DOI: 10.13137/978-88-8303-913-3/18479

2. Distanza intertestuale e lingua fonte: premesse teoriche, compilazione di un corpus e procedure di analisi

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

Abstract

This chapter illustrates the theoretical background of the implementation of computational linguistic methods to probe the translation universals hypothesis. Starting from the assumption that both the translation process and the source language impact the linguistic features of translations, we use Labbé's method for calculating intertextual distance to check whether it can distinguish translated from non-translated texts and proves successful in grouping together texts translated from the same language within a corpus of translations. In addition to compiling a balanced corpus of newspaper articles (both originally written in Italian and translated from several languages), ad hoc procedures are necessary to offset the impact of different text lengths and contents on intertextual distance values. The selection of text chunks of equal length and different language tokens (grammar words, multi-words etc.), along with POS-tagging procedures to identify additional useful linguistic features, provide a promising approach to evaluate different methods to calculate the intertextual distance between translated and non-translated texts (cosine similarity, machine learning, stylometry).

Keywords

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

1. Premesse teoriche

Questo lavoro[1], presentato in occasione delle giornate di studio *Language, Translation, Corpora: Comparing Research Methods and Traditions* (Trieste, 1-2 dicembre 2016), rappresenta la continuazione di una linea di ricerca iniziata qualche anno fa (Ondelli 2008, Ondelli & Viale 2010, Ondelli 2013a) tesa a verificare con strumenti quali-quantitativi la validità degli assunti teorici noti come "universali traduttivi" (Baker 1993 e 1996). Nel momento in cui gli studiosi di traduttologia allontanano la loro attenzione dal rapporto col testo fonte (e quindi dal riferimento a una presunta "fedeltà" della resa: Toury 1980 e 1995) per concentrarsi sulla ricezione della traduzione stessa nel sistema linguistico e culturale di arrivo, dal punto di vista della lingua ci si chiede quali siano le conseguenze del processo traduttivo *in primis* e del contatto interlinguistico in seconda battuta.

In particolare, l'ipotesi degli universali traduttivi postula che, a prescindere dalle lingue in gioco, i traduttori seguirebbero comportamenti condivisi che comprendono la tendenza a semplificare il lessico e la sintassi del testo di partenza; esplicitare informazioni lasciate implicite; conformarsi maggiormente alla norma riconosciuta nella lingua di arrivo, "normalizzando" le particolarità stilistiche del testo fonte; realizzare testi che presentano un grado di somiglianza reciproca maggiore rispetto a testi prodotti direttamente da scriventi nativi nella lingua data. A queste tendenze generali, che dipendono dal processo della traduzione in sé, a tutti i livelli di analisi si aggiunge, seppur in gradi diversi a seconda dell'esperienza del traduttore e del prestigio culturale della lingua e del testo di partenza, l'inevitabile interferenza (o *transfer*) delle strutture della lingua fonte sulla lingua di arrivo.

La teoria degli universali traduttivi è stata oggetto di critiche e precisazioni (Halverson 2010, Malmkjær 2011, Mauranen 2004, Tirkkonen-Condit 2004), soprattutto per la difficoltà nello stabilire univocamente a quale delle varie tendenze è possibile ricondurre la gamma dei tratti linguistici rilevati, anche in base ai diversi tipi testuali in gioco (Ondelli & Viale 2010; Ondelli 2013b). In particolare, oltre a tenere distinti fenomeni misurabili in base al confronto con il testo fonte (*S-Universals,* indagabili per mezzo di corpora paralleli; Chesterman 2004) dai fenomeni che distinguono le traduzioni da testi analoghi prodotti direttamente

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, 2 e 3 e Paolo Nadalutti i paragrafi 4, 5 e 6. Gli autori ringraziano Arjuna Tuzzi per la preziosa consulenza.

da parlanti nativi (*T-Universals*, secondo un approccio che prevede l'impiego di corpora monolingui paragonabili), i problemi che sono emersi nella valutazione degli universali traduttivi si ricollegano innanzitutto alla necessità di lavorare con corpora di ampie dimensioni, come immediatamente riconosciuto da Baker (1993). Poiché i comportamenti linguistici tenuti nella stesura di traduzioni sarebbero governati da leggi probabilistiche piuttosto che da necessità cognitive (Toury 2004 e 2012: part 4), occorre infatti analizzare una mole notevole di dati che permetta di apprezzare quantitativamente le tendenze enucleate sopra. Una conseguenza immediata per la raccolta di dati linguistici che siano reali e rappresentativi del "traduttese" è la necessità di tenere conto del ruolo dominante dell'inglese (sia diretto che indiretto; cfr. il concetto di "traduzioni invisibili" in Grasso 2007) nell'odierno mondo globalizzato (Mauranen 2008).

