Università degli Studi di Padova

FACOLTÀ DI SCIENZE STATISTICHE
Corso di Laurea in Scienze Statistiche Demografiche e Sociali

Tesi di Laurea Specialistica

Prendere le distanze

Misure e approfondimenti sul *corpus* EASIEST

Relatore:
Ch.mo Prof. Lorenzo Bernardi

Laureando:
Matteo Passoni

Anno Accademico 2009-2010

Indice

Introduzione

Il lavoro che segue è un approfondimento relativo al progetto EASIEST [1](Espressione Autistica: Studio Interdisciplinare con Elaborazione Statistica e Testuale) in cui sono già stati affrontati studi statistici sui testi prodotti con la Comunicazione Facilitata dai ragazzi autistici, volti prinicipalmente alla verifica dell'originalità dello stile e alla valutazione delle differenze nel costrutto grammaticale, nel lessico, nella semantica con i testi prodotti dai facilitatori o da ragazzi "normali".

L'obiettivo di questo lavoro, basato sui testi prodotti dai tredici ragazzi autistici, costituenti il Gruppo 1 del progetto, è di confrontare gli autori tra di loro, per trovare somiglianze (o verificare le differenze) nel modo di scrivere, seguendo la metodologia proposta negli articoli di *Labbé & Labbé (2001, 2007)*; il cui scopo è proporre una metrica di distanze tra testi scritti basata sul "conteggio" dei vocaboli.

Nel capitolo 2, dopo aver presentato brevemente le caratteristiche principali degli autori e dei testi prodotti, abbiamo calcolato la distanza di Labbé tra autori (paragrafo 2.2.3) e tra perio-

[1]Per conoscere le diverse componenti del progetto: *L. Bernardi (a cura di); Il delta dei significati. Uno studio interdisciplinare sull'espressione autistica, Ed: Carocci, 2008*

di di facilitazione (paragrafo 2.2.4). L'ultimo paragrafo del capitolo (paragrafo 2.2.5) presenta il calcolo di altri due indici di confronto tra testi: l'indice di connessione lessicale e l'indice di indipendenza lessicale.

Nel capitolo 3, a partire dal vocabolario utilizzato da tutti gli autistici, abbiamo costruito 1000 campioni di 1000 parole per analizzare le distribuzioni campionarie delle distanze di Labbé tra coppie di autori, sintetizzate nelle matrici di medie e varianze. L'ultima parte di questo capitolo (paragrafo 3.2) riguarda la "scoperta" di un caso particolare tra gli autori, sulla base delle analisi fin qui fatte.

Il capitolo 4 utilizza lo strumento statistico dell'analisi di raggruppamento per verificare l'eventuale presenza di gruppi (costituiti per somiglianza) e fornisce uno spunto per analisi future più approfondite sui periodi di facilitazione.

I risultati ottenuti, i commenti e tutto questo lavoro esulano da considerazioni nosologiche e terapeutiche, ma vogliono dare un piccolo contributo alla conoscenza di eventuali specificità dell'"universo autistico" per cui non esistono ancora modelli di comprensione o di spiegazione causale, né tantomeno metodi terapeutici o educativi capaci di garantire una "guarigione" da tale disturbo.

Capitolo 1

Autismo e Comunicazione Facilitata: scoprire un nuovo universo

"I beni più grandi ci vengono dalla follia, purché la follia ci sia data per dono divino."

Socrate

1.1 La Sindrome Autistica

Il termine autismo deriva dal greco *autòs* (io stesso) e fu inizialmente introdotto dallo psichiatra svizzero Eugen Bleuler nel 1911 per indicare un sintomo comportamentale della schizofrenia indicante la perdita di contatto con la realtà circostante e la conseguente concentrazione di tutta l'attività mentale sul mondo interiore.

(Paola Venuti, 2003)

1.1.1 Fisiologia

La comunità scientifica internazionale considera l'autismo come la conseguenza di un disturbo cerebrale. In particolare, grazie a numerose ricerche e studi multidisciplinari, è stata messa in evidenza una disfunzionalità nella formazione del reticolo neuronale situato nel tronco encefalico che riceve gli input sensoriali; è l'apparato che condiziona la regolazione e l'equilibrio delle attività del Sistema Nervoso Centrale (SNC) e dei cambiamenti di stato fisio-psicologici.

Il disturbo si costituisce molto prima della nascita del bambino, in una fase in cui il cervello umano è ad uno stadio molto primitivo della sua formazione.

Una volta nato, un "bambino normale" ri-conosce subito la madre attraverso un procedimento innato; successivamente, mediante l'esperienza e il contatto fisico, si proietta nella realtà circostante fatta di persone (altre rispetto alla madre) e oggetti. Nel corso di vita il bambino "diventa grande" imparando a conoscere il mondo esterno attraverso i sensi e provando diverse emozioni nei confronti degli altri.

Per una persona che soffre di autismo, interagire con il mondo non è piacevole e può diventare fonte di angoscia e dolore; per questo motivo è possibile riscontrarne i sintomi già nei primi anni di vita. In particolare, i segnali di tale disfunzione si manifestano con gravi alterazioni in 3 aree:

- della comunicazione verbale e non;
- dell'interazione sociale
- dell'immaginazione o repertorio d'interessi.

L'autismo si trova associato anche ad altri disturbi del SNC come epilessia, sclerosi tuberosa, sindrome di Lett, sindrome di Down, sindrome dell'X fragile, rosolia congenita. L'incidenza è di 6 bambini su 1000, ma in caso di comorbilità[1] si può arrivare ad 1 su 250; cifre che variano ulteriormente a seconda dei criteri diagnostici utilizzati. Questo sottolinea quanto poco si conosce di tale sindrome, oltre al fatto che non sono ancora state trovate spiegazioni scientifiche che motivino la diversa incidenza tra i sessi: 4 a 1 in favore degli uomini.

1.1.2 Sintomatologia

Come detto nel paragrafo precedente, i sintomi di una possibile sindrome autistica si possono vedere subito nei primi anni di vita:

Area della comunicazione

Il bambino utilizza il linguaggio in modo non convenzionale, bizzarro oppure appare muto, ripete frasi o parole sentite da altri (**acolalia**). Nonostante le capacità imitative siano integre, queste persone non riescono a gestire le "imitazioni" in situazioni diverse; spesso vi è una componente di "ritardo mentale".

[1] Per comorbilità s'intende la presenza congiunta di due (o più) disfunzioni, nello stesso individuo.

Interazione sociale

Il soggetto sembra non avere interesse agli altri. Pare proiettato nel suo mondo: evita il contatto visivo, sembra insensibile o iper-eccitabile agli stimoli che vengono da fuori, fatica ad incominciare una conversazione o a rispettare i turni di parola.

Area degli interessi

Di solito vengono ripetuti, al limite dell'ossessione, pochi limitati movimenti; gli autistici possono manifestare un interesse eccessivo per oggetti o parti di essi, in particolare se hanno forme tondeggianti o possono rotolare. Viene sovente riscontrata una resistenza al cambiamento che per alcuni diventa vero e proprio terrore fobico: in questi casi, il soggetto può scoppiare in crisi di pianto o riso; può diventare autolesionista, iperattivo ed aggressivo verso le persone e gli oggetti. Alcuni mostrano invece una totale passività nei confronti degli eventi, tanto che risultano impermeabili a qualsiasi stimolo.

> In sintesi, l'autismo è un disturbo del quale non è ancora chiara l'eziologia e la diagnosi viene effettuata sulla base di indicatori comportamentali secondo la modalità classificatoria utilizzata dai manuali diagnostci DSM IV e ICD 10 in cui il sintomo e il comportamento coincidono.
>
> (Paola Venuti 2003)

Questa citazione permette di giustificare il titolo del capitolo. Si pensi all'universo: quanto conosciamo sulla storia di ogni stella?

Poco più di nulla. Lo stesso vale per la sindrome autistica: la letteratura specializzata si occupa principalmente di fisiologia e sintomatologia. Nel primo caso indaga sulle conseguenze che il disturbo ha sul corpo e sulla mente dell'uomo, nel secondo studia i sintomi, ovvero i segnali che possono far pensare alla presenza (o meno) della sindrome. L'eziologia, la causa primordiale del disturbo, è sconosciuta e difficilmente può essere studiata in modo rigoroso.

Questo progetto si inserisce in un ambito della fisiologia poco esplorato dagli studi sull'autismo; quello **comunicativo**.

1.2 La Comunicazione Facilitata

La comunicazione facilitata, d'ora in poi C.F., è un metodo complesso che, attraverso strumenti (per lo più *ausili tecnici*) e persone formate e competenti (*i facilitatori*) consente un giusto ed equo rapporto tra facilitato e facilitatore.

Il suo utilizzo, mediante un intervento educativo graduale, permette a persone con problemi di comunicazione di esprimere il proprio pensiero, altrimenti bloccato a causa di una comunicazione verbale *atipica*.

L'obiettivo della C.F. è quindi quello di rendere via via più autonomo il soggetto autistico nell'esprimere i propri pensieri, proponendo un rapporto tra facilitatore e facilitato volto ad un progressivo distacco. La forte interazione all'inizio del "trattamento" è comunque indispensabile per aiutare l'autistico a prendere confidenza con questo mondo, per lui ancora sconosciuto.

1.2.1 Il metodo

Noi, esseri umani *tipici*, siamo portati a dare per scontato che la realtà esterna sia uguale per tutti, in particolare per coloro che non presentano alcun tipo di disturbo psico–fisico. Per chi lavora con la C.F. questo non è un dato di fatto, non è la *normalità*: di ogni "allievo" che inizia a rapportarsi con la C.F. è fondamentale capire *chi è e come "funziona"*, ma anche tentare di ipotizzare la mappa del *suo mondo e dei suoi pensieri*. Solo conoscendolo in profondità si può costruire un progetto *ad hoc* evitando stimoli fuori misura, tenendo sotto controllo i progressi e le informazioni acquisite nel tempo; solo così si può entrare in confidenza e *condividere* con lui tempi e modi di assimilazione delle informazioni.

L'utilizzo del computer sembra essere d'aiuto per imparare il processo comunicativo; è il punto di arrivo di un cammino che parte dal riconoscimento delle lettere su una tastiera di carta e il punto di partenza per il *dialogo* con il facilitatore. Le informazioni scritte, in particolare le domande, se sono visualizzate sullo schermo del computer vengono capite e interpretate in modo migliore rispetto a quelle poste verbalmente (probabilmente per una decifrazione errata del tono e della cadenza). Vedere ciò che si scrive permette un ulteriore controllo e una verifica nei confronti del pensiero che si vuole esprimere. L'utilizzo di questi strumenti impone lentezza, pazienza, ritmo e cadenza concedendo il tempo necessario per l'attuazione corretta del processo *pensiero-movimento*.

Le persone autistiche subiscono forti impulsi emotivi che sfociano in ansia e irrequietezza; l'utilizzo di intermediari oggettivi – il computer – permette di ridurre l'emotività individuale degli interlocutori, uno dei principali freni per l'interazione con soggetti autistici.

1.2.2 I facilitatori

Il Facilitatore, partendo con un contatto fisico, ponendo la mano su quella dell'allievo, passando poi al gomito, alla spalla, alla testa e, infine, solamete con la sua presenza, consente al facilitato di superare le difficoltà del processo comunicativo. Attraverso la scrittura e tramite il riconoscimento di immagini, l'allievo comunica ciò che pensa e le scelte che compie.
Il facilitatore ha diverse funzioni:

- offre un *supporto fisico*; aiuta il soggetto ad isolare ed estendere il dito indice, a controllare il movimento del puntare il dito e a ritirare la mano dopo ogni selezione. Consente al facilitato di superare difficoltà fisiche specifiche come la coordinazione (occhio-pensiero-mano) o l'irregolare tono muscolare che, in alcuni casi, risulta essre o troppo alto o troppo basso.

- garantisce la perseveranza nel portare a termine un compito dato fornendo un controllo sull'impulsività (Crossley 1990)

- offre un *supporto emotivo*: questo è fondamentale per instaurare una relazione di fiducia, nella quale è più facile che si sviluppi la comunicazione; ne è una prova il fatto che le produzioni diminuiscano in termini di *qualità* quando cambia il facilitatore.

Basilare rimane comunque la capacità di lasciarsi andare serenamente alla scoperta della storia, dei vissuti, delle emozioni e della realtà di precedenti apprendimenti sconosciuti ed insondabili della persona *atipica*.

1.3 Dove e come procedere?

A partire dalla diversità di opinioni e dalle controversie che riguardano la comunicazione di soggetti autistici attraverso l'uso della C.F. (Green 1994 - Skeptic, v.2, n.3: pag. 68-76; Jakobson et al. 1995 - American Psycologist, v.50, n.9: pag. 750-765); questo progetto si propone di studiare la produzione dei testi dal punto di vista statistico-linguistico; un approccio che non entra nel merito del dibattito sulla correttezza (o meno) del metodo della C.F., perché ne utilizza il "prodotto finale": le parole in sé.

1.3.1 Il progetto

EASIEST è un acronimo: Espressione Autistica Studio Interdisciplinare con Elaborazione Statistico-Testuale.

Nel capitolo precedente si è parlato del funzionamento della comunicazione facilitata e del modo corretto di utilizzo; con questa metodologia sono stati raccolti testi di diverso tipo a seconda degli obiettivi raggiungibili,del grado di facilitazione necessario e del tempo trascorso dall'inizio dell'uso della C.F.:

- **Copiare, Nominare** (Es: "Scrivi Albero"; "Che cos'è questo?" di fronte ad un'immagine);
- **Scelta multipla, Domande chiuse, Completamento** (Es: "Vuoi giocare a carte o con la palla?"; "Sai in che anno è iniziata la Seconda Guerra Mondiale?"; "Per mangiare si usano le...."");

- **Domande su contesti noti** conosciuti dal facilitatore e quindi con una gamma di risposte limitata (Es: “Cosa hai mangiato a pranzo?”);

- **Conversazione aperta** (Es: “Di cosa vuoi parlare oggi?”).

Per ottenere degli elaborati studiabili dal punto di vista statistico, il progetto ha preliminarmente provveduto alla formazione dei facilitatori e alla preparazione dei facilitati. Questo punto è sottoposto a maggior critica da chi è in disaccordo con l’utilizzo della C.F. nella restituzione di **testi propriamente comunicativi**; infatti, perchè questa fase sia metodologicamente corretta e l’effetto facilitatore sia nullo, il dialogo sarebbe dovuto avvenire tra i ragazzi autistici e “sconosciuti” adeguatamente formati.
Data la complessità e l’eterogeneità di comportamenti, reazioni e atteggiamenti dei soggetti interessati si è preferito mantenere un protocollo di lavoro il più possibile rigido, mantenendo i facilitatori abituali.

1.3.2 Il protocollo

Si elencano di seguito i punti fermi su cui si basa tutto il progetto:

- **Soggetti**: ragazzi con Disturbo Generalizzato dello Sviluppo (anche in presenza di comorbilità)

- **Modalità di selezione**: tutti i soggetti dovevano avere esperienze di facilitazone con almeno tre facilitatori differenti. Coloro che avevano comunicato con meno di tre, o di cui non è stato possibile recuperare i testi prodotti, non sono stati considerati.

- **Il campione**: Gruppo 1 (nella totalità del progetto sono stati formati 3 gruppi), formato dai soggetti che hanno prodotto testi di *alta qualità* dal punto di vista dei contenuti, delle abilità linguistiche, della lunghezza e complessità. Di questi si conosce tutta la storia dalle prime esperienze di comunicazione facilitata in poi (apprendimento e miglioramento fino all'autonomia).

- **Criterio temporale di selezione dei testi**: l'esperienza espressa in anni di pratica è stata suddivisa in 5 fasce; primo semstre, secondo semestre, secondo anno, terzo anno, oltre il terzo anno.

- **Numerosità dei testi per soggetto**: per ogni fascia temporale dovevano essere fornite un numero di sedute tale da ricoprire 15 pagine standard di testo, per arrivare ad un totale di 70-80 pagine per soggetto[2].

- **Scelta degli intermediari**: la produzione è stata sviluppata totalmente grazie a supporti informatici quali il PC o la macchina da scrivere elettronica.

- **Preferenza dei testi**: per poter valutare oggettivamente la *non influenza* del facilitatore, sono stati scelti i testi scritti dai soggetti che avevano raggiunto il massimo livello di autonomia all'interno del processo comunicativo.

[2]Questo criterio poteva non essere rispettato per le prime due fasce, dove la produzione è stata necessariamente più limitata, ma si è cercato comunque di avvicinarsi alle dimensioni richieste per poter apprezzare nel tempo l'evoluzione del linguaggio.

- **Scelta delle coppie facilitato/facilitatore**: riguarda la fluidità della comunicazione, al fine di sviluppare dialoghi ”ritmati” e continuativi nella coppia; per questo si è tenuto conto di quelli avvenuti con un parente stretto, come la madre e/o il padre.

- **Qualità delle sedute**: sono state scelte le sedute in base al contenuto e non al numero di parole scritte. Il materiale testuale prodotto dal soggetto ha contribuito alla produzione del *vocabolario* del linguaggio autistico; pertanto sedute troppo brevi non sono state prese in considerazione. Gli scritti del facilitatore sono importanti e servono per confutare o affievolire l'idea che, in questo tipo di comunicazione, il testo sia influenzato e non prodotto autonomamente dal soggetto autistico.

Tenendo conto di queste scelte metodologiche-progettuali volte a rendere la raccolta e la quantità dei dati *omogenea*, si è proceduto alla formazione dei facilitatori attraverso un protocollo di lavoro.

Capitolo 2

Analisi Esplorative

2.1 Il Gruppo 1

Come detto nel paragrafo 1.3.2 in questo lavoro si tiene conto solamente della produzione scritta dei ragazzi apparteneti al Gruppo 1 del progetto.

In questo campione sono rientrati 13 giovani autistici, 12 ragazzi e 1 ragazza, provenienti da tre centri diversi che hanno partecipato al progetto EASIEST. La tabella 2.1 mostra l'anno di nascita, la provenienza e il genere dei soggetti: come si può notare, in alcuni casi gli anni di nascita sono molto differenti tra loro; situazione da tenere sempre presente nella lettura dei risultati che seguiranno, ricordando anche che si tratta di uno studio di *fattibilità* per l'analisi della produzione scritta.

Soggetto	Provenienza	Genere	Anno	Soggetto	Provenienza	Genere	Anno
AF	Genova	Maschio	1978	LB	Genova	Maschio	1987
AN	Genova	Maschio	1989	LP	Roma	Maschio	1990
CM	Roma	Maschio	1974	MO	Genova	Maschio	1989
DDL	Genova	Maschio	1996	MV	Genova	Maschio	1992
DR	Genova	Femmina	1996	OP	Genova	Maschio	1985
DV	Genova	Maschio	1987	PCM	Padova	Maschio	1981
FP	Genova	Maschio	1993				

Tabella 2.1: Gruppo 1

Per capire meglio come avviene il processo di facilitazione spiegato nel par 1.2.2, si riporta l'andamento nel tempo del livello di facilitazione per i diversi soggetti (Tabella 2.2).

Dalla tabella si può notare come da un livello massimo di facilitazione (contatto mano su mano) si passa ad uno inferiore (mano-spalla, mano-schiena) fino ad arrivare alla quasi totale autonomia del soggetto: sembra, allora, esserci un miglioramento nelle capacità relazionali.

I ragazzi imparano *l'arte della comunicazione*?

A questo non si può ancora rispondere. Solamente 2 su 13 raggiungono l'autonomia completa; tuttavia in tutti si nota un progressivo miglioramento e quindi una riduzione del contatto fisico indice del livello di facilitazione. Sembra che i ragazzi imparino a riconoscere e a reagire agli stimoli propri dell'interazione sociale e comunicativa.

Soggetto	I semestre	II semestre	II anno	III anno	oltre III anno
AF	Massimo	Massimo	Massimo	Medio	Medio
AN	Massimo	Medio	Medio	Medio	Basso
CM	Medio	Medio	Medio	Medio	Medio
DDL	Massimo	Massimo	Medio	Medio	Basso
DR	Massimo	Medio	Basso	Basso	Autonomo
DV	NNN [1]	NNN	Basso	Basso	Basso
FP	Massimo	Medio	Basso	Basso	Basso
LB	Massimo	Medio	Medio	Basso	Basso
LP	Medio	Medio	Medio	Medio	Basso
MO	Medio	Basso	Basso	Medio	Basso
MV	Medio	Basso	Medio	Basso	Basso
OP	Massimo	Massimo	Medio	Basso	Autonomo
PCM	Massimo	Massimo	Medio	Basso	Basso

Tabella 2.2: Livello di facilitazione per periodo

2.2 Analisi della produzione scritta

2.2.1 Il Corpus

Come detto in precedenza, la popolazione oggetto di analisi è formata dagli scritti dei ragazzi autistici, considerati sia nella totalità (tutta la produzione di ogni individuo), sia divisi per periodo di facilitazione, a seconda del tipo di analisi effettuata.

