Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea

Michele A. Cortelazzo[1], Paolo Nadalutti[2], Arjuna Tuzzi[3]

[1]Università di Padova – cortmic@unipd.it
[2]Università di Padova – paolo.nadalutti@unipd.it
[3]Università di Padova – arjuna.tuzzi@unipd.it

Abstract

The main aim of this study is testing the performance of Labbe's intertextual distance in the case of a corpus of novels of the Italian literature; moreover an iterative revised version of intertextual distance is introduced. The revised procedure is based on repeated measures of the intertextual distance between pairs of text-chunks of equal size. It is intended to counterbalance the size factor and assess the similarity of texts of very different sizes. In order to test if and to what extent intertextual distance is suitable to assess authorship attribution in the case of Italian novels, we took into account the rankings obtained with reference to each novel ordering the others by means of the values of intertextual distances. Opportunities and limits of the available procedures are discussed.

Riassunto

L'obiettivo principale del contributo è valutare il funzionamento della distanza intertestuale di Labbé nel caso di un corpus di opere della letteratura italiana e presentarne una nuova versione iterativa. La nuova procedura si basa su misure ripetute della distanza intertestuale per coppie di porzioni di testo della stessa dimensione, mira a compensare l'effetto delle dimensioni e a determinare la similarità tra testi di dimensioni molto diverse. Per valutare se, e in che misura, la distanza intertestuale può essere impiegata per l'attribuzione d'autore nel caso di opere della narrativa italiana, si considerano le graduatorie che, data un'opera, si ottengono ordinando le altre in base al valore della distanza intertestuale. Vengono, infine, discussi i limiti e i vantaggi delle diverse procedure.

Keywords: attribuzione d'autore, corpora, distanza intertestuale, letteratura italiana, graduatorie

1. Introduzione

Non esiste in letteratura una definizione univoca del concetto di similarità (e dissimilarità) tra testi e il problema di ridurre un insieme di caratteristiche possedute da una coppia di testi in un'unica misura quantitativa in grado di rappresentarne la distanza cambia a seconda dell'oggetto di studio, dell'ambito applicativo e della disciplina di riferimento.

Il concetto di similarità è di grande importanza in tutte le applicazioni di *text clustering* in quanto la maggior parte delle procedure per l'analisi dei gruppi si basano su una matrice di vicinanza (o distanza) che mette in relazione tutte le coppie di testi. In letteratura le proposte

sono numerose e difficilmente sintetizzabili: sono stati elaborati diversi approcci alle misure della distanza, che variano a seconda delle diverse scuole di pensiero che le hanno generate, del tipo di dati su cui si basano, delle prestazioni che garantiscono, ecc. Tuttavia, è interessante osservare che, a fronte di una grande disponibilità di misure (un migliaio secondo Rudmann, 1998) e di procedure di raggruppamento (Everitt, 1980), gli effetti e l'efficacia delle diverse scelte rispetto alla struttura dei *cluster* ottenuta e rispetto agli scopi perseguiti non sono ancora noti e assodati (Berry, 2004; Huang, 2008).

In ambito testuale le misure di similarità e le procedure di *text clustering* talvolta si prefiggono di riconoscere l'autorialità delle opere (Love, 2002). È abbastanza difficile immaginare una misura quantitativa in grado di gestire le numerose dimensioni che possono determinare il grado di similarità tra due testi (alcune anche latenti e non osservabili) e, tra queste, distinguere un presunto "stile inconfondibile" dell'autore, tuttavia la questione resta interessante, molto dibattuta in letteratura (Stamatatos, 2009) e discussa dal punto di vista di diverse discipline (qualche esempio: Lockers e Witten, 2010; Bagavandas e Manimannan, 2008, Koppel *et al.*, 2008).

Tralasciando sia la tradizione degli studi di stilistica, che si avvale di criteri di tipo qualitativo, sia gli studi che si basano su informazioni di tipo morfo-sintattico e semantico, in questo lavoro[1] abbiamo scelto di operare sul piano lessicale nell'ambito dei *bag-of-words approaches*, cioè con le forme presenti nel corpus e le loro frequenze. In particolare, abbiamo preso in considerazione alcune proposte di calcolo della distanza intertestuale avanzate da Labbé (Labbé e Labbé, 2001, 2003; Labbé, 2007; Labbé, 2010) e ne abbiamo ricavato una nuova versione iterativa basata su misure ripetute della distanza intertestuale per coppie di porzioni di testo di uguali dimensioni. Gli scopi perseguiti sono:

1. valutare il funzionamento della distanza intertestuale su un corpus in lingua italiana;
2. limitare gli effetti delle dimensioni dei testi per poter arrivare a una misura affidabile della distanza anche in presenza di testi di dimensioni significativamente diverse;
3. appurare l'utilizzabilità della distanza intertestuale ai fini dell'attribuzione d'autore.

2. Il corpus

Il corpus è costituito da 92 opere di narrativa pubblicate tra il 1941 e il 2009 che include 7,8 milioni di occorrenze (N) e poco meno di 150mila forme diverse (V). Il Type-Token Ratio V/N è pari a 1,9%, corrispondente a una frequenza media di 52 occorrenze per forma. La percentuale di *hapax* nel vocabolario è del 38,7%.

