Quaderni di Statistica
Vol. 1, 1999

Una distribuzione mistura per la lunghezza delle parole nella lingua italiana

Angela D'Elia
Dipartimento di Scienze Statistiche, Università di Napoli Federico II
e-mail: angdelia@unina.it

Summary: In this paper we focuse on words' length frequencies in written Italian language. A compound probabilistic model is proposed to explain the mechanism of words' choice. Maximum Likelihood and Minimum Chi-square methods are used in the estimation procedure. Furthermore, a likelihood ratio test is developed to verify the structural homogeneity among different authors. Some samples from the main pieceworks of three Italian novelists, from $19th$ and $20th$ century, are studied to get an empirical check of the appropriateness of the proposal. The results confirm the existence and co-existence of several words categories in the structure of Italian written communication and their homogeneous use among the authors.

Key words: Words length, Mixture of distributions, Homogineity Test.

1. Introduzione

Il linguaggio nelle sue diverse forme di espressione può essere oggetto di analisi statistiche rivolte all'interpretazione del contenuto dei testi (*statistica testuale*) o allo studio di specifiche caratteristiche della lingua (*statistica linguistica*) a prescindere dal loro significato. All'interno di tali analisi statistiche è possibile individuare diversi

filoni di ricerca: individuazione di leggi e schemi generatori di regolarità presenti nella lingua (Sichel, 1975; Witzum *et al.*, 1994; Zipf, 1935); analisi descrittiva della distribuzione delle sillabe, dei segni di interpunzione, dei monogrammi, ecc. (Rizzi, 1985); analisi della distribuzione delle diverse categorie grammaticali (Rizzi, 1994); analisi multivariata (generalmente, analisi delle corrispondenze) dei dati testuali, per individuare le strutture latenti presenti nel vocabolario e realizzare mappe interpretative del linguaggio (Balbi, 1998; Bolasco, 1993; Lebart e Salem, 1994; Lebart, Morineau e Piron, 1995); valutazione della "vicinanza" linguistica di Autori diversi, attraverso tecniche di *clustering* o di *scaling* (Piccolo, 1991); analisi della successione della lunghezza delle parole mediante modelli dinamici (Piccolo, 1991) o, più in generale, analisi delle componenti del linguaggio (Corduas, 1995).

Il presente lavoro è rivolto all'analisi statistica della lunghezza delle parole presenti in un componimento scritto. Infatti, mentre la lunghezza dei testi deriva principalmente dal genere letterario cui essi appartengono (poesia, prosa giornalistica, prosa narrativa, ecc.), la variabile "Lunghezza delle parole" può essere considerata indicatrice dell'impronta stilistica dell'Autore. D'altra parte, è innegabile che, qualunque sia lo stile dell'Autore, l'alternarsi di vocaboli di diversa lunghezza nei componimenti scritti sia anche il risultato di strutture e vincoli che caratterizzano la lingua mediante cui ci si esprime. In questo lavoro viene proposto un modello probabilistico per descrivere e spiegare il meccanismo che determina la scelta delle parole, più o meno lunghe, attraverso le quali veicolare il messaggio. Inoltre, mediante l'analisi di testi di Autori diversi, si intende valutare la vicinanza delle diverse impronte stilistiche al modello proposto.

La struttura dell'articolo è la seguente: nel Paragrafo 2 sono discussi alcuni elementi caratterizzanti i componimenti scritti, con riferimento a quanto avviene per la lingua italiana. Nel Paragrafo 3 è proposto un modello probabilistico (una mistura di distribuzioni) che descriva la presenza dei differenti nuclei individuati nell'articolarsi della lingua italiana. Gli aspetti inferenziali volti alla stima sia dei parametri delle distribuzioni che al peso con cui esse intervengono

nel modello probabilistico sono oggetto del Paragrafo 4: in particolare, per la natura dei dati si considerano due diversi metodi di stima (Massima Verosimiglianza e Minimo Chi quadrato); nel paragrafo 5 si sviluppa, poi, un test del rapporto di verosimiglianza per verificare l'omogeneità della struttura linguistica in Autori differenti. Nel Paragrafo 6 si procede ad una verifica empirica dell'adeguatezza dello schema proposto e dell'adattamento ad un campione di testi scritti. Alcune considerazioni finali concludono il lavoro.

2. Considerazioni strutturali sulla lingua italiana

Un componimento scritto può essere considerato la traduzione di un pensiero o di una informazione, più o meno estesi, mediante il quale le persone comunicano tra di loro. La comunicazione, nel senso più vasto del termine, è infatti l'obiettivo prioritario per l'Autore di un testo scritto, sia esso un componimento letterario, una relazione scientifica, un articolo di un quotidiano, ecc.

Qualunque sia la natura dello scritto, all'Autore si impone una scelta tra vocaboli, locuzioni, costrutti verbali, che gli permettano di esprimersi in modo efficace e di divulgare un messaggio che sia intellegibile e stimolante per il lettore: il realizzarsi di tali condizioni è infatti un requisito fondamentale per la dialettica della comunicazione.

Evidentemente, esistono da un canto esigenze morfologiche relative alla flessione delle parole, e dall'altro vincoli sintattici nella formazione delle frasi, che influiscono sull'elaborazione del testo (Tabossi, 1999). In particolare, riteniamo che sia possibile distinguere un momento statico ed uno dinamico nell'elaborazione di un messaggio scritto. Il primo riguarda strettamente la scelta delle parole da utilizzare, con i soli vincoli imposti dalle regole ortografiche e grammaticali, presenti in ogni lingua. Il secondo, invece, attiene alla collocazione di ciascun vocabolo all'interno di una sequenza di parole, che deve essere coerente sul piano logico ed intellegibile in rapporto al lettore, oltre che grammaticalmente corretta.

