elementi a favore dell'attribuzione a Gramsci come autore del testo e degli elementi a sfavore.

Potremmo scegliere di basarci solo sulla presenza di molte figure retoriche di ripetizione o ad esempio sulla presenza di costruzioni che si trovano in Gramsci, come ad esempio la definitio al testo 33; oppure pensare di non attribuire questi testi a Gramsci proprio perché sono presenti elementi che si discostano dal suo profilo generale, come ad esempio la bassa probabilità di una esclamativa all'interno del testo o l'alto uso di parentetiche aperte da lineetta rispetto alle parentetiche aperte da parentesi tonda.

Il secondo approccio è probabilmente il più vincente; i due testi infatti non sono di Gramsci: il testo 11 è di Ottavio Pastore, mentre il testo 33 è di Omero Concetto.

Se in questo caso siamo riusciti a non attribuire testi non gramsciani a Gramsci, quante sono le probabilità che Gramsci differisca dal suo profilo generale?

Noi ad esempio non avremmo mai attribuito il testo 14 a Gramsci, un testo altamente polemico che racconta del suicidio di tale Sperindio Tagliani.

Questo testo ha infatti caratteristiche che abbiamo trovato raramente nei testi Gramsciani da noi analizzati: l'uso insistito dei tre punti di sospensione, presenti ben 4 volte all'interno del testo (dove all'interno di tutto il sub-corpus gramsciano dei 50 testi sono presenti solo 4 volte).

I puntini di sospensione vengono usati anche per creare una battuta per mezzo di un gioco di parole:

> L'affinità dei due studi deve essere determinata specialmente e forse unicamente dal fatto che a Strasburgo sono celeberrime le oc...he. (Il signor Censore gusterà molto questa bottata e me ne renderà merito in indulgenza per qualche capestreria: ci conto e lo spero).

Ma oltre all'uso insistito dei puntini di sospensione è proprio la struttura del testo ad averci spinto a non attribuirlo a Gramsci: gran parte del testo è infatti strutturato come un racconto in cui si narrano i motivi che hanno portato Sperindio al suicidio, in maniera molto simile al testo 30 del corpus dei 50 testi non gramsciani, un testo di Giacomo Menotti.

Riassumendo possiamo quindi affermare che l'attribuzione per il linguista o il filologo non è impossibile; tuttavia quando si incontrano autori che hanno figure retoriche e costruzioni simili a Gramsci e più o meno nella stessa quantità, l'attribuzione diventa incerta e le probabilità di attribuire un testo non gramsciano a Gramsci si alzano.

Un altro problema che rende l'attribuzione difficile è la divisione all'interno dei corpus fra testi didattico-argomentativi e testi polemici: da una parte abbiamo due modalità di costruzione del testo che rendono molto simili gli autori; questo avviene soprattutto nel testo di tipo polemico, dove i periodi sono sempre molto corti e paratattici, dove le figure di ripetizione sono molto presenti ed è quindi più difficile differenziare tra diversi autori.

Lo stile polemico tende però anche all'esagerazione: come nel caso prima illustrato è possibile anche trovare costrutti o tratti stilistici molto diversi dal profilo generale dell'autore che avevamo invece precedentemente stilato.

Nonostante in questo caso lo scopo dell'attribuzione sia discriminare fra testi non gramsciani e testi gramsciani, una attribuzione di tipo qualitativo risulta ancora più difficile quando all'interno del corpus vi sono molti autori rappresentati da un solo testo o da pochi testi.

Durante l'analisi qualitativa infatti si cercano i tratti più salienti dell'autore, quelli più visibili ad occhio nudo e allo stesso tempo anche quelli più frequenti; per questo, per quanto sembri controintuitivo l'attribuzione qualitativa funziona meglio con un campione ben rappresentativo di testi per autore.

Allo stesso tempo, quando all'interno dei testi gramsciani troviamo strutture o segni grafici che raramente appaiono in Gramsci, saremo più tentati di non attribuire il testo a lui.

Come vedremo nel prossimo paragrafo però, anche gli *n*-grammi e l'entropia relativa possono non attribuire testi di Gramsci a Gramsci proprio perché molto probabilmente si discostano da un suo profilo generale.

3.4.1 I testi non attribuiti a Gramsci da n-grammi e entropia

Abbiamo riportato al paragrafo 3.1.3 i risultati dell'analisi con metodi matematici di Basile e altri.

Sebbene, come abbiamo già precedentemente spiegato, non ci sia stato nessun falso positivo, il metodo non ha però attribuito tutti i testi di Gramsci a Gramsci.

Il metodo degli *n*-grammi non infatti ha riconosciuto i testi di Gramsci numero 1, 2, 7, 8, 14, 35, 50; mentre l'entropia non ha riconosciuto i testi 1, 2, 14 e 50.

Ci siamo chiesti: che cos'hanno in comune questi testi e in che cosa si differenziano dagli altri testi del corpus?

I testi non riconosciuti dagli n-grammi, a nostro parere, hanno in comune il fatto di avere al loro interno alcune parole che sono sottoutilizzate in Gramsci rispetto al sub-corpus degli autori non gramsciani, questo dato è stato confermato da una analisi quantitativa da noi eseguita sul corpus.

Il corpus dei 100 testi gramsciani e non gramsciani è composto da un totale di 104569 word token e da un vocabolario di 14547 word type; calcolando la Type Token Ratio possiamo vedere come questa sia del 13.9%; è quindi al di sotto del 20%, ovvero il corpus ha una dimensione adatta per essere indagato con metodi statistici, anche se la percentuale di hapax risulta invece leggermente alta in quanto è del 55,8%.