Nonostante i limiti e le difficoltà (abbondantemente sottolineati da House 2008), ci sembra che la teoria degli universali traduttivi offra spunti interessanti per provare a valutare l'impatto delle traduzioni sulla percezione linguistica dei parlanti. La costante esposizione a testi che non esplicitano la loro natura di traduzioni (cfr. il concetto di *covert translations* di House 1977 e 1997) ma che sono caratterizzati da una lingua in qualche modo diversa da quella prodotta in condizioni analoghe da scriventi nativi (il *traduttese*, variamente indicato come *lingua ibrida* in Trosborg 1997 o *terzo codice* in Frowley 2000) potrebbe avere ricadute apprezzabili su una comunità di consumatori di prodotti paraletterari, di informazione e di intrattenimento (sul traduttese come caso particolare di contatto linguistico cfr. le considerazioni svolte in McEnery et al. 2006: 93-94). La traduzione massificata di testi secondo ritmi e procedure industriali è infatti suscettibile di risentire maggiormente del processo traduttivo (si pensi al caso del doppiaggio: cfr. Rossi 2006) e di intervenire sugli equilibri linguistici di una comunità di parlanti che, nel caso dell'Italia, ha solo di recente completato (nella migliore delle ipotesi) il processo di alfabetizzazione e appropriazione dell'idioma nazionale.

2. Un nuovo approccio alle traduzioni

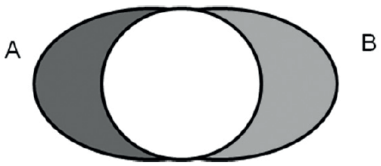
Lo studio condotto da Ondelli e Viale (2010: 59) sulle traduzioni di tipo giornalistico si concludeva con una prospettiva di ricerca incentrata sulla distanza intertestuale come strumento utile a "cercare di mettere ordine in un quadro oltremodo confuso a causa della pluralità di fattori concomitanti che entrano in gioco" nella valutazione dell'assetto di un testo tradotto. Ed è da questa premessa che prende le mosse il nostro contributo.

Il concetto di distanza intertestuale rientra nei tentativi di ridurre l'informazione contenuta nei testi di un corpus a una singola dimensione quantitativa al fine di misurare efficacemente la somiglianza o dissomiglianza dei testi stessi. Tra le varie proposte avanzate, quella di Labbé e Labbé (2001; una versione in italiano è disponibile in Labbé 2010) ha già trovato applicazione su corpora di testi in lin-

gua italiana di tipo letterario e politico/istituzionale (cfr. Cortelazzo et al. 2013), in particolare ai fini dell'attribuzione d'autore (*authorship attribution*).

Secondo il modello di Labbé, la misura della similarità di due testi sfrutta la frequenza della parola come unità di misura e si basa sulla differenza tra le frequenze osservate nei due testi oggetto del confronto per pervenire a una misura sintetica della similarità (o, più precisamente, di dissimilarità perché di tratta di una distanza). Si considera l'insieme di tutte le parole presenti in A e in B con le relative frequenze assolute ($f_{i,A}$ e $f_{i,B}$). Se i testi sono di uguale dimensione ($N_A = N_B$) si può procedere direttamente con il calcolo della differenza in termini di frequenza assoluta per ciascuna parola. Se, viceversa, i testi sono di dimensione diversa (per es. $N_A \leq N_B$), si possono ottenere frequenze confrontabili riconducendo la frequenza ($f^*_{i,B}$) di ogni parola del testo più grande B alla dimensione del testo più piccolo A attraverso una semplice proporzione:

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



La distanza tra A e B:

$$d_{(A,B)} = \frac{\sum_{i \in (A,B)}^{V_{(A,B)}} \left| f_{iA} - f^*_{iB} \right|}{N_A + N_{B'}}$$

risulta un valore compreso tra 0 (i due testi contengono le stesse parole con le stesse frequenze) e 1 (i due testi non hanno alcuna parola in comune). Secondo Labbé (2010: 123) la distanza intertestuale viene influenzata principalmente da quattro fattori esterni. Il più importante (che viene definito il "genere") pare corrispondere non solo alla variazione diamesica, ma anche a tipi testuali aventi convenzioni stilistiche più o meno evidenti (si può ben comprendere lo scarto tra prosa e poesia, più difficile definire intuitivamente quello tra commedia e tragedia). In seconda battuta interviene la dimensione diacronica (e anche questo è abbastanza intuitivo), quindi l'autore e, infine, i temi trattati (dunque gli aspetti più propriamente semantici).

Nell'interesse di Labbé, se si mantengono quanto più possibile invariati gli altri fattori (quindi tramite il confronto di tipi testuali analoghi prodotti nello stesso periodo storico e che non presentino eccessive variazioni tematiche), sarebbe possibile stabilire delle soglie al di sotto delle quali due o più testi sono attribuibili alla stessa penna. In questo alveo si sono mosse anche le ricerche di

Cortelazzo et al. (2013) su un corpus di romanzi italiani, che hanno sviluppato una rivisitazione della procedura di calcolo proposta da Labbé al fine di tenere conto dell'impatto della diversa lunghezza dei testi oggetto di analisi (vedi anche Tuzzi 2010).

Ora, se il metodo sviluppato da Labbé per l'attribuzione d'autore richiede di disinnescare una serie di fattori di disturbo fondamentalmente riconducibili alla variazione sociolinguistica in diacronia, diamesia e diafasia (ma, nella situazione italiana, Cortelazzo et al. 2013 aggiungono considerazioni relative alla diatopia), nel caso l'analisi riguardi testi tradotti il quadro si complica ulteriormente (cfr. Bernardini 2016). Il quesito principale riguarda ovviamente l'identità stessa dell'autore a cui si intende attribuire un testo: se, all'interno di un ampio corpus di traduzioni a opera di persone diverse, il traduttore A ha tradotto più testi, in parte opere dell'autore 1 e in parte dell'autore 2, a chi saranno attribuiti questi testi se calcoliamo la distanza intertestuale? Chi avrà il ruolo preminente, gli autori delle opere fonte (anche se poi il corpus comprende opere tradotte da persone diverse) o il loro traduttore (o i loro traduttori se più di uno), che però potrebbe aver tradotto più autori all'interno del corpus?

