DOI: 10.13137/978-88-8303-913-3/18480

3. Distanza intertestuale e lingua fonte: analisi di un corpus giornalistico

STEFANO ONDELLI
Università di Trieste
PAOLO NADALUTTI
Gruppo Interdisciplinare di Analisi Testuale

Abstract

This chapter illustrates the results of a revised method for calculating the intertextual distance between newspaper articles originally written in Italian and translated from other languages. Starting from the theoretical background provided by the translation universals hypothesis, we have used five different parsing and token-selection criteria to check whether intertextual distance measures can distinguish native texts from translations and group together texts translated from the same source language. Although the combined impact of several factors (source language, contents, author and translator) needs to be taken into account, results show that translations tend to be mutually closer, while intertextual distance measures are greater between translations and non-translated texts (and vice versa). In addition, although the distinction is not as clear-cut, translations from the same language tend to group together since they are intertextually closer than translations from other source languages. However, further research is necessary to explain the erratic results obtained when we have used grammar words to calculate the intertextual distance.

Keywords

Computational linguistics, corpus linguistics, intertextual distance, translation universals, translation studies

1. Introduzione

Questo capitolo[1] illustra i risultati dell'applicazione del metodo della distanza intertestuale (Labbé & Labbé 2001, rivisto secondo le considerazioni contenute in Cortelazzo et al. 2013 e Tuzzi 2010) a due subcorpora di articoli di giornale scritti originariamente in italiano (d'ora in avanti "subcorpus nativo" e "testi nativi") e tradotti in italiano da diverse lingue (d'ora in avanti "subcorpus tradotto" e "traduzioni" o "testi tradotti"). La composizione dei subcorpora e i metodi di trattamento sono descritti nel capitolo precedente del presente volume. In questa prima fase ci limiteremo a considerare i trattamenti *a*, *b*, *c*, *d* ed *e*, rimandando ad altra occasione le indagini che possono essere svolte a partire da testi sottoposti a *POS-tagging* e lemmatizzazione.

Le domande a cui si cerca risposta con gli esperimenti che seguono sono principalmente tre:

a) la distanza intertestuale permette di distinguere i testi nativi dai testi tradotti?
b) La distanza intertestuale permette di identificare i diversi gruppi di testi tradotti in base alle lingue di partenza?
c) Tra tutti i metodi di trattamento dei corpora proposti, quale risulta il più efficace?

A ciascuna di queste domande sarà dedicato uno dei paragrafi che seguono: nel §2 indaghiamo le differenze tra articoli tradotti e articoli nativi, mostrando come la distanza intertestuale si riveli uno strumento efficace per operare tale distinzione. Nel §3 usiamo la distanza intertestuale per verificare, invece, come testi tradotti in italiano a partire da diverse lingue mostrino delle somiglianze riconducibili all'influenza della lingua di origine. Nel §4, infine, traiamo le conclusioni di queste prime applicazioni della distanza intertestuale alle traduzioni e prospettiamo ulteriori ricerche tese a valutare il ruolo dei vari fattori in gioco. Prima di procedere oltre, è bene ricordare che il corpus in esame mal si presta a rispondere al quesito di ricerca numero 3 esposto al §4 del capitolo precedente, e cioè quale variabile, tra lingua di partenza, autore del testo fonte e traduttore, sia preminente nel calcolo della distanza intertestuale tra traduzioni. Tale diffi-

1 La ricerca e i testi che la illustrano sono il frutto di un approccio interdisciplinare che ha visto la piena collaborazione di entrambi gli autori sotto tutti i punti di vista. A soli fini dell'attribuzione di questo capitolo, specifichiamo che Stefano Ondelli ha redatto i paragrafi 1 e 2 e Paolo Nadalutti i paragrafi 3 e 4.

coltà discende dal fatto che diversi testi attribuiti allo stesso traduttore in realtà sono opera non di persone fisiche ma di agenzie che si avvalgano di collaboratori diversi per traduzioni non solo da lingue diverse ma anche dalla stessa lingua.

2. Macrotesti tradotti e macrotesti nativi

La domanda da cui prende le mosse questa prima ricerca è la seguente: se il "traduttese" (Trosborg 1997 e Frawley 2000) presenta caratteristiche tendenziali che lo distinguono dall'italiano prodotto direttamente da parlanti nativi, se ne può rilevare l'impatto sulla distanza tra i testi? In pratica, avendo a disposizione due subcorpora (traduzioni e testi nativi), se calcoliamo la distanza tra i vari testi che li compongono, quando mettiamo a confronto due traduzioni o due testi nativi, rileveremo sempre valori inferiori rispetto a quelli ottenuti dal confronto tra una traduzione e un testo nativo?

La Tabella 1 offre i primi dati utili per rispondere al nostro quesito. Ricordiamo che, poiché la distanza intertestuale è sensibile alle dimensioni e ai contenuti dei testi, non è stato possibile utilizzare direttamente i singoli articoli, ma abbiamo confrontato tra loro 30 macrotesti ottenuti aggregando diverse traduzioni (con lingue di partenza e traduttori diversi) e 30 macrotesti ottenuti aggregando diversi articoli nativi (di autori diversi) secondo la procedura di campionamento descritta al capitolo 2:§4 in questo volume (200 campionamenti di *chunks* di 3.500 occorrenze ciascuno), calcolando poi la distanza sia sull'intero vocabolario sia su sottoinsiemi di parole grammaticali.

Tabella 1. Media e deviazione standard della distanza intertestuale campionata.

Trattamento	*Distanza campionata media tra testi del subcorpus tradotto*	*Distanza campionata media tra testi del subcorpus nativo*	*Deviazione standard media subcorpus tradotto*	*Deviazione standard media subcorpus nativo*	*Distanza campionata media tra testi nativi e tradotti*	*Deviazione standard media per distanze tra testi nativi e tradotti*
a) Normalizzazione leggera	0,529	0,523	0,007	0,009	0,536	0,009
b) Polirematiche	0,559	0,553	0,007	0,009	0,567	0,008
c) Locuzioni	0,558	0,552	0,007	0,009	0,566	0,008
d) Grammaticali	0,138	0,138	0,008	0,010	0,144	0,010
e) Grammaticali + Locuzioni	0,138	0,132	0,024	0,018	0,154	0,012

Come possiamo vedere, la media delle distanze intertestuali calcolate tra i macrotesti nativi e tradotti risulta sempre (e significativamente) maggiore rispetto alla media calcolata internamente ai due subcorpora, il che indica una maggiore vicinanza reciproca tra le traduzioni (e tra i testi nativi) rispetto a quando il confronto avviene tra i due subcorpora (cfr. anche la Figura 2 sotto). Nel dettaglio, la media delle distanze tra i macrotesti tradotti non si allontana molto dalla media riferita ai macrotesti nativi, anche se quest'ultima è sistematicamente minore (sorprendentemente, la differenza è costante in *a*, *b* e *c*, pari a 0,006). Al contrario, le deviazioni standard sono sistematicamente più basse nel subcorpus delle traduzioni (con l'eccezione del trattamento *e*). Da queste osservazioni possiamo concludere che i macrotesti nativi presentano una somiglianza reciproca più marcata rispetto ai macrotesti tradotti. Ciò potrebbe essere dovuto all'influenza delle lingue fonte: mentre per i macrotesti nativi questo fattore è assente, i campioni estratti dai macrotesti tradotti e sottoposti a confronto potrebbero comprendere di volta in volta lingue di partenza diverse, che determinano una distanza reciproca leggermente maggiore.