I testi del corpus sono stati normalizzati attraverso l'uso del software Taltac dove è stata applicata una normalizzazione leggera che ha eliminato in particolare tutte le maiuscole non rilevanti; sono poi stati estratti e selezionati i poliformi da noi ritenuti più interessanti.

Il corpus così normalizzato con all'interno i poliformi è stato processato con il software di statistica testuale Iramuteq.

Abbiamo quindi calcolato le specificità con Iramuteq, dove Iramuteq calcola attraverso il modello ipergeometrico[22] quanto una parola è associata ad una modalità di una variabile (in

[22] Il modello ipergeometrico calcola la probabilità che una forma grafica compaia in un dato gruppo, cioè calcola la probabilità che una data forma grafica *wi* compaia in un dato gruppo *pj* zero volte, una volta, due volte ecc. Le forme grafiche che presentano frequenze vicine alla media della distribuzione prendono il nome di "forme banali" in quanto il numero di presenze è vicino a quello atteso. Il confronto della frequenza con il valore medio è utile per stabilire se una forma grafica è specifica per il gruppo, in quanto sovrautilizzata

questo caso all'appartenere ad un testo di Gramsci) in maniera positiva (sarà quindi sovrautilizzata) o in maniera negativa (sottoutilizzata):

al testo 1 troviamo 6 occorrenze di *guerra* (-3,75 in Gramsci); al testo 2 troviamo 1 occorrenza di *guerra* e 2 occorrenze per *compagni* (-4,1).

Al testo 7, 1 occorrenza per *compagni*, al testo 14, 2 occorrenze per *guerra*, al testo 35, 5 occorrenze per *pace (-6.6)*, infine al testo 50, 2 occorrenze per *guerra*.

Dal punto di vista qualitativo abbiamo notato invece che in alcuni di questi testi sono presenti delle catene polisindetiche con *e*, rare nei testi di Gramsci, come ad esempio in Gramsci 7: «*e* per gli uni *e* per gli altri sempre presente».

In Gramsci 8: «che l'uomo ha fatto per liberarsi *e* dai privilegi *e* dai pregiudizi *e* dalle idolatrie», «non debba sapere come *e* perché *e* da chi sia stato preceduto, *e* quale giovamento possa trarre da questo sapere».

In Gramsci 35: «essi sono ora la calamita che muta la disposizione caotica delle molecole umane, *e* chiarifica gli aggregati, *e* pone nel primo piano le maggioranze effettive».

In Gramsci 1: «*e* riconoscendo per il momento la propria immaturità ad assumere il timone dello Stato, *e* permettesse che nella storia fossero lasciate operare quelle forze che il proletariato, non sentendosi di sostituire, ritiene più forti. *E* il sabotare una macchina».

In Gramsci 2: «conflitti sanguinosi avvennero tra operai francesi *e* spagnuoli *e* vi furono molti feriti».

In Gramsci 14: «Leumann mise a disposizione un grandioso fabbricato *e* sei palazzine per ospedale *e* case di convalescenza».

In Gramsci 50: «*e* possa trovare un pubblico che la sostenga *e* la migliori con la collaborazione del suo fervore».

Se guardiamo ai testi che l'entropia non ha riconosciuto come gramsciani, possiamo notare come non abbia riconosciuto i testi 1 e 2, i due testi gramsciani più vecchi (31 ottobre 1914 e 13 novembre 1915) all'interno del corpus dei 100 testi.

È possibile che in questo caso l'entropia abbia notato una differenza di stile causata dallo sviluppo successivo dello stile gramsciano.

Nell'analisi ci sembra in particolare che la ripresa anaforica di preposizioni che introducono sintagmi preposizionali sia molto poco presente, se non addirittura assente all'interno dei due testi.

(caratteristica positiva) o sottoutilizzata (caratteristica negativa) rispetto alla frequenza attesa. (Tuzzi, 2003, p. 133)

.

Il testo 14, oltre ad essere molto corto, è un tripudio di esclamative mai presenti in maniera così massiccia all'interno dei 100 testi gramsciani.
Il testo 50 è il più corto fra tutti i testi gramsciani, ci sono poche riprese anaforiche.
Nel secondo test il metodo degli n-grammi non ha riconosciuto invece il testo 20 e il testo 25.
Non abbiamo trovato particolarità dal punto di vista lessicale se non al testo 20 con l'uso di alcuni termini francesi come: *sales hommes, ordure, marchè des dupes, jusqui au bont*.
Sono anche presenti in due occorrenze i puntini di sospensione, assai rari in Gramsci.
Il testo 25 è invece un testo molto corto, dove le figure retoriche di ripetizione sono assenti.

3.4.2 Costruzione del corpus

Ci siamo chiesti se la divisione fra testi argomentativi e testi politici potesse influire in qualche modo con l'attribuzione, cioè se ci fosse uno scompenso fra testi argomentativi e testi politici.
Se guardiamo a come sono stati strutturati i 100 testi all'interno del corpus possiamo vedere come questi siano divisi fra testi di tipo didattico-argomentativo e testi di tipo polemico (al paragrafo 3.1 abbiamo spiegato come all' interno dei testi aspetti didattico-argomentativi e polemici possano mescolarsi; abbiamo quindi etichettato come testi didattico-argomentativi i testi che secondo noi rientravano maggiormente in questo tipo di costruzione del discorso e abbiamo fatto lo stesso per quelli polemici).
Dalla nostra analisi qualitativa abbiamo identificato nel corpus gramsciano come testi didattico argomentativi 36 testi e 14 come testi polemici, mentre nei testi non gramsciani abbiamo identificato come testi didattici 33 testi e 17 come polemici; in altre parole, in Gramsci i testi didattici rappresentano il 72% del corpus, mentre i testi polemici il 28%, nei testi non gramsciani i testi didattici rappresentano il 66% del corpus, mentre i testi polemici rappresentano il 34% del corpus.
Possiamo dedurre che nel corpus dei 100 testi non c'è uno squilibrio fra testi di tipo didattico-argomentativo e testi di tipo polemico[23].