Le 92 opere sono di 33 scrittori italiani contemporanei, ciascuno dei quali contribuisce con almeno due opere: Ammaniti (4), Arbasino (3), Baricco (3), Berto (2), Bevilacqua (2), Bufalino (2), Buticchi (5), Calvino (8), Chiara (2), Eco (2), Faletti (4), Fenoglio (2), Levi C. (2), Levi P. (4), Magris (2), Maraini (3), Mazzantini (2), Morante (3), Moravia (4), Ortese (2), Parise (2), Pasolini (2), Pavese (2), Piovene (2), Pratolini (2), Rigoni Stern (2), Romano (2), Sciascia (3), Tabucchi (3), Tamaro (3), Veronesi (2), Vittorini (2), Volponi (4). In appendice è riportato l'elenco completo delle opere con autore, titolo e anno di pubblicazione.

1 Questo lavoro fa parte delle attività condotte nell'ambito del GIAT - Gruppo Interdisciplinare di Analisi Testuale (http://www.giat.org/).

Per la valutazione dell'omogeneità (Popescu *et al.*, 2009) si deve tenere conto che nella costruzione di questo corpus è stato possibile ridurre alcune dimensioni di variazione. Per esempio non è rilevante la lingua, perché si tratta di opere scritte tutte originariamente in italiano. Il genere testuale identificabile nella narrativa (romanzi e racconti) riduce la variazione diafasica e diamesica e, dato che si tratta di opere prodotte da scrittori di professione per un pubblico adulto di lettori non specialisti, anche quella diastratica. L'arco di tempo, inoltre, è tale da rendere irrilevante la variabile diacronica. Permangono, viceversa, la variazione diatopica, dovuta alla provenienza regionale dell'autore, e la varietà dei temi affrontati nello sviluppo narrativo.

Il corpus è stato elaborato attraverso il software *Taltac2* (Bolasco, 2010) e i calcoli statistici sono stati svolti con il software *R* (R Development Core Team, 2011). In tutte le analisi sono state prese in considerazione le forme grafiche ottenute attraverso una fase di *parsing* e una fase di normalizzazione per la riduzione delle maiuscole.

3. La distanza intertestuale

Nata sulla scia dei lavori di Muller (Muller, 1968, 1977) e ispirata al concetto di connessione lessicale (Brunet, 1988; Muller e Brunet, 1988; Cortelazzo e Tuzzi, 2008), la distanza intertestuale nella versione[2] proposta da Labbé rientra nell'ambito dei *bag-of-words approaches* per la misura della dissimilarità. Si basa su un insieme di indicatori semplici contenuti nel vocabolario di frequenza di due testi (presenza, assenza, frequenza di parole) e ne deriva un indicatore composito che ne rappresenta la distanza:

$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f^*_{i,B} \right|}{2N_A}$$

dove A e B rappresentano la coppia di testi di dimensioni N_A e N_B con $N_A \leq N_B$ e $V_{A \cup B}$ è il vocabolario dell'unione dei due testi, cioè l'insieme di tutte le parole presenti in almeno uno dei due con le rispettive frequenze $f_{i,A}$ e $f_{i,B}$. Al fine di rendere i due testi confrontabili, la frequenza osservata $f_{i,B}$ della forma *i*–esima nel testo più ampio B viene ridotta in ragione della dimensione del testo più breve A attraverso la stima

$$f^*_{i,B} = f_{i,B} \frac{N_A}{N_B}$$

ottenendo $\sum_{i \in V_A} f_{i,A} = \sum_{i \in V_B} f^*_{i,B} = N_A$. La misura proposta possiede i requisiti per essere considerata una distanza in senso statistico (non-negatività, identità degli indiscernibili, simmetria, disuguaglianza triangolare).

3.1. Ulteriori aggiustamenti

Nonostante sia prevista una correzione, la distanza intertestuale nella sua forma base resta dipendente dalle dimensioni in gioco come, peraltro, buona parte delle misure lessicometriche

2 Per una descrizione si consiglia http://images.math.cnrs.fr/La-classification-des-textes.html

disponibili in letteratura (Baayen, 2001; Strauss *et al.*, 2008; Tweedie e Baayen, 1998). Quando vengono confrontati testi di dimensioni significativamente diverse $N_A < N_B$, la stima $f^*_{i,B}$ produce una lunga coda di basse frequenze in *B* che squilibrano la distanza. Per migliorare la *performance* della distanza intertestuale, in letteratura sono disponibili diverse soluzioni di aggiustamento della misura. Una delle soluzioni proposte consiste nell'escludere dal calcolo tutte le forme del vocabolario di *B* con $f^*_{i,B} < 1$, che risponde al criterio di escludere tutte quelle parole che, se *B* avesse la dimensione di *A*, non arriverebbero nemmeno a essere degli *hapax*. Lo stesso aggiustamento risulta controproducente se applicato a due testi di dimensioni simili.

La distanza intertestuale può essere calcolata sia sulle forme che sui lemmi e anche su un sottoinsieme di forme o lemmi. Si possono, infatti, immaginare procedure di calcolo della distanza basate solo sulle forme grammaticali (articoli, congiunzioni, preposizioni, pronomi), oppure su una categoria decisa a priori della *part-of-speech* (per es. sui verbi lemmatizzati o sui sostantivi ad alta frequenza), oppure ancora su un insieme di forme di contenuto scelte in maniera arbitraria.