Una spiegazione alternativa (o aggiuntiva, poiché i due fattori non si escludono a vicenda) potrebbe fare riferimento ai contenuti: oltre all'impatto della lingua di partenza, le traduzioni potrebbero essere caratterizzate da contenuti più variabili e condurre a un piccolo incremento della distanza intertestuale. Questa ipotesi sembra essere corroborata dai risultati ottenuti con i due trattamenti che dovrebbero riuscire meglio a limitare l'effetto dei contenuti: *d* (grammaticali) ed *e* (grammaticali + locuzioni). Come si può vedere, la normalizzazione leggera (*a*) e ancor più le polirematiche (*b*) colgono maggiormente la composizione lessicale dei macrotesti che, probabilmente proprio in virtù dell'incidenza dei contenuti, risultano reciprocamente più distanti rispetto a quando calcoliamo la distanza intertestuale considerando solo le parole grammaticali. Anzi, in questo caso traduzioni e testi nativi risultano equidistanti all'interno dei propri subcorpora. Le locuzioni (*c*) risultano invece essere un elemento di disturbo: da una parte il relativo trattamento produce distanze intertestuali inferiori solo alle polirematiche (quindi sembrerebbero risentire dell'effetto dei contenuti), dall'altra, unitamente ai grammaticali, ottengono una distanza media pari ai soli grammaticali tra i macrotesti tradotti e addirittura inferiore tra i macrotesti nativi. Come già evidenziato (capitolo 2:§5), la natura composita di questa componente delle risorse statistico-linguistiche disponibili nel software *Taltac*[2] non ci permette di offrire spiegazioni valide per un simile comportamento ondivago.

Anche per quanto concerne le deviazioni standard delle distanze intertestuali, notiamo che la differenza tra i valori riferiti alle traduzioni e ai macrotesti nativi è molto ridotta, sebbene risulti costantemente maggiore nel secondo dei due subcorpora e, seppure in misura minore, nel confronto tra macrotesti tradotti e nativi (con l'eccezione del trattamento *e*, che addirittura produce il valore più basso nel confronto tra subcorpora). Ciò significa che c'è minore omogeneità tra i valori delle distanze intertestuali tra gli articoli scritti direttamente in italiano,

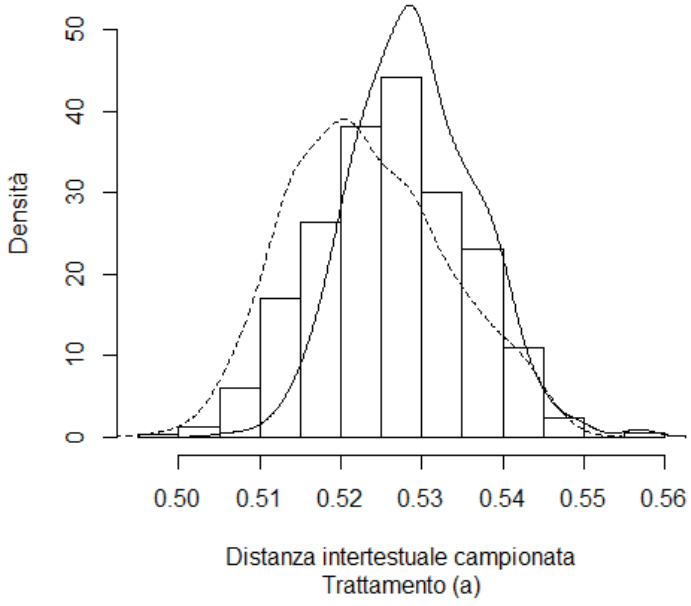
un risultato che parrebbe confermare l'universale traduttivo della convergenza o *levelling out*: "steering a middle course between any two extremes, converging towards the center, with the notion of center and periphery being defined from within the translation corpus itself" (Baker 1996: 184). In altre parole, nei nostri subcorpora i macrotesti tradotti tendono a essere mediamente meno simili tra loro rispetto ai macrotesti nativi (tra loro), ma ci sono meno traduzioni che sono molto differenti dalle altre; di converso i macrotesti nativi tendono a essere più omogenei, ma presentano alcuni casi che si discostano marcatamente dalla media. Insomma, la più alta deviazione standard tra i macrotesti nativi sta a indicare che le rispettive distanze intertestuali sono più "perturbate" rispetto alle distanze intertestuali tra macrotesti tradotti.

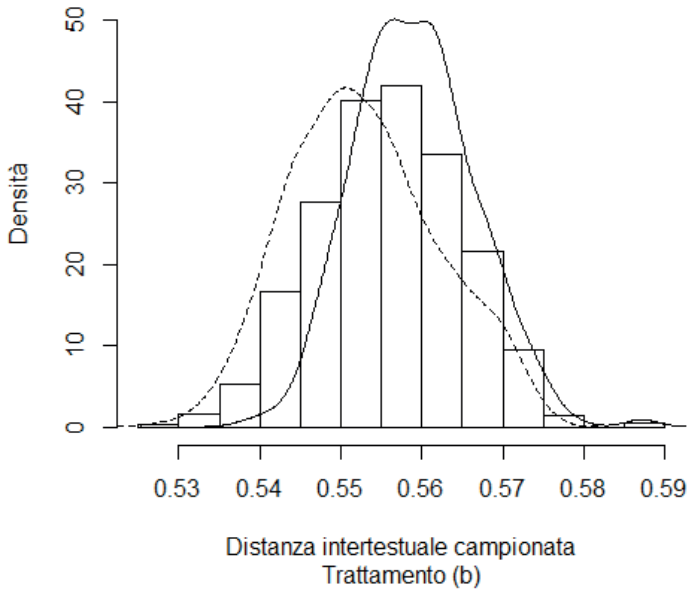
Queste conclusioni valgono per tutti i trattamenti a cui è stato sottoposto il corpus, con un leggero incremento della deviazione standard della distanza intertestuale calcolata considerando solo le parole grammaticali, ma anche con la notevole eccezione del trattamento *e*. In questo caso i valori all'incirca triplicano (per le traduzioni) o raddoppiano (per i macrotesti nativi), così ribaltando la situazione descritta sopra: sono le distanze intertestuali delle traduzioni ad avere una distribuzione più perturbata e non è chiaro il motivo per cui questo trattamento conduca a risultati così eccentrici rispetto agli altri.

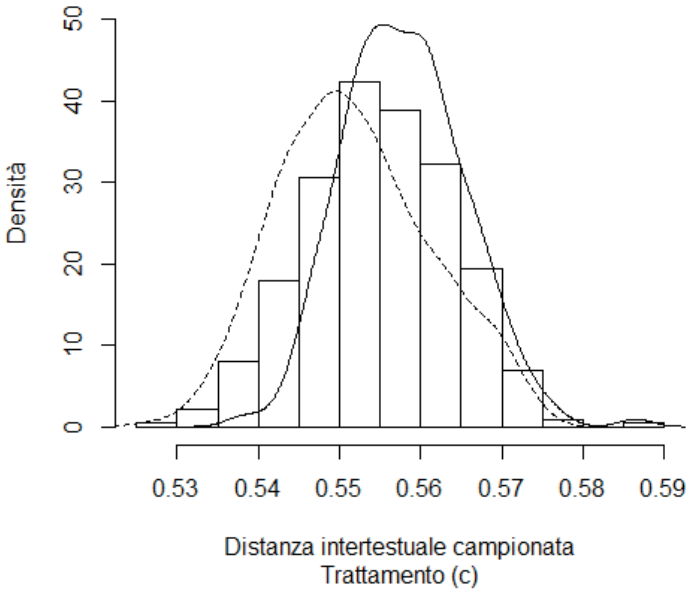
Finora abbiamo considerato la distanza intertestuale media e la sua variazione standard all'interno di ciascun subcorpus (tradotto e nativo), ma ciò che ci preme è, ovviamente, sapere se la distanza intertestuale tra le traduzioni risulti (sempre) minore rispetto alla distanza intertestuale tra una traduzione e un testo nativo, e viceversa. Grazie alla Figura 1 qui sotto possiamo visualizzare la distribuzione della distanza intertestuale per i cinque trattamenti e notare come le distanze tra i macrotesti tradotti (linea continua) abbiano distribuzioni diverse dalle distanze tra macrotesti non tradotti (linea tratteggiata).