[23] Poiché abbiamo due mutuabili dicotomiche possiamo calcolare l'indice φ tetracorico: questo indice calcola il grado di associazione tra due variabili dicotomiche con valori che vanno da +1 o a -1 attraverso la formula:

$$\varphi = \frac{ad - bc}{\sqrt{(a+b)\,(c+d)\,(a+c)(b+d)}}$$

Vi è però all'interno del corpus dei 100 testi quello che secondo noi è uno squilibrio di tipo tematico, questo squilibrio potrebbe spiegare anche la sottoutilizzazione o la sovrautilizzazione di alcune delle parole precedentemente descritte.

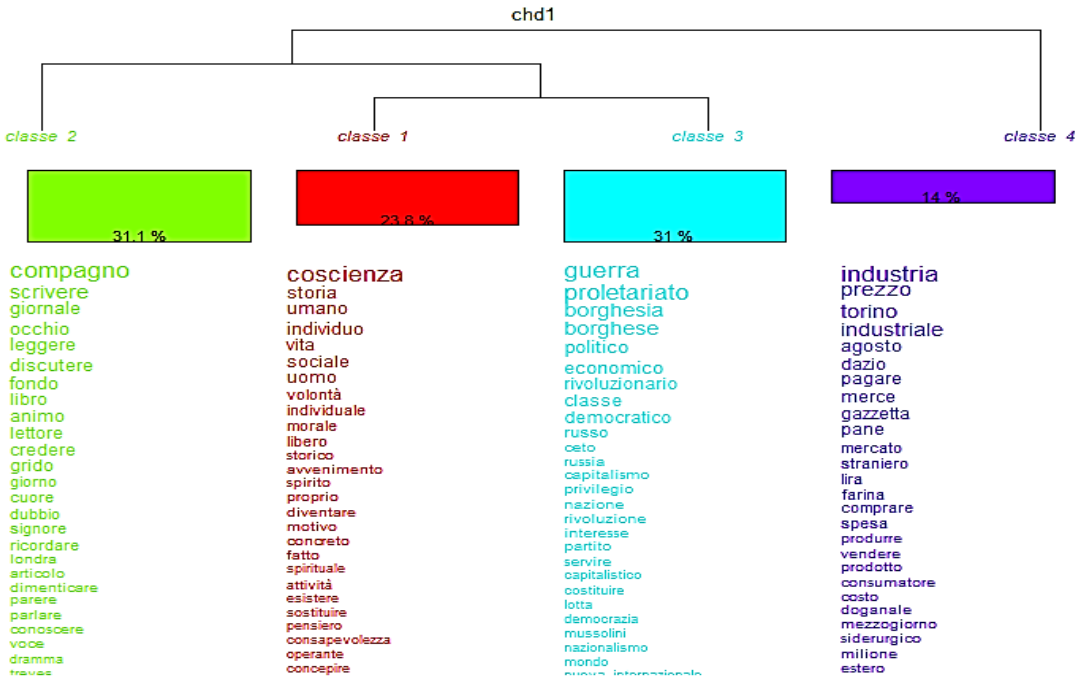


Figura 18 Analisi Reinert sul corpus di 100 testi gramsciani e non gramsciani

In figura 18 possiamo vedere i risultati dell'analisi basata sul metodo Reinert da noi eseguita sul corpus dei 100 testi con il software Iramuteq.

L'analisi suddivide ogni testo in paragrafi rispettando la punteggiatura: le porzioni di testo vengono definite Unità di Contesto Elementari; basandosi sulle occorrenze e le co-occorrenze delle parole piene ed eliminando le parole vuote viene costruita una matrice di parole per Unità di Contesto Elementari.

Viene applicata una cluster analysis che traduce il concetto di similarità lessicale: ogni classe è composta dalle Unità di Contesto Elementari che presentano gli stessi contenuti.

L'analisi basata sul metodo Reinert divide il corpus in diverse classi lessicali: in questo caso l'analisi ha trovato quattro grandi aree lessicali o tematiche: la classe 2 la cui parola tematica principale è data da *compagno*, la classe 1 la cui parola tematica principale è data da

Applicando la formula ai testi gramsciani e non gramsciani e alla divisione fra testi polemici e didattico argomentativi abbiamo come risultato 0,06 percui non è presente una associazione fra testi gramsciani e non gramsciani e fra testi polemici e testi politico-didattici.

coscienza, nella classe 3 ritroviamo invece *guerra*, mentre nella classe 4 troviamo la parola *industria*.