Prima di essere analizzato, il corpus è stato *pulito* da tutti i simboli di interpunzione, sono stati rimossi gli spazi in eccesso e sostuiti gli accenti, non riconosciuti dal programma utilizzato (**R**), con un raddoppiamento della vocale (*es: perché* diventa *perchee*, *papà—papaa*).

Soggetto	Parole	Soggetto	Parole
AF	4285	LB	2287
AN	2977	LP	5446
CM	6258	MO	3174
DDL	4629	MV	2664
DR	3077	OP	4144
DV	2350	PCM	5527
FP	1760		

Tabella 2.3: Numero di parole scritte

Nella tabella 2.3 sono riportate le dimensioni dei corpus di ogni singolo autistico; si nota subito che hanno lunghezze molto differenti tra loro, dal minimo di 1760 parole al massimo di 6258. Va precisato che per il soggetto che ha usato il numero più basso di parole, FP, sono

state riportate solamente le conversazioni fino al terzo anno di facilitazione: non compaiono le conversazioni relative all'ultimo periodo (oltre il terzo anno).

L'ampiezza del range delle produzioni può essere imputata alla natura del campione utilizzato: età, provenienza e tipologia di disturbo autistico sono diversi per ogni soggetto. Vent'anni di differenza (come nel caso di CM con FP) si riflettono sia nella lunghezza, sia nei contenuti dei dialoghi con i rispettivi facilitatori: il vocabolario è più ampio per il soggetto più vecchio e, di conseguenza, gli stessi contenuti sono diversi.

2.2.2 La Term Document Matrix

A partire dai corpus di ogni singolo autistico si è costruito il vocabolario, formato da tutte le 9756 parole scritte almeno una volta. Da questo è stata costruita una matrice, chiamata **Term Document Matrix (TdM)**, che ha sulle *i–righe* le parole del vocabolario in ordine lessicografico e sulle *j–colonne* i 13 ragazzi autistici (gli autori). In ogni cella è contenuto il numero di volte che quella parola è stata usata da ogni autistico (vedi Tabella 2.4).

La TdM è uno strumento molto duttile ed efficace per l'analisi testuale perché contiene tutte le informazioni presenti nei testi; è l'oggetto principale da cui partire per qualsiasi studio sul corpus. Ogni elemento della matrice rappresenta la frequenza assoluta di ogni parola in ogni autore (da cui possiamo ricavare hapax, dislegomena, ecc.); sulle righe possiamo individuare gli eventuali errori di scrittura e chi li ha commessi; confrontando le colonne (i corpus individuali) ricaviamo i profili lessicali di ogni autore.

Il profilo lessicale altro non è che l'insieme delle parole utilizzate da un autore; il confronto tra

Parole	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
ordino	0	0	0	0	0	1	0	0	0	0	0	0	1
ore	1	0	0	0	1	1	0	0	2	0	0	0	0
orecchi	0	0	0	0	0	0	0	0	1	0	0	0	0
orecchie	0	0	1	6	4	0	0	0	0	1	1	0	0
orecchio	1	0	4	0	0	0	1	0	2	0	0	0	0
organi	0	0	0	0	0	0	0	0	0	1	0	0	0

Tabella 2.4: Term Document Matrix

due profili si basa sulla frequenza assoluta di ogni parola presente in entrambi i testi: l'obiettivo di questa procedura è calcolare una misura di *distanza* che dia un'idea della somiglianza tra i testi considerati.

2.2.3 La distanza inter-testuale di Labbé

L'obiettivo appena descritto di fornire una misura della somiglianza, si risolve ricorrendo ad una metrica con determinate proprietà:

- invarianza rispetto alla lunghezza dei testi comparati
- adattabilità a molti testi
- i valori devono essere compresi tra 0 (i due testi hanno lo stesso vocabolario e la stessa frequenza nei type) e 1 (non ci sono type in comune)
- Dati due testi A e B, la distanza δ è simmetrica: $\delta_{(A,B)} = \delta_{(B,A)}$

- "robustezza"(per quanto possibile): ad un piccolo cambiamento in uno dei due testi corrisponde un piccolo cambiamento nella distanza

Dati due testi A e B, chiamiamo

V_A e V_B il numero di types nei testi A e B (il vocabolario)

F_{iA} la frequenza dell' i-esimo type in A

F_{iB} la frequenza dell'i-esimo type in B

N_A e N_B il numero di tokens nei due testi (la lunghezza totale di A e B), con

$N_A = \sum_{V_A} F_{iA}$ e $N_B = \sum_{V_B} F_{iB}$.

La distanza relativa viene calcolata come:

$$\delta_{(A,B)} = \frac{\sum_{i \in V_A} |F_{iA} - F_{iB}| + \sum_{i \in V_B} |F_{iB} - F_{iA}|}{N_A + N_B} \, . \tag{2.1}$$

Nel caso in cui i testi siano completamente diversi per i type usati il risultato è 1 (anche in caso di lunghezze molto diverse), tuttavia il minimo teorico 0 viene raggiunto solo nel caso di uguale lunghezza dei due testi. A questo si aggiunge il fatto che in 2.1 l'intersezione viene contata due volte, dando molta importanza ai type in comune piuttosto che a quelli specifici. Labbé e Labbé, nel loro articolo del 2001 [2], propongono una modifica all'impianto metodologico alla base di 2.1 proprio per dare una soluzione al problema della lunghezza dei due testi e al raggiungimento del minimo teorico. Dati A e B, supponiamo $N_A \leq N_B$, la frequenza di ogni i–esimo type presente nel testo più lungo (F_{iB}) viene ridotta sulla base della grandezza del testo più corto.

[2]Per ulteriori approfondimenti sulla distanza intertestuale: *Brunet(1988)*, *Labbé e Labbé (2001)*, *Merriam (2002)*, *Labbé (2007)*.

La stima F^*_{iB} si ottiene sulla base della proporzizone per cui

$$F^*_{iB} : F_{iB} = N_A : N_B \Rightarrow F^*_{iB} = F_{iB}\frac{N_A}{N_B} \tag{2.2}$$

quindi, $N_A = N^*_B = \sum_{V_B} F^*_{iB}$.

Ora si può sostituire F_{iB} con F^*_{iB} e N_B con N^*_B nella 2.1; la *nuova misura* raggiunge lo 0 quando tutti i type di a sono presenti in b con frequenza $F_{iA} = F^*_{iB}$, cioè quando il testo più corto è una specie di *modello* di quello più lungo.

Risolto l'inconveniente della lunghezza con la *riscalatura* appena descritta, gli autori considerano il problema dei type in comune tra i due testi, calcolando la distanza assoluta in due passi: per primi i V_A types (vengono contati una volta anche quelli in comune), successivamente solo i V^*_B, quelli relativi al solo testo riscalato, in cui $F_{iA} = 0$. La distanza assoluta risulta quindi

$$d_{V_{A,B^*}} = \sum_{V_A, V^*_B} |F_{iA} - F^*_{iB}| \ ;$$

quando A e B non hanno type in comune è uguale alla somma dei tokens nei due testi – $N_A + N^*_B$ –; la distanza relativa raggiunge il massimo teorico, 1. Nel caso in cui i due testi abbiano type in comune il risultato ha un valore compreso tra 0 e 1, calcolato come:

$$d_{(A,B)} = \frac{\sum_{V_A, V^*_B} |F_{iA} - F^*_{iB}|}{\sum_{V_A} F_{iA} + \sum_{V_B} F^*_{iB}} = \frac{\sum_{V_A, V^*_B} |F_{iA} - F^*_{iB}|}{N_A + N^*_B} \ . \tag{2.3}$$

Si noti che le F_{iA} sono numeri interi, mentre le F^*_{iB} – essendo delle *stime* – includono valori decimali che influiscono nel calcolo della distanza; a questo proposito si può aggiustare il valore della 2.3 considerando solo le parole la cui frequenza stimata è maggiore di uno.

La **soglia** ci permette di applicare il calcolo della distanza solamente alle parole del testo più lungo (B) che hanno frequenza tale da comparire almeno una volta nel testo riscalato; la 2.3

verrebbe aggiornata sulla base della condizione $F^*_{iB} \geq 1$.

Esempio 1:

Prendiamo i primi due autori esaminati, AF e AN; dalla TdM calcoliamo:

- $N_{AN} = 2977$, il numero di type utilizzati da AN;
- $N_{AF} = 4285$ il numero di type di AF;
- $F_{i,AN}$ e $F_{i,AF}$ le frequenze dei types nei rispettivi testi;
- $F^*_{i,AF}$ le frequenze stimate nella riscalatura del corpus.

Possiamo vedere il risultato della riscalatura del testo più lungo nella Tabella 2.5:

	$F_{i,AN}$	$F_{i,AF}$	$F^*_{i,AF}$
trovare	2	7	4.86
trovo	5	2	1.39
tua	2	5	3.47
tutti	11	8	5.56
tutto	5	3	2.08
una	25	29	20.15
volta	5	3	2.08
vorrei	3	2	1.39

Tabella 2.5: Esempio di “riscalatura” del corpus più lungo

i valori nella terza colonna sono le stime calcolate moltiplicando il corrispondente valore di $F_{i,AF}$ per la costante di normalizzazione $N_{AN}/N_{AF} = 0.695$; la distanza di Labbé calcolata per questi due testi è pari a 0.76, se si usa la 2.3 senza alcuna soglia d'inclusione per le $F^*_{i,AF}$. Nel caso imponessimo il calcolo solo alle parole con $F^*_{i,AF} \geq 1$ la distanza sarebbe inferiore e pari a 0.602, la lettura di questo dato è molto semplice: i due testi condividono il 40% delle parole utilizzate mentre, nel caso senza alcuna soglia, solo il 34%.

Data la scarsa numerosità campionaria a disposizione e la lunghezza dei corpus abbastanza ridotta rispetto a quella su cui la letteratura ha applicato questa metodologia – studi su novelle e romanzi di autori storici con un numero di type ben più elevato –, utilizzare un livello di soglia per escludere alcune parole ridurrebbe ulteriormente la dimensione dei corpus in esame.

Nell'esempio precedente il corpus AF, una volta riscalato con la 2.2, ha la stessa dimensione di quello più corto (2977 types); se considerassimo le parole che *effettivamente* costituirebbero il nuovo testo, quelle con frequeza attesa ($F^*_{i,AF} \geq 1$), avremmo un testo di 1945.298 types con una perdita di 1031.702 parole (sono cifre decimali perché stiamo ragionando sulle frequenze attese).

Calcolando la 2.3 con la soglia, rischiamo di eliminare dal conteggio gli hapax (parole che compaiono una volta sola) del corpus più grande; questo non comporterebbe alcun problema se i testi fossero abbastanza lunghi e la percentuale sul totale non fosse alta, come invece succede nel nostro campione: il range di variazione è compreso tra il 20% e il 40% e, se si considera che le lunghezze vanno da 1760 a 6258 (Tabella 2.3), la proporzione è troppo elevata.

La relazione tra hapax e lunghezza del corpus è rappresentata nella Figura 2.1, in cui si identificano due gruppi abbastanza distinti: uno formato dai corpus di lunghezza compresa tra i 2000

e i 3000 types e percentuale di hapax intorno al 30%, l'altro dai testi più lunghi e proporzione minore.

Due autori si discostano notevolmente dal resto del campione: si tratta di FP (ma l'elevato numero di hapax può essere imputato all'esigua lunghezza del suo testo) e AF che, pur avendo un corpus abbastanza lungo, ha una proporzione di hapax molto alta.

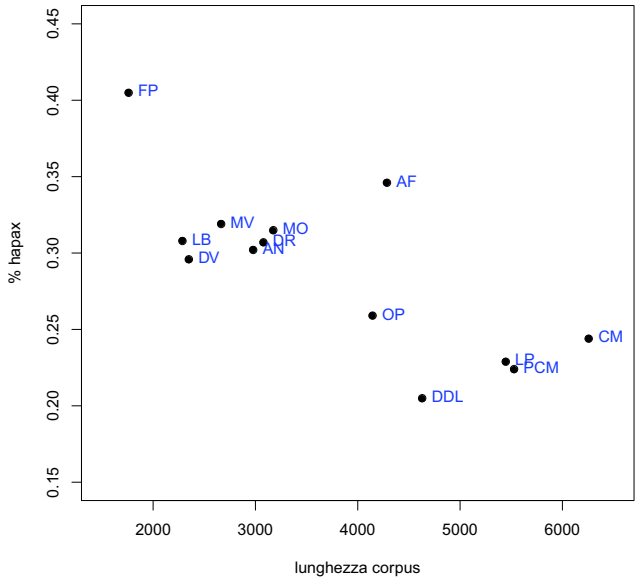

Figura 2.1: Percentuale hapax nei corpus

La maggior parte dei testi analizzati è formata da hapax, come se ogni parola fosse pesata e pregna di significato all'interno della frase.

Quindi, per rispettare questa particolarità, la distanza tra i testi è stata calcolata senza considerare alcun livello di soglia per la frequenza attesa dei type e, partendo dalle colonne della TdM, abbiamo ottenuto la matrice delle distanze di Labbé tra coppie di autori (Tabella 2.6), che risulta essere:

- quadrata, di ordine 13x13
- triangolare, infatti $d(a, b) = d(b, a)$
- ha traccia nulla, perché $d(a, a) = 0$.

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
AF	0												
AN	0.76	0											
CM	0.67	0.62	0										
DDL	0.74	0.62	0.56	0									
DR	0.73	0.62	0.55	0.57	0								
DV	0.73	0.64	0.58	0.61	0.62	0							
FP	0.75	0.65	0.61	0.62	0.63	0.65	0						
LB	0.75	0.59	0.58	0.57	0.60	0.62	0.59	0					
LP	0.74	0.63	0.54	0.60	0.60	0.65	0.65	0.58	0				
MO	0.72	0.61	0.55	0.56	0.56	0.61	0.60	0.55	0.58	0			
MV	0.76	0.63	0.59	0.63	0.62	0.63	0.65	0.58	0.60	0.60	0		
OP	0.74	0.66	0.57	0.59	0.61	0.63	0.64	0.61	0.64	0.59	0.62	0	
PCM	0.73	0.62	0.52	0.55	0.57	0.60	0.62	0.55	0.56	0.53	0.57	0.59	0

Tabella 2.6: Matrice delle distanze di Labbé tra coppie di autistici

Le distanze tra profili lessicali sono molto elevate e questo si deve al fatto che stiamo analizzando parole utilizzate da persone diverse[3]; tuttavia, i valori relativi ad AF risultano molto

[3]Nell'articolo di *Labbé & Labbé (2001)*, viene presentata una scala standardizzata per la distanza intertesuale dove il valore 0.65 rappresenta la distanza massima per testi scritti nello stesso linguaggio da autori diversi.

elevati: la media delle sue distanze è 0.74, valore molto più grande rispetto agli altri che variano tra 0.58 e 0.64 (Tabella 2.7).

AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
0.74	0.64	0.58	0.60	0.61	0.63	0.64	0.60	0.61	0.59	0.62	0.62	0.59

Tabella 2.7: Media delle distanze tra un autistico e gli altri

La proporzione di hapax presente in ogni corpus è una causa delle differenze così marcate. Infatti, sembra esserci una correlazione tra distanza media e proporzione di parole usate una volta sola (Figura 2.2); l'unico autore per cui non sembra valere questo ragionamento è AF, che si comporta come un outlier.

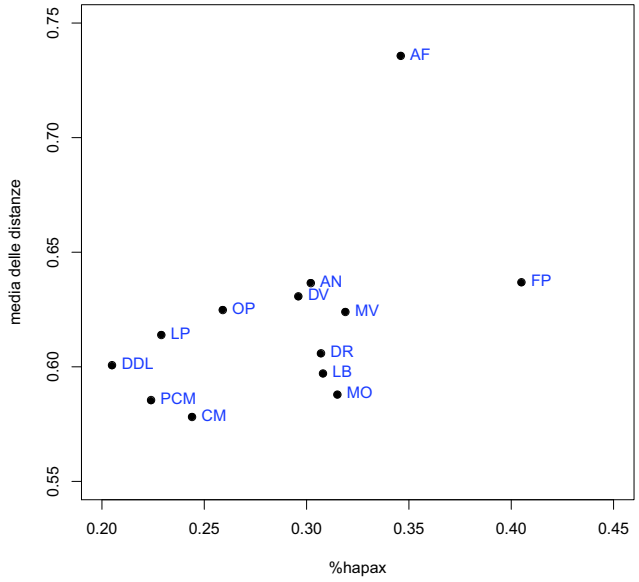

Figura 2.2: Relazione tra distanza media e percentuale di hapax

Ulteriore conferma a queste ultime osservazioni si trova incrociando i dati relativi alla lunghezza dei testi con la distanza media: sembra delinearsi una relazione inversa tra le due quantità e, anche in questo grafico, la produzione scritta di AF risulta essere un punto a parte rispetto agli altri (Figura 2.3).

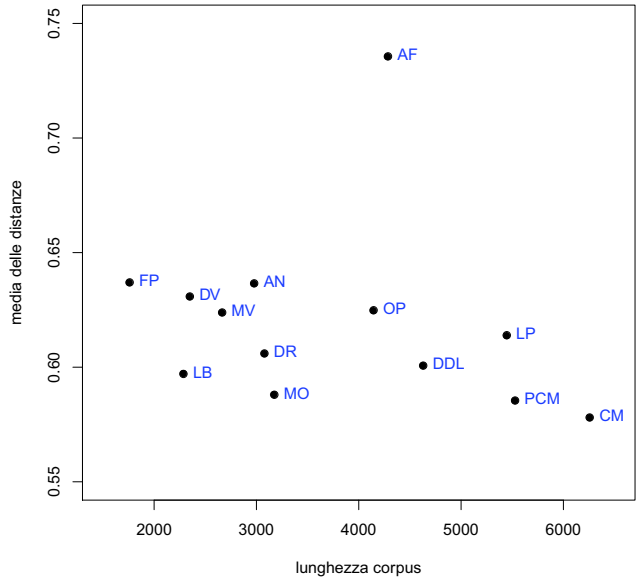

Figura 2.3: Relazione tra distanza media e lunghezza dei corpus

2.2.4 I periodi di facilitazione

Dai dati in nostro possesso è stato possibile recuperare gli scritti relativi ai periodi di facilitazione a cui i ragazzi si sono sottoposti (Tabella 2.2) e costruire una TdM che sulle colonne (profili lessicali) riporta i diversi momenti per ogni autistico. Ogni autore ha cinque profili lessicali diversi (tranne FP, di cui mancano gli scritti relativi all'ultimo periodo) che rappresentano il "cammino" individuale nell'uso della comunicazione facilitata; una sorta di evoluzione in termini di uso di parole, pensieri e scrittura.

La tabella 2.8 riporta le distanze intertestuali, calcolate con la 2.3, tra i testi prodotti nei diversi periodi da un solo autistico: è un blocco di dimensioni 5×5 ottenuto dalla matrice contenente le distanze di Labbé calcolate per ogni periodo di ogni autore (di dimensioni 64×64) da cui otteniamo, sezionando lungo la diagonale, 12 matrici quadrate di dimensione 5×5 e una 4×4 (FP).

	LB1sem	LB2sem	LB2year	LB3year	LB>3year
LB1sem	0				
LB2sem	0.64	0			
LB2year	0.67	0.61	0		
LB3year	0.70	0.65	0.66	0	
LB>3year	0.77	0.73	0.74	0.77	0

Tabella 2.8: Matrice di distanze intertestuali per periodi di facilitazione, autistico LB

In questo modo possiamo valutare, in termini di distanza intertestuale, come si differenzia il modo di scrivere nell'arco temporale della facilitazione. Perché ci sia un apprendimento da parte dell'autistico dovremmo trovare una relazione diretta tra valori delle distanze e periodi di tempo; piccoli tra tempi adiacenti, elevati tra tempi distanti.