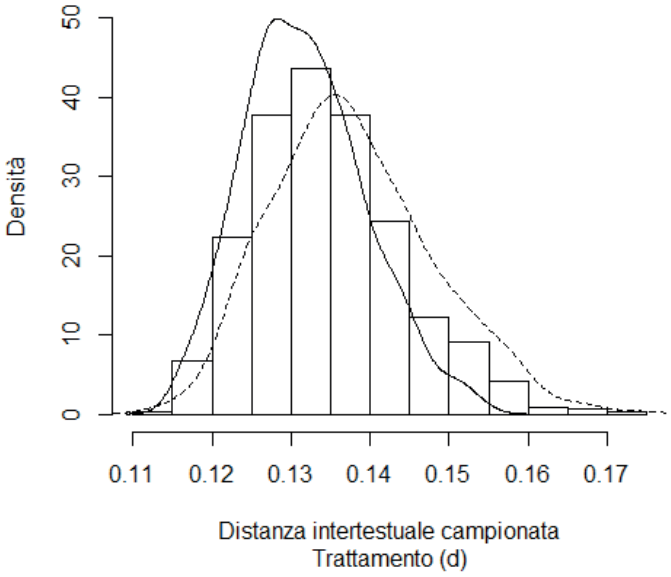
Rispetto alla distribuzione generale (i rettangoli al centro), la linea continua tende a formare curve più alte e caratterizzate da andamenti più "ripidi" e basi più "strette" della linea tratteggiata, che però si posiziona quasi sempre "a sinistra" delle linee continue (fa eccezione il trattamento *d*): ciò indica che i macrotesti tradotti sono tutti più distanti tra loro, ma lo sono in maniera costante, mentre i macrotesti nativi risultano reciprocamente più vicini ma c'è maggiore variabilità interna. Infine, i grafici dimostrano che le medie riportate in Tabella 1 non sono solamente frutto del caso o di picchi che inficiano le distribuzioni.

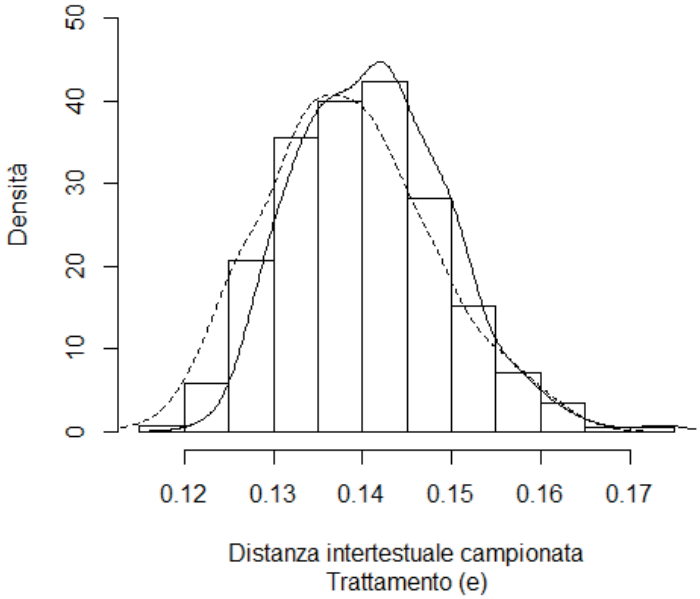
Figura 1. Distribuzione delle distanze intertestuali tra macrotesti nativi e tradotti.





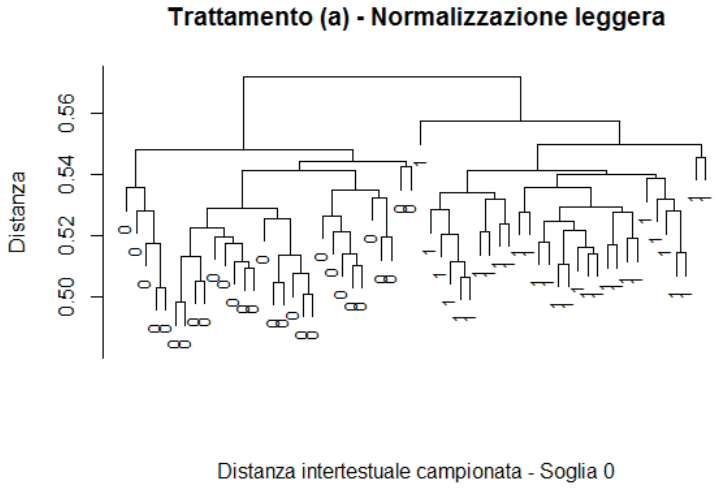

Densità
0
10
20
30
40
50
0.53
0.54
0.55
0.56
0.57
0.58
0.59
Distanza intertestuale campionata
Trattamento (c)

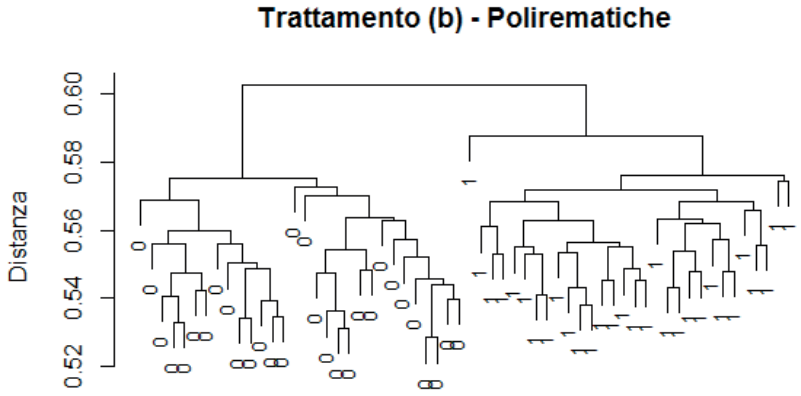

Densità
0
10
20
30
40
50
0.11
0.12
0.13
0.14
0.15
0.16
0.17
Distanza intertestuale campionata
Trattamento (d)