In questo articolo ci soffermiamo sul primo dei due aspetti delineati, svolgendo, per semplicità, un'analisi statica della lunghezza dei vocaboli utililizzati nei componimenti scritti.

La scelta di utilizzare un vocabolo piuttosto che un altro risponde a due esigenze spesso contrapposte: economicità e comprensibilità. In molte situazioni (messaggi pubblicitari, slogan politici, ecc.) la comunicazione del messaggio risulta tanto più efficace, quanto più è breve il testo in cui esso è contenuto: ciò, infatti, richiede un minore sforzo da parte del lettore e, quindi, aumenta l'impatto e la probabilità di suscitare un'elevata soglia di attenzione. D'altra parte, la comunicazione di un messaggio ha senso se risulta comprensibile per il destinatario: tale esigenza richiede che l'Autore arricchisca il testo con vocaboli, aggettivi, avverbi, che servano a specificare e chiarire il messaggio in esso contenuto.

Queste contrapposte esigenze hanno determinato (almeno per la lingua italiana, la cui evoluzione dura da diversi secoli) una struttura del discorso articolata sull'alternarsi di *concetti*, espressi mediante vocaboli di media-lunga dimensione, e di *connettivi* di breve lunghezza. Ciò deriva da una parte dall'elevata frequenza con cui i connettivi appaiono in un discorso, il che rende necessaria la loro breve lunghezza; d'altra parte, c'è la necessità che le idee e i concetti principali siano contenuti in vocaboli isolatamente ed autonomamente comprensibili, il che richiede parole di media-lunga dimensione.

La distinzione da noi effettuata sintetizza, in effetti, concetti abituali negli studi linguistici, in particolare quelli connessi alla morfologia ; in tale ambito, infatti, si parla di parole-contenuto ("parole di classe aperta") e parole-funzione ("parole di classe chiusa") che, rispettivamente, evocano le nostre definizioni di concetti e connettivi. L'evoluzione della lingua registra costantemente e con una certa velocità modifiche ed aggiunte alle parole-contenuto, mentre occorrono secoli prima che si registri una modifica sostanziale nell'uso delle parole-funzione. Poiché queste ultime esprimono una relazione concettuale stabile tra altre parole esse sono costitutive per l'uso di una lingua in un determinato contesto. Ne consegue che esse sono necessariamente molto usate e generalmente di breve lunghezza, per il

principio del minimo sforzo a parità di informazione veicolata che presiede ad ogni comunicazione umana (Akmajian *et al.*, 1996, 29-30).

L'alternanza di queste due categorie, concetti e connettivi, nel linguaggio conduce ad ipotizzare che la lunghezza delle parole adoperate nei componimenti scritti derivi dal sovrapporsi e combinarsi di due diversi nuclei di vocaboli di differente lunghezza media.

A tali categorie va aggiunto, in modo peculiare per la lingua italiana, il gruppo di vocaboli caratterizzati da lunghezza unitaria: (*a, e, i, o*). Esso infatti contiene non solo la principale congiunzione (*e*) ed il più diffuso articolo plurale maschile (*i*), ma anche la terza voce singolare presente del verbo essere (*è*) che, oltre a godere di significato autonomo, svolge anche l'importante ruolo di verbo ausiliare.

L'esistenza e l'importanza di tale gruppo di vocaboli all'interno della lingua italiana porta, quindi, a considerare una combinazione di tre nuclei *1)* parole di una sola lettera, *2)* altri connettivi, *3)* concetti, ognuno dei quali è necessario per un'efficace formulazione e trasmissione del messaggio contenuto nel testo scritto.

3. Il modello probabilistico

Le considerazioni e le motivazioni svolte nel paragrafo precedente inducono a formulare una proposta di modello probabilistico per la variabile $\mathcal{X}$ "Lunghezza delle parole" che si concretizzi nella mistura finita di 3 variabili casuali discrete.

Tale mistura, con distribuzione di probabilità

$$\Pr(\mathcal{X} = x) = \pi_1 p_1(x) + \pi_2 p_2(x) + \pi_3 p_3(x) \qquad (x \in \Re),$$

è caratterizzata dalle singole distribuzioni di probabilità che la compongono $(p_1(\cdot), p_2(\cdot), p_3(\cdot))$ e dai corrispondenti pesi (π_1, π_2, π_3), dove

$$\pi_j > 0, \qquad j = 1, 2, 3; \qquad \pi_1 + \pi_2 + \pi_3 = 1.$$

Per quanto concerne la specificazione di un'opportuna forma parametrica per le componenti $(p_1(\cdot), p_2(\cdot), p_3(\cdot))$, è utile fare alcune riflessioni:

1. le parole di lunghezza unitaria costituiscono una categoria a parte, autonoma e svincolata dagli altri nuclei della lingua. E' possibile, quindi, assumere in corrispondenza di tale gruppo la presenza di una componente degenere, con l'intera massa di probabilità sul valore $x = 1$;

2. i connettivi che non appartengono al gruppo 1, sono comunque caratterizzati da breve lunghezza con moda in $x = 2$: si pensi alla frequenza di utilizzo degli articoli (il, la, le, lo, un), delle preposizioni semplici (di, da, in, su) o composte (al, ai), delle congiunzioni (ma, se), dei pronomi (me, te, mi, ti, ci, vi), ecc. Secondo l'assunzione – ampiamente accertata negli studi linguistici – che un vocabolo (e quindi anche un connettivo) è tanto più usato quanto più è corto, risulta naturale ipotizzare una decadenza esponenziale nella distribuzione di frequenza di connettivi di lunghezza maggiore di 2;

3. infine, il nucleo dei vocaboli destinati ad esprimere concetti è per sua natura costituito da parole di dimensione media-lunga, essendo queste portatrici di un significato autonomo. Evidentemente, anche all'interno di questa categoria, risulta naturale assumere una legge di decadenza esponenziale della frequenza di utilizzo al crescere della lunghezza dei singoli vocaboli. Tale decadenza può essere assunta a partire dal valore 5, empiricamente determinato per la lingua italiana.