Quest'ipotesi deriva direttamente dalla rigida gradualità del protocollo a cui sono sottoposti i facilitatori per quanto riguarda sia l'insegnare a scrivere, sia le domande poste al facilitato (vedi il paragrafo 1.3.1); tuttavia le differenze dipendono anche dall'individualità stessa degli autori che, per la prima volta, vengono messi di fronte alla possibilità di poter comunicare.

Nel nostro caso, pare non esserci alcuna relazione tra periodo di facilitazione e lunghezza del testo scritto: la tabella 2.9 mette in evidenza la variabilità della dimensione dei corpus, in-

dipendentemente dal tempo trascorso, come se fossero altre variabili ad influenzare il flusso comunicativo degli autistici.
Oltretutto, in termini di calcolo delle distanze, le lunghezze dei testi nei diversi periodi non sono adatte per il calcolo della 2.3: stiamo confrontando testi molto al di sotto della soglia consigliata dalla letteratura (Tabella 2.9), situazione che crea valori distorti delle distanze intertestuali. Allora la tabella 2.8 va letta in un'ottica puramente indicativa, considerando che il valore della distanza tra gli ultimi due periodi (0.77) è dovuto alla differenza di lunghezza tra i due testi, 120 contro 419 parole.

Maggiore è la differenza nella lunghezza tra i due testi in considerazione, maggiore è il valore della distanza intertestuale. Ad esempio, per gli autori che hanno il primo testo al di sotto delle 100 parole le distanze tra i periodi raggiungono valori troppo elevati (≥ 0.80), come se i corpus fossero scritti in lingua diversa (vedi Tabella 2.10).

Bisogna quindi prendere con cautela l'analisi per periodi di facilitazione; le lunghezze dei testi estremamente variabili portano distorsioni nel calcolo delle distanze, ottenendo valori che non rispecchiano la realtà: un valore come quello tra il primo semestre e il terzo anno di facilitazione per DDL (0.96) indica che i due testi condividono solamente il 4% dei types utilizzati, un valore come quello precedente si trova nel confronto tra testi scritti in due lingue diverse (in letteratura la soglia minima per testi scritti nello stesso linguaggio è il 35%, che corrisponde ad una distanza pari a 0.65). Di conseguenza, anche tenendo conto degli eventuali errori, del disturbo di cui soffrono gli autori, della difficoltà nell'imparare a comunicare attraverso un computer, i valori ottenuti sono totalmente influenzati dalle diverse lunghezze dei testi; risulta difficile quindi poter descrivere in modo corretto l'andamento delle distanze nel tempo.

	periodo facilitazione				
	1 sem	2 sem	2 anno	3 anno	>3 anno
AF	921	570	595	756	1443
AN	260	598	529	1040	550
CM	1382	1191	1174	1109	1402
DDL	39	154	1386	1694	1356
DR	226	331	788	985	750
DV	313	762	233	459	583
FP	634	147	704	276	634
LB	598	559	591	419	120
LP	846	1242	920	1148	1290
MO	78	614	241	541	1700
MV	115	146	147	651	1605
OP	76	527	1999	604	949
PCM	645	842	1289	1381	1369

Tabella 2.9: Lunghezza dei testi per periodo di facilitazione

	1 sem	2 sem	2 anno	3 anno	oltre 3 anno
DDL1sem	0.00	0.86	0.97	0.96	0.96
MO1sem	0.00	0.85	0.80	0.82	0.86
OP1sem	0.00	0.87	0.87	0.90	0.90

Tabella 2.10: Distanze tra periodi. Autori che nel primo periodo hanno scritto meno di 100 parole

2.2.5 Indici di confronto tra testi

Oltre alla *distanza di Labbé* possiamo calcolare altri indici di somiglianza tra due testi: l'**indice di connessione lessicale** e l'**indice di indipendenza lessicale**.
Queste due misure sono più *grezze* rispetto a quella proposta da Labbé: si basano, infatti, sul confronto tra i vocabolari utilizzati e non sulle frequenze assolute delle parole; verificando quali e quanti sono i types presenti in entrambi i testi o in uno solo dei due.

Indice di connessione lessicale

Questo indice è stato proposto per risolvere il problema di attribuzione di un'opera anonima ad un autore noto, basandosi sul vocabolario comune tra i testi.
Prendiamo due testi A e B con il rispettivo vocabolario V_A e V_B; la parte comune ai due scritti si indica con $V_{A\cap B}$, il vocabolario totale (cioè del corpus che unisce i due testi) è $V_{A\cup B}$ e i vocabolari propri di A e B vengono indicati rispettivamente come $V_{A\cap\bar{B}}$ e $V_{\bar{A}\cap B}$ ($\bar{A}$ rappresenta, infatti, le parole che non sono presenti nel testo A). L'indice di connessione lessicale (**C**) corrisponde al rapporto tra la parte comune e il totale del vocabolario:

$$C_{V_{A,B}} = \frac{V_{A\cap B}}{V_{A\cup B}}. \tag{2.4}$$

Ovviamente, varia tra 0 e 1: nel primo caso ci troviamo di fronte a due testi completamente differenti – nessuna parola in comune –, nel secondo i due testi sono assolutamente identici. In questo contesto, allora, $C_{V_{A,B}}$ rappresenta la percentuale di parole in comune tra due autori; un'ulteriore strumento di verifica per risultati fin qui ottenuti. La tabella A.1 in Appendice,

contiene il valore dell'indice calcolato sia tra coppie di autori, sia tra uno e tutti gli altri (è riportato solo un valore per coppia, infatti $C_{V_{A,B}} = C_{V_{B,A}}$); in entrambi i casi la connessione tra i testi è molto bassa. Gli autori utilizzano vocabolari molto diversi e mediamente non arrivano a condividere 20 parole su 100 utilizzate: ognuno ha un modo differente di esprimersi, con le proprie particolarità che incidono in grande misura nel confronto tra coppie.

Per eliminare l'influenza dell'"individualità espressiva" nel calcolo dell'indice, abbiamo eliminato gli hapax che, comparendo una volta sola, rappresentano le specificità dei corpus esaminati. I risultati sono riportati per intero in Appendice (Tabella A.3), qui sotto riportiamo il valore medio per ogni autore, con e senza hapax:

AUT	hap	no hap	AUT	hap	no hap
AF	0.131	0.157	LB	0.165	0.217
AN	0.154	0.212	LP	0.166	0.212
CM	0.171	0.231	MO	0.165	0.232
DDL	0.179	0.231	MV	0.158	0.212
DR	0.163	0.210	OP	0.162	0.214
DV	0.168	0.215	PCM	0.171	0.240
FP	0.146	0.197			

Tabella 2.11: Media dell'indice di Connessione Lessicale con e senza hapax, tra coppie, per autore

La crescita dell'indice è dovuta all'aumento della proporzione di parole in comune sul totale; eliminando gli hapax abbiamo reso "meno particolari" i testi analizzati, ma la connessione rimane bassa e il vocabolario comune non raggiunge il 30% di quello totale, in tutte le coppie di autori.

La figura qui sotto rende più immediata la verifica di quanto appena detto, oltre a risaltare la particolarità rappresentata da AF che, in entrambi i casi, utilizza un vocabolario molto diverso dagli altri.

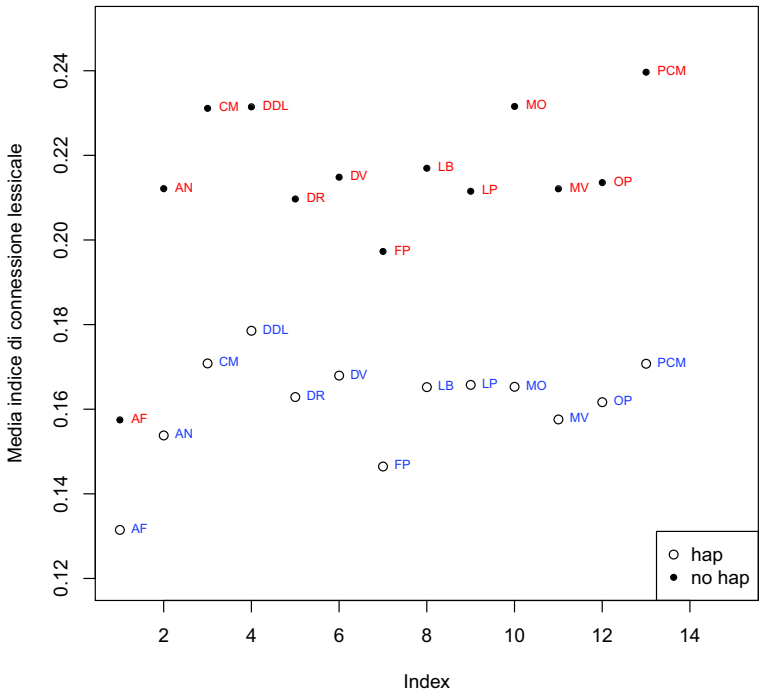

Figura 2.4: Media dell'indice di Connessione Lessicale con e senza hapax

Indice di Indipendenza lessicale

Questo indice viene utilizzato per valutare quanto due testi (A e B) sono dipendenti l'uno dall'altro in termini di vocabolario; in particolare si misura quanto "pesa" la parte propria di un testo sul vocabolario (del testo stesso).

L'indipendenza lessicale di A da B si indica con $I_{V_A(B)}$ e si ottiene dalla formula:

$$I_{V_A(B)} = \frac{V_{A \cap \bar{B}}}{V_A} \tag{2.5}$$

Di conseguenza, il grado d'indipendenza di B da A si calcola con:

$$I_{V_B(A)} = \frac{V_{\bar{A} \cap B}}{V_B}. \tag{2.6}$$

L'indice così calcolato raggiunge la massima indipendenza ($I_{V_A(B)} = 1$) nel caso di due testi completamente diversi, mentre la perfetta dipendenza si ha quando $I_{V_A(B)} = 0$ e i due testi sono identici: la tabella A.2 in Appendice contiene, sulle righe, i valori calcolati tra coppie di autori, mentre l'ultima colonna rappresenta il grado di indipendenza tra il corpus di un autore e quello formato dai testi di tutti gli altri.

I valori molto elevati, in tutti i casi superiori a $0,5$, rendono più chiara l'importanza delle parole diverse tra autori. L'indice ci dice quante parole del proprio testo *non* vengono condivise oppure, dal secondo punto di vista, quante ne ha in comune con l'altro. La lettura per riga ci indica il grado di indipendenza *di* quell'autore dagli altri (l'indice calcolato con la 2.5), sulle colonne il grado di indipendenza *da* quell'autore (testo B): possiamo affermare, quindi, che AF è il più "autonomo", mentre FP è quello con cui *gli altri* condividono meno vocaboli.

I valori nell'ultima colonna, in cui il confronto avviene con un corpus formato dall'unione dei restanti dodici, sono così bassi, perché, da un punto di vista matematico, la probabilità di trovare

parole diverse in due testi diminuisce all'aumentare della lunghezza di uno dei due.

Come nel paragrafo precedente vediamo la relazione tra hapax e indipendenza: eliminando le parole che compaiono una volta sola, l'indice diminuisce perché non vengono considerate alcune parole appartenenti al vocabolario di un solo testo. In Appendice si trova la tabella con i valori medi d'indipendenza lessicale per ogni autore (A.5), di seguito riportiamo il grafico della relazione tra la diminuzione dell'indice (calcolata come $I_{V_A(B)} - I_{V_{A^*}(B^*)}$, dove A^* è il testo A senza hapax) e la percentuale di hapax nei corpus. Come si pensava, sembra esserci una relazione diretta: infatti, all'aumentare del numero di hapax aumenta la differenza tra i due indici; solamente AF, pur avendo un alto numero di parole "singolari" è caratterizzato da un valore basso per la differenza. Quest'ultimo fatto ci porta a pensare che non siano solamente gli hapax ad influenzare i valori di AF, ma che egli abbia un modo totalmente diverso di esprimersi con le parole rispetto a tutti gli altri.

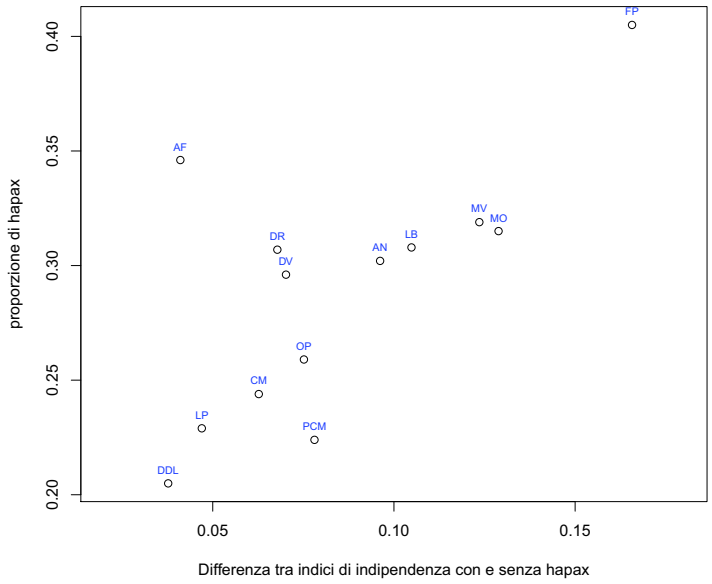

Figura 2.5: Relazione tra hapax e differenza tra indici di indipendenza

Capitolo 3

Le distribuzioni di distanze

Nel capitolo precedente abbiamo calcolato la distanza intertestuale tra coppie di autistici e descritto il fenomeno in modo *statico*, considerando come unità di analisi i corpus individuali; in questo capitolo le analisi utilizzano esclusivamente la TdM, la matrice che ha sulle righe i types utilizzati e sulle colonne gli autori stessi.

3.1 Il procedimento

Per rendere *dinamica* l'analisi abbiamo estratto in modo casuale campioni di mille types dalle righe, costruendo per ognuno di questi una matrice di dimensioni 1000×13 (parole per autori) le cui celle riportano il numero di volte in cui è stata utilizzata l'i–esima parola dal j–esimo autore; le colonne di queste matrici–campioni rappresentano i sub–corpora individuali, cioè le unità di riferimento per le prossime analisi.

E' importante sottolineare che il campionamento è avvenuto senza reinserimento per evitare di

considerare la stessa parola due o più volte e che la numerosità è un compromesso suggerito dalla letteratura[1], dovuto all'elevata percentuale di hapax presente in ogni corpus e alla lunghezza dei testi piuttosto esigua.

Grazie al software utilizzato per le analisi abbiamo costruito una *array* formata dalle 1000 matrici estratte – la dimensione è $1000 \times 13 \times 1000$ (types, autore, campione) – a cui abbiamo applicato il calcolo della distanza 2.3 e ottenuto una nuova array costituita da tante matrici quadrate di distanze intertestuali (come la Tabella 2.6) quanti sono i campioni estratti.

La *tridimensionalità* di quest'oggetto ci permette di determinare l'evoluzione della distanza intertestuale tra due autori nei 1000 campioni: basta "tagliare" l'array lungo le righe o le colonne (essendo matrici di distanze il valore non cambia) per ottenere le distribuzioni di distanze tra l'i–esimo (o j–esimo se si lavora sulle colonne) autistico e tutti gli altri. Nel caso si voglia analizzare una particolare coppia di autori bisogna isolare il procedimento ad una sola cella delle matrici selezionando l'*i–esima* riga e la *j–esima* colonna (con $i \neq j$), lasciando la terza dimensione (il *k–esimo* campione) libera: così facendo otteniamo la **distribuzione** della distanza intertestuale tra coppie di autori, da cui possiamo ricavare media e varianza.

Sono state calcolate 78 distribuzioni – da un gruppo di n individui si ottengono $n(n-1)/2$ coppie diverse –, i cui grafici sono riportati in Appendice, e di ognuna si è provveduto al calcolo di media e varianza, come si può vedere nell'esempio seguente(Figura 3.1): i due autistici sono OP e PCM, sull'asse delle x abbiamo le classi di distanze di Labbé, sulle y il numero di

[1] *Labbé & Labbé (2001)*: [...] l'accuratezza della metrica proposta dalla 2.3 viene ridotta a causa dei valori decimali delle $F^*_{i,b}$; questo effetto aumenta nel caso le parole poco frequenti siano una parte importante del testo, come avviene nei testi piccoli. Per evitare tutto ciò non è conveniente applicare il calcolo a testi troppo esigui (meno di mille *tokens*) [...]

campioni; media e varianza sono uguali a 0.596 e 0.006.

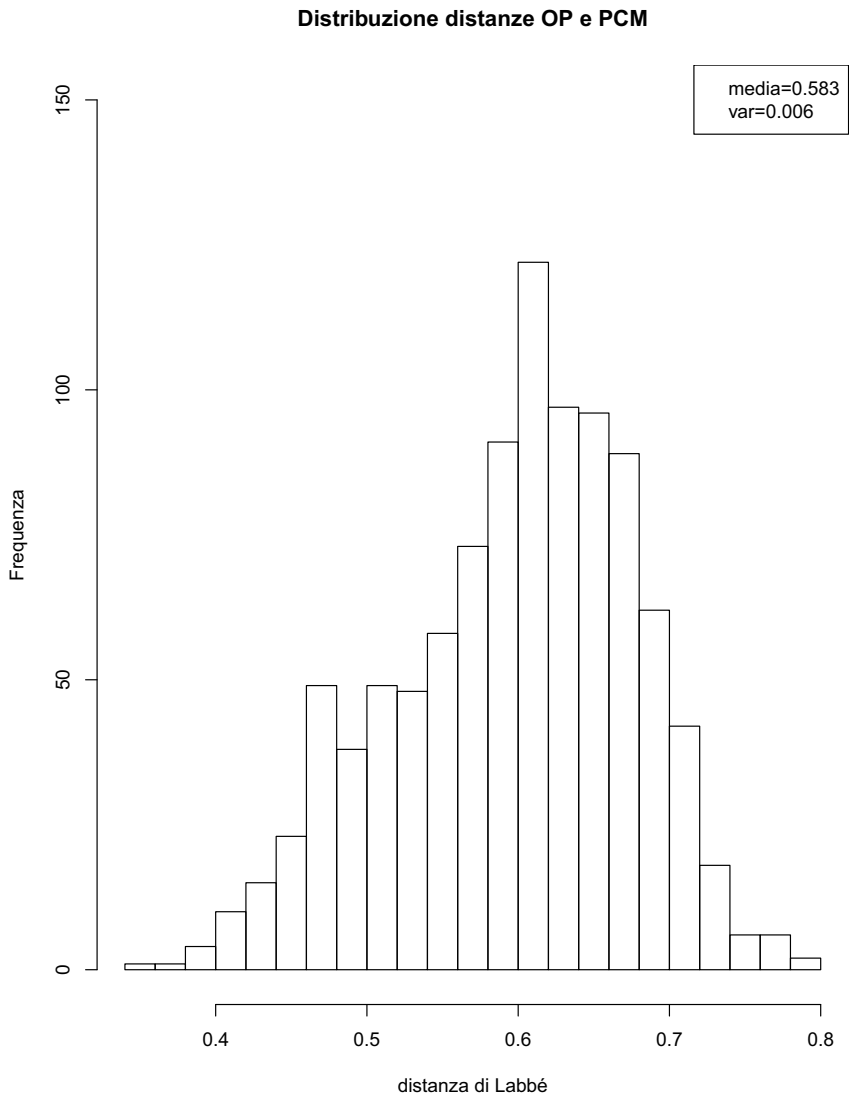

Figura 3.1: Distanza di Labbé tra OP e PCM nei 1000 campioni: $\mu = 0.596$, $\sigma^2 = 0.006$

Per sintetizzare l'array delle distribuzioni abbiamo costruito due matrici; una delle medie e

una delle varianze campionarie, in cui la media è

$$\mu_{a,b} = \left(\sum_{k=1}^{n} d_{a,b,k} \right) / n, \text{ con } n = 1000 \text{ e } a \neq b$$

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
AF	0.00												
AN	0.76	0.00											
CM	0.68	0.62	0.00										
DDL	0.74	0.62	0.57	0.00									
DR	0.74	0.62	0.56	0.58	0.00								
DV	0.74	0.65	0.59	0.61	0.62	0.00							
FP	0.75	0.65	0.62	0.62	0.64	0.65	0.00						
LB	0.76	0.60	0.58	0.57	0.61	0.63	0.60	0.00					
LP	0.74	0.64	0.55	0.60	0.61	0.65	0.66	0.59	0.00				
MO	0.73	0.61	0.55	0.57	0.57	0.61	0.61	0.55	0.58	0.00			
MV	0.76	0.65	0.60	0.63	0.63	0.63	0.66	0.60	0.61	0.61	0.00		
OP	0.74	0.66	0.58	0.59	0.62	0.63	0.65	0.62	0.64	0.59	0.63	0.00	
PCM	0.73	0.63	0.53	0.56	0.58	0.61	0.63	0.56	0.56	0.54	0.58	0.60	0.00

Tabella 3.1: Matrice delle medie (delle distribuzioni di distanze, tra coppie di autistici)

e la varianza è

$$\sigma^2_{a,b} = \sum_{k=1}^{n} \left(d_{a,b,k} - \mu_{a,b} \right)^2 / \left(n - 1 \right), n = 1000 \text{ e } a \neq b.$$

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
AF	0												
AN	0.00301	0											
CM	0.00255	0.00529	0										
DDL	0.00216	0.00621	0.00503	0									
DR	0.00290	0.00651	0.00720	0.00623	0								
DV	0.00311	0.00616	0.00500	0.00527	0.00592	0							
FP	0.00300	0.00786	0.00780	0.00752	0.00942	0.00623	0						
LB	0.00275	0.00874	0.00622	0.00715	0.00894	0.00550	0.00905	0					
LP	0.00227	0.00760	0.00519	0.00563	0.00684	0.00512	0.00881	0.01000	0				
MO	0.00301	0.00600	0.00650	0.00583	0.00789	0.00490	0.00866	0.00698	0.00774	0			
MV	0.00281	0.00924	0.00655	0.00677	0.00835	0.00494	0.00983	0.01210	0.00994	0.00827	0		
OP	0.00216	0.00511	0.00559	0.00484	0.00685	0.00421	0.00766	0.00996	0.00612	0.00649	0.00805	0	
PCM	0.00284	0.00670	0.00572	0.00701	0.00835	0.00554	0.00984	0.00786	0.00656	0.00723	0.00774	0.00615	0

Tabella 3.2: Matrice delle varianze (delle distribuzioni di distanze, tra coppie di autistici)

Confrontando la Tabella 2.6 con la 3.1 possiamo sottolineare che la distanza media nel campione rispecchia totalmente quanto detto nel capitolo precedente a proposito del modo di scrivere degli autistici analizzati: i valori medi delle distribuzioni tra AF e gli altri sono molto elevati e le rispettive varianze sono tra le più basse delle 78 calcolate.