Non sono peraltro infrequenti casi di traduzioni realizzate dalla stessa persona a partire da lingue diverse. Se poi, sempre all'interno del nostro ipotetico corpus, il traduttore A è responsabile di diversi testi le cui lingue di partenza sono il francese e l'inglese, la misura della distanza intertestuale identificherà sempre la stessa penna oppure distribuirà le opere tra un traduttore A dall'inglese e un traduttore A dal francese (per tacere del ruolo degli autori dei testi fonte)? Infine, se l'ipotesi del traduttese è fondata, e cioè se è vero che in una traduzione resta sempre una qualche traccia del processo traduttivo a prescindere dalle lingue in gioco, in un corpus comprendente traduzioni (da lingue e autori diversi e a opera di traduttori diversi) e testi non tradotti (che, per brevità, d'ora innanzi definiremo "nativi") di autori diversi, la distanza intertestuale suddividerà i testi in due macrogruppi riconducibili a un astratto "autore-traduttore" *vs* un altrettanto astratto "autore nativo"?

3. Compilazione del corpus

Come abbiamo già visto, per l'italiano sono stati fatti alcuni tentativi di applicazione del metodo della distanza intertestuale di Labbè da parte di Cortelazzo et al. 2013 e, specificatamente per le traduzioni, di Bernardini (2016) Quest'ultimo, però, effettua le sue misurazioni su un numero ridotto di testi (48 traduzioni di 16 romanzi da 4 lingue diverse) che risultano poco utili a valutare efficacemente l'incidenza dei fattori "traduttore" e "lingua di partenza": è inevitabile che le diverse traduzioni di una stessa opera risultino meno "distanti". Inoltre questi esperimenti hanno sempre coinvolto testi di tipo letterario, che forse sono i meno adatti a rilevare le conseguenze linguistiche del processo traduttivo in

sé. È probabile, infatti, che un'opera letteraria, soprattutto se di prestigio tale da meritare più traduzioni successive, imponga la propria "impronta" individuale; il traduttore si sforzerà di renderne lo stile, magari dandone la propria resa interpretativa, ma più difficilmente si abbandonerà agli automatismi traduttivi che invece potrebbe applicare nel caso di traduzioni più "dozzinali" e di minor prestigio (ma cfr. le osservazioni di Gallitelli 2016: cap. III sul mercato della traduzione editoriale nell'era della globalizzazione).

Insomma, per tentare di cogliere le tendenze del traduttese, ci pare più opportuno considerare gli esiti di una produzione più rapida e "automatica", come potrebbe essere la letteratura di consumo o il giornalismo. In particolare la stampa periodica, proprio per la rapidità di esecuzione di traduzioni che non puntano a esibire la loro origine esogena (*covert*), per il ruolo attualmente ricoperto di modello di rifermento per un italiano scritto di media formalità e per l'ampio pubblico raggiungibile, sembra offrire il materiale più adatto ad analisi intese a confermare o smentire la teoria degli universali traduttivi (cfr. le considerazioni svolte in Ondelli 2008: 87 e segg.).

I quesiti che abbiamo posto in questa nuova fase delle nostre ricerche sono i seguenti:

1) se applichiamo il metodo della distanza intertestuale di Labbé a un corpus comprendente testi tradotti in italiano a partire da diverse lingue e testi scritti direttamente in italiano, riusciamo a distinguere le traduzioni dai testi nativi?
2) Con la distanza intertestuale riusciamo a raggruppare chiaramente i testi compresi in un corpus di sole traduzioni secondo la loro lingua fonte?
3) Se in un corpus di traduzioni sono presenti testi a firma di traduttori diversi ma il cui autore e lingua di partenza sono gli stessi, che risultati produce il calcolo della distanza intertestuale? Dominano la lingua e lo stile dell'autore o lo stile del traduttore (Zanettin 2012: § 2.2.2.; per una panoramica, cfr. Li 2017: §2)?

In realtà si potrebbe continuare: esistono traduzioni di testi di autori che non scrivono nella loro lingua madre o casi di incertezza della lingua fonte (soprattutto tra i testi prodotti, per es., dalle istituzioni dell'Unione Europea: cfr. Ondelli 2003 e Ondelli 2013c: § 2.1) ma, viste le difficoltà insite nella compilazione di un corpus adatto ai nostri scopi, abbiamo preferito rimandare eventuali approfondimenti a momenti successivi.