3.2. Una versione iterativa della distanza intertestuale basata su campionamenti

Per limitare l'effetto delle dimensioni sul valore della distanza intertestuale, in questo contributo ne viene proposta una versione iterativa che, invece di rendere i due testi confrontabili correggendo le frequenze, usa misure ripetute della distanza su campioni di porzioni di testo della stessa dimensione. Lo scopo è quello di riuscire a confrontare due testi di dimensioni significativamente diverse, dove per significativamente diverse si intende il caso in cui uno dei due testi ha dimensioni inferiori all'altro di un ordine di grandezza (per un esempio cfr. Tuzzi, 2011).

Dati *p* testi si parte dalla matrice quadrata $(p \times p)$ che ha come elemento generico d_{ij} ($i, j = 1 \ldots p$) la distanza calcolata per la coppia generica di testi (i, j). In questa matrice si considerano solo le $p(p-1)/2$ distanze per coppie non identiche ($d_{ii} = 0$) e non ridondanti ($\forall i, j \;\; d_{ij} = d_j$). Fissata la dimensione *n* delle porzioni di testo e il numero *m* di iterazioni (numero di campioni), per ogni iterazione $k = 1 \ldots m$ viene estratto un campione di *p* porzioni di dimensione *n*, uno per ciascun testo. A partire da queste *p* porzioni viene calcolata la matrice di distanze che contiene la distanza intertestuale per tutte le coppie di porzioni all'iterazione *k*. Siccome le porzioni hanno la stessa dimensione *n*, nel calcolo della distanza intertestuale il passaggio alla frequenza stimata $f^*_{i,B}$ è inutile e la formula per la coppia costituita dalle porzioni di testo *A* e *B* si semplifica in:

$$d(A,B) = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f_{i,B} \right|}{N_A + N_B} = \frac{\sum_{i \in V_{A \cup B}} \left| f_{i,A} - f_{i,B} \right|}{2n}.$$

Al termine delle *m* iterazioni si ottiene una matrice tridimensionale di dimensioni $(p \times p \times m)$ con elemento generico d_{ijk} costituita da $mp(p-1)/2$ distanze per coppie di testi non identiche e non ridondanti.

Le distanze osservate sono ipoteticamente distribuite secondo una distribuzione normale $N(\mu,\sigma)$ che sappiamo essere dipendente dalla dimensione *n* delle porzioni ma, grazie alla procedura, potenzialmente indipendente dalla dimensione dei testi. A partire dalla matrice tridimensionale viene calcolata una matrice bidimensionale dove il generico elemento è calcolato come media campionaria dei valori sulle *m* iterazioni:

$$\hat{d}_{ij} = \frac{\sum_{k=1}^{m} d_{ijk}}{m}$$

e viene assunto come stima della distanza per la coppia di testi *i*, *j* condizionata alla dimensione *n* delle porzioni di testo del campione.

4. Risultati

Per l'analisi del vocabolario del corpus è stata innanzitutto presa in considerazione la distanza nella versione base, calcolata su tutte le forme. Le distanze sono state ricalcolate in altre tre versioni applicando di volta in volta uno dei seguenti aggiustamenti:

1. calcolo basato sulle sole forme *i* con $f^{*}_{i,B} \geq 1$ (distanza con taglio a soglia 1);

2. calcolo basato sulle forme grammaticali (430 forme costituite da articoli, congiunzioni, preposizioni e pronomi che assicurano una copertura del vocabolario dello 0,3% e una copertura del corpus del 45,0%);

3. calcolo basato su un insieme di forme di contenuto (657 forme costituite da sostantivi non ambigui appartenenti alle alte e medie frequenze, che assicurano una copertura del vocabolario dello 0,4% e una copertura del corpus del 9,6%).

Per quanto riguarda la versione iterativa, la procedura è stata applicata quattro volte estraendo *m* = 200 campioni di 92 porzioni di testo di dimensione *n* pari a 5.000, 10.000, 15.000 e 20.000 occorrenze. Il numero di coppie su cui è stata calcolata la distanza intertestuale è elevato: per ogni iterazione *k* si lavora con $(92 \times 91)/2 = 4.186$ coppie e, con l'estrazione dei 200 campioni, si arriva a 837.200 osservazioni della distanza intertestuale. Nel caso delle porzioni da 20.000 occorrenze il numero di testi si riduce di una unità in quanto è stata esclusa dal calcolo un'opera di Antonio Tabucchi di lunghezza prossima a 20mila occorrenze. La stessa opera è stata esclusa dal computo quando è stato necessario confrontare i risultati ottenuti con diverse procedure di calcolo.

La figura 1 mostra la distribuzione della distanza intertestuale nella versione base (linea continua) e nelle tre versioni con aggiustamento, cioè calcolata con il criterio di taglio a soglia 1, sulle forme grammaticali o sull'insieme dei 657 sostantivi ad alta frequenza. La versione base e quella con il taglio a soglia 1 risultano abbastanza simili; la versione basata sui grammaticali si attesta su valori mediamente più bassi e mostra una variabilità inferiore; quella basata sui sostantivi valori mediamente più elevati e una variabilità maggiore.

La figura 2 mostra la distribuzione delle distanze intertestuali nel caso della versione base (linea continua) e la distribuzione delle distanze intertestuali medie ottenute con la procedura iterativa basata sul campionamento. La distanza intertestuale ha un valore che decresce al crescere delle dimensioni delle porzioni di testo usate nella fase di campionamento. Inoltre, nel caso in cui

venga usata la formula base con distanze calcolate su coppie di testi di dimensioni diverse, la distribuzione mostra maggiore variabilità e minore regolarità.