La Figura 3.2 è una parte del grafico che rappresenta la posizione delle diverse distribuzioni, in termini di media e varianza; è facile notare che la nuvola di punti, prodotta dai valori delle distribuzioni di distanze tra AF e gli altri, si posiziona quasi completamente nella parte destra del riquadro, caratterizzata da valori medi elevati. L'unico punto estraneo a questo insieme, rappresenta la distribuzione della distanza tra AF e CM che, pur avendo valore piccolo in ascissa, è comunque superiore a tutti quelli calcolati per altre coppie di autori: quindi

$$min(d_{AF,.}) > max(d_{a,b}) \text{ con } a \neq b \neq AF;$$

per lo stesso motivo, negli altri tre riquadri, il puntino che rappresenta AF si trova nell'angolo in basso a destra del quadrante, a cui corrispondono valori in ascissa elevati (media) e piccoli in ordinata (varianza).

La Tabella 3.2 riporta i valori delle varianze delle distribuzioni: l'intervallo in cui sono compresi ha come estremi 0.00216 e 0.01210; i valori molto piccoli, quasi prossimi allo 0, evidenziano il fatto che ogni distanza tra due autori nei 1000 campioni si discosta di poco dal rispettivo valore medio; come se fosse indipendente dai types estratti. Questo fatto può essere dovuto all'esigua numerosità di ogni campione; ma, per verificare empiricamente quest'affermazione, bisognerebbe disporre di testi più lunghi in modo da poter confrontare le varianze relative a numerosità crescenti di types estratti (1000,2000,...,10000,...).

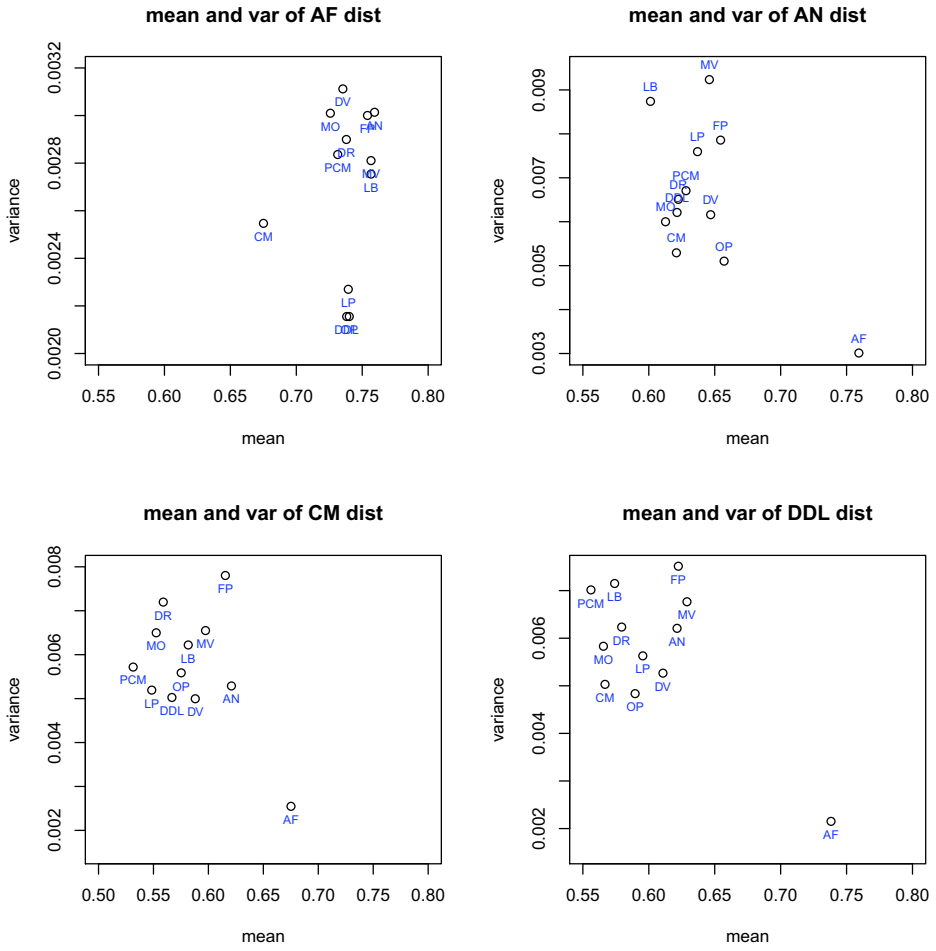

Figura 3.2: Sintesi delle distribuzioni tra coppie di autori: in ascissa μ, in ordinata σ^2 (parte 1)

3.2 Il caso AF

Nei paragrafi precedenti si è accennato ad AF come caso particolare del gruppo esaminato; alla fine del secondo capitolo abbiamo messo in evidenza le peculiarità che evidenzia il calcolo della distanza di Labbé sul suo modo di scrivere:

1. I valori elevati nella Tabella 2.6 – e di conseguenza nella Tabella 2.7 – ci danno percezione di come questo autore si discosti dal resto del gruppo,

2. Figura 2.2 e Figura 2.3 confrontano il valore della distanza media in funzione di due quantità (la proporzione di hapax nel testo e la lunghezza del corpus) riscontrando il comportamento singolare di AF,

3. La Tabella 3.1 conferma la tesi del punto 1: il valore medio delle distribuzioni campionarie di AF è superiore a quello degli altri.

Nell'ultimo punto è importante sottolineare che il valore medio è ottenuto dalle distanze intertestuali calcolate all'interno dei mille campioni; la diversità di AF emerge, quindi, nonostante l'estrazione casuale dei types in ogni campione. Se poi confrontiamo le varianze di tali distribuzioni (Tabella 3.2) notiamo che, oltre ad essere le più basse, sono molto "concentrate" intorno a $\cong 0.0025$.

La Tabella 3.3 ci mostra il range della varianza per le distribuzioni riferite ad un autore e viene calcolato come:

$$range(x) = max(x) - min(x)$$

dove x rappresenta ogni singola colonna della Tabella 3.2, escluso lo 0.

Il valore per AF, così calcolato, è il più piccolo tra tutti e sottolinea come le varianze delle distribuzioni tra questo autore e gli altri siano molto simili; situazione dovuta a scostamenti dal valor medio esigui in ogni campione estratto.

Autore	Range	Autore	Range
AF	0.00096	LB	0.00934
AN	0.00622	LP	0.00773
CM	0.00525	MO	0.00565
DDL	0.00536	MV	0.00928
DR	0.00652	OP	0.00780
DV	0.00312	PCM	0.00700
FP	0.00684		

Tabella 3.3: Range delle varianze delle distribuzioni di distanze

Una spiegazione plausibile è quella che identifica AF come un “outlier” nel modo di scrivere dei tredici autistici in esame; rappresenta cioè un soggetto “egualmente diverso” (in termini di distanza intertestuale) da tutti gli altri, i quali hanno un “comportamento” molto simile nei suoi confronti.

Questa supposizione trova conferma nei grafici che rappresentano la relazione tra media e varianza delle distribuzioni campionarie, come la Figura 3.2 e la restante parte riportata in seguito (Figura 3.3). In tutti i riquadri il punto che indica la distanza tra l’autore considerato e AF si posiziona nell’angolo in basso a destra, esattamente all’opposto della nuvola formata dagli altri.

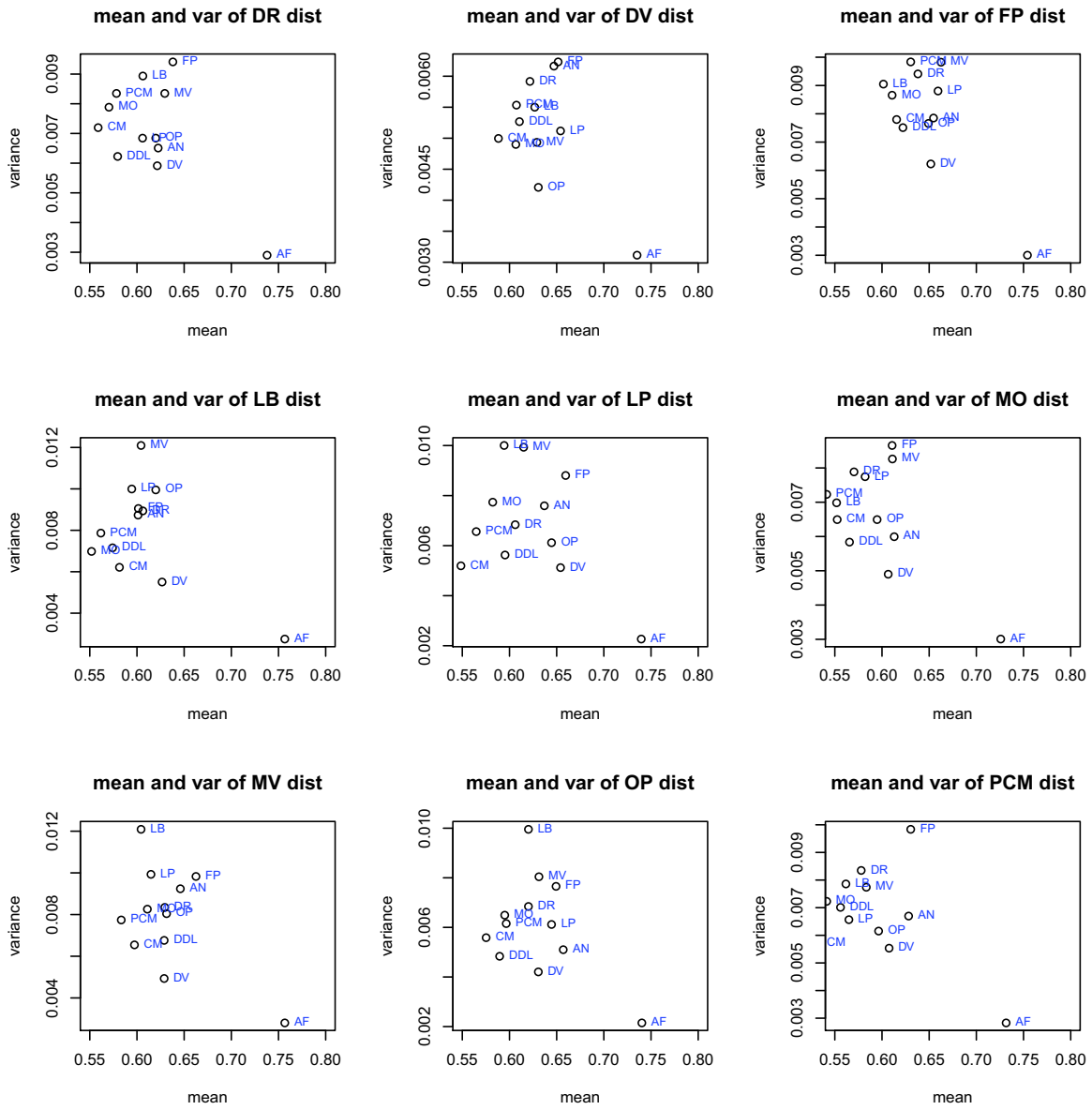

Figura 3.3: Sintesi delle distribuzioni tra coppie di autori: in ascissa μ, in ordinata σ^2 (parte 2)

Un'ulteriore conferma di quanto appena detto si ha confrontando gli indici di connessione e indipendenza lessicale che, "contando" il numero di vocaboli presenti nei diversi corpus, rappresentano una prima misura (molto grezza) della diversità tra gli autori.

Nella Tabella A.2, l'ultima colonna indica il grado di indipendenza tra un autore e i restanti dodici: è quindi una misura di quante parole *non* vengono condivise con gli altri. Ebbene, il valore 0.478 (relativo ad AF) indica un grado molto elevato; infatti, nonostante le dimensioni differenti dei vocabolari, condivide solamente 53 parole su 100 con tutti gli altri (questa lettura fa riferimento alla Formula 3.1).

Come conseguenza, ci si aspetterebbe un valore piccolo di $C_{V_{AF,altri}}$; questo non accade perché, nel calcolo dell'indice di connessione lessicale tra un autore e i restanti (ultima colonna della Tabella A.1), l'importanza del vocabolario in comune viene ridimensionata dalla grandezza di quello totale, che rappresenta il denominatore della frazione ed è formato dai 9756 types utilizzati da tutto il gruppo di ragazzi. Per questo, nel confronto con gli altri, è meglio ricavare la **dipendenza lessicale**, calcolata come:

$$D_{V_A(B)} = 1 - I_{V_A(B)}. \tag{3.1}$$

La Formula 3.1 indica, in opposizione a $I_{V_A(B)}$, la proporzione di vocaboli condivisi da un autore sul totale, rappresentato dal *suo* vocabolario: non risente, quindi, della numerosità elevata di types che caratterizza il denominatore della Formula 2.4.

Ovviamente, le misure di indipendenza e di connessione sono differenti nella sostanza; si rifanno, cioè, a quantità diverse e difficilmente comparabili (in termini numerici).

Capitolo 4

Analisi di raggruppamento

L'idea originale di classificare e, successivamente, raggruppare secondo una sistematica precisa viene fatta risalire al *Systema Naturae* del naturalista svedese Linneo, che ha classificato gli esseri viventi in gruppi (in base a caratteristiche generali comuni) per poi ripartirli in sottogruppi sempre più specializzati e localizzati. Con i legami esistenti tra le categorie animali, Linneo costruì un albero dalle cui basi partono ramificazioni principali, sulle quali si innestano varie ramificazioni secondarie. Ad esempio, l'*uomo* è collocabile su un ramo secondario dell'albero: fa parte, prima di tutto, dei *primati* che, a loro volta, appartengono agli *amnioti*, i quali sono un sottogruppo degli *animali vertebrati*.

La classificazione a partire dall'*osservazione* della realtà, senza aver definito a priori le classi, è l'obiettivo delle tecniche di *analisi di raggruppamento* o *cluster analysis* che cercano di assegnare entità multivariate a poche categorie, non ancora definite. Una volta costituito un gruppo, non è necessario che le entità appartenenti abbiano le stesse caratteristiche; anzi, quanto più numerose sono le variabili osservate, tanto meno riconoscibili sono le modalità che lo identifi-

cano.

Le unità che fanno parte dello stesso gruppo sono allora “simili”, o “somiglianti”.

4.1 Metodologia

L’albero delle specie naturali è un esempio di classificazione di tipo **gerarchico**. In un’analisi gerarchica dei gruppi, ogni classe fa parte di una più ampia e così via, fino a quella che contiene tutte le entità analizzate.

Le tecniche di analisi gerarchica si possono suddividere in:

- *agglomerative*, se l’analisi parte considerando ogni unità iniziale, delle *n* considerate, come un gruppo a sé stante, fino ad arrivare al passo *n-1* nel quale si forma il gruppo che le contiene tutte.

- *divisive*, quando si parte dal gruppo che contiene tutte le entità e, ad ogni passo dell’analisi, lo si ripartisce in un sottogruppo fino a che ogni gruppo è formato da una sola entità (stadio *n-1*).

Nel nostro caso abbiamo utilizzato due metodi appartenenti al primo gruppo[1].

L’analisi (gerarchica) agglomerativa, indipendentemente dal metodo utilizzato, segue un preciso ordine di operazioni, che qui riportiamo:

1. data una matrice simmetrica di distanze tra n entità, si trovano le due che sono più vicine e, con queste, si forma un gruppo. A questo punto, assumendo distanza nulla al suo

[1] Per una spiegazione approfondita dell’analisi di raggruppamento e dei suoi metodi si rimanda a:

Fabbris L.; Statistica Multivariata Ed: McGraw-Hill (1997)

interno, si calcolano le distanze tra il gruppo appena formato e le rimanenti unità: in questo momento si decide la strategia agglomerativa da utilizzare.

2. Nella nuova matrice di distanze, di dimensione $n-1$, si individuano le unità più vicine e si forma un nuovo gruppo. Successivamente si ricalcolano le distanze tra il gruppo formato e le rimanenti entità.

3. Si ripete il procedimento $n-1$ volte finché tutte le unità fanno parte di un unico gruppo.

In questa sequenza è fondamentale il punto 1, in cui, dopo aver unitò le due unità più vicine, si ricalcolano le distanze. Il metodo per calcolare le nuove misure influisce sulla struttura dei gruppi finali: a seconda di quale scegliamo uniremo un'unità precisa, che non è sempre la stessa per tutti i metodi.

Supponiamo, quindi, di aver aggregato le due entità *i* e *j* (possono essere singole o rappresentare un gruppo) e di voler calcolare la distanza con una delle rimanenti, *k*; avremo tre misure di distanza tra le entità: $d_{i,k}$, $d_{j,k}$ e $d_{i,j}$ per cui vale sicuramente $d_{i,j} < d_{i,k} < d_{j,k}$ (ma può anche essere $d_{i,j} < d_{j,k} < d_{i,k}$).

La distanza tra il gruppo e l'entità esterna, $d_{(i,j)k}$, si calcola combinando le tre distanze appena viste con pesi diversi a seconda del metodo utilizzato. Nel caso vengano generate delle partizioni tali per cui vale la disugualgianza:

$$d_{i,j} \leq max\left\{d_{i,k}, d_{j,k}\right\} \text{ con } i, j, k = 1, ..., n$$

si dice che tale metodo genera un'*ultrametrica*.

Questa disuguaglianza assicura che le distanze alle quali i gruppi si uniscono assumano valori progressivamente crescenti o decrescenti, a seconda che l'analisi si basi su misure di dissomiglianza o di somiglianza. La rappresentazione grafica di questo tipo di analisi è un diagramma ad albero, *dendrogramma*, su assi cartesiani che riportano in ascissa le n entità analizzate, in ordinata i livelli di aggregazione delle unità.

4.2 Le strategie agglomerative utilizzate

La matrice di partenza per la cluster analysis è la matrice di distanze di Labbé (Tabella 2.6); una volta uniti i due autori più vicini, CM e PCM, a distanza 0.52, si ricalcolano le nuove distanze in base al metoro agglomerativo scelto. Nel nostro caso abbiamo optato per il confronto tra il metodo di Ward e il metodo del legame completo.