Il corpus già utilizzato per la ricerca illustrata da Ondelli e Viale (2010) si prestava alle nostre nuove analisi limitatamente alla componente dei testi nativi (1.008 articoli comparsi tra il 2001 e il 2008 su *Corriere della Sera*, *Repubblica* e *Unità*), mentre le traduzioni erano viziate dalla dominanza quantitativa quasi assoluta dell'inglese come lingua fonte (ivi: 56). Si è reso dunque necessario compilare un nuovo subcorpus di 516 articoli tradotti da 22 lingue straniere e pubblicati

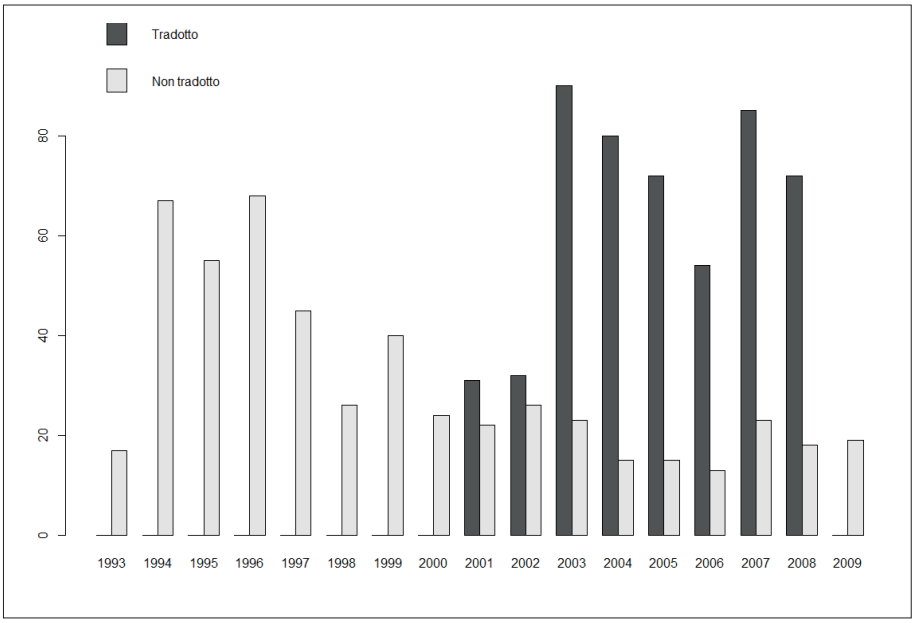
sulla rivista *Internazionale* tra il 1993 e il 2009. Al termine della selezione alcune lingue di partenza, come naturalmente l'inglese, risultavano comunque più presenti e ne abbiamo limitato l'incidenza a circa 50 articoli (Tabella 1):

Tabella 1 – Lingue fonte del subcorpus traduzioni: totale 516 testi.

Lingua	Numero di testi	Lingua	Numero di testi
francese	50	ungherese	22
spagnolo	49	norvegese	16
inglese	49	sloveno	12
russo	48	bulgaro	12
tedesco	48	romeno	10
polacco	38	coreano	9
ceco	37	finlandese	8
neerlandese	26	albanese	7
danese	25	estone	2
svedese	23	lituano	2
cinese	22	lettone	1

Anche per quanto riguarda il periodo di pubblicazione (grafico 1), la distribuzione dei testi tradotti presenta discrepanze, soprattutto nel confronto tra testi nativi. Tuttavia, data la limitatezza dell'arco temporale (17 anni), difficilmente si può pensare che si siano verificati degli sviluppi significativi della lingua in diacronia; tutt'al più si potrebbero ipotizzare delle differenze importanti in termini contenutistici. In effetti, anche se gli articoli nativi sono stati selezionati in modo tale che i contenuti fossero quanto più possibile analoghi alle traduzioni (società, costumi, economia, eventi internazionali ecc.; cfr. Ondelli & Viale 2010 e, sull'importanza della rappresentatività di un corpus, cfr. Baroni & Bernardini 2006: §3), poiché si concentrano tutti nel secondo decennio del periodo considerato, con tutta probabilità presenteranno argomenti assai diversi rispetto alle traduzioni degli anni 1993-2009. Ciononostante, con le adeguate procedure di bilanciamento e di campionamento che abbiamo adottato (v. § 4 e 5) è possibile mitigare l'effetto dei contenuti sulla misurazione della distanza intertestuale.

Grafico 1 – Distribuzione temporale dei due subcorpora: articoli per anno.



Dopo una normalizzazione leggera (v. sotto § 5) il subcorpus degli articoli nativi comprende 1.008 testi composti da 93 autori per un totale di 997.047 occorrenze, mentre quello delle traduzioni conta 516 testi a nome di 67 traduttori per un totale di 632.059 occorrenze. Naturalmente vi sono sia autori che traduttori che hanno firmato più articoli (il limite massimo è 65 per i giornalisti e 93 per i traduttori, il minimo è un solo testo). Tuttavia, nel caso dei traduttori, il fatto che ad alcune sigle siano attribuiti numerosi testi tradotti da un notevole numero di lingue straniere diverse (per es. le iniziali CP identificano 93 testi tradotti da ben 17 lingue che vanno dal cinese all'ungherese) impedisce di pensare che le iniziali identifichino una sola persona fisica: si tratta evidentemente di un'agenzia che si avvale di collaboratori diversi, con tutte le ovvie conseguenze sulla possibilità di stabilire la preminenza dell'autore originale o del traduttore nel calcolo della distanza intertestuale. Infatti, poiché per svariati testi non siamo in grado di identificare con certezza il traduttore, il nostro corpus non si dimostra adatto allo scopo. Restano comunque casi in cui la stessa persona traduce verosimilmente da lingue diverse ma imparentate: è il caso di Gronberg, che firma 13 articoli dal danese e 2 ciascuno da norvegese e svedese.