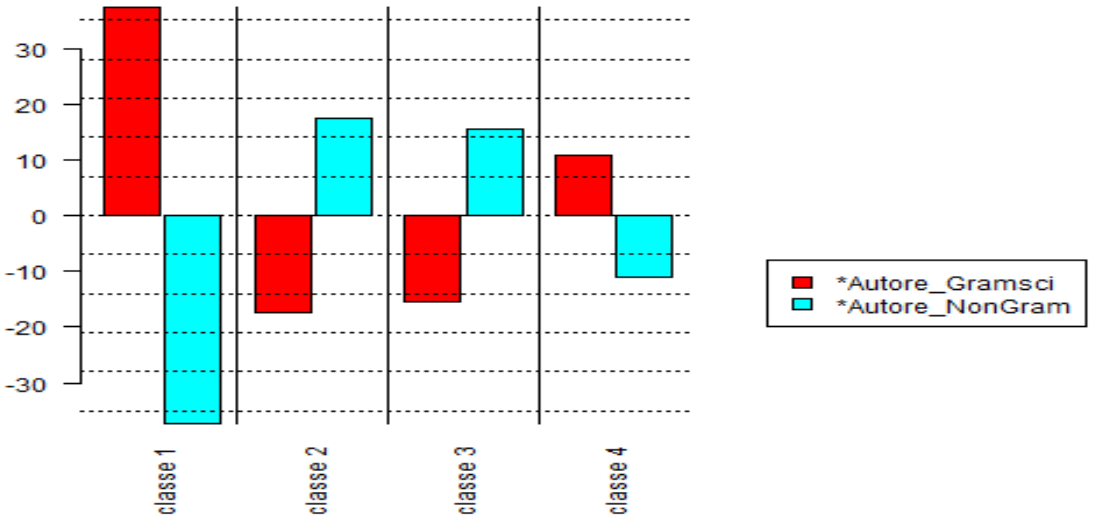


Figura 19 X^2 per autore e classe lessicale

La figura 19 rappresenta il calcolo del X^2 [24] per classe lessicale e autore, rappresenta cioè quanto i testi gramsciani o non gramsciani siano associati a una specifica classe lessicale.
In questo caso possiamo vedere come Gramsci sia molto associato alla classe lessicale 1, ovvero quella con parola tematica legata a *coscienza*.
Potremmo definire la classe lessicale 1 come classe lessicale legata alla filosofia, e in effetti, come abbiamo visto al paragrafo 3.2.1, questo lessico è molto presente in Gramsci; allo stesso modo questa classe lessicale è poco associata nel subcorpus dei 50 testi non gramsciani.
Allo stesso modo però la classe 2 e la classe 3 (rispettivamente la classe lessicale *compagno* e la classe lessicale *guerra*) sono poco associate a Gramsci, mentre la classe 4 (*industria*) è più legata a Gramsci che non agli autori non gramsciani.
Questo risultato potrebbe forse spiegare anche il perché alcune parole come *compagni*, *pace* e *guerra* siano presenti all'interno di gran parte dei testi non riconosciuti dagli *n*-grammi e dall'entropia.
Sarebbe interessante capire se effettivamente queste sovrautilizzazioni o sottoutilizzazioni di determinate classi lessicali siano legate a delle differenze stilistiche, cioè se effettivamente Gramsci abbia una tendenza ad occuparsi di ambiti più filosofici e in tal caso tralasci ad esempio discorsi sulla guerra o se invece questo risultato sia dipeso da una cattiva

[24] Il X^2 indica se vi è un'associazione esistente fra due mutuabili.

costruzione del corpus, creando così uno sbilanciamento di tematiche fra Gramsci e gli altri autori.

CONCLUSIONE

Abbiamo visto come i metodi di tipo quantitativo possano essere dei validi strumenti per l'attribuzione d'autore, sia che questi siano a base statistica o a base matematica.

Abbiamo dimostrato come nello specifico caso gramsciano un'attribuzione di tipo qualitativo non sia impossibile ma abbiamo dimostrato anche come questa sia dipesa non tanto da elementi di tipo qualitativo ma da elementi di tipo quantitativo e da elementi che sono legati più ad aspetti di tipo grafico che non ad aspetti di tipo linguistico (ad esempio l'uso di aprire le parentetiche con lineetta invece che con parentesi tonda).

Un metodo come quello sviluppato da Basile ed altri diventa sicuramente uno strumento prezioso nelle mani del filologo o del linguista, in particolare in casi simili a quello presentato dove il lessico, le strutture retoriche, la semantica, la sintassi sono molto uniformati fra i vari autori.

All'interno dell'ambito dell'attribuzione d'autore questi metodi sono inoltre certamente economici in termini di tempo (un essere umano impiegherà molto tempo per completare una analisi stilistica su un corpus ampio) e come abbiamo detto possono confermare o smentire le supposizioni del linguista o del filologo.

Non possiamo fare a meno di notare però che, se si vuole uscire dall'ambito della pura e semplice attribuzione, questi metodi, soprattutto quando si usano misure di distanza, non dicono però poi molto sui testi.

Possono dirci ad esempio che Gramsci è più vicino e quindi più simile ad un altro autore ma non spiegano il perché lo sia, e anche quando osserviamo un dato di tipo statistico, come ad esempio un maggiore presenza di *e* per un determinato autore, il metodo non ci spiega però in che modo questa maggiore presenza di *e* sia usata all'interno del testo.

A noi pare che sia proprio questo il lavoro del linguista dal punto di vista dell'attribuzione d'autore quando si avvale di metodi quantitativi in ottica esplorativa: prendere il dato quantitativo e interpretarlo all'interno del testo.

Per il momento sono ancora lontani i giorni in cui il linguista o il filologo verrà soppiantato dalla macchina ed anzi, lo studioso in ambito letterario o linguistico non deve preoccuparsi di essere rimpiazzato.