4.2.1 Il metodo di Ward

Con questo metodo, la coppia di entità da aggregare ad un certo gradino dell'analisi è quella che minimizza la devianza tra i centroidi dei possibili gruppi.

La distanza a cui si aggrega un'entità *k* al gruppo di nuova formazione *(i,j)* è:

$$d_{k,(i,j)} = \sqrt{\frac{(n_i + n_k)d^2_{i,k} + (n_j + n_k)d^2_{j,k} - n_k d^2_{i,j}}{n_i + n_j + n_k}}; \tag{4.1}$$

dove: n_i, n_j, n_k rappresentano le numerosità delle entità e, identificando degli autori, sono uguali a 1 (sono > 1 nel caso si tratti già di un gruppo e non più di un singolo); $d_{i,j}, d_{k,j}, d_{k,i}$ sono le distanze di Labbé della Tabella 2.6.

Il metodo di Ward

- è stato pensato per distanze euclidee, ma può essere usato per ogni tipo di distanze;
- ha il difetto di unire *outliers* nei primi passi del processo di aggregazione.

Il risultato dell'applicazione di questo algoritmo ai nostri dati è sintetizzato nella tabella qui sotto:

Stadio	Entità 1	Entità 2	Distanza	Stadio	Entità 1	Entità 2	Distanza
1	CM	PCM	0.524	7	MV	OP	0.624
2	MO	stadio 1	0.545	8	DV	stadio 7	0.630
3	DDL	LB	0.569	9	AN	FP	0.647
4	DR	stadio 2	0.570	10	stadio 6	stadio 8	0.649
5	stadio 3	stadio 4	0.581	11	stadio 9	stadio 10	0.661
6	LP	stadio 5	0.588	12	AF	stadio 11	0.851

Tabella 4.1: Algoritmo del metodo di Ward

Prendiamo ad esempio il secondo passo del processo: la lettura per riga ci dice che l'autore MO si aggrega all'entità formata allo stadio 1, generando un nuovo gruppo di cui fanno parte CM, PCM e MO. Allo stadio 5, ad una distanza pari a 0.581, si forma un gruppo unendo le entità formate ai passi 3 e 4; la nuova entità è costituita da CM, PCM, MO, DR, DDL e LB.
L'algoritmo continua finché tutti gli autori vengono riuniti in un unico gruppo (passo 12): è in questo stadio del processo che AF si unisce a tutti gli altri.

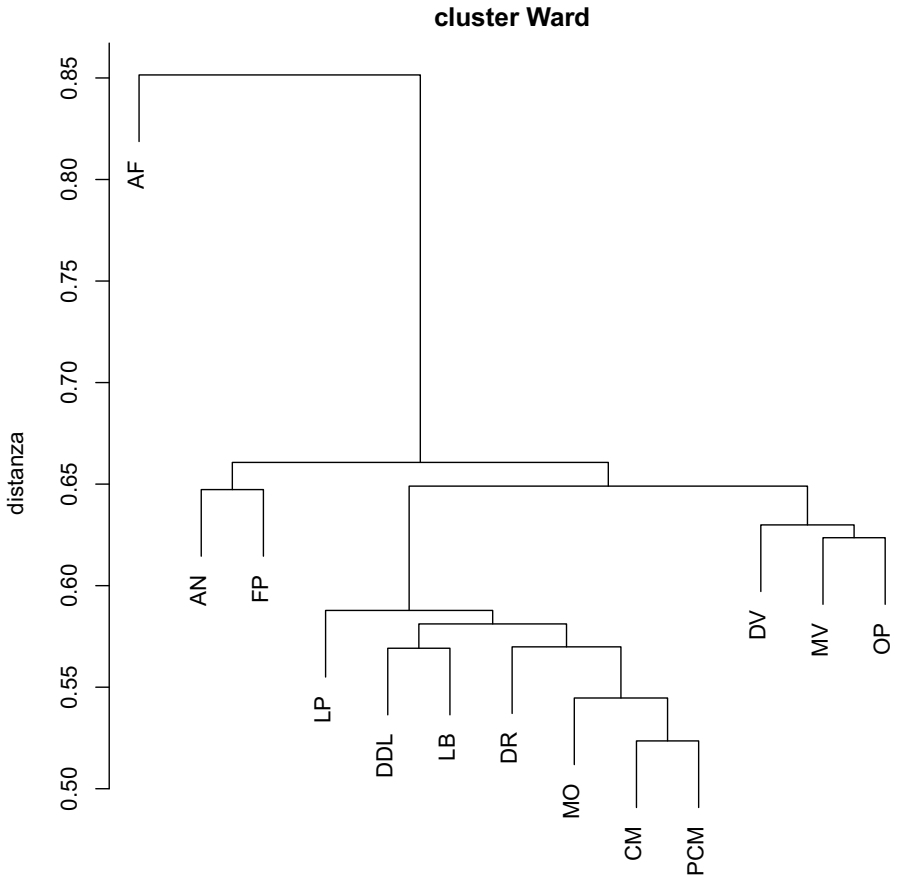

Figura 4.1: Dendrogramma, metodo di Ward

La rappresentazione grafica della Tabella 4.1 è il dendrogramma riportato in Figura 4.1. Da questo si possono individuare quattro gruppi ben definiti: il primo si forma ad una distanza pari a 0.588 (stadio 6); il secondo, costituito da DV, MV e OP nel passo 8 del processo e il terzo, di cui fanno parte AN e FP che si aggregano a distanza 0.647.

AF, che viene considerato come un gruppo a sé, entra a far parte della classificazione ad una distanza molto elevata (0.851), a conferma di quanto detto nel paragrafo 3.2.

4.2.2 Il metodo del legame completo

Viene anche chiamato del "vicino più lontano": tra l'entità esterna *k* e il gruppo di formazione *(i,j)*, la distanza è data dal valore più elevato tra $d_{i,k}$ e $d_{j,k}$, cioè

$$d_{k,(i,j)} = max\left\{d_{i,k}, d_{j,k}\right\}, \text{ con } i \neq j \neq k = 1, ..., n. \tag{4.2}$$

Questo criterio produce gruppi di forma circolare caratterizzati da forte somiglianza interna: l'entità candidata all'unione è sempre la più vicina, ma è la distanza a determinare le caratteristiche del nuovo gruppo; scegliendo quella più elevata ci assicuriamo della maggior prossimità (o "somiglianza") con le altre.

Il metodo del legame completo può essere utilizzato con qualunque misura di distanza e genera un'ultrametrica.

Come per il paragrafo precedente, vediamo nello specifico come funziona il metodo del legame completo sugli autori esaminati. La tabella sottostante riporta, per ogni stadio del processo, le due entità che si aggregano e la distanza calcolata:

Stadio	Entità 1	Entità 2	Distanza	Stadio	Entità 1	Entità 2	Distanza
1	CM	PCM	0.524	7	MV	OP	0.624
2	MO	stadio 1	0.545	8	AN	stadio 6	0.628
3	DDL	stadio 2	0.563	9	DV	stadio 7	0.631
4	DR	stadio 3	0.574	10	FP	stadio 9	0.650
5	LB	LP	0.581	11	stadio 8	stadio 10	0.656
6	stadio 4	stadio 5	0.598	12	AF	stadio 11	0.757

Tabella 4.2: Algoritmo del metodo del legame completo

Confrontando le due tabelle fin qui presentate, si può notare che non ci sono molte differenze tra i valori delle distanze relative ad ogni stadio del processo; sono le entità aggregate che differiscono tra i due metodi e generano un diverso raggruppamento degli autori (vedi Figura 4.2).

Il numero di gruppi non è ben definibile a prima vista; per fare ciò, dobbiamo immaginare una linea orizzontale posta sull'asse delle ordinate ad altezza 0.65 (la distanza massima proposta dalla letteratura per testi scritti nello stesso linguaggio). Questa retta, tagliando il dendrogramma, mette in evidenza due gruppi distinti: uno formato da AN, DR, DDL, MO, CM, PCM, LB, LP e l'altro, a cui appartengono FP, DV, MV e OP.

AF rimane isolato e si unisce per ultimo a distanza 0.757.

Rispetto ai nostri obiettivi, cioè valutare le differenze nel modo di scrivere tra i ragazzi autistici, il legame completo è il metodo più adatto. I gruppi così generati, sono caratterizzati da una forte somiglianza interna e, tenendo conto che la matrice di partenza è costituita dalle distanze intertestuali, la classificazione ottenuta rappresenta proprio il fenomeno analizzato.

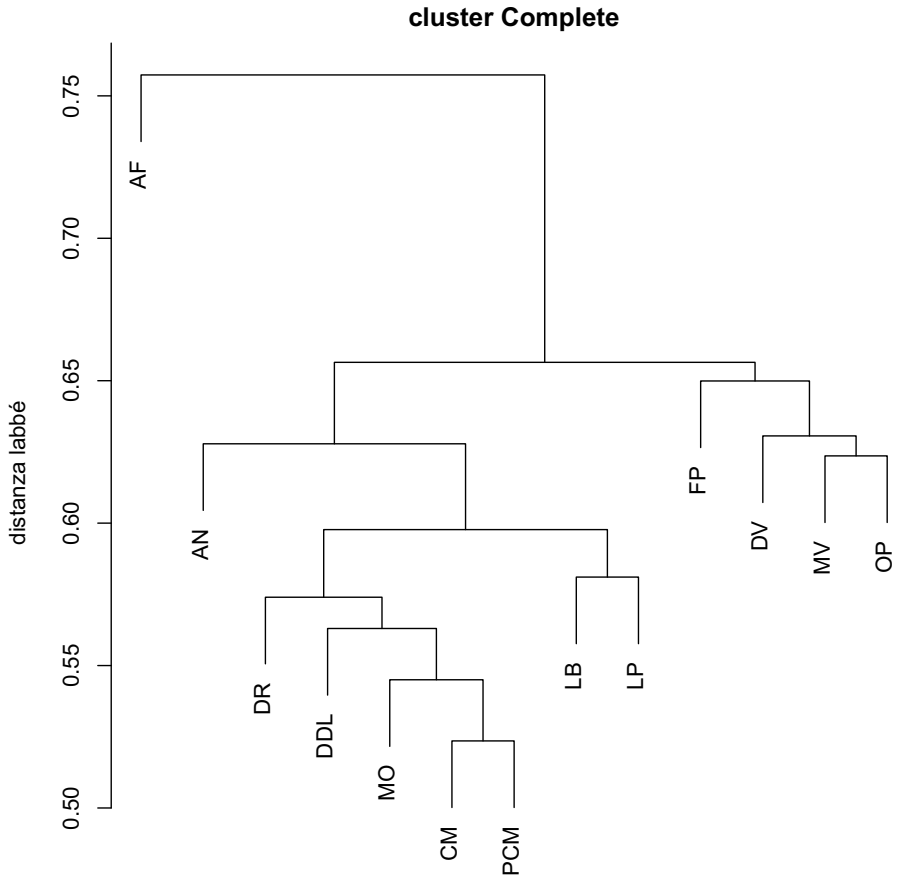

Figura 4.2: Dendrogramma, metodo del legame completo

4.3 Cluster e periodi di facilitazione

Nel paragrafo 2.2.4 sono stati introdotti i periodi di facilitazione e le problematiche relative al calcolo della distanza di Labbé per questi testi; tuttavia, abbiamo voluto applicare i metodi di raggruppamento alla corrispondente matrice di prossimità. La lettura del dendrogramma relativo al metodo del legame completo (Figura 4.5) non è immediata: gli eventuali gruppi non sono identificabili a prima vista; sembrano delinearsi due macro gruppi (al centro del dendrogramma), costituiti principalmente da testi scritti dal secondo anno di facilitazione in poi. Infatti, gli scritti relativi ai primi periodi (primo e secondo semestre) rimangono abbastanza isolati fino a distanze prossime a 1; un esempio è il gruppo che si forma per ultimo, ad una distanza pari a 0.982, dall'unione di *OP1sem, DDL1sem, DDL2sem, AF2sem, AF2year*.

Una caratteristica comune a tutti gli autori sembra essere la *prossimità temporale* nei gruppi: difficilmente vengono aggregati due testi relativi a periodi di facilitazione non adiacenti, come se ci fosse un filo conduttore (la maggiore confidenza con gli strumenti e i modi della comunicazione facilitata?) a legare i testi in tutto l'arco temporale.

Il dendrogramma prodotto dal metodo di Ward (Figura 4.6) è differente da quello appena visto; innanzitutto, la distanza calcolata con la 4.1 produce livelli di raggruppamento – i valori in ordinata – maggiori di 1 poiché nella formula vengono contate le unità che costituiscono le entità candidate all'aggregazione. I gruppi sono abbastanza delineati e identificabili: un primo gruppo è quello formato dai cinque testi di LP, che si colloca a sinistra nel dendrogramma; negli altri gli autori si mischiano casualmente. Tenendo conto del gruppo *mono autore*, abbiamo immaginato di tagliare il grafico ad un altezza tale da considerarlo come gruppo a sé. Si sono individauati

nove gruppi le cui numerosità sono molto diverse tra loro (e non permettono confronti oggettivi tra gruppi), ma la *prossimità temporale* sembra ancor più accentuata: il quarto gruppo è formato dai testi in sequenza temporale di *PCM, CM, DR, DDL e OP*, lo stesso vale per il sesto costituito interamente dagli scritti di *DV* e per gli ultimi due, in cui si aggregano testi relativi ai primi periodi di facilitazione.

4.4 Cluster e distribuzioni

Anche le distribuzioni del capitolo 3 si prestano alla cluster analysis; possiamo utilizzare, infatti, le distanze medie di Labbé tra coppie di autistici contenute nella tabella 3.1 e applicare i due metodi agglomerativi (Figura 4.3 e Figura 4.4).

Il dendrogramma del legame completo è differente da quello ottenuto a partire dalla matrice di distanze di Labbé tra coppie di autori (Figura 4.2): i due gruppi non sono più identificabili, a causa dell'aggregazione di unità diverse ad ogni stadio del processo. AF è sempre un autore a sé stante, anche nell'analisi sulle distribuzioni campionarie.

Il metodo di Ward, invece, genera un dendrogramma molto simile a quello in Figura 4.1; i tre gruppi sono ancora ben definiti e identici (in termini di unità costituenti), anche se cambiano i valori delle distanze di aggregazione e la successione delle entità candidate all'unione.