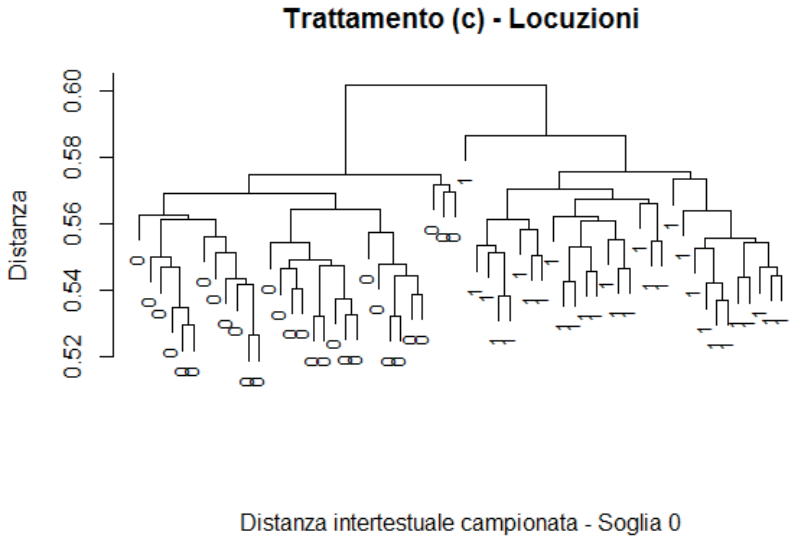


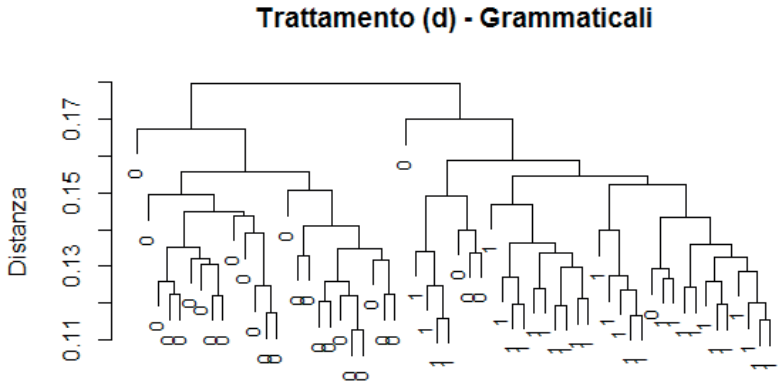
Per illustrare più efficacemente i risultati dei nostri calcoli uno strumento molto utile è il dendrogramma. Il dendrogramma consente di visualizzare in maniera intuitiva gruppi diversi di elementi (nel nostro caso: i macrotesti) e come questi siano collegati tra loro in base a una misura di distanza reciproca (nel nostro caso la distanza intertestuale campionata). Ogni "foglia" dell'albero rovesciato che costituisce il dendrogramma rappresenta uno di questi elementi. Gli elementi stessi sono collegati da linee e finiscono per formare dei raggruppamenti, mentre l'asse delle ordinate indica la distanza a cui due macrotesti vengono collegati.

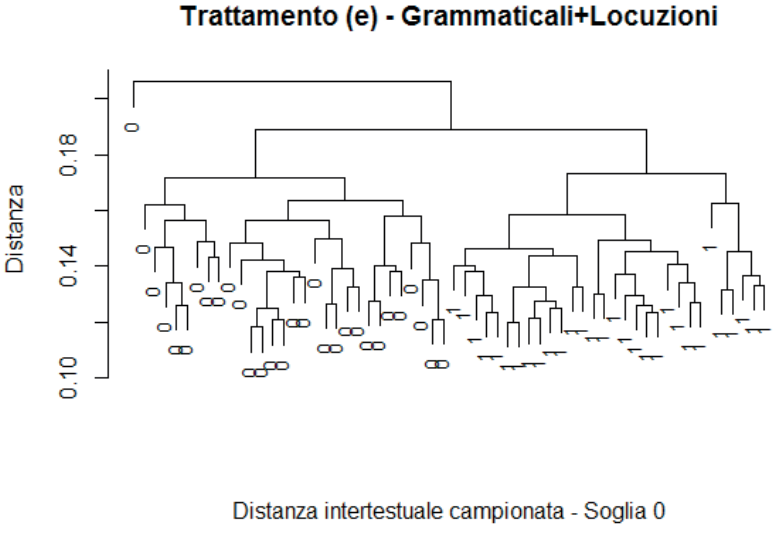
Figura 2. Dendrogrammi per i 60 macrotesti (0 = nativi, 1 = tradotti).




Trattamento (b) - Poliremantiche
Distanza
0.52
0.54
0.56
0.58
0.60
Distanza intertestuale campionata - Soglia 0


Trattamento (c) - Locuzioni
Distanza
0.52
0.54
0.56
0.58
0.60
Distanza intertestuale campionata - Soglia 0


Trattamento (d) - Grammaticali
Distanza
0.11
0.13
0.15
0.17
Distanza intertestuale campionata - Soglia 0


Trattamento (e) - Grammaticali+Locuzioni
Distanza
0.10
0.14
0.18
Distanza intertestuale campionata - Soglia 0

La Figura 2 presenta i dendrogrammi relativi ai diversi trattamenti a cui sono stati sottoposti i nostri subcorpora. I grafici sono stati generati secondo un processo agglomerativo, tramite il metodo del legame completo. Ricordiamo che i processi di *clustering* agglomerativi mirano a raggruppare un insieme di unità statistiche di base (nel nostro caso i macrotesti) in un insieme di gruppi meno numerosi delle unità statistiche di base stesse. Nella fattispecie della tecnica usata in questo capitolo, il *clustering* viene definito gerarchico, in quanto l'insieme di gruppi individuati è caratterizzato da relazioni di appartenenza univoche a gruppi più ampi. Nel dettaglio, la tecnica usata segue questo processo: individua, nell'insieme di macrotesti, i due macrotesti che hanno la minore distanza reciproca (cioè individua i macrotesti che sono più simili tra loro) e li associa, formando un gruppo (ed è per questo che la tecnica è definita "agglomerativa"). Il nuovo gruppo appena creato viene ora trattato come se fosse una delle unità statistiche di base, perciò bisogna procedere a ricalcolare la distanza tra questo nuovo macrotesto frutto dell'unione di due unità di base e tutte le altre unità di base. Qui entra in gioco il metodo del legame completo: per calcolare la distanza tra il neo-gruppo e tutte le altre unità statistiche di base viene presa la più alta tra le distanze delle unità statistiche che fanno parte del neo-gruppo e le unità statistiche di base. Il processo viene dunque ripetuto fino a quando tutti i gruppi e le unità statistiche di base confluiscono in un gruppo unico.

Come possiamo vedere, con i primi tre trattamenti (*a, b* e *c*) i macrotesti tradotti (contrassegnati con il numero 1) sono sempre più vicini agli altri macrotesti tradotti, mentre i macrotesti nativi (contrassegnati con 0) sono più vicini agli altri macrotesti nativi. A titolo di esempio, per leggere il dendrogramma relativo alla normalizzazione leggera, occorre fare riferimento alla scala riportata sull'asse delle ordinate. I 30 macrotesti tradotti presentano una distanza intertestuale che va da poco meno di 0,50 a poco meno di 0,55, valori deducibili dalla lettura dell'asse delle ordinate in corrispondenza delle singole aggregazioni (linee orizzontali di collegamento tra macrotesti). Per meglio comprendere questi numeri, è utile tornare a consultare la Tabella 1, da cui rileviamo che la distanza intertestuale campionata media del subcorpus tradotto è pari a 0,529, cioè circa a metà strada tra 0,50 e 0,55.

Il fatto che i macrotesti trattati con le procedure *a, b* e *c* vengano posizionati nell'albero secondo una netta divisione tra traduzioni e testi nativi è un segnale evidente della reale esistenza di un "effetto traduzione": se dalla Tabella 1 si possono notare solamente i dati aggregati, con la Figura 2 invece vediamo come, in maniera sistematica, il macrotesto "più vicino" a un macrotesto nativo sia composto da testi nativi, e la stessa distribuzione vale per i macrotesti tradotti. Inoltre, tale vicinanza emerge anche per i vari raggruppamenti di macrotesti, fino ad una divisione in due parti del corpus tra traduzioni e testi nativi. Occorre infatti ricordare che il metodo agglomerativo, una volta associati due elementi, ricalcola le distanze tra tutti gli altri elementi e la coppia appena formata.

