Impaginazione
Verena Papagno

EUT Edizioni Università di Trieste, 2017.

ISBN 978-88-8303-912-6 (print)
ISBN 978-88-8303-913-3 (online)

EUT - Edizioni Università di Trieste
Via E. Weiss, 21 – 34128 Trieste
eut@units.it
http://eut.units.it
https://www.facebook.com/EUTEdizioniUniversitaTrieste

Testi, corpora, confronti interlinguistici: approcci qualitativi e quantitativi

a cura di
Giuseppe Palumbo

EUT EDIZIONI UNIVERSITÀ DI TRIESTE

Sommario

DOI: 10.13137/978-88-8303-913-3/18477

Introduzione

GIUSEPPE PALUMBO
Università di Trieste

Lo sviluppo di tecniche e metodi di archiviazione e interrogazione di database testuali di medie e grandi dimensioni ha rappresentato una vera e propria svolta per la ricerca linguistica. Per i linguisti un tale database prende generalmente il nome di *corpus* e benché non vi sia consenso unanime su come esattamente si possa definire questo tipo di raccolta, vi è tuttavia accordo sul fatto che con questo termine si intende una collezione di testi che abbia le seguenti principali caratteristiche: i testi devono essere in formato elettronico; le dimensioni complessive del corpus devono essere tali da rendere preferibile l'interrogazione e l'analisi attraverso strumenti automatici; i testi, infine, devono essere stati raccolti secondo criteri che li rendano (in una certa misura) rappresentativi di un dato genere o tipo testuale, di una data varietà o modalità di produzione linguistica o di una determinata categoria di usi linguistici.

Grazie ai corpora la ricerca linguistica ha potuto dotarsi di un nuovo, efficace metodo di verifica empirica delle ipotesi avanzate in sede teorica. Gli approcci e le prospettive adottate dai linguisti che si servono di corpora per le loro analisi rimangono in ogni caso molto vari, sia per quel che riguarda i metodi sia per le tematiche di ricerca.

Per alcuni studiosi il ricorso a un corpus serve essenzialmente a fornire repertori di riscontri autentici dei tratti o fenomeni indagati. In questi casi il corpus o i corpora impiegati fungono fondamentalmente da strumento di convalida di

ipotesi o teorie, le quali possono tuttavia essere discusse o indagate anche al di là del corpus considerato. La base empirica offerta dal corpus può in tali casi offrire lo spunto per un'analisi qualitativa; con gli opportuni accorgimenti, l'indagine può anche arrivare ad adottare metodi quantitativi, in tal modo rendendo ancora più significativo il ricorso al corpus.

Per altri linguisti il corpus arriva ad essere vero e proprio oggetto di indagine, nel senso che i fenomeni indagati sono consustanziali alla raccolta di testi considerata, ovvero osservabili principalmente in rapporto ad essa. È ovvio che anche in questi casi il punto di partenza sarà costituito da ipotesi "esterne" al corpus (nel senso che l'analisi del dato testuale non parte mai alla cieca): il punto è che si tratterà di studi difficilmente eseguibili se non sulla base di un corpus da sottoporre a indagine tramite strumenti di analisi automatica. L'indagine, in tali casi, non può prescindere dal corpus, ferma restando la necessità di proporre interpretazioni per i dati empirici e quantitativi che l'analisi propone. La scelta di fare affidamento esclusivamente sui dati testuali darà spesso alle indagini un carattere essenzialmente esplorativo e non di rado l'analisi porterà non tanto a risposte conclusive quanto all'affinamento e alla revisione delle ipotesi iniziali.

I due approcci generali sopra descritti costituiscono evidentemente una semplificazione dei due poli estremi entro cui può muoversi, metodologicamente, la ricerca linguistica basata su corpora testuali. Nel mezzo ricade un'ampia varietà di approcci che combinano metodi qualitativi e metodi quantitativi, o diverse tipologie degli uni e degli altri.