4. Bilanciamento e campionamenti

Per quanto riguarda le misure generali del corpus, alla luce della diversità delle dimensioni complessive risulta inutile fare confronti tra i subcorpora delle traduzioni e dei testi nativi. Nella tabella 2 ci limitiamo a presentare i dati generali e quelli relativi ai singoli articoli, ottenuti dopo una normalizzazione leggera (v. sotto § 5). Come si può vedere, emergono grandi differenze a livello dei singoli testi per quanto riguarda sia la lunghezza (da un minimo di 216 occorrenze a un massimo di oltre 6.000) sia la ricchezza lessicale. Sebbene la formula del calcolo della distanza di Labbé preveda un correttivo per tenerne conto, la misurazione della distanza intertestuale resta sensibile alle dimensioni dei testi considerati e, di conseguenza, non sarà possibile procedere a confronti diretti tra i testi bensì occorrerà procedere a campionamenti secondo il modello già proposto da Cortelazzo et al. 2013.

Tabella 2 – Misure lessicometriche del corpus.

Dati complessivi	Dati dei singoli articoli
• 1.524 testi • N = 1.629.106 • V = 82.835 • V/N = 5% • Hapax = 38.675 • % Hapax = 47%	• N min = 216 • N max = 6.697 • N media = 1.069 • V min = 157 • V max = 2.600 • V media = 526 • V/N media = 51% • % Hapax media = 52%

Per realizzare un confronto tra testi nativi e traduzioni occorre procedere a un campionamento in grado di sterilizzare l'effetto dell'autore e del traduttore, della lingua di partenza e dei contenuti. Consideriamo quindi tutti i 516 testi tratti da *Internazionale* (tradotti dunque da 22 lingue diverse) e una selezione casuale di 516 articoli nativi, effettuata tramite un campionamento stratificato per testata e autore per evitare che si verificasse una qualsiasi preminenza e assicurare al tempo stesso la rappresentatività di tutte le categorie coinvolte. Abbiamo così deciso di riunire i testi tradotti e nativi rispettivamente in 30 + 30 sottoinsiemi per ottenere dei macrotesti composti da 16 o 17 articoli ciascuno, cercando di evitare che si verificassero concentrazioni anomale di articoli dello stesso autore o traduttore o tradotti dalla stessa lingua di partenza. Poiché la lunghezza media degli articoli è di poco superiore alle mille occorrenze, ogni macrotesto risulta sufficientemente grande da permettere di eseguire 200 campionamenti di segmenti (*chunks*) di 3.500 occorrenze ciascuno.

Per semplificare, è come se confrontassimo 30 opere di un astratto "giornalista nativo modello" con 30 opere di un altrettanto astratto "traduttore modello".

Per evitare che i contenuti e le diverse dimensioni dei testi incidano sul calcolo della distanza intertestuale, confrontiamo tra loro dei brani di 3.500 parole estratti casualmente da ciascuna opera, ripetendo l'estrazione per 200 volte in ciascuna opera. L'assunto è che, se la distanza intertestuale di Labbé è in grado di accoppiare i testi (o gli estratti) di uno stesso autore, nel nostro caso dovrebbe essere in grado di distinguere nettamente testi nativi e traduzioni.

Il secondo dei nostri quesiti riguarda invece la possibilità di individuare la lingua di partenza misurando la distanza intertestuale tra diverse traduzioni. A questo scopo abbiamo selezionato dal nostro corpus tratto da *Internazionale* solo le lingue fonte che presentavano almeno 20 articoli (Tabella 3), ottenendo un totale di 436 testi e 554.052 occorrenze che permette di coprire una discreta varietà di famiglie linguistiche:

Tabella 3 – Corpus per il confronto tra le lingue fonte.

Lingua	Frequenza	Lingua	Frequenza
francese	50	ceco	37
inglese	49	neerlandese	26
spagnolo	49	danese	25
russo	48	cinese	23
tedesco	48	svedese	23
polacco	38	ungherese	22

Sempre per ovviare al possibile impatto dovuto alla diversa dimensione, ai contenuti e al traduttore, gli articoli attribuiti a ciascuna lingua di partenza sono stati fatti confluire in 3 macrotesti in cui l'ordinamento dei singoli articoli è casuale: ciò significa che le lingue più rappresentate producono raccolte di oltre 15 articoli, l'ungherese di appena 7. Quindi procediamo al calcolo della distanza intertestuale secondo il metodo di campionamento già visto sopra nel confronto tra traduzioni e testi nativi. In altre parole, in questo caso è come se avessimo tre opere di dodici "traduttori modello" corrispondenti alle lingue fonte (tre opere del traduttore francese, tre opere del traduttore inglese e così via), e calcoliamo le distanze intertestuali all'interno del corpus per vedere se i campioni estratti dai testi tradotti dalla stessa lingua fonte risultano più vicini.

5. Trattamento del corpus

Per preparare i subcorpora all'analisi ci siamo serviti di *Taltac*[2] (www.taltac.it), un software che permette diversi livelli di intervento per l'individuazione degli elementi lessicali che compaiono nei testi (di seguito: *trattamenti*). Procediamo a

diversi trattamenti per testarne le conseguenze sul calcolo della distanza intertestuale secondo le considerazioni esposte qui di seguito.

a) Normalizzazione leggera

Con questo trattamento il software si limita a trasformare in accenti gli apostrofi posizionati erroneamente (per cui *liberta'* → *libertà*) e a trasformare in minuscole le maiuscole dovute esclusivamente al contesto sintattico, per cui *Non* e *non* verranno considerati un'unica forma grafica se la maiuscola è dovuta alla presenza di un segno di interpunzione forte, mentre *Franco* e *franco* non saranno considerati equivalenti se la maiuscola è dovuta al fatto che si tratta di un nome proprio e non al contesto sintattico.

b) Polirematiche

Taltac[2] è in grado di riconoscere unità lessicali superiori (es. *forze dell'ordine*), locuzioni varie (*fra l'altro, riguardo a*), nomi propri di vario tipo, che verranno trattati come forme a sé stanti all'interno del vocabolario.