Questi strumenti al contrario aiutano il linguista e allo stesso modo questi strumenti non possono fare a meno del linguista durante alcune operazioni.

Ricordiamo infatti che operazioni come la lemmatizzazione non possono essere del tutto automatizzate; inoltre il linguista o il filologo è di fondamentale importanza per la costruzione del corpus, fase preliminare importantissima per la buona riuscita di un lavoro di attribuzione.
Come il linguista ha bisogno dell'informatico o dello statistico è valido anche il contrario: l'approccio quantitativo non esclude quello qualitativo.
L'ideale a nostro parere è, qualora la situazione lo consenta, di unire i due approcci di tipo quanti e qualitativo creando quello che viene definito approccio quanti qualitativo.
Usando un approccio di questo tipo la parte quantitativa dà validità alle supposizioni dello studioso che potrebbero essere tacciate di poca scientificità e allo stesso modo l'approccio qualitativo interpreta il dato quantitativo o prepara il dato quantitativo per essere interpretato, per poi collocarlo all'interno di un discorso più ampio.
L'unione fra questi due approcci risulta secondo noi vincente e permetterebbe così di superare anche le antiche diatribe fra metodo quantitativo e metodo qualitativo.
Tuzzi infatti dice:

> Tuttavia, la distinzione dei metodi di ricerca in qualitativi e quantitativi sembra in via di superamento e chi ancora crede si tratti di una vera e propria scelta alternativa rischia di risultare anacronistico. Nel caso dell'analisi del contenuto la contrapposizione "quantitativo" *versus* "qualitativo" è un problema, se non proprio falso, sicuramente mal posto, perché ogni approccio di tipo statistico deve operare mediante strumenti di tipo quantitativo. Non è però quantitativo l'oggetto di studio e per poter trattare statisticamente le informazioni si passa attraverso una forma di codifica. Nell'analisi testuale convivono contesti e significati di parole, di natura puramente qualitativa, con ranghi, frequenze e distribuzioni di probabilità, che sono invece quantitativi, nel rispetto della natura di entrambi.
>
> (Tuzzi, 2003, p. 28)

Dal punto di vista linguistico inoltre i metodi quantitativi non sono soltanto utili come strumenti per l'analisi testuale ma a nostro parere potrebbero anche aiutarci meglio a comprendere, attraverso il machine learning e il NLP, il linguaggio umano o in questo specifico caso la scrittura.
Quello che ci chiediamo dopo la nostra analisi è: perché gli *n*-grammi e l'entropia riescono a selezionare in maniera così precisa lo stile gramsciano?

Quanto della nostra scrittura (e per nostra scrittura intendiamo la scrittura di noi tutti e non soltanto di chi è scrittore di mestiere) è regolato da un set di regole predicibili (o comunque esistenti) e quanto si discosta da questa regolarità?

Perché di fatto gli *n*-grammi e l'entropia non hanno creato nessun falso positivo e perché ne hanno invece creati di negativi?

Ci rendiamo contro che questa non è forse la sede adatta per rispondere a queste domande e che certo noi in questo momento non possiamo avere le risposte, ma questo ci sembra uno dei motivi per cui riteniamo che questi metodi, combinati certamente con l'analisi qualitativa, rappresentino il futuro dell'attribuzione d'autore.

Esattamente come il linguista, anche questi metodi analizzano quella catena di informazioni potenzialmente infinta che è il linguaggio umano e cercano di darle una regolarità e un senso.

Ci auspichiamo quindi che sempre un numero maggiore di studiosi nell'ambito letterario si interessino ai metodi quantitativi e che lo stesso avvenga tra chi invece è più pratico dei metodi quantitativi e non di analisi testuale, in modo tale da creare un nuovo e interessante campo interdisciplinare.

BIBLIOGRAFIA

Bibliografia generale

Andorno, C. (2003). *Linguistica Testuale* . Roma: Carrocci.

Argamon S, Whitelaw C, Chase P, & Hota S. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology, 58 (6),* 802-822.

Baayen R., Van Halteren H. & Tweedie F. (1996). Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing 11 (3),* 121-131.

Basile C, Benedetto D, Degli Esposti M. & Caglioti E. (2010). L'attribuzione dei testi gramsciani: metodi e modelli matematici. *La matematica nella società e nella cultura, 3*, 235-269.

Brandwood L. & Cox D. R. (1959). On a Discriminating Problem Connected with the Works of Plato. *Journal of the Royal Statistical Society B, 21*, 195-200.

Burroghs, J. (1987). *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method.* Oxford: Clarendon Press.

Burroghs, J. (1987). Word Patterns and Story Shapes, The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing, 2,* 61-70.

Canter, D. (1992). An Evaluation of the 'Cusum' Stylistic Analysis of Confessions . *Expert Evidence, 1*, 93-99.

Cavalli, A. (2001). *Incontro con la sociologia.* Bologna: Il Mulino.

Cortelazzo, M. A. (2012). L'attribuzione d'autore delle traduzioni. In M. A. Cortelazzo, *I sentieri della lingua. Saggi sugli usi dell'italiano tra passato e presente.* (p. 87-92). Padova: Esedra Editrice.

Cortelazzo, M. A. (2013). Metodi quantitativi e qualitativi di analisi dei testi. *Contemporanea : rivista di studi sulla letteratura e sulla comunicazione* , 299-310.

Cortelazzo, M. A. (2016). Il linguaggio della politica . In *L'italiano, conoscere e utilizzare una lingua formidabile.* Roma : Gruppo Editoriale L'Espresso.