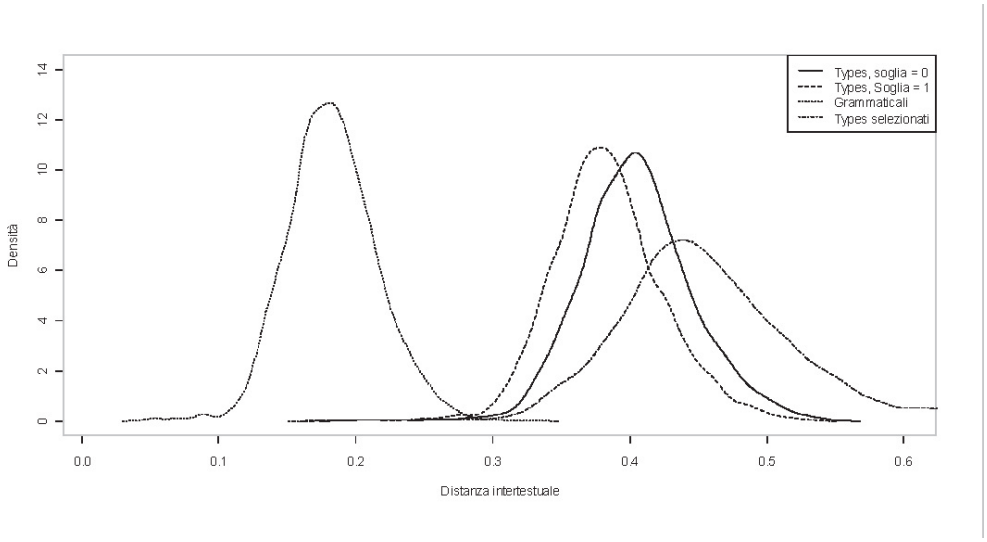

Figura 1 : Distribuzione della distanza intertestuale nella versione base (soglia = nea continua) e nelle versioni con aggiustamenti.

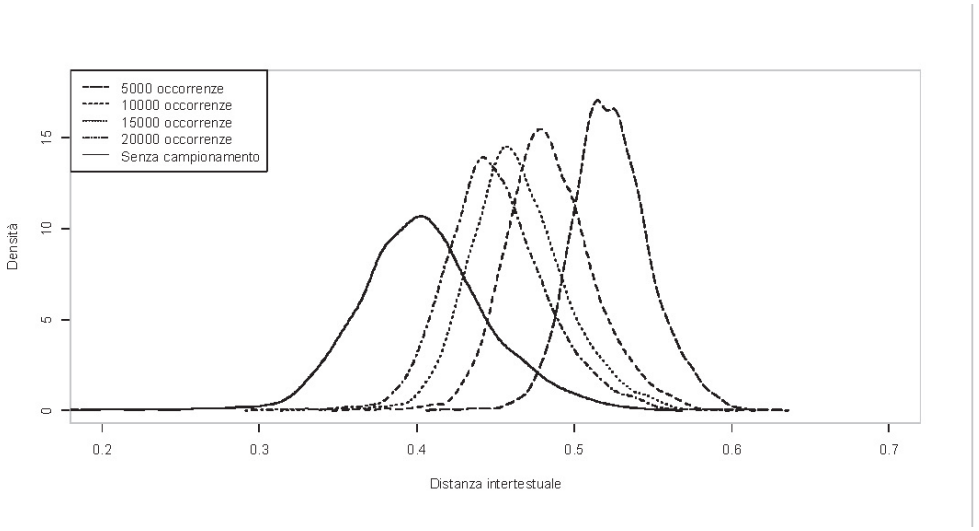

Figura 2 : Distribuzione della distanza intertestuale nella versione base (senza campionamento, linea continua) e nelle versioni con campionamento.

4.1. Valutare le graduatorie create per ogni testo dalle diverse procedure di calcolo

Se la distanza intertestuale è sensibile alla mano dell'autore, ci si può attendere che opere scritte dallo stesso autore siano tra loro vicine. A partire dai dati contenuti nelle matrici di distanza ottenute con le diverse procedure di calcolo, per ogni opera si possono mettere in ordine secondo il valore della distanza intertestuale tutte le altre, a partire da quella più vicina fino a quella più lontana. L'assetto delle 92 graduatorie che si ottengono per ciascuna procedura di calcolo adottata, offre un'indicazione sulla capacità della distanza intertestuale di riconoscere come "più vicine" le opere scritte dallo stesso autore.

Data un'opera e la relativa graduatoria costruita con le distanze intertestuali di quell'opera dalle restanti 91, si può osservare il rango occupato dalle opere scritte dallo stesso autore. Si supponga di avere all'interno del corpus $p+1$ opere scritte da un dato autore. Prendendo in considerazione la graduatoria della prima, in una situazione ottimale la distanza intertestuale dovrebbe collocare le altre *p* nelle prime *p* posizioni. Se tutte le opere dell'autore dessero questo risultato, tutti i $p(p+1)$ confronti sarebbero associati a ranghi *k* tra 1 e *p*. Se, viceversa, i ranghi si allontanano dalle prime *p* posizioni, significa che la distanza non è in grado di cogliere correttamente l'effetto attribuibile all'autore.