Con i trattamenti *a* e, soprattutto, *b* aumenta l'incidenza dell'argomento dei testi perché aumenta la precisione con cui le forme grafiche individuate riflettono diverse accezioni semantiche: se la forma *forze* può comparire in un testo di fisica o sociologia, l'inserimento nel sintagma *forze dell'ordine* ne rispecchia più chiaramente il significato. A di là del campionamento, che dovrebbe averlo almeno in parte disinnescato, l'effetto dei contenuti non va trascurato, come ha dimostrato un (seppur limitato) primo esperimento in cui con tutta probabilità il calcolo della distanza intertestuale raggruppava gli articoli tradotti dal tedesco perché trattavano invariabilmente di economia e finanza (Albertini 2011).

Per cercare di annullare il più possibile l'impatto degli argomenti dei testi abbiamo pensato di eseguire le procedure per la misurazione della distanza intertestuale prendendo in considerazione esclusivamente le parole vuote estratte dai nostri subcorpora. In effetti, seppure con diversa metodologia d'indagine, Baroni e Bernardini (2006) hanno già ottenuto riscontri positivi da un tentativo di misurare con metodi quantitativi la differenza tra testi tradotti e non tradotti prendendo in considerazione le *function words* (nello specifico, i pronomi clitici; cfr. anche Argamon & Levitan 2005; Argamon et al. 2007; Binongo 2003; Stamatatos 2009; Zhao & Zobel 2005). Nel nostro caso abbiamo selezionato i trattamenti che seguono.

c) Locuzioni

In questo caso si procede a estrarre le polirematiche classificate come "locuzioni grammaticali" nel database delle risorse statistico-linguistiche dispo-

nibili in *Taltac*[2]. Si tratta di costrutti di vario tipo, per es. congiunzioni come *dato che*, ma anche sintagmi la cui classificazione da parte del software lascia qualche perplessità (per es. a *in rovina* viene assegnata funzione aggettivale, mentre *un insieme di* viene annoverato tra i sintagmi preposizionali), come pure appare talvolta difficile includere tra le *function words* avverbiali come *senza dubbio* o *in verità*. Tuttavia, l'elenco fornito da *Taltac*[2] comprende grosso modo collocazioni frequenti della lingua italiana suscettibili di avere funzione grammaticale e abbiamo deciso di includere questo trattamento tra le nostre procedure.

d) Grammaticali

Questo trattamento si basa su un elenco di parole grammaticali ottenuto tramite il *tagging* realizzato con *Taltac*[2] sul corpus di 160 romanzi italiani utilizzato in Cortelazzo et al. 2013 (ringraziamo gli autori per avercelo messo a disposizione). L'elenco comprende articoli, preposizioni, congiunzioni e pronomi ma non gli avverbi. Infatti, aggettivi e avverbi sono trasversali alle categorie di parole piene e vuote e presentano casi alquanto dubbi: in questo studio abbiamo preferito attenerci a un approccio massimalista secondo il quale consideriamo grammaticali solo le classi chiuse del lessico italiano.

e) Locuzioni + grammaticali

Con quest'ultimo trattamento estraiamo dai nostri subcorpora i dati linguistici combinati di entrambi i trattamenti *c* e *d*.

f) Lemmatizzazione e *POS-tagging*

Con il programma di *POS-tagging Treetagger* (http://www.cis.uni-muenchen.de/ -schmid/tools/TreeTagger/) è possibile lemmatizzare il corpus, attribuire ciascuna forma grafica a una classe grammaticale e ottenere informazioni di tipo morfologico (per es. modi e tempi verbali).

6. Analisi e obiettivi

Nel capitolo 3 di questo volume, a partire dai materiali linguistici ottenuti tramite i trattamenti esposti al §5, calcoliamo la distanza intertestuale secondo il metodo di Labbé modificato con la procedura di campionamento sui 60 macrotesti (30 + 30) che compongono i due subcorpora di confronto tra italiano nativo e italiano delle traduzioni. L'attesa è che i macrotesti di ciascun gruppo risultino più vicini tra loro, come se venissero attribuiti a due autori distinti ("autore modello" e "traduttore modello"). Nel caso della normalizzazione leggera e

dell'identificazione delle polirematiche dovremmo essere in grado di vedere se la procedura di campionamento riesce a mitigare o annullare l'effetto dei contenuti. Nel caso dei trattamenti *c*, *d* ed *e*, il problema dei contenuti non si pone perché la distanza intertestuale viene calcolata considerando solo sottoinsiemi di parole vuote; la procedura di campionamento assicura comunque di disinnescare l'impatto delle differenze dimensionali tra i macrotesti. In conclusione saremo in grado di valutare se tutte le procedure portano ai medesimi risultati o, nel caso di differenze, potremo stabilire quale sia più indicata per distinguere le traduzioni dai testi nativi.