Sulla base di tali considerazioni, riteniamo che un modello che possa descrivere adeguatamente la distribuzione della lunghezza delle parole in un testo scritto sia una mistura di 3 variabili casuali Geometriche. In particolare, ci sembra appropriato assumere che la prima componente X_1 sia degenere, la seconda componente X_2 sia troncata in $x = 1$ e la terza X_3 sia troncata in $x = 4$, cioè:

$$
\begin{aligned}
X_1 \sim Geo(\theta_1) \rightarrow \quad & p_{X_1}(x;\theta_1) = \theta_1 = 1, && x = 1; \\
X_2 \sim Geo(\theta_2) \rightarrow \quad & p_{X_2}(x;\theta_2) = \theta_2(1-\theta_2)^{x-2}, && x = 2, 3, ...; \\
X_3 \sim Geo(\theta_3) \rightarrow \quad & p_{X_3}(x;\theta_3) = \theta_3(1-\theta_3)^{x-5}, && x = 5, 6, ...
\end{aligned}
$$

Ne consegue che la distribuzione mistura risulterà:

$$\Pr(\mathcal{X} = x) = (1 - \pi_2 - \pi_3)p_{X_1}(x) + \pi_2 p_{X_2}(x; \theta_2) + \pi_3 p_{X_3}(x; \theta_3).$$

Poiché per qualsiasi variabile casuale $X \sim Geo(\theta)$, vale la seguente relazione ricorsiva:

$$\Pr(X = x + 1) = \Pr(X = x)(1 - \theta) \qquad x = 1, 2, ...$$

è immediato dedurre che la seconda (X_2) e la terza componente (X_3) della mistura hanno rispettivamente moda in $x = 2$ e $x = 5$. Tale caratteristica le rende adeguate a descrivere il comportamento della variabile "Lunghezza delle parole" in vocaboli che esprimono, rispettivamente, connettivi e concetti[1].

La bimodalità è la diretta espressione della compresenza di due sottopopolazioni non degeneri con distribuzioni Geometriche, rappresentate dalle componenti X_2 e X_3 della mistura. Ciascuna di queste sottopopolazioni è caratterizzata da un parametro θ_j $(j = 2, 3)$ che esprime una misura di occorrenza di parole brevi all'interno della rispettiva categoria.

Ora, per una variabile casuale Geometrica troncata in $x = c$ il valor medio e la varianza risultano, rispettivamente:

$$E(X) = \sum_{x=c+1}^{\infty} x\theta(1 - \theta)^{x-(c+1)} = c + \frac{1}{\theta},$$

$$Var(X) = \frac{1 - \theta}{\theta^2}.$$

Pertanto, per le due componenti non degeneri della mistura si ha che:

$$E(X_2) = 1 + \frac{1}{\theta_2}, \qquad (c = 1)$$

[1] Diverse verifiche empiriche (Piccolo, 1991; 1998), volte ad un'analisi esplorativa della lunghezza delle parole in componimenti italiani di natura discorsiva, hanno, infatti, confermato la presenza di distribuzioni di frequenza bimodali, con mode in $x = 2$ e $x = 5$.

$$E(X_3) = 4 + \frac{1}{\theta_3}, \qquad (c = 4)$$

mentre

$$Var(X_i) = \frac{1 - \theta_i}{\theta_i^2}, \qquad (i = 2, 3).$$

Quindi

se $\theta_2 \to 1$: $\quad E(X_2) \to 2 \quad Var(X_2) \to 0$

se $\theta_3 \to 1$: $\quad E(X_3) \to 5 \quad Var(X_3) \to 0.$

Questo significa che, all'interno della categoria dei connettivi (X_2), all'aumentare di θ_2 la lunghezza media dei vocaboli tende ad essere pari a 2: quindi θ_2 diviene misura dell'occorrenza di connettivi brevi. Analogamente, nella categoria dei vocaboli che esprimono concetti (X_3), al crescere di θ_3 la lunghezza media delle parole tende a 5: θ_3 è quindi misura dell'occorrenza di concetti espressi tramite vocaboli di breve lunghezza. Infatti per θ_2 e/o $\theta_3 \to 1$, le due distribuzioni tendono esse stesse a divenire degeneri (la varianza è nulla), con massa di probabilità completamente addensata in corrispondenza di $x = 2$ e $x = 5$, rispettivamente.

Viceversa, in corrispondenza di valori bassi di θ_2 e θ_3, nelle due categorie aumenta la lunghezza media dei vocaboli, sia che essi esprimano connessione, sia che si riferiscano a concetti. Aumenta, d'altra parte, anche la variabilità delle distribuzioni della "Lunghezza delle parole" intorno ai rispettivi valori medi.

4. Stima parametrica del modello

Dalle considerazioni precedenti emerge come l'analisi linguistica di un testo e dell'impronta stilistica in esso presente possa essere condotta anche mediante la misura della minore o maggiore occorrenza di vocaboli brevi all'interno delle diverse categorie che abbiamo delineato.