Un esempio può essere utile per spiegare questo modo di osservare l'efficacia della misura adottata: consideriamo il caso dello scrittore Niccolò Ammaniti, che contribuisce al nostro corpus con quattro opere ($p+1=4$). Nella tabella 1 sono riportate le prime 10 posizioni delle quattro graduatorie basate sui valori della distanza intertestuale calcolate con la procedura iterativa e porzioni di 10.000 occorrenze. Se la misura funzionasse in maniera ottimale, le quattro graduatorie vedrebbero sempre la presenza di tre testi di Ammaniti nelle prime tre posizioni ($k \leq 3$) per un totale di $p(p+1)=12$ collocazioni corrette. Le graduatorie ottenute attraverso la versione iterativa della distanza intertestuale portano a un risultato abbastanza vicino a quello ottimale per Ammaniti (tab. 2): nove collocazioni su 12 sono corrette (nelle prime tre posizioni) e le restanti tre sono nelle immediate vicinanze.

	Ammaniti							
k	*1996*		*1999*		*2001*		*2006*	
1	**Ammaniti 2006**	0,406	**Ammaniti 2006**	0,388	Tamaro 1991	0,418	**Ammaniti 1999**	0,388
2	**Ammaniti 1999**	0,408	**Ammaniti 1996**	0,408	Maraini 1972	0,430	**Ammaniti 1996**	0,406
3	**Ammaniti 2001**	0,447	**Ammaniti 2001**	0,431	**Ammaniti 1999**	0,431	Faletti 2006	0,426
4	Fenoglio 1963	0,448	Baricco 1999	0,441	**Ammaniti 2006**	0,439	Faletti 2009	0,428
5	Faletti 2006	0,448	Faletti 2006	0,441	Levi P. 1978	0,440	Faletti 2004	0,431
6	Faletti 2009	0,450	Fenoglio 1963	0,442	**Ammaniti 1996**	0,447	Faletti 2002	0,437
7	Baricco 1999	0,450	Faletti 2009	0,444	Mazzantini 2002	0,450	**Ammaniti 2001**	0,439
8	Faletti 2004	0,451	Tamaro 1991	0,445	Fenoglio 1954	0,453	Fenoglio 1963	0,439
9	Faletti 2002	0,453	Baricco 2005	0,447	Maraini 1999	0,458	Baricco 1999	0,442
10	Mazzantini 2002	0,458	Fenoglio 1954	0,449	Pavese 1950	0,463	Parise 1982	0,447

Tabella 1 : Graduatorie delle quattro opere di Ammaniti basate sulla versione iterativa della distanza intertestuale con campioni di porzioni da 10.000 occorrenze. Valori della distanza intertestuale e opere collocate nelle prime 10 posizioni.

	Ammaniti				
k	1996	1999	2001	2006	Tot. confronti
$k \leq 3$	3	3	1	2	9
$4 \leq k \leq 6$	0	0	2	0	2
$7 \leq k \leq 9$	0	0	0	1	1
$k > 9$	0	0	0	0	0
Tot. confronti	3	3	3	3	12

Tabella 2 : Sintesi delle collocazioni delle opere di Ammaniti nelle sue quattro graduatorie. Numero di confronti per fascia di rango (k).

	versioni con aggiustamento			
	base	taglio soglia 1	grammaticali	sostantivi alta freq.
$k \leq p$	125	120	131	107
$p < k \leq 2p$	15	21	15	13
$2p < k \leq 3p$	13	7	12	13
$3p < k \leq 4p$	6	7	8	5
$4p < k \leq 5p$	11	4	9	7
$k > 5p$	42	49	37	67
Totale	212	212	212	212
Purezza	58,96%	56,60%	61,79%	50,47%
Impurità	19,87%	23,11%	17,45%	31,60%
	versione iterativa con campionamento			
	5.000 occ.	10.000 occ.	15.000 occ.	20.000 occ.
$k \leq p$	112	122	124	125
$p < k \leq 2p$	24	23	27	26
$2p < k \leq 3p$	15	13	9	9
$3p < k \leq 4p$	11	5	7	6
$4p < k \leq 5p$	10	10	8	6
$k > 5p$	40	39	37	40
Totale	212	212	212	212
Purezza	52,83%	57,55%	58,49%	58,96%
Impurità	18,87%	18,40%	17,45%	18,86%

Tabella 3 : Sintesi delle collocazioni nelle graduatorie delle opere scritte da uno stesso autore. Numero di confronti per fascia di rango.

La tabella 3 riporta i dati complessivi delle collocazioni per tutte le opere e tutti gli autori del corpus distinti per il caso di applicazione della distanza intertestuale nella sua forma base (senza campionamento e senza aggiustamenti), per i tre casi di calcolo con aggiustamento (distanza con taglio a soglia 1, grammaticali, sostantivi ad alta frequenza) e per le quattro diverse parametrizzazioni della procedura di calcolo della distanza intertestuale con campionamento (porzioni di testo da 5.000, 10.000, 15.000 e 20.000 occorrenze).

Le ultime due righe della tabella 3 si possono leggere come indici sintetici per una valutazione complessiva della bontà dei *ranking* ottenuti. L'indice di purezza è la percentuale di confronti che hanno dato la giusta collocazione (cioè p opere nelle prime p posizioni rispetto alle $p+1$ opere scritte dallo stesso autore), l'indice di impurità è la percentuale di confronti che hanno dato una collocazione "molto lontana" da quella giusta (in questo caso si è scelto di considerare molto lontane le opere collocate oltre $5p$ posizioni).