Se la divisione tra macrotesti nativi e traduzioni è netta con i primi tre trattamenti, la situazione si complica leggermente quando nel calcolo della distanza

tra macrotesti entrano in gioco le parole grammaticali. In combinazione con le locuzioni (trattamento *e*), c'è un solo macrotesto che risulta totalmente eccentrico, accoppiandosi a grande distanza (superiore a 0,20) con tutti gli altri macrotesti (nativi e tradotti). È difficile dire che cosa renda questo macrotesto così diverso dagli altri; quel che è certo è che il fattore di disturbo risiede nelle parole grammaticali: nel dendrogramma relativo al trattamento *d*, cinque macrotesti nativi "invadono" il campo delle traduzioni a diverse distanze. Anche in questi casi è difficile ipotizzarne la causa: oltre alla procedura di campionamento seguita, è proprio la selezione delle parole grammaticali che dovrebbe garantire il minimo impatto dei contenuti sul calcolo della distanza. Delle due l'una: o in questi cinque campionamenti si è creata qualche combinazione particolare (per es. delle lingue di partenza, ma è molto difficile), oppure si deve concludere che le sole parole grammaticali sono meno precise delle altre risorse linguistiche nel cogliere le specificità del "traduttese" (sul perché terneremo in sede di conclusioni). Resta il fatto che, in ultima analisi, sia all'interno del gruppo degli articoli tradotti sia all'interno degli articoli nativi, la distanza tra i macrotesti è tendenzialmente minore rispetto alla distanza con quelli dell'altro gruppo e, dopotutto, nei dendrogrammi si creano due gruppi (*cluster*) di macrotesti ben distinti, pur con qualche *misclassification*.

3. L'effetto della lingua di partenza

Nelle considerazioni presentate qui di seguito la distanza intertestuale viene utilizzata per verificare se emergano differenze tra macrotesti tradotti da lingue diverse. Come già illustrato nel §2, anche in questo caso i testi tradotti sono stati uniti e raggruppati in tre macrotesti per ogni lingua considerata, così da poter generare segmenti di dimensioni sufficienti a consentire il calcolo della distanza intertestuale secondo il campionamento descritto al capitolo 2 in questo volume. Per poter limitare quanto più possibile l'eventuale influenza dello stile individuale del traduttore, i tre macrotesti per ogni lingua sono stati generati in modo da suddividere i testi attribuiti agli stessi traduttori in modo casuale. Come sopra, ci affidiamo a due diversi approcci per verificare se i macrotesti originati dalla stessa lingua di partenza risultino reciprocamente più vicini rispetto agli altri macrotesti tradotti da altre lingue: prima l'analisi aggregata delle distanze intertestuali, poi un'analisi più puntuale illustrata tramite dendrogrammi.

Già in Tabella 2 possiamo notare come le distanze tra macrotesti tradotti a partire dalla stessa lingua siano mediamente inferiori rispetto ai valori relativi ai macrotesti tradotti da lingue diverse. Stavolta la differenza è costante in tutti i trattamenti, con uno scarto minimo nel caso di *e* e massimo quando si prendono in considerazione le poliremiatiche (che dovrebbe essere più efficaci nel cogliere i contenuti).

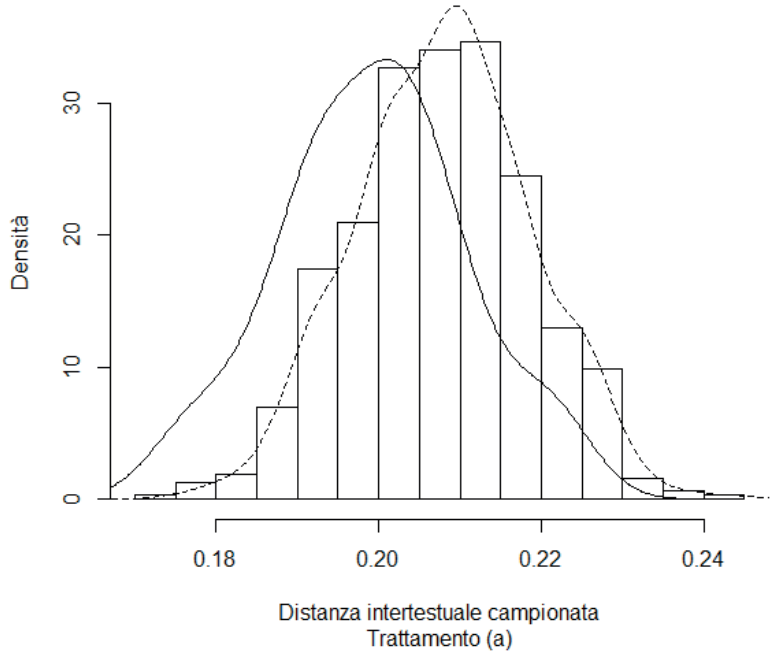
Tabella 2. Media della distanza intertestuale campionata e deviazione standard per macrotesti tradotti dalla stessa lingua e macrotesti tradotti da lingue diverse

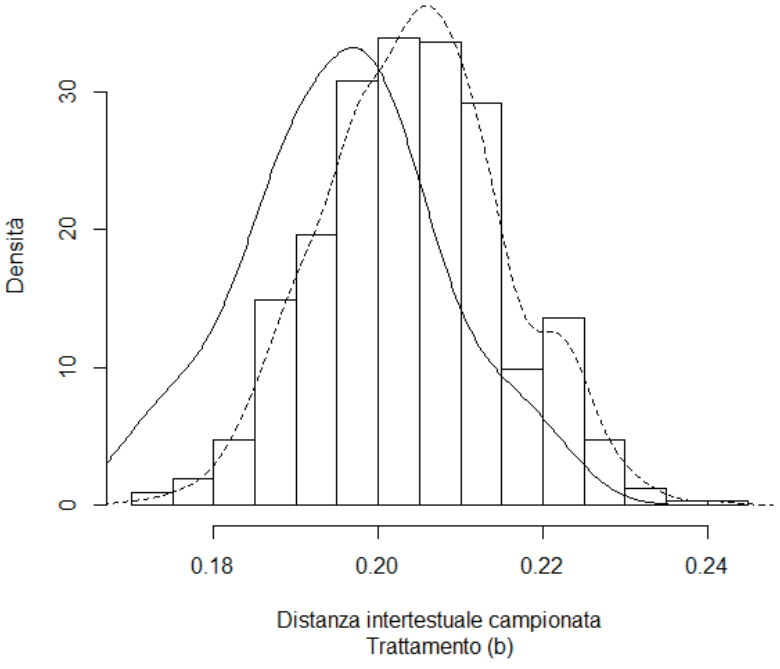
Trattamento	*Distanza campionata media infra-lingua*	*Deviazione standard infra-lingua*	*Distanza campionata media extra-lingua*	*Deviazione standard extra-lingua*
a) Normalizzazione leggera	0,199	0,012	0,208	0,011
b) Locuzioni	0,196	0,012	0,204	0,011
c) Polirematiche	0,196	0,012	0,204	0,011
d) Grammaticali	0,125	0,009	0,134	0,010
e) Grammaticali + Locuzioni	0,238	0,010	0,239	0,009