In particolare, al fine di una completa conoscenza della mistura di distribuzioni, che assumiamo descriva l'alternarsi di parole lunghe

e brevi nei testi scritti in lingua italiana, è necessario stimare sia i parametri θ_2 e θ_3 caratterizzanti le componenti (X_2, X_3) che il peso che ciascuna di essa ha nella mistura. Poiché $\pi_1 + \pi_2 + \pi_3 = 1$, è sufficiente stimare i soli pesi π_2, π_3 delle componenti X_2, X_3 ed ottenere, poi, $\hat{\pi}_1 = 1 - \hat{\pi}_2 - \hat{\pi}_3$.

Allo scopo di sottolineare la sostanziale stabilità dei risultati che si ottengono con il modello proposto, discuteremo, di seguito, due differenti metodi di stima dei parametri: il metodo della Massima verosimiglianza (par. 4.1) e il metodo del minimo Chi quadrato (par. 4.2). Questo ultimo metodo risulta computazionalmente più agevole (soprattutto nella sua versione modificata) e conduce a stimatori con le medesime proprietà asintotiche di quelli di Massima verosimiglianza (Harris e Kanji, 1983).

4.1. Metodo della massima verosimiglianza

Si consideri la distribuzione di frequenza della variabile "Lunghezza delle parole" in un testo. Siano $\mathbf{x} = (x_1, x_2, ..., x_k)'$ le modalità che tale variabile assume nel campione considerato e $n_1, n_2, ..., n_k$ le corrispondenti frequenze assolute, tali che $n_1 + n_2 + \ldots + n_k = n$.

La funzione di log-verosimiglianza è

$$\mathcal{L}(\theta, \pi; \underline{x}) = \sum_{i=1}^{k} \mathcal{L}_i(\theta, \pi; x_i).$$

Il contributo della modalità x_i $(i = 1, 2, ..., k)$ alla funzione di log-verosimiglianza della mistura risulta:

$$\mathcal{L}_i(\theta, \pi; x_i) = \begin{cases} n_i log\{\pi_2\theta_2(1-\theta_2)^{x_i-2} + \pi_3\theta_3(1-\theta_3)^{x_i-5}\} & i = 5, 6, ... \\ n_i log\{\pi_2\theta_2(1-\theta_2)^{x_i-2}\} & i = 2, 3, 4 \\ n_i log\{(1-\pi_2-\pi_3)\cdot 1\} & i = 1. \end{cases}$$

Gli stimatori di massima verosimiglianza Ψ_n del vettore dei parametri $\psi = (\pi_2, \pi_3, \theta_2, \theta_3)'$ sono, dunque, le soluzioni del sistema di equazioni:

$$\frac{\partial}{\partial \theta_j}\dot{\mathcal{L}}(\theta, \pi; \underline{x}) = 0, \quad j = 2, 3;$$

$$\frac{\partial}{\partial \pi_j}\mathcal{L}(\theta, \pi; \underline{x}) = 0, \quad j = 2, 3.$$

La soluzione delle precedenti equazioni può essere raggiunta per via numerica, mediante un opportuno algoritmo di ottimizzazione (ad es. il metodo di Newton-Raphson). A tal fine, è opportuno individuare adeguati valori iniziali ψ^0 per i parametri, per l'influenza che essi hanno sul raggiungimento e sulla velocità della convergenza. Si noti che, nel caso in esame, è opportuno esplicitare i vincoli:
$0 \leq \pi_j \leq 1$, $0 \leq \theta_j \leq 1$, $\quad j = 2, 3$.

Essendo rispettate le condizioni di regolarità, gli stimatori di Massima Verosimiglianza così ottenuti godono delle proprietà asintotiche di Normalità, efficienza e non distorsione (BAN):

$$\sqrt{n}(\Psi_n - \psi_0) \to N(\mathbf{0}, I(\psi_0)^{-1}),$$

per $n \to \infty$, dove $I(\psi_0)$ è la matrice di informazione attesa di Fisher in corrispondenza dei veri valori dei parametri ψ_0, (Titterington *et al.*, 1985).

4.2. Metodo del minimo Chi quadrato

Siano $n_1, n_2, ..., n_k$ le frequenze delle modalità $x_1, x_2, \ldots, x_k$; mediante la distribuzione teorica assunta $\Pr(\mathcal{X} = x)$, cioè per la mistura di v.c. geometriche, è possibile calcolare le frequenze teoriche np_i delle k modalità:

$$np_i(\psi) = n \Pr(\mathcal{X}_i = x_i) \qquad i = 1, 2, \ldots, k.$$

Si noti che

$$\Pr(\mathcal{X}_i = x_i) = \begin{cases} \pi_2\theta_2(1-\theta_2)^{x_i-2} + \pi_3\theta_3(1-\theta_3)^{x_i-5} & i = 5, 6, ... \\ \pi_2\theta_2(1-\theta_2)^{x_i-2} & i = 2, 3, 4 \\ (1-\pi_2-\pi_3)\cdot 1 & i = 1. \end{cases}$$

Le stime del minimo Chi quadrato sono i valori che minimizzano la seguente espressione:

$$Chi(\psi) = \sum_{i=1}^{k} \frac{(n_i - np_i(\psi))^2}{np_i(\psi)}$$

o nella versione modificata, proposta da Neyman,

$$MChi(\psi) = \sum_{i=1}^{k} \frac{(n_i - np_i(\psi))^2}{n_i}.$$

La seconda espressione è asintoticamente equivalente alla prima se il modello per $\mathcal{X}$ è ben specificato; evidentemente, essa risulta computazionalmente più agevole, in quanto richiede la stima dei parametri solo al numeratore. In tal caso, gli stimatori del metodo del minimo Chi quadrato Modificato sono le soluzioni delle equazioni:

$$\frac{\partial MChi(\psi)}{\partial \psi_j} = 2 \sum_{i=1}^{k} \frac{np_i(\psi)}{n_i} \frac{\partial np_i(\psi)}{\partial \psi_j} = 0, \qquad j = 1, 2, 3, 4.$$

Gli stimatori così ottenuti sono consistenti, asintoticamente Normali e convergenti agli stimatori di Massima verosimiglianza, anche se godono di efficienza asintotica di ordine inferiore.