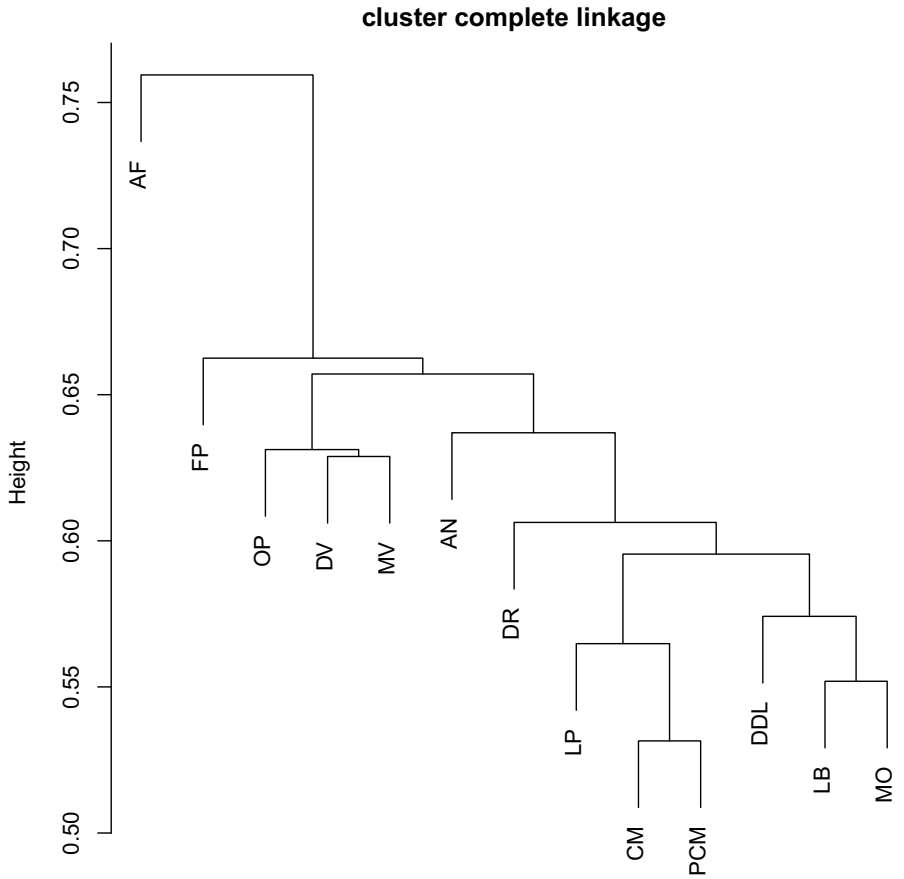

Figura 4.3: Dendrogramma, medie delle distribuzioni, metodo del legame completo

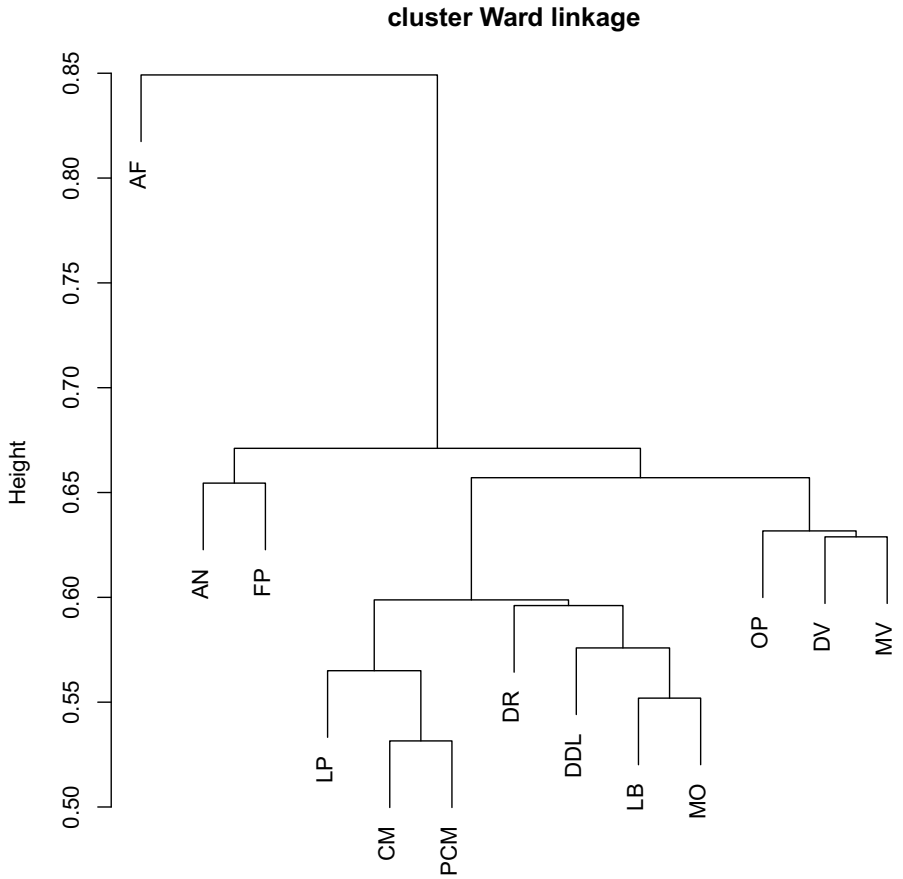

Figura 4.4: Dendrogramma, medie delle distribuzioni, metodo di Ward

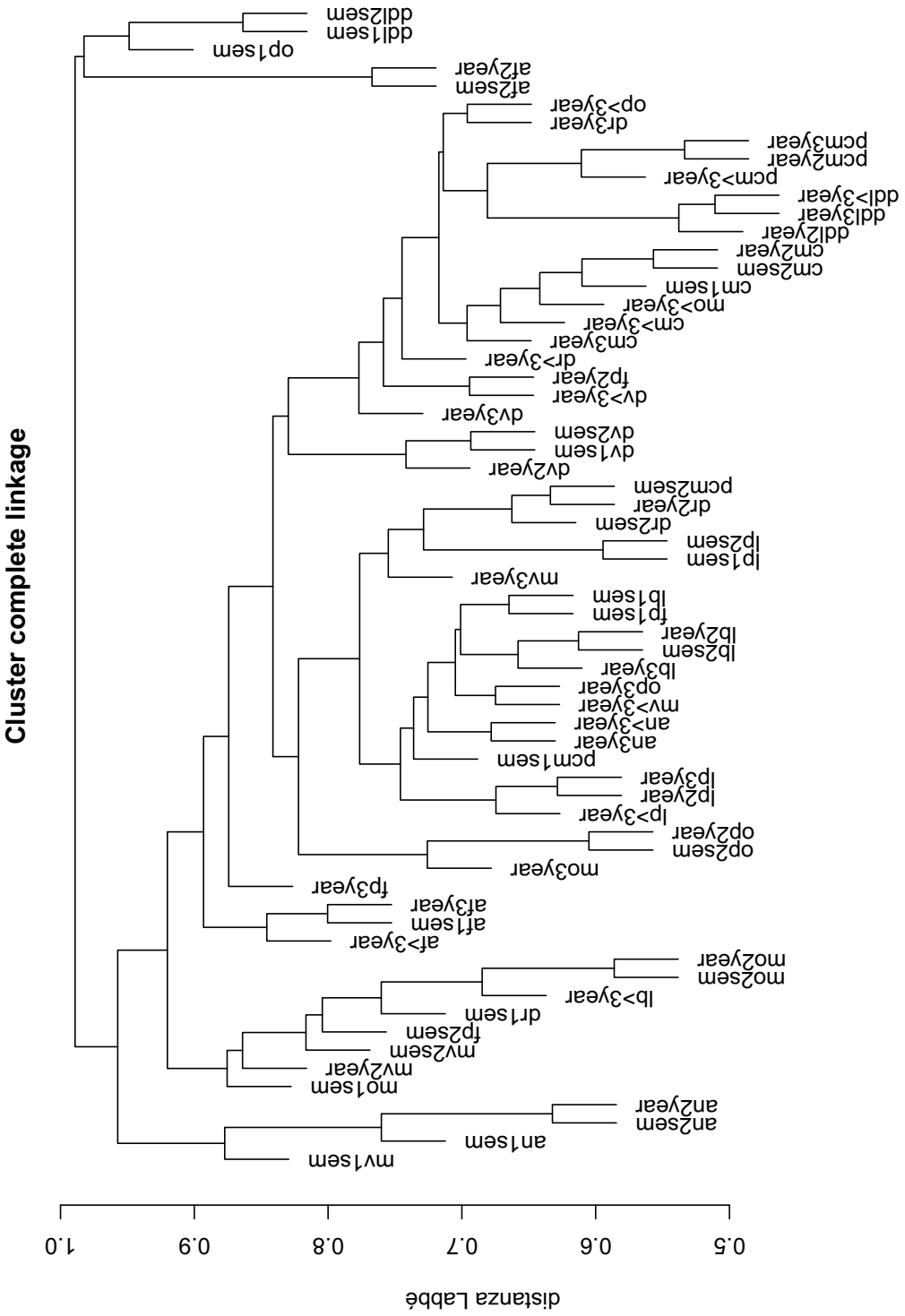

Figura 4.5: Dendrogramma, periodi di facilitazione, metodo del legame completo

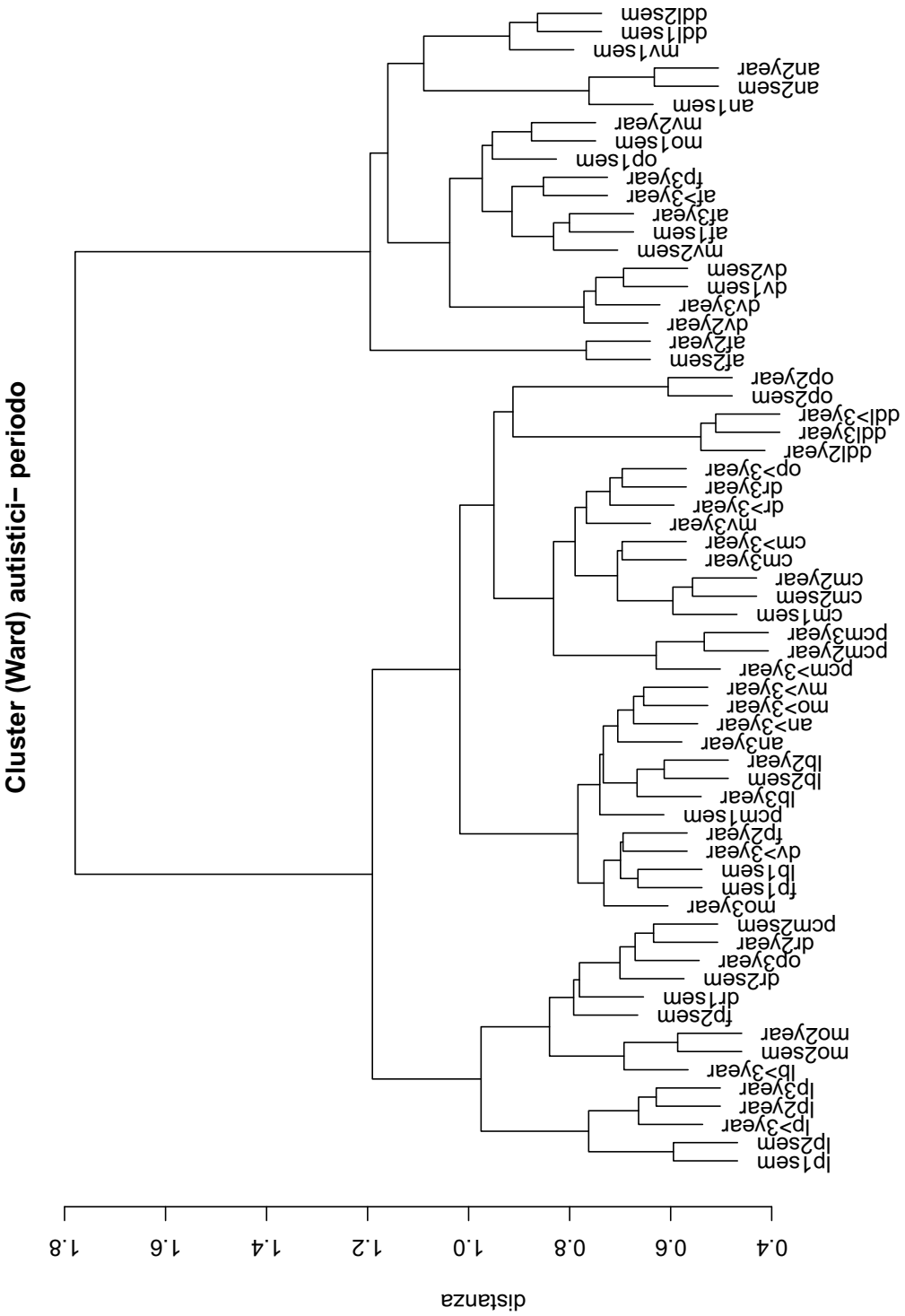

Figura 4.6: Dendrogramma, periodi di facilitazione, metodo di Ward

Conclusioni

Le analisi svolte in questo lavoro hanno messo in luce alcune particolarità del corpus "Gruppo 1", appartenente al progetto EASIEST. Prima fra tutte l'alta percentuale di hapax nei diversi corpora ha sia un significato linguistico (un'alta percentuale di parole che compaiono una volta sola nel testo indica un linguaggio originale), sia un significato statistico: è uno dei motivi per cui abbiamo ottenuto valori così elevati delle distanze calcolate nel capitolo 2. A tal proposito, non dobbiamo dimenticare che i tredici autori hanno età ed esperienze di vita diverse, ma, soprattutto, che la letteratura propone valori soglia per autori *tipici*.

La matrice delle distanze intertestuali tra coppie di autori (Tabella 2.6) riporta sì valori elevati, ma molto simili per dodici autori; l'unico a differenziarsi completamente (da tutti) è AF che, in tutte le analisi, ha un comportamento da *outlier*; il problema che si presenta ora è stabilire *se* esiste un "linguaggio autistico", un modo di usare le parole comune alle persone con questa patologia. Per trovare una risposta a quest'ipotesi, bisognerebbe disporre di un gruppo di "controllo" formato da ragazzi "normali", con caratteristiche socio-demografiche identiche a quelle del Gruppo 1; oltrettutto, i risultati ottenuti dal confronto tra i due gruppi potrebbero fornire ulteriori strumenti di conoscenza relativi all'universo autistico.

La cluster analysis utilizzata nel quarto capitolo è un ottimo strumento per visualizzare e iden-

tificare le somiglianze tra gli autori. In questo lavoro è stata utilizzata sia come strumento di verifica delle congetture fatte durante le analisi – la particolarità di AF viene sottolineata in tutti i dendrogrammi riportati –, sia come metodo d'indagine nell'analisi per periodi di facilitazione. Il problema principale dello studio di questi testi è rappresentato dalla loro lunghezza; nonostante ciò, sia il metodo di Ward, sia il metodo del legame completo raggruppano ai primi passi dell'algoritmo testi relativi a tempi adiacenti e, generalmente, appartenenti ad uno stesso autore. Sembra, quindi, esistere un filo conduttore tra i periodi di facilitazione, che può essere dovuto alla gradualità temporale del livello comunicativo (vedi paragrafo 1.3.1) e alla confidenza tra facilitatore e facilitato – anche questo è un rapporto che si forma, ma soprattutto cresce, nel tempo –, oppure ad un effettivo miglioramento delle capacità comunicative del soggetto (intuibile già nella progressiva diminuzione del livello di facilitazione nei periodi).

Al di là dei procedimenti statistici utilizzati e dei risultati ottenuti, rimane la bellezza del progetto EASIEST (i cui dati sono stati la materia prima di questo lavoro): studiare il modo di scrivere di ragazzi autistici è il punto di arrivo di un percorso che parte dal presupposto di voler conoscere l'altro, sebbene *diverso*, senza i pregiudizi che caratterizzano il nostro vivere quotidiano e che portano ad una patologia ben più grave dell'autismo; essere *gravemente normodotati*.

Per concludere, non ci sono parole migliori se non quelle di uno dei ragazzi, Pier Carlo:

Appreso ho tramite cara seria CF a menti e cuori incontrare. Comprendo immensa paura di certezze abbandonare ma voglio critici intelligentemente disponi-

bili a loro pregiudizi sostituire con curiosi studi su come molto uso di CF oggettivamente molto migliora vita di noi ineducabili marchiati e di nostre famiglie incapaci ingiustamente decretate.

Vi entusiasticamente ringrazio, vostro leggere mie parole per me è iniziare a insieme veloci navigare.

Appendice A

A.1 Tabelle e figure del capitolo 2

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM	TUTTI ALTRI
AF	1	0.119	0.160	0.143	0.123	0.136	0.121	0.121	0.138	0.131	0.117	0.130	0.138	0.114
AN	0.119	1	0.152	0.169	0.152	0.159	0.147	0.169	0.159	0.159	0.153	0.150	0.157	0.087
CM	0.160	0.152	1	0.188	0.169	0.166	0.145	0.161	0.199	0.166	0.163	0.182	0.200	0.148
DDL	0.143	0.169	0.188	1	0.197	0.181	0.163	0.188	0.190	0.187	0.165	0.180	0.193	0.112
DR	0.123	0.152	0.169	0.197	1	0.169	0.145	0.158	0.174	0.171	0.164	0.166	0.167	0.098
DV	0.136	0.159	0.166	0.181	0.169	1	0.155	0.182	0.160	0.176	0.185	0.171	0.177	0.080
FP	0.121	0.147	0.145	0.163	0.145	0.155	1	0.167	0.132	0.154	0.142	0.146	0.141	0.069
LB	0.121	0.169	0.161	0.188	0.158	0.182	0.167	1	0.164	0.180	0.163	0.158	0.171	0.076
LP	0.138	0.159	0.199	0.190	0.174	0.160	0.132	0.164	1	0.168	0.157	0.161	0.189	0.128
MO	0.131	0.159	0.166	0.187	0.171	0.176	0.154	0.180	0.168	1	0.156	0.161	0.175	0.098
MV	0.117	0.153	0.163	0.165	0.164	0.185	0.142	0.163	0.157	0.156	1	0.160	0.167	0.082
OP	0.130	0.150	0.182	0.180	0.166	0.171	0.146	0.158	0.161	0.161	0.160	1	0.176	0.109
PCM	0.138	0.157	0.200	0.193	0.167	0.177	0.141	0.171	0.189	0.175	0.167	0.176	1	0.123

Tabella A.1: Indice di Connessione Lessicale ($C_{V_{A,B}}$) tra coppie di autori.

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM	TUTTI ALTRI
AF	0	0.829	0.715	0.786	0.819	0.822	0.844	0.841	0.772	0.810	0.838	0.799	0.776	0.478
AN	0.717	0	0.636	0.685	0.727	0.753	0.778	0.743	0.662	0.716	0.747	0.707	0.675	0.348
CM	0.732	0.793	0	0.736	0.768	0.793	0.821	0.800	0.696	0.772	0.788	0.738	0.700	0.365
DDL	0.699	0.732	0.606	0	0.685	0.742	0.773	0.737	0.642	0.700	0.750	0.687	0.646	0.281
DR	0.722	0.745	0.619	0.655	0	0.748	0.787	0.764	0.650	0.709	0.740	0.692	0.670	0.308
DV	0.632	0.691	0.544	0.621	0.662	0	0.742	0.697	0.609	0.652	0.668	0.626	0.586	0.245
FP	0.649	0.697	0.570	0.635	0.688	0.719	0	0.706	0.650	0.672	0.723	0.657	0.640	0.293
LB	0.661	0.668	0.547	0.602	0.674	0.688	0.722	0	0.593	0.638	0.696	0.645	0.590	0.263
LP	0.743	0.770	0.634	0.713	0.744	0.787	0.825	0.784	0	0.752	0.781	0.744	0.689	0.344
MO	0.705	0.733	0.622	0.669	0.707	0.739	0.774	0.736	0.658	0	0.750	0.700	0.654	0.306
MV	0.703	0.720	0.588	0.674	0.692	0.707	0.775	0.739	0.645	0.706	0	0.673	0.635	0.317
OP	0.733	0.764	0.628	0.703	0.734	0.760	0.798	0.778	0.698	0.743	0.762	0	0.682	0.336
PCM	0.736	0.768	0.624	0.703	0.748	0.764	0.812	0.773	0.676	0.738	0.765	0.719	0	0.340

Tabella A.2: Indice di Indipendenza Lessicale ($I_{V_{A(B)}}$) tra coppie di autori.

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
AF	1	0.138	0.187	0.187	0.148	0.159	0.137	0.144	0.157	0.171	0.136	0.152	0.171
AN	0.138	1	0.214	0.219	0.235	0.216	0.212	0.246	0.209	0.224	0.204	0.202	0.225
CM	0.187	0.214	1	0.266	0.230	0.216	0.187	0.221	0.278	0.229	0.207	0.249	0.289
DDL	0.187	0.219	0.266	1	0.242	0.218	0.195	0.234	0.242	0.245	0.199	0.261	0.270
DR	0.148	0.235	0.230	0.242	1	0.219	0.172	0.191	0.206	0.234	0.216	0.192	0.230
DV	0.159	0.216	0.216	0.218	0.219	1	0.217	0.211	0.199	0.243	0.237	0.218	0.226
FP	0.137	0.212	0.187	0.195	0.172	0.217	1	0.247	0.162	0.224	0.220	0.197	0.197
LB	0.144	0.246	0.221	0.234	0.191	0.211	0.247	1	0.193	0.249	0.226	0.198	0.243
LP	0.157	0.209	0.278	0.242	0.206	0.199	0.162	0.193	1	0.214	0.211	0.205	0.262
MO	0.171	0.224	0.229	0.245	0.234	0.243	0.224	0.249	0.214	1	0.236	0.236	0.275
MV	0.136	0.204	0.207	0.199	0.216	0.237	0.220	0.226	0.211	0.236	1	0.210	0.244
OP	0.152	0.202	0.249	0.261	0.192	0.218	0.197	0.198	0.205	0.236	0.210	1	0.244
PCM	0.171	0.225	0.289	0.270	0.230	0.226	0.197	0.243	0.262	0.275	0.244	0.244	1

Tabella A.3: Indice di Connessione Lessicale senza hapax, tra coppie di autori.

	AF	AN	CM	DDL	DR	DV	FP	LB	LP	MO	MV	OP	PCM
AF	0	0.805	0.661	0.661	0.783	0.791	0.836	0.816	0.730	0.770	0.822	0.760	0.724
AN	0.678	0	0.487	0.558	0.596	0.670	0.721	0.652	0.543	0.642	0.693	0.607	0.546
CM	0.705	0.731	0	0.629	0.703	0.743	0.793	0.747	0.595	0.720	0.756	0.660	0.603
DDL	0.705	0.697	0.516	0	0.655	0.716	0.770	0.711	0.585	0.674	0.742	0.603	0.573
DR	0.680	0.639	0.494	0.551	0	0.683	0.776	0.730	0.578	0.649	0.694	0.646	0.567
DV	0.598	0.615	0.429	0.518	0.586	0	0.698	0.672	0.518	0.586	0.627	0.541	0.500
FP	0.545	0.532	0.340	0.438	0.579	0.566	0	0.549	0.477	0.523	0.574	0.464	0.430
LB	0.600	0.543	0.367	0.447	0.603	0.630	0.647	0	0.490	0.550	0.620	0.543	0.427
LP	0.728	0.721	0.529	0.632	0.712	0.748	0.810	0.763	0	0.721	0.740	0.690	0.607
MO	0.602	0.626	0.443	0.504	0.589	0.629	0.703	0.642	0.523	0	0.647	0.541	0.454
MV	0.636	0.621	0.426	0.536	0.577	0.605	0.687	0.643	0.473	0.583	0	0.539	0.448
OP	0.705	0.707	0.518	0.569	0.705	0.707	0.762	0.741	0.622	0.673	0.722	0	0.448
PCM	0.689	0.691	0.485	0.577	0.670	0.708	0.769	0.703	0.561	0.644	0.696	0.696	0

Tabella A.4: Indice di Indipendenza Lessicale senza hapax, tra coppie di autori.