Successivamente, i cinque trattamenti e il calcolo della distanza intertestuale con campionamento verranno applicati ai 36 macrotesti (3 per 12 lingue diverse) compresi nel subcorpus di testi tradotti estratti da *Internazionale*. Naturalmente restano valide le considerazioni svolte per il confronto tra macrotesti nativi e tradotti, ma in questo caso ci aspettiamo che i tre macrotesti di ciascuna lingua risultino reciprocamente più vicini, come se fossero il prodotto della stessa penna.

In applicazioni successive, i dati ottenuti da lemmatizzazione e *POS-tagging* potranno confermare o affinare le nostre conoscenze in merito alle differenze tendenziali tra traduzioni e testi nativi in relazione e nozioni come la ricchezza lessicale e densità lessicale. In particolare, anche se tradizionalmente la ricchezza lessicale è calcolata come V/N%, soprattutto nel caso di lingue morfologicamente ricche come l'italiano questo dato è solo parzialmente indicativo del bagaglio lessicale di uno scrivente. Se, infatti, l'universale traduttivo della semplificazione ipotizza che i traduttori ricorrano a un lessico meno vario rispetto a scriventi nativi, questa relativa povertà lessicale dovrebbe essere colta più precisamente a livello dei lemmi piuttosto che delle forme grafiche: la presenza di svariate forme flesse è infatti conseguenza del contesto sintattico e non dell'inventiva lessicale di chi scrive.

Lo studio delle classi grammaticali e delle informazioni rese disponibili dal *tagset* utilizzato per l'italiano da *Treetagger*, oltre a misurare la densità lessicale (teoricamente minore nelle traduzioni), ci permetterà estrarre informazioni importanti sulla frequenza di alcuni elementi che possono essere rivelatori del processo traduttivo in sé o dell'interferenza esercitata dalle lingue fonte. A titolo di esempio, una maggior presenza di pronomi personali soggetto e aggettivi e pronomi dimostrativi e possessivi potrebbe essere il segnale dell'universale traduttivo dell'esplicitazione o il risultato dell'interferenza di lingue di partenza che utilizzano questi elementi più frequentemente dell'italiano. Anche l'analisi delle voci verbali (per es. frequenza del perfetto semplice o della perifrasi *stare* + gerundio), del tasso di nominalizzazione, della presenza di connettivi, della frequenza di certe strutture (per es. l'anteposizione dell'aggettivo al nome) ecc. può fornire informazioni utili per valutare sia l'assetto del traduttese nel confronto con i testi nativi sia le eventuali peculiarità dovute alla singola lingua fonte.

Le analisi potranno contribuire da un lato a gettare luce sull'effettiva esistenza di un "italiano delle traduzioni" (almeno per quanto concerne il tipo testuale con-

siderato, cioè l'articolo giornalistico), dall'altro ad affinare le procedure di calcolo della distanza intertestuale e della conseguente attribuzione d'autore. Sondaggi successivi potranno mettere a confronto il metodo qui esposto con altri metodi di calcolo della similarità/dissimilarità dei testi come il coseno di similitudine (che risulta insensibile alla lunghezza dei testi; cfr. Huang 2008), il *supervised learning* (Baroni & Bernardini 2006; Joula & Mikros 2016) e altri approcci stilometrici (Rybicki 2012).

RIFERIMENTI BIBLIOGRAFICI

Albertini S. (2011) *L'italiano nelle traduzioni di "Internazionale": analisi di un corpus*, tesi di laurea triennale in Comunicazione Interlinguistica Applicata, Trieste, Università degli studi di Trieste.

Argamon, S. & Levitan, S. (2005) "Measuring the usefulness of function words for authorship attribution", in *Proceedings of the 2005 ACH/ALLC Conference*, Victoria, BC, Canada, June 2005.

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

Baker M. (1993) "Corpus linguistics and translation studies – Implications and applications", in *Text and Technology. In Honour of John Sinclair*. Ed. by Baker M., Francis G., Tognini-Bonelli E., Amsterdam/Philadelphia, John Benjamins, pp. 233-250.

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.

Baroni M. & Bernardini S. (2006) "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text", in *Literary and Linguistic Computing* 21(3), pp. 259-274.

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

Binongo J. N. G. (2003) "Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution", in *Chance*, 16(2), pp. 9-17.

Chesterman A., (2004) "Beyond the Particular", in *Translation Universals. Do they exist?* Ed. by Mauranen A. & Kujamäki P., Amsterdam/Philadelphia, John Benjamins, pp. 33-49.

Cortelazzo M.A., Nadalutti P., Tuzzi A. (2013) "Improving Labbé's Intertextual Distance: Testing a Revised Version on a Large Corpus of Italian Literature", in *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.

Gallitelli E. (2016) *Il ruolo delle traduzioni in Italia dall'Unità alla globalizzazione. Analisi diacronica e focus su tre autori di lingua inglese: Dickens, Faulkner e Rushdie*, Roma, Aracne.

Grasso D. E. (2007) *Innovazioni sintattiche in italiano (alla luce della nozione di calco)*, thèse de doctorat, Univ. Genève, no. L. 629, disponibile online all'indirizzo http://archiveouverte.unige.ch/unige:475.