Cortelazzo M. A, Tuzzi A. & Nadalutti P. (2013). Improving Labbé's Intertextual Distance: Testing a Revised version on a Large Corpus of Italian Literature. *Journal of Quantitative Linguistics 20 (2),* 125-152 (DOI: 10.1080/09296174.2013.773138, ISSN: 0929-6174).

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

Dell'Anna, M. (2010). *Lingua italiana e lingua politica* . Roma : Carrocci editore .

Desideri, P. (1984). *Teoria e prassi del discorso politico. Strategie persuasive e percorsi comunicativi.* Roma: Bulzoni.

Eder, Maciej & Rybicki. (2011). Stylometry with R. *Proceedings of the 2011 Digital Humanities conference* (p. 308-311). Stanford: Stanford University.

Foster, D. (1989). *An Elegy by W.S.: A Study in Attribution.* Newark: University of Delaware press.

Frantzeskou G, Stamatatos E, Gritzalis S. & Katsikas, S. (2006). Effective identification of source code authors using byte-level information. *Proceedings of the 28th International Conference on Software Engineering*, (p. 893-896).

Hann P. D & Schils E. (1993). The Qsum Plot Exposed. *Proceedings of the 14th ICAME Conference.* Amsterdam: Rodopi.

Hirst G. & Feiguina O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and linguistics computing*.

Holmes, D. (1994). Autorship attribuition. *Computers and the Humanities, vol 2*, 87-106.

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, Volume 13, Issue 3, 1 Spetember*, 111-117.

Holmes D. I. & Forsyth R. (1995). The Federalist Revisited: New Directions in Autorship Attribuition. *Literary and Liguistic Computing, 10*, 111-127.

Joula , P. (2013). How a Computer Program Helped Show J.K. Rowling write A Cuckoo's Calling. *Scientific American*, https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/. (ultimo accesso 10/02/2018)

Joula , P. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities, Vol. 30, Supplement 1*, 100 - 113 .

Joula, P. (2006). Autorship attribuition. *Foundation and Trends in Information Retrival Vol 1, No.3 ,* 233-334.

Karlgren J & Eriksson G. (2007). Authors, genre, and linguistic convention. *Proceeding of the SIGIR Workshop on Plagiarism Analysis, Authorship Attribution, and Near-Duplicate Detection*, (p. 23-28).

Keselj V, Peng F, Cercone N, & Thomas C. (2003). N-gram-based author profiles for authorship attribution. *Proceedings of the Pacific Association for Computational Linguistics,* (p. 255-264).

Koppel M, Schler J. & Bonchek-Dokow E. (2007). Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research, 8,* 1261-1276.

Koppel M & Schler J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis,* (p. 69-72).

Koppel M & Schler J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis,* (p. 69-72).

Koppel M, Akiva N. & Dagan, I. (2006). Feature instability as a criterion for selecting potential style markers. *Journal of the American Society for Information Science and Technology 57 (11),* 1519-1525.

Laan, N. M. (1995). Stylometry and Method. The Case of Euripides. *Literary and Linguistic Computing, Volume 10, Issue 4, 1 November*, 271-278.

Labbè D. & Labbè C. (2001). Inter-Textual Distance and Authorship Attribution. Corneille and Molière. *Journal of Quantitative Linguistics, 8 (3)*, 213-23111.

Lana, M. (2010). Come scriveva Gramsci? Metodi matematici per riconoscere scritti gramsciani anonimi. *Informatica Umanistica, 3*, 31-56.

Lana, M. (2011). Individuare scritti gramsciani anonimi in un" corpus" giornalistico. Il ruolo dei metodi quantitativi. *"Studi storici: rivista trimestrale dell'Istituto Gramsci", 52 (4)* , 859-880.

Lutoslawski (A), W. (1897). On stylometry. *The Classical Review, vol 11, No 6, July*, 284-286.

Lutoslawsky (B), W. (1897). *The origin and growth of Plato's logic; with an account of Plato's style and of the chronology of his writings.* London, New York: Longmans, Green and co.

Manning C. D, Raghavan P & Schütze H. (2008). *Introduction to information retrival.* Cambridge: Cambridge University Press.

Matthews, R. (1994). A Bard by Any Other Name. *New Scientist, 1909*.

Mendenhall, T. C. (1887). The Characteristic Curves of composition. *Science, 11*, 237-249.

Mengaldo, P. V. (2007). *Prima lezione di stilistica* . Laterza .

Merriam, T. (1979-1980). What Shakespeare Wrote in Henry VIII, Part 1 e Part 2. *The Bard 2*, 81-94, 111-118.

Merriam, T. (1993). Neural Computation in Stilometry. I. An application to the works of Shakespeare and Fletcher,. *Literary and Linguistic Computing, 8*, 203-209.

Miller, G. (1996). *The science of words.* New York: Freeman.

Morgan, A. d. (1851/1882). "Letter to Rev. Heald 18/08/1851,". In S. E. Morgan, *Memoirs of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with Selections from his Letters.* London: Longman's Green and Co.

Mortara Garavelli, B. (2003). *Manuale di Retorica* . Milano: Bompiani.

Morton, A. Q. (1978). *Literary Detection* . New York : Scribners.

Morton, A. Q. (1986). Once A Test of Authorship Based on Words Which Are Not Repeated in the Sample. *Literary and Linguistic Computing*, 1-8.

Morton, A. Q. (1991). *Proper Words in Proper Places.* Computing Science Department: University of Glasgow.