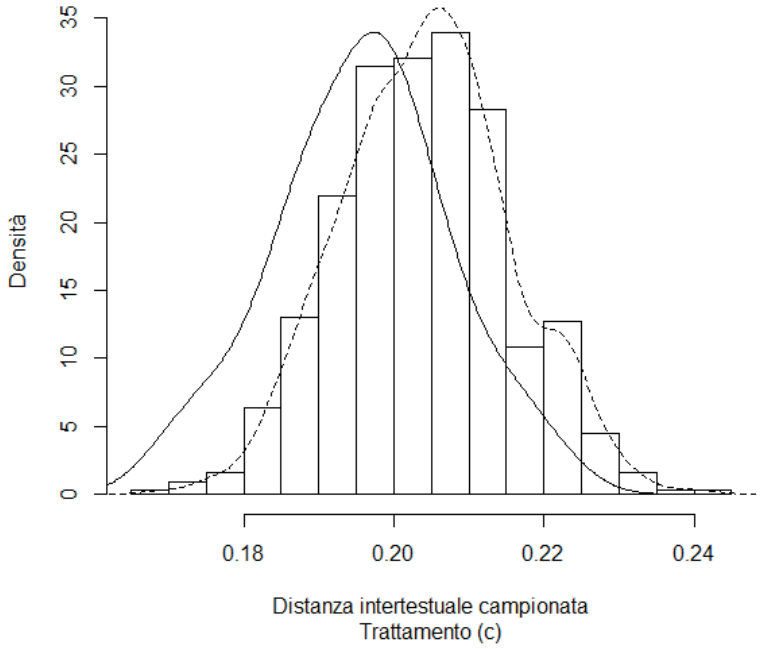
Come abbiamo già fatto nel caso del confronto tra testi nativi e testi tradotti (Figura 1), grazie ai grafici possiamo visualizzare nel dettaglio la distribuzione delle distanze intertestuali per ogni trattamento del corpus (Figura 3). Stavolta le linee continue rappresentano le distanze tra traduzioni dalla stessa lingua, mentre le linee tratteggiate rappresentano distanze tra macrotesti tradotti da lingue diverse. Analogamente a quanto avveniva in Figura 1, possiamo vedere come le linee continue si posizionino quasi sempre "a sinistra" delle linee tratteggiate (anche stavolta fa eccezione il trattamento *e*): ciò significa che nel primo subcorpus la distribuzione delle distanze è sbilanciata verso "il basso", e che le medie riportate in Tabella 2 non sono solamente frutto del caso o il risultato di picchi che inficiano la regolarità delle distribuzioni, a conferma della possibilità che esista un "effetto lingua fonte" rilevabile con la misura della distanza intertestuale.

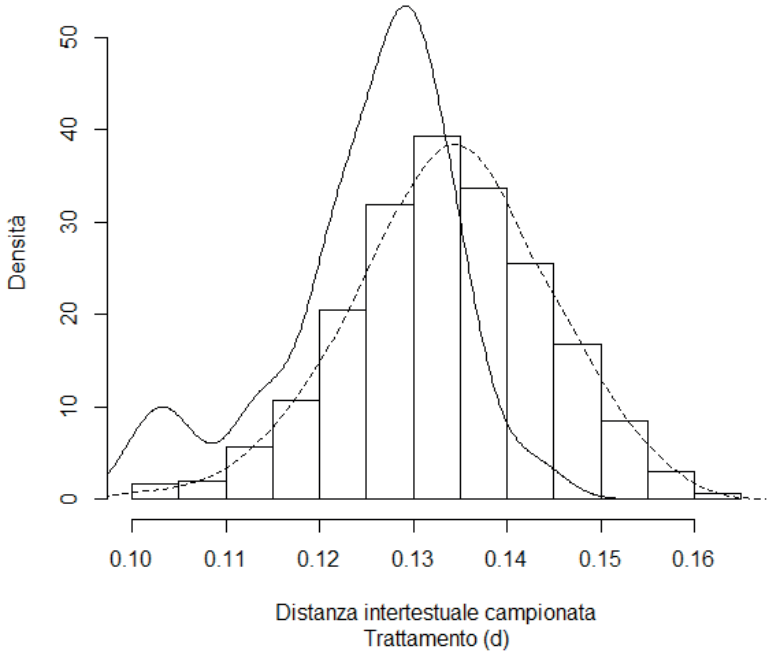
Può essere interessante notare come per i diversi trattamenti siano presenti alcune irregolarità nelle distribuzioni, che non sempre assumono la forma delle classiche "campane" normali. Soprattutto possiamo fare riferimento ai trattamenti *b*, *c* e *d*. Nei primi due è presente una "gobba" alla destra della distribuzione; tale discontinuità è dovuta a un gruppo di distanze particolarmente alto. In questi due casi le distanze maggiori originano da coppie di testi tradotti da lingue diverse, infatti possiamo notare come la gobba non sia presente nella linea continua. Nel caso *d* invece, notiamo una gobba nella parte bassa della distribuzione, e soprattutto che tale gobba è presente sulla linea continua, quella riferita alla distribuzione delle distanze tra testi provenienti dalla stessa lingua. Non sappiamo stabilire perché si verifichino tali perturbazioni, ma si tratta di un indizio del fatto che certi trattamenti del corpus risultano più efficaci di altri nell'evidenziare differenze o similarità tra i testi.

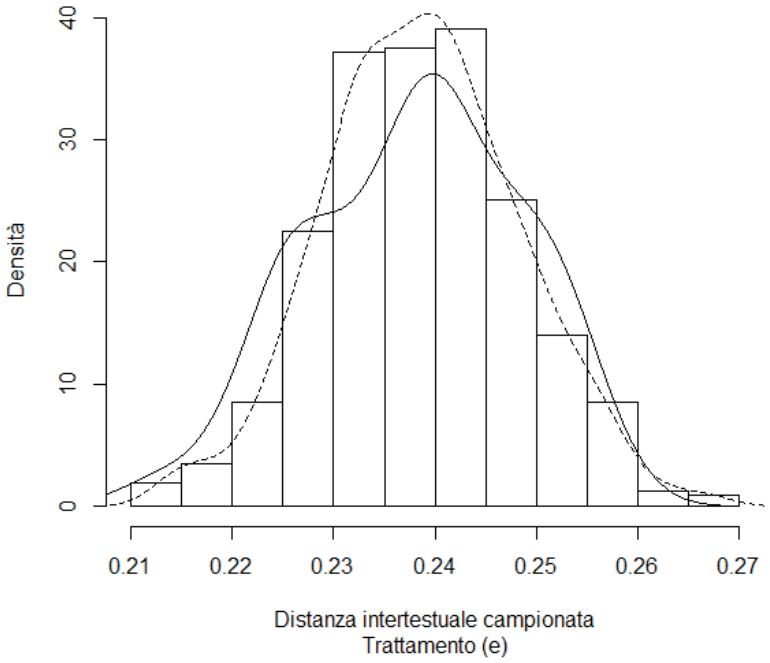
Figura 3. Distribuzione delle distanze intertestuali tra macrotesti tradotti dalla stessa lingua e da lingue diverse.





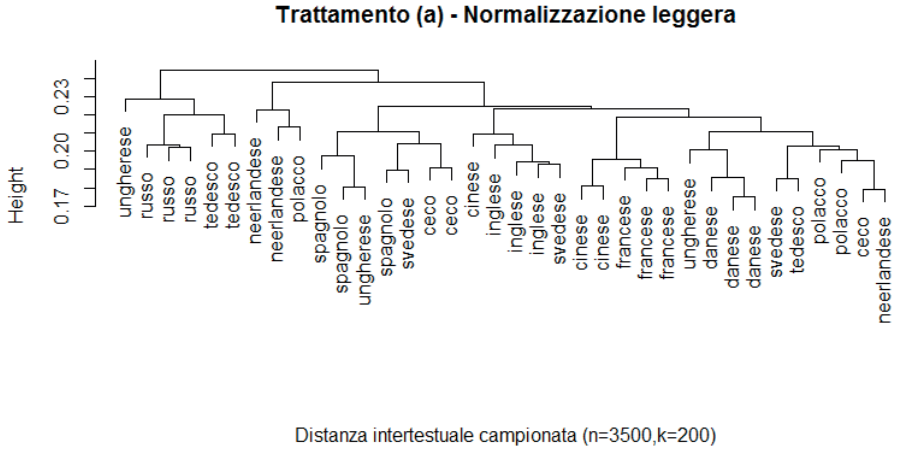

Densità
0
5
10
15
20
25
30
35
0.18
0.20
0.22
0.24
Distanza intertestuale campionata
Trattamento (c)