AUT	hap	no hap	AUT	hap	no hap
AF	0.804	0.763	LB	0.644	0.538
AN	0.712	0.616	LP	0.747	0.701
CM	0.761	0.698	MO	0.704	0.575
DDL	0.690	0.662	MV	0.688	0.564
DR	0.708	0.641	OP	0.732	0.656
DV	0.644	0.574	PCM	0.735	0.657
FP	0.667	0.501			

Tabella A.5: Media dell'indice di Indipendenza Lessicale con e senza hapax, tra coppie, per autore

A.2 Tabelle e figure del capitolo 3

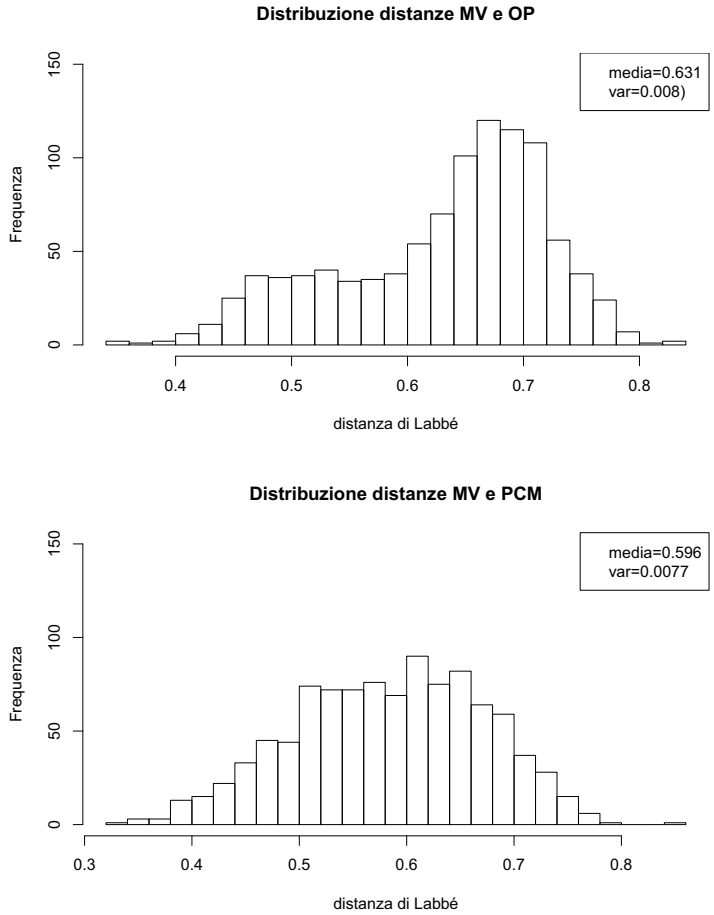

Figura A.1: Distribuzioni campionarie delle distanze intertestuali tra MV e gli altri autistici

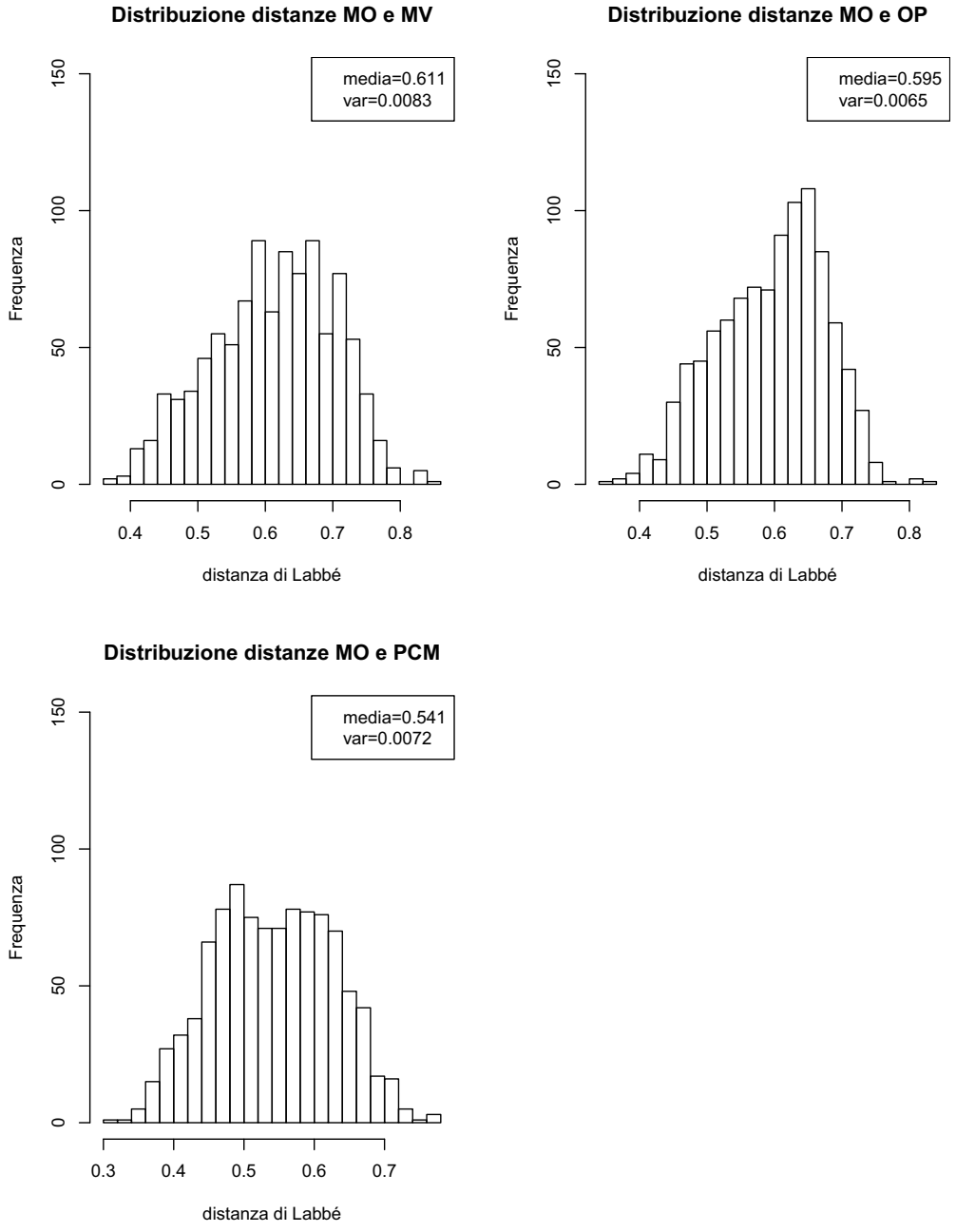

Figura A.2: Distribuzioni campionarie delle distanze intertestuali tra MO e gli altri autistici

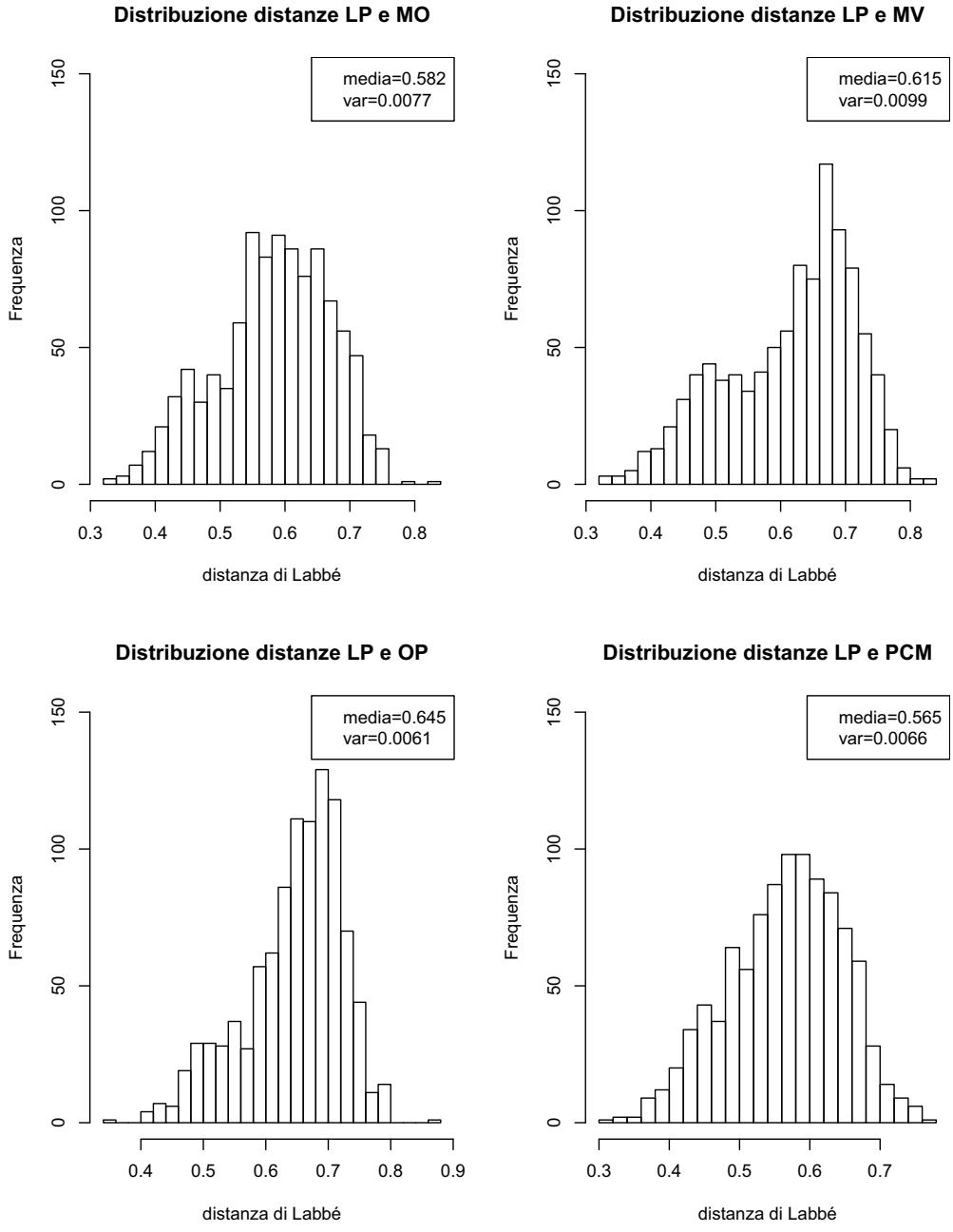

Figura A.3: Distribuzioni campionarie delle distanze intertestuali tra LP e gli altri autistici

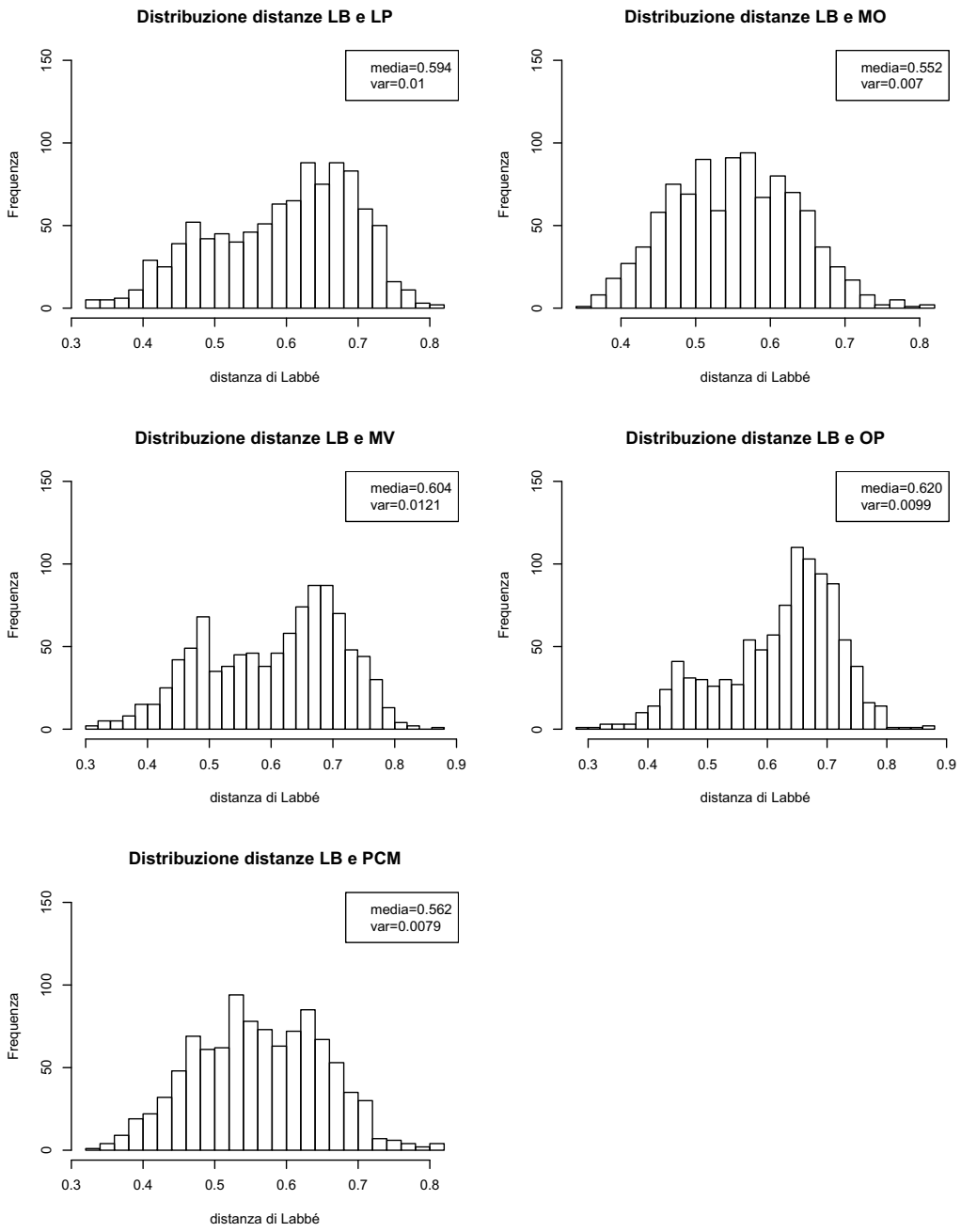

Figura A.4: Distribuzioni campionarie delle distanze intertestuali tra LB e gli altri autistici

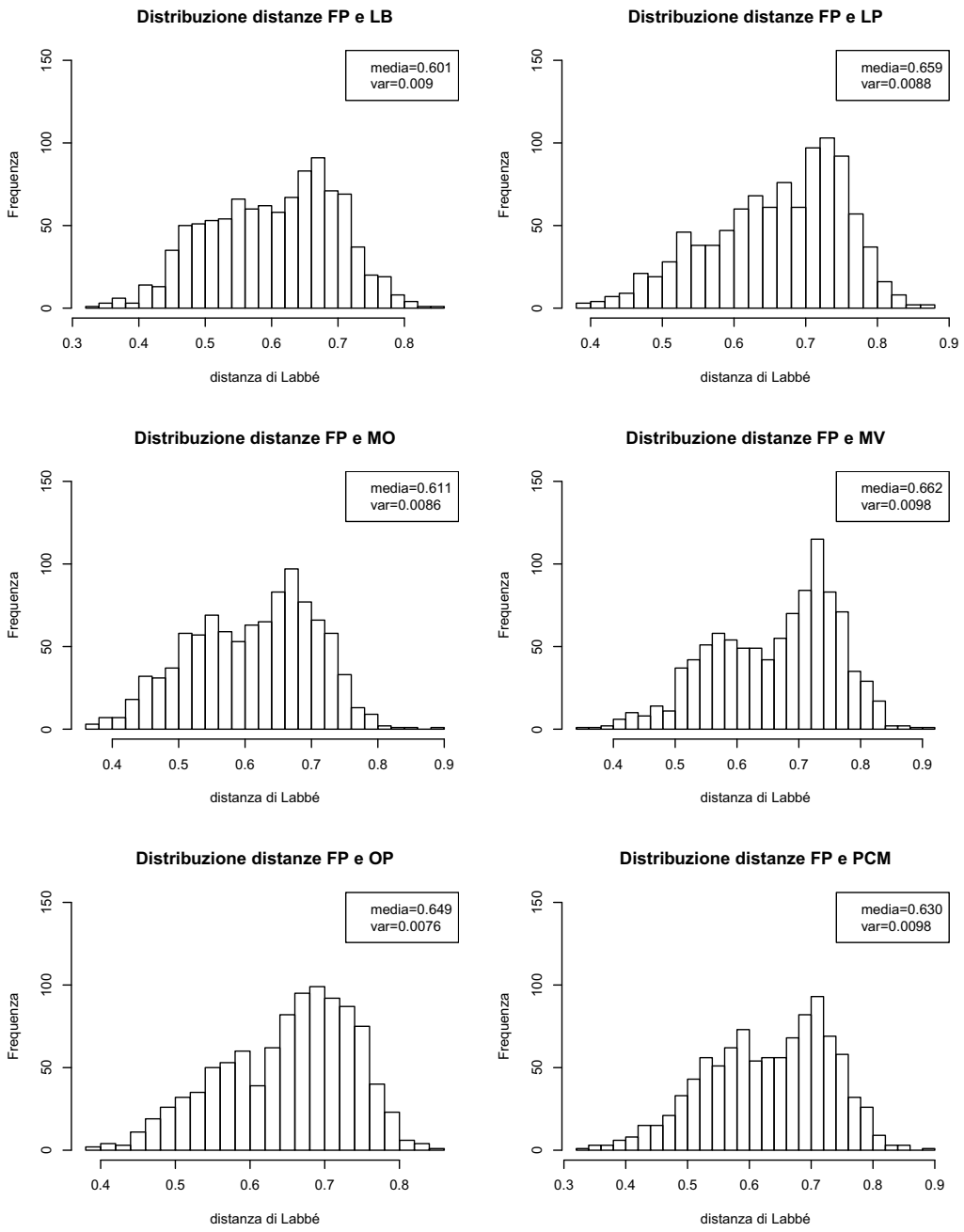

Figura A.5: Distribuzioni campionarie delle distanze intertestuali tra FP e gli altri autistici

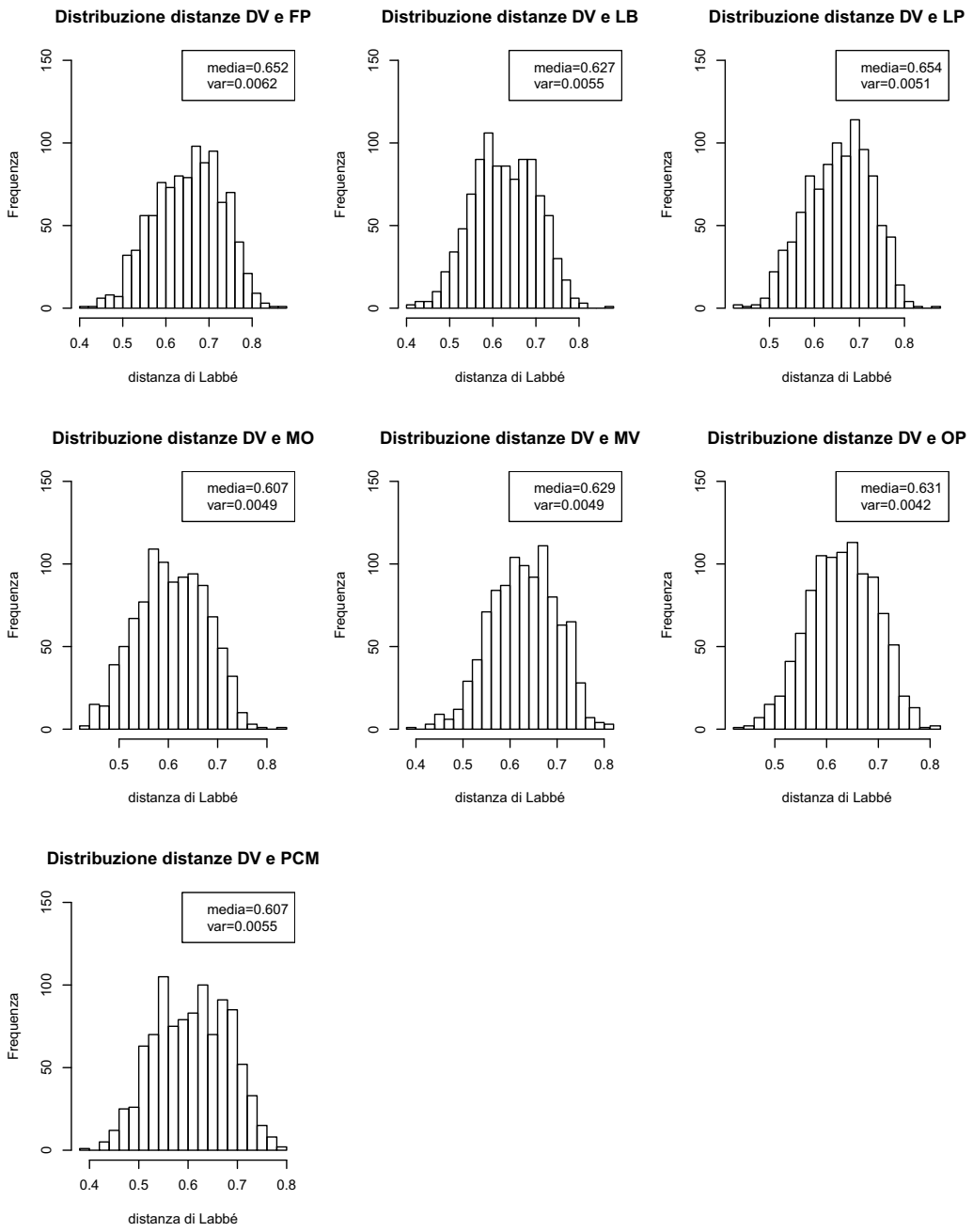

Figura A.6: Distribuzioni campionarie delle distanze intertestuali tra DV e gli altri autistici