Attraverso i valori della tabella 3 e i grafici con le cumulate (figg. 3 e 4) si osserva come la versione iterativa della distanza intertestuale, con l'esclusione dei campioni di dimensioni troppo limitate (5.000 occorrenze), raggiunga risultati migliori rispetto alla versione base, avvicinando le opere scritte dallo stesso autore (fig. 4). Il risultato che si ottiene con altre modalità di aggiustamento della distanza intertestuale mostra come l'uso delle sole forme grammaticali migliori i risultati ottenibili con la distanza nella versione base mentre sia l'uso di un insieme limitato di forme di

contenuto (nel nostro caso i sostantivi ad alta frequenza) sia il taglio sulle basse frequenze non riescano a garantire buone *performance* (fig.3).

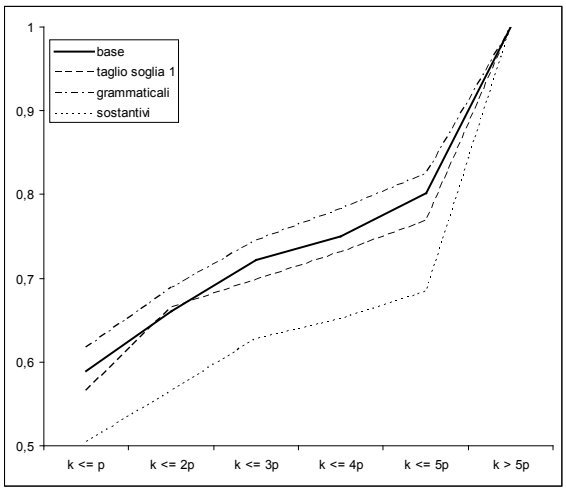

Figura 3 : Cumulata delle collocazioni per la distanza intertestuale base (linea continua) e per le tre versioni con aggiustamento.

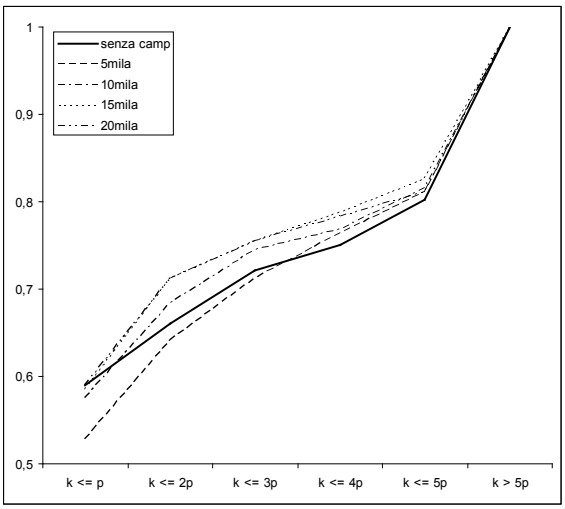

Figura 4 : Cumulata delle collocazioni per la distanza intertestuale base (linea continua) e per le quattro versioni con campionamento.

5. Conclusioni

Alla luce della sperimentazione svolta su questo corpus di 92 opere della letteratura italiana, riteniamo che abbia senso proseguire con gli approfondimenti su quali siano limiti e vantaggi, opportunità e criticità (Viprey e Ledoux, 2007) della distanza intertestuale e del suo uso nei casi in cui cambino gli obiettivi perseguiti, la lingua, il genere testuale, le unità di analisi coinvolte, le strategie di calcolo, ecc.

Come si può vedere da un esempio di dendrogramma basato su un'analisi gerarchica agglomerativa con metodo del legame completo e distanze intertestuali calcolate nella versione iterativa su campioni di porzioni di testo da 20.000 occorrenze (fig. 5), per le applicazioni di *text clustering* la distanza intertestuale sembra garantire buone *performance* in termini di riconoscimento di gruppi omogenei nei quali l'autore ricopre un ruolo predominante. Precedenti esperienze hanno dimostrato che la versione iterativa della distanza intertestuale risulta determinante nella classificazione in gruppi omogenei di testi di lunghezza significativamente diversa (Tuzzi, 2011). Nel caso di questo studio la versione iterativa funziona meglio della versione base e permette una migliore attribuzione d'autore, anche se non in modo risolutivo. Una spiegazione per questo risultato va cercata nel fatto che i testi scritti dallo stesso autore possono avere anche nelle dimensioni un elemento che li accumuna e che va a vantaggio delle versioni della distanza intertestuale non basate sul campionamento. D'altro canto il campionamento riduce la variabilità e, in qualche modo, potrebbe ridurre le potenzialità discriminatorie.