Mosteller F. & Wallace D. L. (1964). *Applied Baesyan and Classical Inference. The Case of Federalist Papers.* Reading (MA): Addison-Wesley.

Rybicki, J. (2012). The great mystery of the almost invisible translator. In M. P. Oakes, & J. Meng, *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research* (p. 231-247). John Benjamins Publishing.

Sanderson C. & Guenter S. (2006). Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation. *Proceedings of the International Conference on Empirical Methods in Natural Language Engineering*, (p. 482 - 491).

Segel, H. B. (1991). Book review. *The Polish review 36 (4)*, 486-495.

Simpson, E. H. (1949). Measurement of Diversity. *Nature, 163*, 68.

Smith, M. W. (1987). Hapax Legomena in Prescribed Positions: An Investigation of Recent Proposals to Resolve Problems of Authorship. . *Literary and Linguistic Computing, 2*, 145-152.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 538 - 556.

Stamatatos, E. (2007). Author identification using imbalanced and limited training texts. *Proceedings of the 4th International Workshop on Text-based Information Retrieval*, (p. 237-241).

Statamatos E, Fatokakis N & Kokkinakis G. (2001). Computer-based authorship attribution without lexical measures. *Computers and Humanities 35 (2)*, 193-214.

Tambouratzis G, Markantonatou S, Hairetakis N, Vassiliou M, Carayannis G & Tambouratzis N. (2004). Discriminating the registers and styles in the Modern Greek language – Part 2 Extending the feature vector to optimize author discrimination. *Literary and Linguistic Computing 19 (2)*, 221-242.

Thisted R & Efron B (1987). Did Shakespeare Write a Newly Discovered Poem? *Biometrika, 74*, 445-455.

Tuzzi, A. (2003). *L'analisi del contenuto.* Roma: Carrocci.

Valenza, R. J. (1990). Are Thisted-Efron Authorship Tests Valid? *Computers and Humanities, 25*, 27- 46.

Valenza R. J & Elliot W. (1996). And then there were none: Winnowing the Shakespeare claimants. *Computers and Humanities, vol 30*, 191- 245.

Van Halteren, H. (2007). Author verification by linguistic profiling: An exploration of the parameter space. *ACM Transactions on Speech and Language Processing, 4*, 1-17.

Villasenor-Pineda P, Montez-y-Gòmez M, Rosso P, & Coyotl-Morales R. (2006). Autorship attribuition using word sequences. *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition* (p. 844-853). Springer.

Wake, W. C. (1957). Sentence-length Distnbutions of Greek Authors. *Journal of the Royal Statistical Society*, 331-46.

Williams, C. B. (1940). A Note on the Statistical Analysis of Sentence-length as a Criterion of Literary Style. *Biometrika 31*, 356-61.

Wouters H, & Ville De Goyet C. (1990). *Molière ou l'auteur imaginaire ?* Bruxelles: Eds Complexe.

Yule, G. U. (1938). On sentence-length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship. *Biometrika, 30*, 363-390.

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

Zipf, G. (1932). *Selected Studies of the Principle of Relative Frquency in Language.* Cambridge,MA: Harvard University.

Testi gramsciani e non gramsciani usati nell'analisi

Testi del primo test

Testi di Gramsci

Pubblicati su «Il Grido del Popolo»

1. *Neutralità attiva e operante*, 31 ottobre 1914.
2. *Dopo il congresso socialista spagnuolo,* 13 novembre 1915.
3. *La luce che si è spenta,* 20 novembre 1915.
4. *L'idea Nazionale*, 27 novembre 1915.
5. *La festuca*, 11 dicembre 1915.
6. *Il Sillabo ed Hegel,* 15 gennaio 1916.
7. *Pietro Gavosto,* 22 gennaio 1916.
8. *Socialismo e cultura*, 29 gennaio 1916.
9. *Armenia,* 11 marzo 1916.
10. *Il Mezzogiorno e la guerra*, 1° aprile 1916.
11. *La paura del "Dumping",* 13 maggio 1916.
12. *Il Dumping germanico,* 20 maggio 1916.
13. *L'eroe*, 17 giugno 1916.
14. *Beneficenza,* 12 agosto 1916.
15. *Contro il feudalismo economico*, 12 agosto 1916.
16. *Monssù Botegari,* 13 gennaio 1917.
17. *Carattere,* 3 marzo 1917.
18. *Note sulla rivoluzione russ*a, 29 aprile 1917.
19. *Il perfido straniero*, 9 giugno 1917.
20. *La scuola di Stenterello,* 15 giugno 1917.
21. *Abbruciamenti*, 21 luglio 1917.
22. *I massimalisti russi*, 28 luglio 1917.
23. *L'orologiaio,* 18 agosto 1917.
24. *Letture*, 24 Novembre 1917.
25. *Intransigenza-intolleranza, intolleranza-intransigenza*, 8 dicembre 1917.
26. *La rivoluzione contro il "Capitale",* 5 gennaio 1918.
27. *La critica critica*, 12 gennaio 1918.

28. *La lega delle nazioni,* 19 gennaio 1918.
29. *Achille Loria,* 19 gennaio 1918.
30. *La funzione sociale del Partito socialista nazionalista,* 26 gennaio 1918.
31. *La famiglia,* 3 marzo 1918.
32. *Il nostro Marx,* 4 maggio 1918.
33. *Libero pensiero e pensiero libero,* 15 giugno 1918.
34. *L'utopia russa,* 27 luglio 1918.