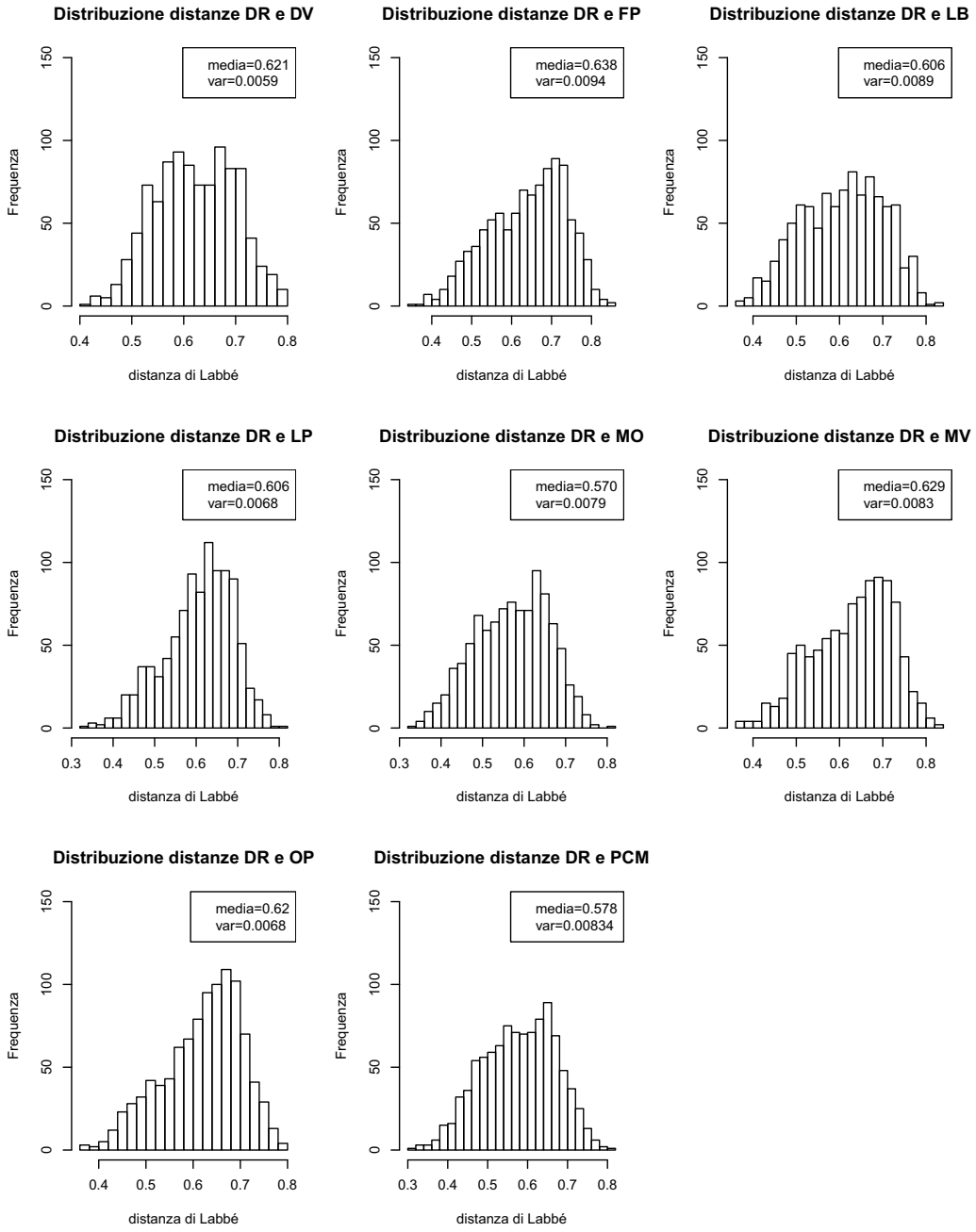

Figura A.7: Distribuzioni campionarie delle distanze intertestuali tra DR e gli altri autistici

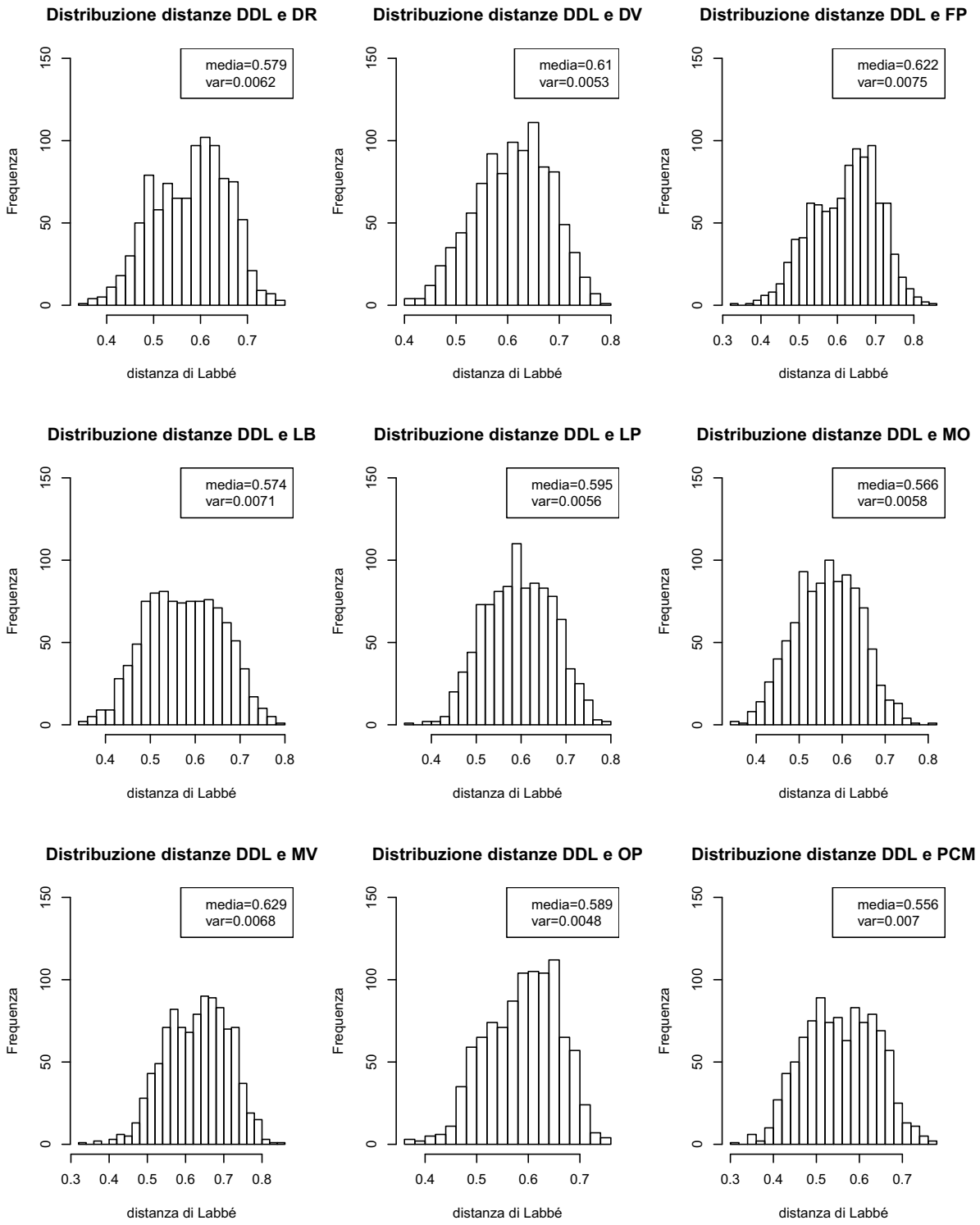

Figura A.8: Distribuzioni campionarie delle distanze intertestuali tra DDL e gli altri autistici

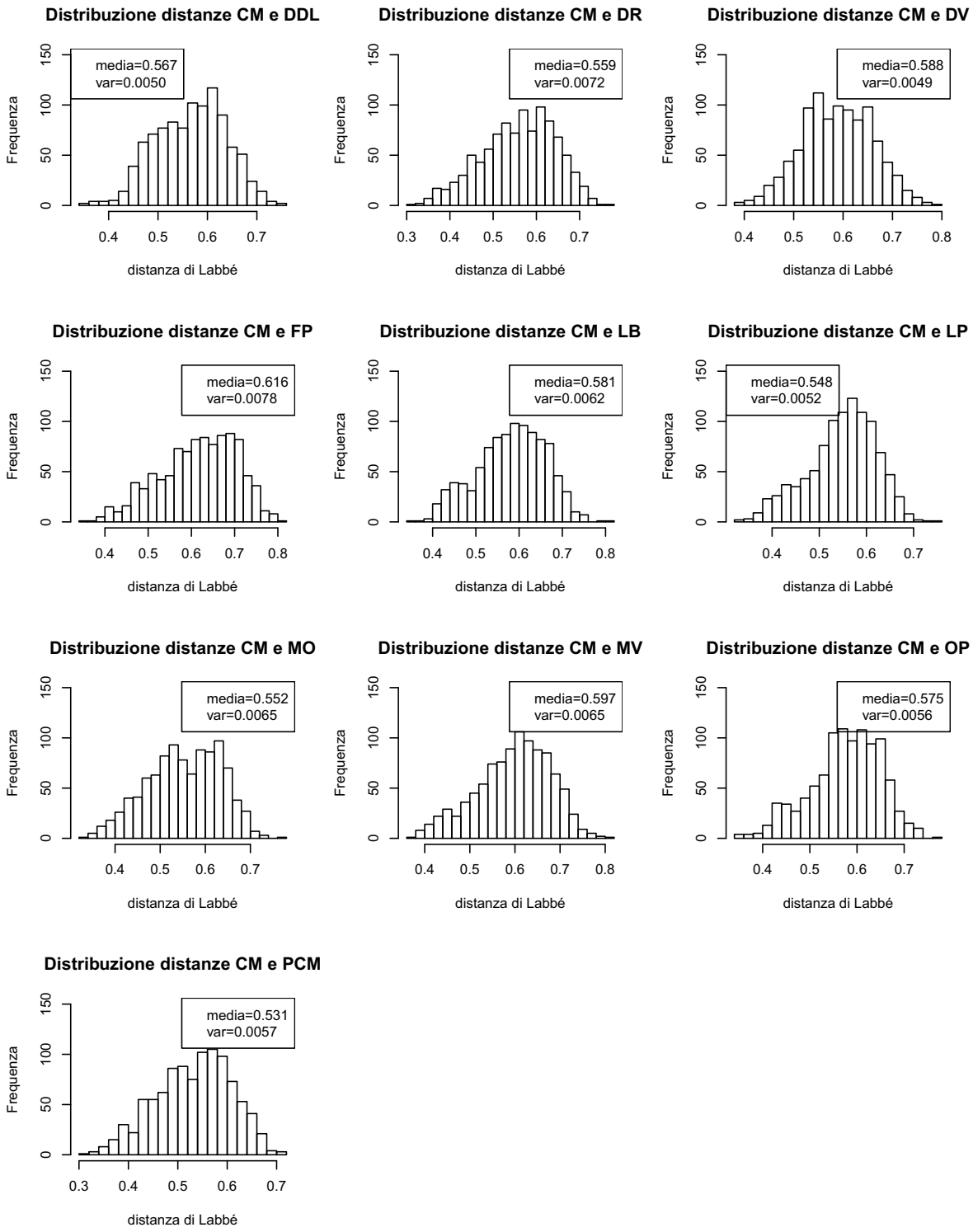

Figura A.9: Distribuzioni campionarie delle distanze intertestuali tra CM e gli altri autistici

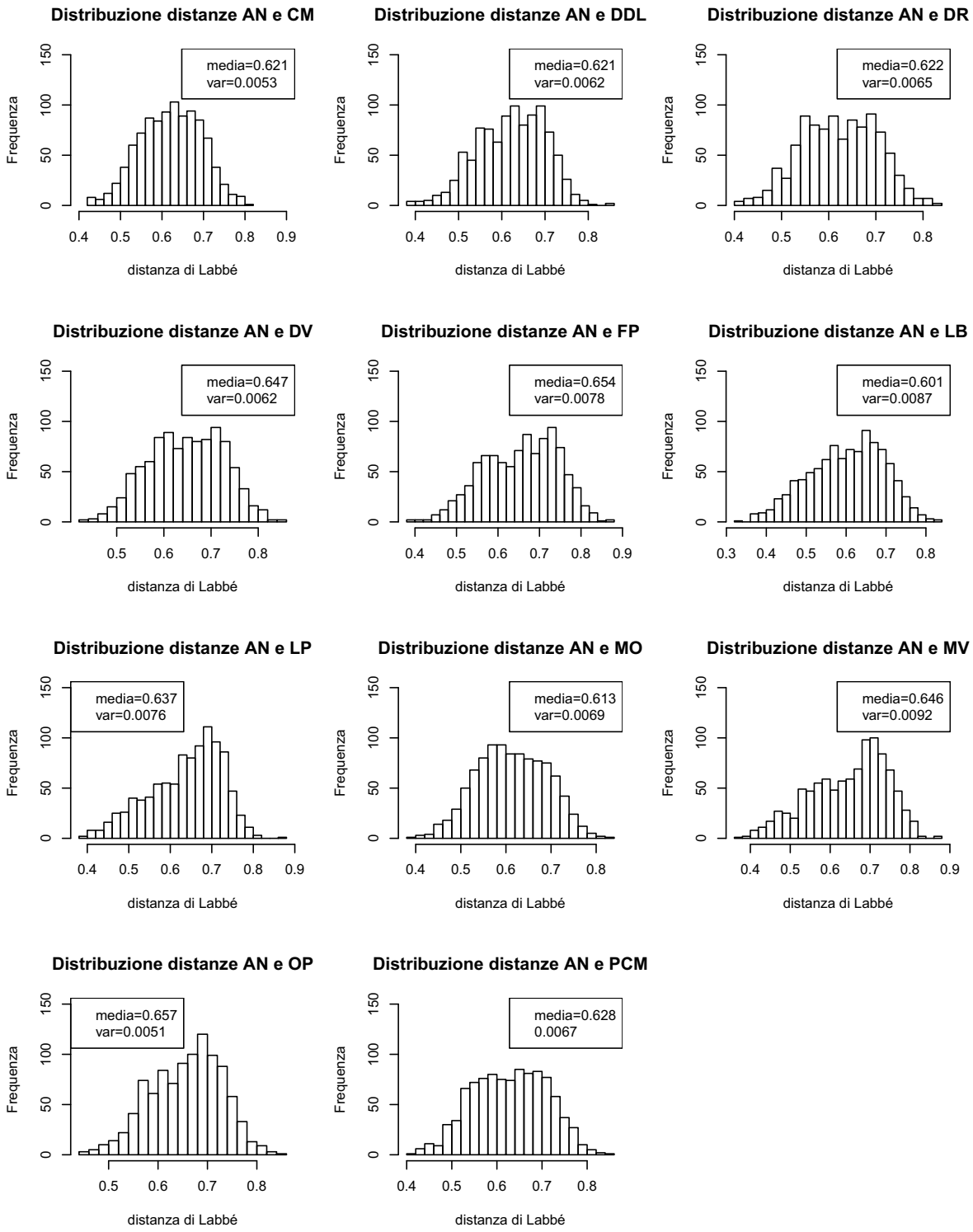

Figura A.10: Distribuzioni campionarie delle distanze intertestuali tra AN e gli altri autistici

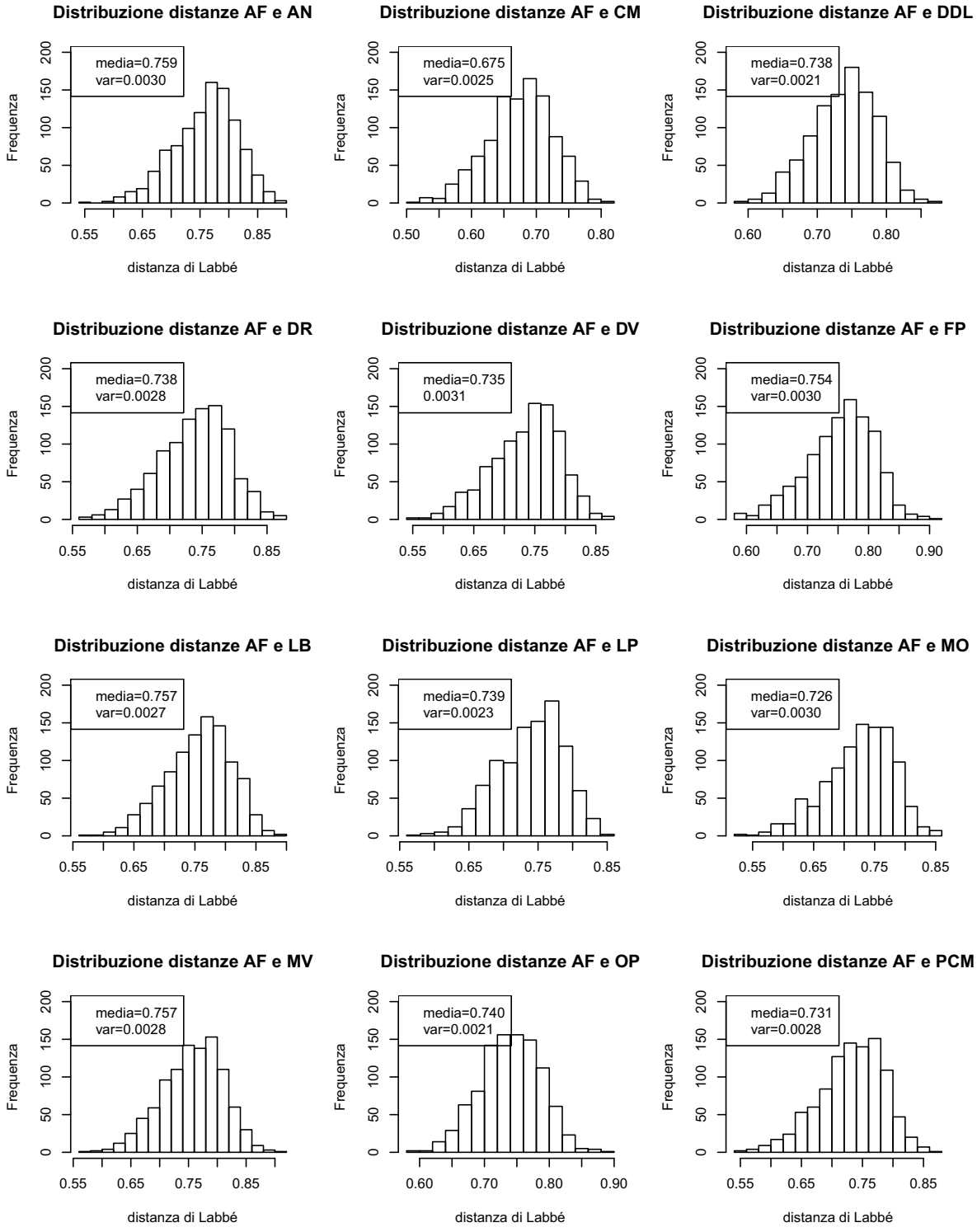

Figura A.11: Distribuzioni campionarie delle distanze intertestuali tra AF e gli altri autistici

Elenco delle tabelle

Elenco delle figure

Bibliografia

[1] R.H. Baayen (2001), *Word frequency distribution*, Kluwer Ac. Publishers, Dordrecht.

[2] L. Bernardi (2008), *Il delta dei significati. Uno studio interdisciplinare sull'espressione autistica*, a cura di, Carocci Faber, Roma.

[3] L. Bernardi (2005, ed), *Percorsi di ricerca sociale*, Carocci, Roma.

[4] M. Cortelazzo, A. Tuzzi (2008), *Metodi statistici applicati all'italiano*, Zanichelli, Bologna.

[5] M. Cortelazzo, A. Tuzzi (2007, eds), *Messaggi dal Colle. I discorsi di fine anno dei presidenti della Repubblica*, Marsilio, Venezia.

[6] L. Fabbris (1997), *Statistica multivariata*, McGraw–Hill, Milano.

[7] C. Labbé, D. Labbé (2001), *Inter–textual distance and Autorship Attribution. Corneille and Molière.*, Journal of Quantitative Linguistics, 8:3, 213–231.

[8] D. Labbé (2007), *Experiments on autorship attribution by intertextual distance*, Journal of Quantitative Linguistics, 14: 1, 33–80.

[9] L. Lamport (1994), *LaTeX: a document preparation system*, Addison–Wesley, Reading, Massachusetts.

[10] D. Piccolo (1998), *Statistica*, Il Mulino, Bologna.

[11] R Development Core Team (2010), *R Foundation for Statistical Computing*, http://www.R-project.org.

[12] A. Tuzzi (2003), *L'analisi del contenuto*, Carocci, Roma.

[13] A. Tuzzi (2005), *Analisi statistica del contenuto*, in L. Bernardi (a cura di), *Percorsi di ricerca sociale. Conoscere, decidere, valutare.* Carocci, Roma.

[14] P. Venuti (2003), *L'autismo*, pp. 17-20 Carocci, Roma.