Dal punto di vista tematico, la varietà è altrettanto grande. In particolare, l'analisi automatica si è dimostrata promettente in molte indagini sui testi che emergono in situazioni di contatto linguistico, spaziando dagli studi sull'interferenza linguistica e sui cosiddetti "universali traduttivi" a quelli sull'inglese come lingua franca. Anche se l'esistenza degli "universali" è stata messa in discussione, software di analisi sempre più raffinati e la possibilità di compilare corpora di dimensioni fino a poco tempo fa impensabili rendono l'approccio della linguistica dei corpora sempre più promettente per il confronto delle traduzioni con testi non tradotti. La definizione del profilo linguistico dei testi tradotti in rapporto ai testi non tradotti può inoltre contare sui metodi di analisi sviluppati nell'ambito della stilometria e dell'attribuzione d'autore.

Il progetto *Attribuzione d'autore, di traduttore e di lingua di partenza: un approccio statistico-linguistico*, realizzato grazie al Finanziamento per la Ricerca di Ateneo – FRA 2014 dell'Università degli Studi di Trieste, costituisce l'ambito nel quale sono nati i contributi inclusi nel presente volume, che allarga la prospettiva anche ad altri studi basati su corpora ma improntati ad approcci di stampo più chiaramente qualitativo.

Filo conduttore dei lavori qui presentati è l'idea del contatto: tra lingue, comunità di parlanti o singoli scriventi. Il contributo di Cortelazzo e Tuzzi, in apertura del volume, applica l'analisi delle corrispondenze e i metodi di calcolo della distanza intertestuale a un corpus di romanzi, con l'obiettivo di risolvere un

"giallo" letterario assurto agli onori delle cronache letterarie internazionali, ossia quello sulla vera identità della scrittrice (o dello scrittore) che si presenta con il nome di Elena Ferrante. Il metodo della distanza intertestuale è anche al centro dei due contributi scritti da Ondelli e Nadalutti, che mirano a valutare l'assetto linguistico dei testi tradotti in italiano in rapporto ai testi che non nascono come traduzioni. Nel primo dei due contributi gli autori illustrano le procedure di compilazione del corpus e le tecniche di analisi utilizzate; nel secondo presentano i risultati dell'analisi, mostrando come i metodi applicati consentano di distinguere i testi tradotti da quelli non tradotti e di raggruppare (entro certi limiti) i testi tradotti in base alla lingua di partenza. Nel contributo di Incelli si mettono a confronto testi di carattere pubblicitario originariamente redatti in italiano e in inglese britannico e americano; lo scopo è quello di identificare – tramite i metodi della *critical discourse analysis* – le specificità dei testi, specie in rapporto alla loro maggiore o minore propensione a coinvolgere "dialogicamente" il destinatario. Lo studio di Peruzzo sposta l'attenzione sul discorso specialistico e mira a identificare, in un corpus di sentenze della Corte Europea dei Diritti dell'Uomo, le "tracce" linguistiche della crescente interazione fra fonti e sistemi giuridici a livello trans-nazionale. Sulla ricerca di "tracce" del contatto linguistico si incentra anche il capitolo di chiusura del volume, in cui Palumbo propone di studiare l'inglese degli articoli di ricerca accademica con gli strumenti della linguistica dei corpora e con gli approcci propri degli studi sulla traduzione: lo scopo è quello di identificare, nell'assetto morfosintattico dei testi, una possibile impronta legata all'influenza della lingua madre degli autori.

DOI: 10.13137/978-88-8303-913-3/18478

1. Sulle tracce di Elena Ferrante: questioni di metodo e primi risultati

MICHELE A. CORTELAZZO
Università di Padova
ARJUNA TUZZI
Università di Padova

Abstract

This chapter illustrates the implementation of quantitative analysis methods on a corpus of modern Italian novels aimed to shed light on the identity of Elena Ferrante, the pen name of a very successful novelist whose real identity is still unknown. After a review of previous attempts conducted according to different approaches (based on lexical, contextual and thematic factors), in order to offset the impact of diatopic varieties of Italian the seven novels written by Elena Ferrante have been compared to 39 novels written by ten authors from Campania (Ferrante's region of origin) according to two methods: correspondence analysis and intertextual distance. Both methods show that Elena Ferrante's novels are more similar to Domenico Starnone's works than to the novels of any other author included in the corpus. In addition, a lexical analysis shows that, compared to the other authors, Ferrante and Starnone share the greatest number of lexical items used exclusively in their novels. Conclusively, the qualitative and quantitative approaches used in this study confirm that a similarity emerges between the novels published by Ferrante and Starnone after the early 1990s and paves the way to further research based on larger corpora of fiction as well as non-fictional texts.