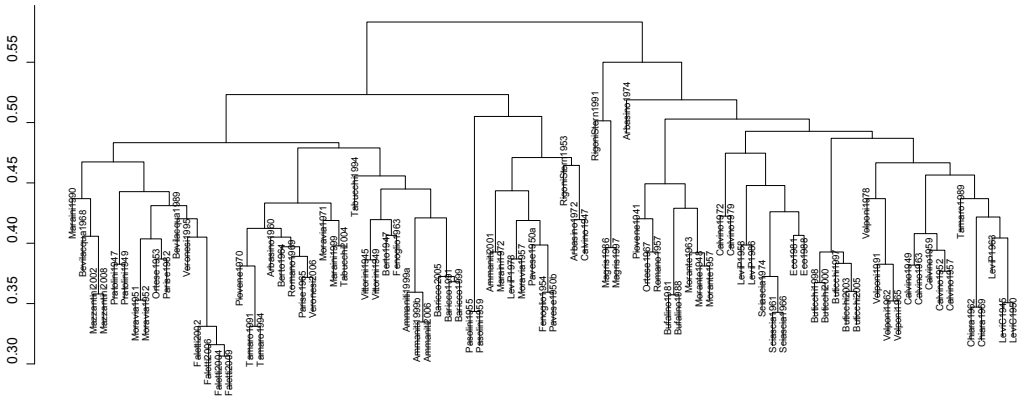

Figura 5 : Dendrogramma della cluster analysis gerarchica agglomerativa con legame completo basata sulla versione iterativa della distanza (200 campioni da 20.000 occorrenze)

La lingua dei corpora è una delle questioni aperte nella sperimentazione della distanza intertestuale perché, sebbene in letteratura si trovino già numerose applicazioni che riguardano il francese e alcune esperienze circoscritte per l'inglese (per es.: Merriam, 2003; Tuzzi 2011) e l'italiano (Cortelazzo *et al.*; Pauli e Tuzzi, 2009; Tuzzi *et al.*, 2010), resta da risolvere il problema della disponibilità di valori di riferimento affidabili e, in ottica comparata multilingue, la ricerca è ancora a uno stadio iniziale.

Per quanto riguarda, infine, le unità di analisi e le strategie di calcolo, questo studio conferma alcuni risultati di studi precedenti, per esempio il fatto che un numero limitato di forme grammaticali possa essere sufficiente per classificare e discriminare i testi (Argamon *et al.* 2007) con un buon grado di affidabilità. Stamatatos (2009) sostiene che possono bastare poche forme grammaticali in stilometria e sono, viceversa, necessarie numerose forme di contenuto nel *text clustering* per *topics* in ambito di *text mining*. Ulteriori sperimentazioni potranno corroborare ulteriormente queste affermazioni.

Riferimenti

Argamon S., Whitelaw C., Chase P., Raj Hota S., Garg N. and Levitan S. (2007). Stylistic Text Classification Using Functional Lexical Features. *Journal of the American Society for Information Science and Technology*, 58(6): 802-822.

Baayen, H.R. (2001). *Word Frequency Distributions. Exploring Quantitative Aspects of Lexical Structure*. Dordrecht: Kluwer Academic Pub.

Bagavandas M. and Manimannan G. (2008). Style Consistency and Authorship Attribution: A Statistical Investigation. *Journal of Quantitative Linguistics*, 15(1): 100-110.

Berry M.W. (2004). *Survey of Text Mining. Clustering, Classification, and Retrieval*. New-York: Springer-Verlag.

Bolasco S. (2010). *Taltac2.10. Sviluppi, esperienze ed elementi essenziali di analisi automatica dei testi*. Milano: LED.

Brunet E. (1988). Une mesure de la distance intertextuelle: la connexion lexicale. *Le nombre et le texte. Revue informatique et statistique dans les sciences humaines*, Université de Liége.

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

Cortelazzo M.A., Nadalutti P. and Tuzzi A., *Improving Labbe's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature*. Manoscritto.

Everitt B. (1980). *Cluster Analysis*. New York: Halsted press.

Huang A. (2008). Similarity Measures for Text Document Clustering. *Proceedings of the NZCSRSC 2008*, April 2008, Christchurch, New Zealand.

Koppel M., Schler J. and Argamon S. (2008). Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology*, 60(1): 9-26.

Labbé C. and Labbé D. (2001). Inter-Textual Distance and Authorship Attribution Corneille and Molière. *Journal of Quantitative Linguistics*, 8(3): 213-231.

Labbé C. and Labbé D. (2003). La distance intertextuelle. *Corpus*, 2:95-118.

Labbé D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics*, 14(1): 33-80.

Labbé D. (2010). Corneille nell'ombra di Molière. Come identificare un autore? *Rivista internazionale di tecnica della traduzione*, 12: 117-138.

Lockers M.J. and Witten D.M. (2010). A comparative study of machine learning methods for authorship attribution. *Literary and Linguistic Computing*, 25 (2): 215-223.

Love H. (2002). *Attributing Authorship: an Introduction*. Cambridge: Cambridge University Press.

Merriam T. (2003). An Application of Authorship Attribution by Intertextual Distance in English, *Corpus*, 2, Decembre 2003 [http://corpus.revues.org/index35.html].

Muller C. (1968). *Initiation à la statistique linguistique*. Paris: Larousse.

Muller C. (1977). *Principes et méthodes de statistique lexicale*. Paris: Hachette.

Muller C. and Brunet E. (1988). La statistique résout-elle les problèmes d'attribution? *Strumenti critici*, n.s., 3(3): 367-387.

Pauli F. and Tuzzi A. (2009). The End of Year Addresses of the Presidents of the Italian Republic (1948-2006): discoursal similarities and differences. *Glottometrics*, 18: 40-51.

Popescu I.-I., Mačutek J. and Altmann G. (2009). *Studies in Quantitative Linguistics 3*. Lüdenscheid: RAM-Verlag.