5. Un test di omogeneità liguistica

I parametri della mistura caratterizzanti l'uso di vocaboli più o meno lunghi da parte di un Autore possono essere oggetto di un test per verificare l'omogeneità di scrittori differenti relativamente alla lunghezza delle parole adoperate.

In particolare, si può sottoporre a test l'ipotesi nulla $H_0 : \psi_a = \psi_b$, di uguaglianza dei parametri in Autori diversi (ad esempio, negli scrittori a e b), mediante il test del rapporto di verosimiglianza.

Come è noto il test del rapporto di verosimiglianza è basato sulla statistica-test:

$$-2 * log(\mathcal{L}_\omega / \mathcal{L}_\Omega) = -2[logL_\omega - logL_\Omega] \sim \chi^2_m$$

dove $\mathcal{L}_\omega, \mathcal{L}_\Omega$ rappresentano il valore della funzione di verosimiglianza massimizzata sotto H_0 e sotto H_1, rispettivamente, mentre m è la differenza tra il numero dei parametri sotto H_1 e quelli vincolati sotto H_0.

Ora, siano $\mathbf{x} = (x_1, x_2, ..., x_k)'$ le medesime modalità osservate per gli Autori a e b, con frequenze $n_{1a,....,}n_{ka}$ e n_{1b}, ..., n_{kb}, rispettivamente.

Per il modello probabilistico $\mathcal{X}$ precedentemente introdotto, la funzione di log-verosimiglianza sotto H_1 è: $\mathcal{L}_\Omega = \sum_{i=1}^{k} \mathcal{L}_{\Omega(i)}$ $(i = 1, ..., k)$, dove

$$\mathcal{L}_{\Omega(i)} = n_{i(a)} * log\{\pi_{2(a)}\theta_{2(a)}(1-\theta_{2(a)})^{(x_i-2)} + \pi_{3(a)}\theta_{3(a)}(1-\theta_{3(a)})^{(x_i-5)}\} +$$

$$n_{i(b)} * log\{\pi_{2(b)}\theta_{2(b)}(1-\theta_{2(b)})^{(x_i-2)} + \pi_{3(b)}\theta_{3(b)}(1-\theta_{3(b)})^{(x_i-5)}\}, \qquad i = 5, 6, ...$$

$$\mathcal{L}_{\Omega(i)} = n_{i(a)} * log\{\pi_{2(a)}\theta_{2(a)}(1 - \theta_{2(a)})^{(x_i-2)}\} +$$
$$n_{i(b)} * log\{\pi_{2(b)}\theta_{2(b)}(1 - \theta_{2(b)})^{(x_i-2)}\}, \qquad i = 2, 3, 4$$

$$\mathcal{L}_{\Omega(i)} = n_{i(a)} * log\{1-\pi_{2(a)}-\pi_{3(a)}\} + n_{i(b)} * log\{1-\pi_{2(b)}-\pi_{3(b)}\}, \qquad i = 1$$

La funzione di log-verosimiglianza sotto H_0, invece, risulta $\mathcal{L}_\omega = \sum_{i=1}^{k} \mathcal{L}_{\omega(i)}$ $(i = 1, ..., k)$, dove

$$\mathcal{L}_{\omega(i)} = n_{i(a)} * log\{\pi_{2(H_0)}\theta_{2(H_0)}(1-\theta_{2(H_0)})^{(x_i-2)} + \pi_{3(H_0)}\theta_{3(H_0)}(1-\theta_{3(H_0)})^{(x_i-5)}\} +$$

$$n_{i(b)} * log\{\pi_{2(H_0)}\theta_{2(H_0)}(1-\theta_{2(H_0)})^{(x_i-2)} + \pi_{3(H_0)}\theta_{3(H_0)}(1-\theta_{3(H_0)})^{(x_i-5)}\} =$$
$$= (n_{i(a)} + n_{i(b)}) * log\{\pi_{2(H_0)}\theta_{2(H_0)}(1-\theta_{2(H_0)})^{(x_i-2)} + \pi_{3(H_0)}\theta_{3(H_0)}(1-\theta_{3(H_0)})^{(x_i-5)}\},$$
$$i = 5, 6, ...$$

$$\mathcal{L}_{\omega(i)} = n_{i(a)} * log\{\pi_{2(H_0)}\theta_{2(H_0)}(1 - \theta_{2(H_0)})^{(x_i-2)}\} +$$
$$n_{i(b)} * log\{\pi_{2(H_0)}\theta_{2(H_0)}(1 - \theta_{2(H_0)})^{(x_i-2)}\} =$$
$$= (n_{i(a)} + n_{i(b)}) * log\{\pi_{2(H_0)}\theta_{2(H_0)}(1-\theta_{2(H_0)})^{(x_i-2)}\}, \qquad i = 2, 3, 4$$

$$\mathcal{L}_{\omega(i)} = n_{i(a)} * log\{1-\pi_{2(H_0)}-\pi_{3(H_0)}\} + n_{i(b)} * log\{1-\pi_{2(H_0)}-\pi_{3(H_0)}\} =$$

$$= (n_{i(a)} + n_{i(b)}) * log\{1 - \pi_{2(H_0)} - \pi_{3(H_0)}\}, \qquad i = 1.$$