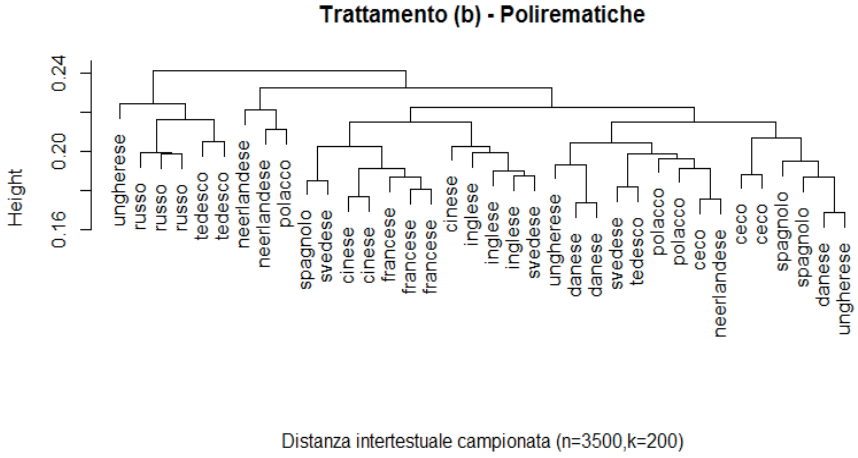

Densità
0
10
20
30
40
50
0.10
0.11
0.12
0.13
0.14
0.15
0.16
Distanza intertestuale campionata
Trattamento (d)

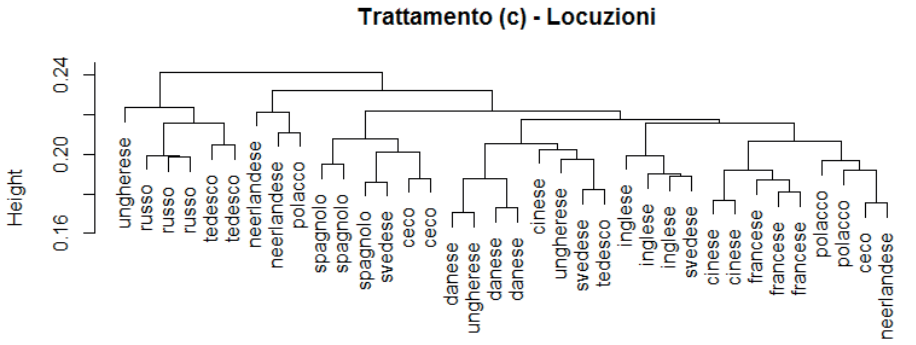


Passando alla rappresentazione dei dati sotto forma di dendrogrammi, la Figura 4 riporta gli accoppiamenti dei macrotesti in base alle lingue fonte secondo i diversi trattamenti a cui abbiamo sottoposto il subcorpus. È possibile notare una certa tendenza dei macrotesti ad aggregarsi per lingua di partenza. Tali accorpamenti mostrano una certa coerenza con le lingue fonte e non sono frutto del caso: infatti a partire dalle 12 lingue straniere presenti in questo subcorpus con 3 macrotesti ciascuna, una volta selezionato un macrotesto, la probabilità che, se viene estratto un altro macrotesto in modo casuale, questo risulti tradotto dalla stessa lingua è pari a 0,08 (8%). Secondo tale ragionamento, la probabilità che nel dendrogramma venga identificata "al primo livello" almeno una terna (come succede nel caso della normalizzazione leggera per danese, francese e russo) è pari a 0,007 (lo 0,7%), ben al di sotto di una soglia ragionevole. La probabilità che al primo livello venga identificata almeno una coppia invece è pari a 0,12 (12%) ma, come possiamo vedere dai grafici, con l'eccezione del trattamento *e*, il numero di coppie correttamente identificate è ben superiore (fino a 6 nel caso del trattamento con i grammaticali).

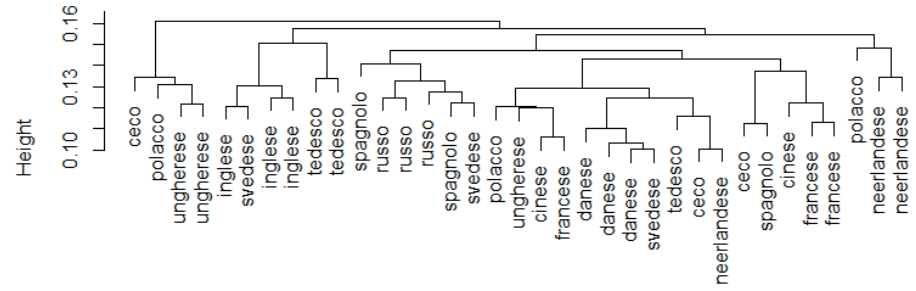
Figura 4. Dendrogrammi dei raggruppamenti di traduzioni dalla stessa lingua.



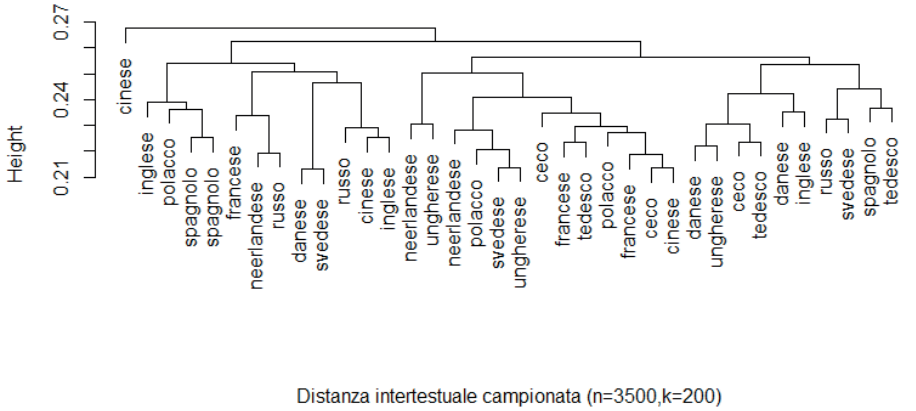




Distanza intertestuale campionata (n=3500,k=200)



Distanza intertestuale campionata (n=3500,k=200)



Distanza intertestuale campionata (n=3500,k=200)

Dai grafici risulta chiaro che il calcolo della distanza intertestuale non riesce a individuare l'effetto della lingua di partenza se vengono considerate le parole grammaticali e le locuzioni individuate da *Taltac*[2] (trattamenti *d* ed *e*). Per il resto (ricordiamo ancora che il metodo agglomerativo, una volta associati due elementi, ricalcola le distanze tra tutti gli altri elementi e la coppia appena formata), il trattamento *a* individua correttamente in base alle lingue fonte tre terne di macrotesti al secondo livello (danese, francese e russo) e al primo livello altre tre coppie (ceco, cinese e tedesco); il trattamento *b* due terne (francese e russo) e quattro coppie (ceco, cinese, danese e tedesco); il trattamento *c* due terne (francese e russo) e cinque coppie (ceco, cinese, danese, spagnolo e tedesco); infine, il trattamento *d* non individua nessuna terna ma ben sei coppie (francese, inglese, neerlandese, russo, tedesco e ungherese).

Notiamo come, con i primi tre trattamenti, le lingue di partenza individuate si ripresentino con costanza: francese e russo sono sempre contenuti nelle terne, il danese compare in una terna e in una coppia; ceco, cinese e tedesco in tre coppie e spagnolo in due. Occorre inoltre notare che una possibile terna di macrotesti tradotti a partire dall'inglese non riesce a formarsi in tutti e tre i casi perché al primo livello c'è sempre un macrotesto tradotto dallo svedese che interviene: poiché si tratta costantemente dello stesso macrotesto, deve esserci qualcosa che lo avvicina particolarmente alle traduzioni dall'inglese.