Pubblicati su «Avanti!»

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

Pubblicati su «La città futura»

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

Testi non gramsciani

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 volontà,* 24 agosto 1918.
14. Gino Castagno, *I pretesi errori confederali*, «Il Grido del Popolo». 24 giugno 1916.
15. C.D, *Per l'ufficio di assistenza popolare, «*L'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*, «L'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*, «L'Avanti!», 3 settembre 1915.
26. Alfonso Leonetti, *Il centenario dalla nascita di Claudio 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,* «L'Avanti!» 8 aprile 1916.
29. Mario Santarosa, *L'eccellenza smentisce,* «Il Grido del Popolo», 1 luglio 1916.
30. Giacinto Menotti Serrati, *Scampoli. Il Primo maggio di Maria*, «L'Avanti!» 8 aprile 1916.
31. Giacinto Menotti Serrati, *Discutendo tra relativisti ed intransigenti,* «L'Avanti!» 9 maggio 1918.
32. Giacinto Menotti Serrati, *Salutatemi la disciplina,* «L'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», 1 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 novembre 1917.
42. Palmiro Togliatti, *La disfatta di A. Lanzillo*, «L'ordine nuovo», 1 maggio 1919.
43. Palmiro Togliatti, *Parole oneste sulla Russia*, «L'ordine nuovo», 1 maggio 1919.
44. Palmiro Togliatti, *Guerra e fede di Giovanni Gentile*, «L'ordine nuovo», 1 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.

Testi del secondo test

1. Gramsci, *La rievocazione di Gelindo*, «Il Grido del Popolo», 25 dicembre1915.
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
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 più chiaramente*, «Il Grido del Popolo», 7 novembre 1914.
11. O.P. [Ottavio Pastore], *Il Papa al congresso della pace*, «Il Grido del Popolo», 15 aprile 1916.
12. Gramsci, *Una verità che sembra un paradosso*, «Avanti!», 3 aprile 1917.
13. G.M.S. [Giacinto Menotti Serrati], *Il più gran terremoto*, «Il Grido del Popolo»,12 agosto 1916.
14. Gramsci, *Con mani di vetro ...*, «Il Grido del Popolo», 13 aprile 1918.
15. Alfonso Leonetti, *Evoluzione e rivoluzione*, «Il Grido del Popolo», 3 agosto 1918.
16. Gramsci, *La lingua unica e l'esperanto*, «Il Grido del Popolo», 16 febbraio 1918.
17. Decio Pettoello, *La dottrina di Norman Angell*, «Il Grido del Popolo», 10 agosto 1918.
18. Gramsci, *Repubblica e proletariato in Francia*, «Il Grido del Popolo», 20 aprile 1918.
19. Zino Zini, *Marx nel pensiero di un cattolico*, «Il Grido del Popolo», 31 agosto 1918.
20. Gramsci, *Due inviti alla meditazione*, «La Città futura», 11 febbraio 1917.
21. A.V. [Andrea Viglongo], *La Costituzione parlamentare inglese*, «Il Grido del Popolo», 5 ottobre 1918.
22. Pietro Gavosto, *Le opinioni dei compagni. Guerra, patria e proletariato*, «Il Grido del Popolo», 9 gennaio 1915.
23. A.T. [Angelo Tasca], *Noterelle di guerra*, «Il Grido del Popolo», 16 gennaio 1915.
24. Gramsci, *Il privilegio dell'ignoranza*, «Il Grido del Popolo», 13 ottobre 1917.
25. Gramsci, *I monaci di Pascal*, «Avanti!», 26 febbraio 1917.
26. Gino [Gino Castagno], *Cinismo*, «Il Grido del Popolo», 20 febbraio 1915.

27. Gramsci, *Disciplina e libertà*, «La Città futura», 11 febbraio 1917.
28. Leo Galetto, *Il proletariato deve servire da «materia anatomica»*, «Il Grido del Popolo», 20 marzo 1915.
29. Gramsci, *Modello e realtà*, «La Città futura», 11 febbraio 1917.
30. Cincali, *Luci ed ombre*, «Il Grido del Popolo», 23 ottobre 1915.
31. Corso Bovio, *ll problema del Mezzogiorno*, «Avanti!», 27luglio 1917.
32. Gramsci, *La Giustizia*, «Il Grido del Popolo», 13 ottobre 1917.
33. Omero Concetto, *Diagnosi interessata*, «Avanti!», 10 agosto 1917.
34. Gramsci, *Letteratura italica: La prosa*, «Avanti!», 17 aprile 1917.
35. Egidio Gennari, *Nazionalisti od internazionalisti?* «Avanti!», 27 agosto 1917.
36. Gramsci, *Rispondiamo a Crispolti*, «Avanti!», 19 giugno 1917.
37. Francesco Ciccotti, *Il reazionario democratico*, «Avanti!», 2 settembre 1917.
38. O.B., *Problemi presenti e futuri*, «Avanti!», 12 settembre 1917.
39. Gramsci, *Spezzatino d'asino e contorno*, «Il Grido del Popolo», 29 aprile 1917.
40. Gramsci, *Analogie e metafore*, «Il Grido del Popolo», 15 settembre 1917.

Software utilizzati per l'analisi quantitativa

TaLTaC^2 (Trattamento automatico Lessicale e Testuale per l'analisi del Contenuto di un Corpus) (www.taltac.it)

IRaMuTeQ (Interfaccia di R per l'Analisi Multidimensionale di Testi e di Questionari) (www.iramuteq.org)