In essa, a causa del vincolo esplicitato sotto l'ipotesi nulla H_0, è richiesta la stima dei soli 4 parametri: $\theta_{2(H_0)}$ $\theta_{3(H_0)}$ $\pi_{2(H_0)}$ $\pi_{3(H_0)}$. Ne risulta, quindi, che

$$-2 * log(\mathcal{L}_\omega / \mathcal{L}_\Omega) \sim \chi^2_4,$$

con regione critica $C_0 = \{\underline{x} : -2 * log(\mathcal{L}_\omega / \mathcal{L}_\Omega) \geq \chi^2_{4,\alpha}\}$.

6. Una verifica empirica

Al fine di ottenere una verifica empirica dell'adeguatezza del modello probabilistico proposto in questo lavoro per la lingua italiana scritta, presentiamo un'analisi svolta su 3 noti romanzieri: Manzoni (XIX secolo), Buzzati e Pavese (XX secolo).

Per tutti gli Autori sono stati considerati testi di prosa contenenti esclusivamente narrazioni; sono stati esclusi, invece, i testi con dialoghi, perché la struttura linguistica in essi presente, dovendo esprimere una forma di comunicazione diretta, risente di esigenze diverse da quelle che abbiamo delineato per i testi scritti in generale.

Il *corpus* di opere considerato è costituito dai capp. V, VI e VII de "Il deserto dei tartari" per Buzzati (3651 parole), dal racconto "Le tre ragazze" per Pavese (3875 parole) e dal cap. XXXII de "I Promessi Sposi" per Manzoni (6381 parole); in quest'ultimo caso si tratta di un lungo Capitolo interamente descrittivo (la descrizione della peste a Milano).

Per tutti gli Autori considerati sono stati stimati, sia con il metodo della Massima Verosimiglianza che del minimo Chi quadrato modificato, i parametri π_2, π_3, θ_2 e θ_3 caratterizzanti la mistura di distribuzioni. Le tabelle seguenti mostrano le stime ottenute[2]:

[2]Le stime sono state ottenute mediante una *routine* di ottimizzazione numerica, presente nel *software* GAUSS (Aptech Systems).

Stime di Massima Verosimiglianza (MV) (*err. stand.*)						
Parametri	Buzzati		Pavese		Manzoni	
π_1	0.058		0.080		0.068	
π_2	0.582	(*0.018*)	0.601	(*0.017*)	0.610	(*0.004*)
π_3	0.360	(*0.018*)	0.319	(*0.017*)	0.322	(*0.004*)
θ_2	0.331	(*0.013*)	0.350	(*0.013*)	0.322	(*0.002*)
θ_3	0.333	(*0.009*)	0.349	(*0.010*)	0.321	(*0.002*)

Stime del minimo Chi quadrato modificato (MChi) (*err. stand.*)						
Parametri	Buzzati		Pavese		Manzoni	
π_1	0.058		0.082		0.069	
π_2	0.593	(*0.014*)	0.606	(*0.011*)	0.611	(*0.004*)
π_3	0.349	(*0.014*)	0.312	(*0.011*)	0.320	(*0.004*)
θ_2	0.328	(*0.010*)	0.359	(*0.009*)	0.336	(*0.002*)
θ_3	0.328	(*0.007*)	0.365	(*0.006*)	0.345	(*0.001*)

Chiaramente il peso π_1 della distribuzione degenere X_1 nella mistura è stato ottenuto mediante la relazione: $\hat{\pi}_1 = 1 - \hat{\pi}_2 - \hat{\pi}_3$.

Si noti che in tutti gli Autori considerati, la distribuzione che ha maggior peso è quella della componente X_2 (categoria dei connettivi) con un peso stimato $\hat{\pi}_2$ oscillante intorno al 60% (i risultati ottenuti con i due metodi di stima sono praticamente coincidenti). Il peso $\hat{\pi}_1$ della distribuzione degenere X_1 relativa ai vocaboli di lunghezza unitaria è, invece, pari al 6–7% in Buzzati e Manzoni, e cresce sino all' 8.2% in Pavese.

Mediante i valori stimati per i parametri $\pi_2, \pi_3, \theta_2, \theta_3$ è possibile inoltre ottenere la distribuzione di probabilità teorica della "Lunghezza delle parole" nei tre Autori, e confrontarla, quindi, con la distribuzione di frequenze osservata.

Nella Figura 1 sono rappresentate per ogni Autore la distribuzione di frequenza osservata e le distribuzioni teoriche calcolate sia mediante le stime di Massima verosimiglianza, sia con le stime del minimo Chi quadrato. In tutti i casi emerge come la mistura di variabili casuali geometriche proposta abbia un buon adattamento complessivo

alla distribuzione osservata della "Lunghezza delle parole". Si noti che in tale situazione, data la numerosità dei vocaboli considerati, non risulta adeguato valutare la bontà di adattamento mediante la statistica X^2, che è influenzata pesantemente dalla numerosità campionaria.