Halverson S. (2010) "Cognitive translation studies: Developments in theory and methods", in *Translation and Cognition*. Ed. by Shreve G. M & Angelone E., Amsterdam/Philadelphia, John Benjamins, pp. 349-369.

House J. (1977) *A Model for Translation Quality Assessment*, Tübingen, Narr.

House J. (1997) *Translation Quality Assessment: A Model Revisited*, Tübingen, Narr.

House J. (2008) *Beyond Intervention: Universals in translation*, in Trans-kom 1(1): 6-19, disponibile online all'indirizzo www.transkom.eu/bd01nr01/transkom__01__01__02__House__Beyond__Intervention.20080707.pdf

Huang A. (2008) "Similarity Measures for Text Document Clustering", in *proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC2008)*, April 2008, Christchurch, pp. 49-56.

Joula, P. & Mikros G. (2016) "Authorship Attribution Using Different Languages", in *Digital Humanities 2016: Conference Abstracts*. Kraków, Jagiellonian University & Pedagogical University, pp. 241-243.

Labbé D. (2010) "Corneille nell'ombra di Molière. Come identificare un autore?", traduzione di Irene Borsato, in *Rivista internazionale di tecnica della traduzione/International Journal of Translation*, 12, pp. 117-138.

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

Li D. (2017) "Translator style: a corpus-assisted approach", in *Corpus Methodologies Explained. An empirical approach to translation studies*. Ed. by Ji M., Oakes M., Li D. & Hareide L., London/New York, Routledge, pp. 103-136.

Malmkjær K. (2011) "Translation Universals", in *The Oxford Handbook of Translation Studies.* Ed. by Malmkjær K. & Windle K., Oxford, Oxford University Press, pp. 83-93.

Mauranen A. (2004) "Corpora, universals and interference", in *Translation Universals. Do they exist?* Ed. by Mauranen A. & Kujamäki P., Amsterdam/Philadelphia, John Benjamins, pp.65-82.

Mauranen A. (2008) "Universal Tendencies in Translation", in *Incorporating Corpora. The Linguist and the Translator.* Ed. by Anderman G. & Rogers M., Clevedon, Multilingual Matters, pp. 32-49.

McEnery T., Ziao R. & Tono Y. (2006) *Corpus-based Language Studies. An Advanced Resource Book*, London/New York, Routledge.

Ondelli S. (2003) "Inglese e 'eurocratese'", in *Italiano e inglese a confronto: problemi di interferenza linguistica*. A cura di Sullam Calimani A.V., Firenze, Franco Cesati, pp. 177-195.

Ondelli S. (2008) "Per un'analisi dell'italiano tradotto nei quotidiani: considerazioni preliminari sulla costituzione di un corpus", in *Rivista internazionale di tecnica della traduzione/ International Journal of Translation*, 10, pp. 81-99.

Ondelli S. (a cura di) (2013a) *Realizzazioni testuali ibride in contesto europeo. Lingue dell'UE e lingue nazionali a confronto*, Trieste, EUT.

Ondelli S. (2013b) "Per una linguistica dei testi", in *Realizzazioni testuali ibride in contesto europeo. Lingue dell'UE e lingue nazionali a confronto*. A cura di Ondelli S., Trieste, EUT, pp. 9-26

Ondelli S. (2013c) "Un genere testuale attraverso i confini nazionali: la sentenza", in *Realizzazioni testuali ibride in contesto europeo. Lingue dell'UE e lingue nazionali a confronto*. A cura di Ondelli S., Trieste, EUT, pp. 67-92.

Ondelli S. & Viale M. (2010) "L'assetto dell'italiano delle traduzioni in un corpus giornalistico. Aspetti qualitativi e quantitativi", in *Rivista internazionale di tecnica della traduzione/International Journal of Translation*,12, pp. 1-62.

Rossi F. (2006) *Il linguaggio cinematografico*, Roma, Aracne.

Rybicki J. (2012) "The great mistery of the (almost) invisible translator. Stylometry in translation, in *Quantitative Methods in Corpus-Based Translation Studies*. Ed. by Oakes M. P. & Ji M., Amsterdam/ Philadelphia, John Benjamins, pp. 231-248.

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

Tirkkonen-Condit S. (2004) "Unique items – over- or under-represented in translated language?", in *Translation Universals. Do they exist?* Ed. by Mauranen A. & Kujamäki P., Amsterdam/Philadelphia, John Benjamins, pp. 177-184.

Toury G. (1980) *In Search of a Theory of Translation*, Tel Aviv, The Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury G. (1995) *Descriptive Translation Studies and Beyond*, Amsterdam/Philadelphia, John Benjamins.

Toury G. (2004) "Probabilistic explanations in translation studies, in *Translation Universals. Do they exist?* Ed. by Mauranen A. & Kujamäki P., Amsterdam/Philadelphia, John Benjiamins, pp. 15-32

Toury G. (2012) *Descriptive Translation Studies and Beyond. Revised edition*, Amsterdam/ Philadelphia, John Benjamins.

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", in *Italian Journal of Applied Statistics/ Statistica Applicata*, 22(1), pp. 77-94.

Zhao Y. & Zobel J. (2005) "Effective authorship attribution using function word", in *Proceedings of the 2nd AIRS Asian information retrieval symposium*, Berlin, Springer, pp. 174-190.

Zanettin F. (2012) *Translation-Driven Corpora*, Manchester, St. Jerome.