R development core team (2010). *R: a language and environment for statistical computing* (ver. 2.13.1). Vienna, Austria: R foundation for statistical computing, http://www.r-project.org.

Rudman J. (1998). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities*, 31: 351-365.

Stamatatos E. (2009). A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60(3): 538-556.

Strauss U., Fan F., and Altmann G. (2008). *Problems in quantitative linguistics 1*. Lüdenscheid: RAM-Verlag.

Tuzzi A. (2011). Reinhard Köhler's Scientific Production: Words, Numbers and Pictures. In: Altmann, G., Grzybek, P., Naumann, S., Vulanovic, R., editors, *Synergetic Linguistics. Text and Language as Dynamic Systems*. Wien: Praesens Verlag, 221-240.

Tuzzi A., Popescu I.-I. and Altmann G. (2010). *Quantitative Analysis of Italian Texts*. Lüdenscheid: RAM-Verlag.

Tweedie F.J. and Baayen R.H. (1998). How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, 32(5): 323-352.

Viprey J.-M. and Ledoux C.N. (2006). About Labbe's "intertextual distance". *Journal of Quantitative Linguistics*, 13(2): 265-283.

Appendice: Elenco delle opere

Niccolò Ammaniti	
1996	*Fango*
1999	*Ti prendo e ti porto via*
2001	*Io non ho paura*
2006	*Come Dio comanda*
Alberto Arbasino	
1960	*L'Anonimo lombardo*
1972	*La bella di Lodi*
1974	*Specchio delle mie brame*
Alessandro Baricco	
1991	*Castelli di rabbia*
1999	*City*
2005	*Questa storia*
Giuseppe Berto	
1947	*Il cielo è rosso*
1964	*Il male oscuro*
Alberto Bevilacqua	
1968	*L'occhio del gatto*
1989	*Il gioco delle passioni*
Gesualdo Bufalino	
1981	*Diceria dell'untore*
1988	*Le menzogne della notte*
Marco Buticchi	
1998	*Menorah*
2000	*Profezia*
2003	*La nave d'oro*
2005	*L'anello dei re*
1997	*Le pietre della luna*
Italo Calvino	
1947	*Il sentiero dei nidi di ragno*
1949	*Ultimo viene il corvo*
1952	*Il visconte dimezzato*
1957	*Il barone rampante*
1959	*Il cavaliere inesistente*
1963	*Marcovaldo, ovvero Le stagioni in città*
1972	*Le città invisibili*
1979	*Se una notte d inverno un viaggiatore*
Piero Chiara	
1962	*Il piatto piange*
1969	*L'uovo al cianuro e altre storie*
Umberto Eco	
1981	*Il nome della rosa*
1988	*Il pendolo di Foucault*
Dacia Maraini	
1972	*Memorie di una ladra*
1990	*La lunga vita di Marianna Ucria*
1999	*Buio*
Margaret Mazzantini	
2002	*Non ti muovere*
2008	*Venuto al mondo*
Elsa Morante	
1948	*Menzogna e sortilegio*
1957	*L'isola di Arturo*
1963	*Lo scialle andaluso*
Alberto Moravia	
1951	*Il conformista*
1952	*I racconti*
1957	*La ciociara*
1971	*Io e lui*
Anna Maria Ortese	
1953	*Il mare non bagna Napoli*
1967	*Poveri e semplici*
Goffredo Parise	
1965	*Il padrone*
1982	*Sillabario n. 2*
Pier Paolo Pasolini	
1955	*Ragazzi di vita*
1959	*Una vita violenta*
Cesare Pavese	
1950a	*La bella estate*
1950b	*La luna e i falò*
Guido Piovene	
1941	*Lettere di una novizia*
1970	*Le stelle fredde*
Vasco Pratolini	
1947	*Cronache di poveri amanti*
1949	*Un eroe del nostro tempo*
Mario Rigoni Stern	
1953	*Il sergente nella neve*
1991	*Arboreto salvatico*
Lalla Romano	
1957	*Tetto murato*
1969	*Le parole tra noi leggere*
Leonardo Sciascia	
1961	*Il giorno della civetta*
1966	*A ciascuno il suo*

Giorgio Faletti
2002 *Io uccido*
2004 *Niente di vero tranne gli occhi*
2006 *Fuori da un evidente destino*
2009 *Io sono Dio*
Beppe Fenoglio
1954 *La malora*
1963 *Una questione privata*
Primo Levi
1958 *Se questo è un uomo*
1963 *La tregua*
1978 *La chiave a stella*
1986 *I sommersi e i salvati*
Carlo Levi
1945 *Cristo si è fermato a Eboli*
1950 *L'Orologio*
Claudio Magris
1986 *Danubio*
1997 *Microcosmi*
1974 *Todo modo*
Antonio Tabucchi
1984 *Notturno indiano*
1994 *Sostiene Pereira*
2004 *Tristano muore. Una vita*
Susanna Tamaro
1989 *La testa tra le nuvole*
1991 *Per voce sola*
1994 *Va dove ti porta il cuore*
Sandro Veronesi
1995 *Venite venite B-52*
2006 *Caos calmo*
Elio Vittorini
1945 *Uomini e no*
1949 *Le donne di Messina*
Paolo Volponi
1962 *Memoriale*
1965 *La macchina mondiale*
1978 *Il pianeta irritabile*
1991 *La strada per Roma*