I valori medi della lunghezza dei vocaboli che esprimono connettivi e di quelli che esprimono concetti, nei tre Autori, sono rispettivamente, per i due metodi di stima:

	MV	MChi	MV	MChi
	$E(X_2)$	$E(X_2)$	$E(X_3)$	$E(X_3)$
Buzzati	4.02	4.05	7.00	7.05
Pavese	3.86	3.78	6.86	6.74
Manzoni	4.10	3.98	7.11	6.90

Tali valori evidenziano una generale omogeneità della struttura linguistica nei tre Autori, per quanto concerne la lunghezza media delle parole che connettono ($E(X_2) \simeq 4$) o che veicolano significati compiuti ($E(X_2) \simeq 7$). In particolare, emerge in Pavese una maggiore tendenza all'uso di vocaboli brevi, sia all'interno dei connettivi, sia nella scelta di parole che esprimono concetti.

L'ipotesi di omogeneità dei tre Autori per quanto concerne l'uso di parole più o meno lunghe, ed il peso che esse rivestono nella distribuzione complessiva della "Lunghezza" delle parole, è stata sottoposta quindi a verifica mediante il test del rapporto di verosimiglianza, sviluppato nel paragrafo 5.

Dal confronto di ciascun Autore con gli altri due sono emersi i seguenti risultati:

Buzzati – Pavese $\quad -2 * log(\mathcal{L}_\omega/\mathcal{L}_\Omega) = 2.08 < 9.49 = \chi^2_{4,0.05}$;

$(p - value = 0.721)$;

Buzzati – Manzoni $\quad -2 * log(\mathcal{L}_\omega/\mathcal{L}_\Omega) = 7.10 < 9.49 = \chi^2_{4,0.05}$;

$(p - value = 0.131)$;

Pavese – Manzoni $-2 * log(\mathcal{L}_\omega/\mathcal{L}_\Omega) = 1.14 < 9.49 = \chi^2_{4,0.05}$;

$(p - value = 0.888)$.

Poiché le differenze risultano non significative in tutti i confronti, ne emerge una sostanziale omogeneità dei tre Autori per quanto concerne la variabile Lunghezza delle parole, e per il peso che le categorie dei connettivi e dei concetti rivestono all'interno dei rispettivi testi.

Il medesimo test è stato quindi utilizzato per verificare l'omogeneità di scelte linguistiche all'interno dello stesso Autore. Per ciascuno dei tre Autori considerati, si è suddiviso l'insieme delle parole, costiuenti le opere considerate, in due metà e si è sottoposta a verifica l'ipotesi nulla che i parametri caratterizzanti la distribuzione della "Lunghezza" delle parole fossero uguali nella prima e nella seconda metà dei testi:

$$H_0 : \psi_1 = \psi_2 \qquad \text{vs.} \qquad H_1 : \psi_1 \neq \psi_2.$$

I risultati ottenuti sono i seguenti:

Buzzati $-2 * log(\mathcal{L}_\omega/\mathcal{L}_\Omega) = 0.25 < 9.49 = \chi^2_{4,0.05}$;

$(p - value = 0.993)$;

Pavese $-2 * log(\mathcal{L}_\omega/\mathcal{L}_\Omega) = 0.09 < 9.49 = \chi^2_{4,0.05}$;

$(p - value = 0.999)$;

Manzoni $-2 * log(\mathcal{L}_\omega/\mathcal{L}_\Omega) = 7.16 < 9.49 = \chi^2_{4,0.05}$;

$(p - value = 0.128)$.

Nel caso quindi di Buzzati e Pavese si ha forte evidenza di omogeneità all'interno dei testi. Nel caso di Manzoni, invece, pur essendo le differenze non significative, emerge una minore omogeneità, che potrebbe essere imputata alle frequenti citazioni di brani del *De peste Mediolani quae fuit anno MDCXXX* dello storico Ripamonti, che appaiono sopratutto nella seconda parte del Capitolo analizzato.