Tra le coppie "malriuscite", non si evidenziano tendenze legate alle famiglie linguistiche. In effetti, data per buona l'ipotesi dell'interferenza linguistica, se la distanza intertestuale avesse funzionato perfettamente come metodo per cogliere l'influenza della lingua di partenza sulle traduzioni, non solo al primo e secondo livello si sarebbero formate coppie e poi terne create dai tre macrotesti per ogni lingua compresi nel nostro subcorpus, ma ai livelli superiori si sarebbero anche dovuti realizzare ulteriori accoppiamenti in base alle famiglie linguistiche (per es. il francese con lo spagnolo; il tedesco con il neerlandese; il ceco con il polacco e il russo). Anche in caso di funzionamento parziale, ci si sarebbe potuti aspettare che il calcolo della distanza intertestuale si sarebbe fatto "ingannare" dai macrotesti tradotti tra lingue della stessa famiglia, per es. accoppiando traduzioni dal tedesco con traduzioni dal neerlandese. Questo invece non si verifica: al primo livello, come a quelli superiori, emergono accoppiamenti tra lingue slave, neolatine e germaniche senza ordine apparente: per es. la terna tradotta dal russo viene collegata con tedesco e ungherese, una coppia di macrotesti tradotti dal cinese con la terna dal francese ecc. Particolarmente problematici risultano inoltre polacco e svedese che, in quanto lingue di partenza, non hanno mai creato accoppiamenti, mentre coppie di macrotesti tradotti dall'inglese e dall'ungherese emergono soltanto a seguito del trattamento *d* (del trattamento *e* si è già parlato in precedenza).

Nonostante tutto, in conclusione, è innegabile che, soprattutto i primi tre grafici riportati in Figura 4, seppure non ideali nella loro distribuzione al fine della conferma dell'ipotesi dell'interferenza della lingua di partenza, forniscono

dati che non possono essere considerati frutto del caso e che confermano l'effetto dell'interferenza linguistica sulle traduzioni.

4. Conclusioni

Per concludere, possiamo tornare ai tre quesiti posti in apertura e tentare di dare una risposta. Il metodo di campionamento per il calcolo della distanza intertestuale da noi proposto al capitolo 2 sembra in grado di cogliere la distinzione tra macrotesti tradotti e nativi. Non emerge alcun errore di classificazione per i trattamenti *a*, *b* e *c*, mentre il trattamento *e* posiziona un solo macrotesto fuori schema, accoppiandolo a tutto il resto del corpus, e il trattamento *d* colloca 5 macrotesti nativi nel ramo del dendrogramma relativo alle traduzioni (Figura 2). In parte, tali discrepanze nei risultati ottenuti tramite gli ultimi due metodi di trattamento del corpus emergono anche dal confronto tra i valori medi e la deviazione standard della distanza intertestuale relativa al subcorpus tradotto e nativo; tuttavia appare evidente che il processo traduttivo comporta delle conseguenze sull'assetto dei testi che viene colto dal calcolo della distanza intertestuale. In qualche modo, sembra dunque che il traduttese effettivamente esista: per paragonare i nostri risultati agli studi di Labbé (2001) sull'attribuzione d'autore, è come se fossimo riusciti ad attribuire una parte dei testi del nostro corpus a un "Traduttore astratto" e una parte a un "Autore nativo astratto", riconoscendone gli stili individuali.

Passando invece all'interferenza linguistica, i risultati che abbiamo ottenuto delineano una situazione più sfumata. Da una parte le traduzioni dalla stessa lingua appaiono reciprocamente più simili di quanto non lo siano nel confronto con traduzioni da lingue diverse (Tabella 2), tuttavia i dendrogrammi che abbiamo ottenuto (Figura 4) riescono solo in parte a collegare tutti e tre i macrotesti tradotti dalla stessa lingua e (di conseguenza) non sembrano in grado di riconoscere le diverse famiglie linguistiche comprese nel corpus. Ancora una volta, i trattamenti *a*, *b* e *c* ottengono risultati migliori e coerenti tra loro, tanto da poter confermare l'esistenza di un effetto della lingua fonte sul testo tradotto, mentre in particolare il trattamento *e* è risultato del tutto inaffidabile.

Resta da capire il perché delle somiglianze e delle differenze nei risultati dei trattamenti. In teoria i trattamenti *a* e *b* sono i più sensibili alle scelte lessicali in genere, e quindi anche ai contenuti, e infatti hanno in genere prodotto risultati simili. Non è chiaro perché il trattamento *c*, teoricamente meno sensibile ai contenuti, non si discosti molto dai primi due; per trovare una spiegazione, ci proponiamo un'analisi approfondita del materiale linguistico che *Taltac*[2] prende in considerazione nell'individuazione automatica delle "locuzioni grammaticali". In questo modo sarà possibile anche ipotizzare le ricadute sul funzionamento (in parte mancato) del trattamento *e* da noi adottato, che mirava ad arricchire il novero delle parole grammaticali in modo da disinnescare l'impatto dei contenuti sulla distanza intertestuale.

Rimane il fatto che anche il trattamento *d* non ha dato risultati particolarmente soddisfacenti. Si potrebbe pensare che le dimensioni dei nostri campionamenti non forniscano materiale sufficiente a permettere un confronto significativo a livello dei grammaticali, che naturalmente sono meno numerosi dell'insieme delle forme grafiche. Negli esperimenti che ci proponiamo di condurre in futuro sarà possibile verificare la variazione dei risultati delle misurazioni in base alle diverse dimensioni dei campioni considerati.

In alternativa, non è improbabile che i grammaticali risentano particolarmente delle idiosincrasie individuali del traduttore, quindi facendo aggio sulle tendenze generali del traduttese. In questo studio, l'influenza dello stile individuale del traduttore (cfr. Bernardini 2016) è effettivamente il convitato di pietra: purtroppo, per i motivi già esposti più volte, il nostro corpus mal si presta a tenere conto di questa variabile e sarà necessario assemblarne un altro ad hoc per procedere alle dovute rilevazioni. Le relative misurazioni, anche su testi nativi, potranno gettare luce su questo importante aspetto del rapporto che si instaura tra testo, lingua fonte, autore e traduttore.

RIFERIMENTI BIBLIOGRAFICI

Baker M. (1996) "Corpus-based Translation Studies: the Challenges that Lie Ahead", in *Terminology, LSP and Translation. Studies in Language Engineering in Honour of Juan C. Sager.* Ed. by Somers H., Amsterdam/Philadelphia, John Benjamins, pp. 175-186.

Bernardini M. (2016) *Originalità della traduzione letteraria: una questione di distanze*, disponibile online all'indirizzo http://www.treccani.it/lingua_italiana/speciali/traduttese/Bernardini.html

Cortelazzo M.A., Nadalutti P., Tuzzi A. (2013) "Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature", *Journal of Quantitative Linguistics*, 20(2), pp. 125-152.

Frawley W. (2000) "Prolegomenon to a Theory of Translation", in *The translation Studies Reader.* Ed. By Venuti L., London/New York, Routledge, pp. 250-263.

Labbé, C. & Labbé, D. (2001) "Inter-textual distance and authorship attribution Corneille and Moliére", *Journal of Quantitative Linguistics*, 8(4), pp. 213–213.

Trosborg A. (1997), "Translating Hybrid Political Texts" in *Text Typology and Translation*. Ed. by Trosborg A., Amsterdam/Philadelphia, John Benjamins, pp 145-158.

Tuzzi A. (2010) "What to put in the bag? Comparing and contrasting procedures for text clustering", *Italian Journal of Applied Statistics/Statistica Applicata*, 22(1), pp. 77-94.