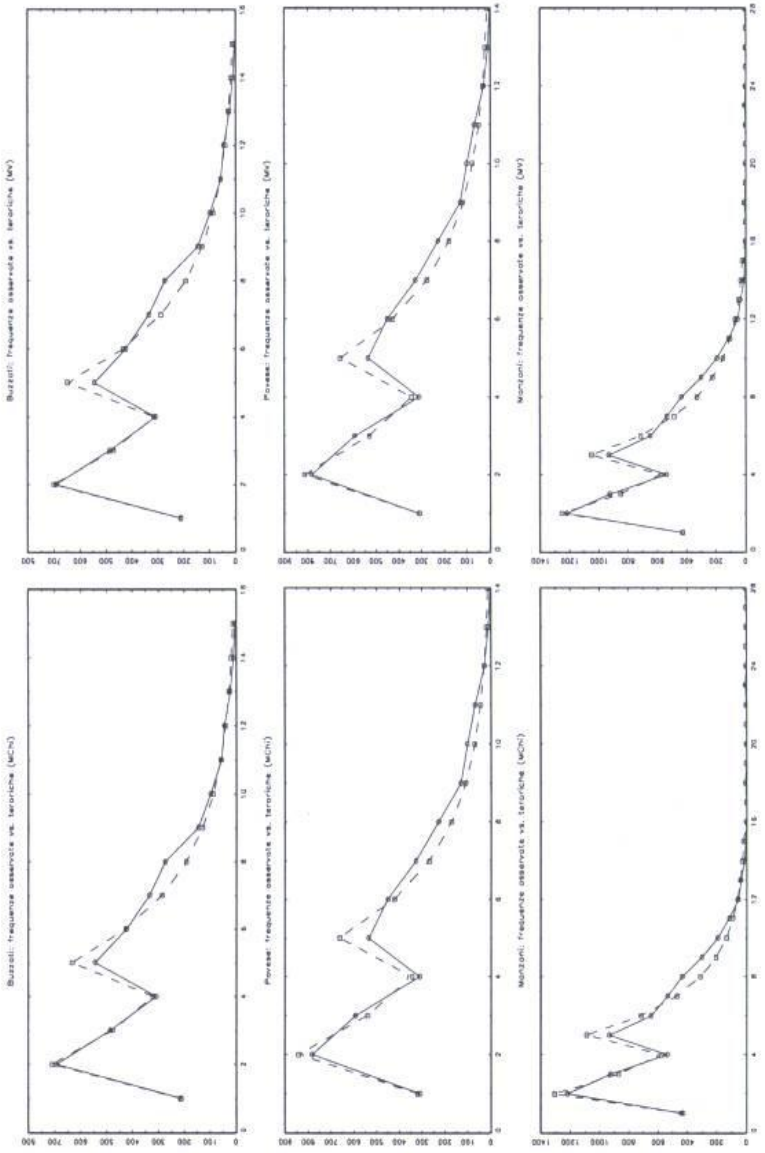

Figura 1. Confronti tra frequenze osservate e frequenze teoriche per i tre Autori

7. Considerazioni finali

In questo lavoro si è proposto un modello probabilistico per descrivere la variabile "Lunghezza delle parole" all'interno di forme di comunicazione scritta in lingua italiana. L'utilizzo di vocaboli più o meno lunghi può, infatti, essere considerato un'indicatore dell'impronta stilistica di un Autore. In particolare, si è ritenuto che la presenza di categorie di vocaboli differenti all'interno di un testo - *connettivi, concetti* e *vocaboli di lunghezza unitaria* potesse essere efficacemente rappresentata attraverso una mistura finita di variabili casuali Geometriche. Si è inoltre messo in luce come i parametri di tale mistura possano essere oggetto di un test delle ipotesi per verificare la presenza o meno di omogeneità tra Autori differenti e all'interno di uno stesso Autore.

L'analisi empirica condotta su tre grandi romanzieri italiani del XIX e del XX secolo ha confermato l'adeguatezza dello schema probabilistico proposto a descrivere l'utilizzo di vocaboli di differente lunghezza nei componimenti scritti. Inoltre, il test di omogeneità ha confermato la presenza di omogeneità negli Autori per quanto concerne l'uso di vocaboli più o meno lunghi, e la sostanziale uguaglianza all'interno dei testi di ciascun Autore.

Ne emerge di conseguenza una conferma all'ipotesi che la lingua italiana scritta sia strutturata in modo sufficientemente regolare sull'articolarsi di tre categorie di vocaboli, il cui utilizzo ed il cui peso sono rimasti pressoché inalterati nella prosa narrativa del XIX e della prima metà del XX secolo.

Ringraziamenti: Questo lavoro ha beneficiato di contributi derivanti da progetti di ricerca MURST, di Ateneo e di interesse nazionale.

Riferimenti Bibliografici

Akmajian A., Demers R. A., Farmer A. K. e Harnish R. M. (1996) *Linguistica*, (nuova edizione), Il Mulino, Bologna.

Balbi S. (1998) Lo studio dei messaggi pubblicitari con l'analisi dei dati testuali, in *Linguistica e Statistica: strategie di lettura*, Quaderni del Dipartimento di Scienze Economiche e Statistiche, 1, 155 - 171.

Bolasco S. (1993) Choix de lemmatisation en vue de reconstructions syntagmatiques du texte par l'analyse des correspondances, in *JADT-93*, 399 - 414.

Corduas M. (1995) La struttura dinamica dei dati testuali, in *JADT-95*, 345 - 352.

Harris R. R., Kanji G. K. (1983) On the use of the Minimum Chi-square estimation, *The Statistician*, 32, 379 - 394.

Lebart L., Morineau A. e Piron M. (1995) *Statistique exploratoire multidimensionelle*, Dunod, Paris.

Lebart L., Salem A. (1994) *Statistique textuelle*, Dunod, Paris.

Mardia K. V., Kent J. T. e Bibby J. M. (1979) *Multivariate Analysis*, Academic Press, London.

Piccolo D. (1991) Metodi statistici per l'analisi testuale, *Quaderni di Statistica ed Econometria*, 13, 1 - 32.

Piccolo D. (1998) Statistica e linguistica: le ragioni metodologiche per un incontro scientifico, in *Linguistica e Statistica: strategie di lettura*, Quaderni del Dipartimento di Scienze Economiche e Statistiche, 1, 13 - 23.

Rizzi A. (1985) Alcune analisi statistiche della lingua italiana, *Statistica*, XLV, 1, 7 - 31.

Rizzi A. (1994) La distribuzione delle forme grammaticali nella lingua italiana, *Statistica*, LIV, 3, 275 - 291.

Sichel H. S. (1975) On a distribution law for word frequencies, *Journal of the American Statistical Association*, 70, 543 - 547.

Tabossi P. (1999) *Il linguaggio*, Il Mulino, Bologna.

Titterington D. M., Smith A. F. M. e Makov U. E. (1985) *Statistical Analysis of finite mixtures distributions*, John Wiley & Sons, Chichester.

Witzum D, Rips E. e Rosenberg Y. (1994) Equidistant letter sequences in the book of Genesis, *Statistical Sciences*, 9, 3, 429 - 438.

Zipf G. K. (1935) *The psycobiology of language*, Houghton-Mifflin, Boston.