Keywords

Authorship attribution, computational linguistics, corpus linguistics, Elena Ferrante, intertextual distance.

1. Premessa

Il caso Elena Ferrante non poteva non attirare l'attenzione degli autori di questo contributo, due studiosi di estrazione diversa, una statistica e un linguista, che da anni si occupano di analisi quantitativa di dati testuali e, in particolare, della misurazione delle affinità lessicali tra testi.

Finora, avevamo giocato a carte scoperte, lavorando su testi di autori certi. L'obiettivo era prima di tutto metodologico e consisteva nel miglioramento degli strumenti utilizzati per misurare la similarità dei testi. A sua volta, la misurazione della similarità tra testi aveva tra i suoi scopi quello di contribuire all'identificazione di autori incerti, nei diversi casi nei quali si può porre un problema di autorialità: perché l'autore è coperto dall'anonimato o da uno pseudonimo, perché l'autore è sospettato di aver firmato opere frutto della mano, o della tastiera, di altri (plagio) o di aver presentato come opere altrui i propri prodotti (falso), perché l'opera risulta frutto della mano di più autori.

Abbiamo valutato l'efficacia delle formule proposte nella letteratura scientifica per riconoscere testi simili tra di loro; abbiamo ragionato sulle caratteristiche più adeguate dei *corpora* necessari per procedere a confronti sensati; in particolare, abbiamo cercato di superare un problema cruciale nella comparazione quantitativa di testi, quello che deriva dalla diversa lunghezza dei testi posti in comparazione (Cortelazzo et al. 2013, Tuzzi 2010).

Sappiamo che il trattamento con metodi quantitativi dei testi, soprattutto dei testi letterari, provoca perplessità e obiezioni sulle quali è bene proporre subito il nostro punto di vista. Nell'ambito degli studi testuali, e in particolare nell'ambito degli studi letterari, viene avanzata una forte obiezione all'uso di strumenti quantitativi: perché mai dobbiamo ricorrere ad aride formule, quando può essere la nostra sensibilità di lettori a darci la migliore analisi di quello che leggiamo? Se lo riteniamo necessario, non possiamo metterci sulle tracce dell'autore nascosto con il nostro fiuto di lettori esperti, addestrati da anni di letture di testi più diversi?

Si possono dare molte risposte a questa obiezione, che ha sicuramente un suo fondamento. La prima è che, a volte, il fiuto inganna. Non è un'ottima risposta, ne siamo consapevoli, perché a volte anche le formule ingannano. E infatti la strada migliore è quella di far interagire l'osservazione quantitativa con l'osservazione qualitativa, quella che ci permette di fiutare stili, idiosincrasie, consuetudini, vezzi, tic degli scrittori che meglio conosciamo. Grazie ai mezzi di analisi automatica dei testi, si possono verificare, in maniera sistematica e quasi sempre incontrovertibile, le intuizioni del ricercatore (in un'ottica di indagine di tipo confermativo). E comunque, un po' di fiuto serve anche a chi utilizza mezzi quantitativi, perché non si possono confrontare testi *ad libitum* o a caso: bisogna avere delle ipotesi che ci guidino nella costituzione del corpus e queste ipotesi sono, inizialmente, per forza di cose qualitative. Ma è vero anche il contrario. Il fiuto è una dote naturale, che può contraddistinguere il critico di razza; ma non è detto

che i cattivi fiutatori non siano poi dei buoni, o anche ottimi, analisti: il supporto dei mezzi automatici di analisi dei testi può sostituire l'intuizione, per indicare le direzioni verso le quali indirizzare la ricerca qualitativa (in questo caso l'analisi quantitativa si pone in un'ottica esplorativa).

In entrambi i casi l'analisi quantitativa può offrire una sicurezza che l'analisi qualitativa non dà con la stessa forza: quella della certezza del dato. Lo studioso può affermare con piena serenità che un fenomeno (per es. una parola particolarmente significativa) è presente, o assente, sulla base di osservazioni sistematiche, che non soggiacciono a discrezionalità soggettive o che non sono minate dalle manchevolezze della memoria umana. In particolare, si possono fare con certezza asserzioni negative, relative, cioè, all'assenza di una forma, di un costrutto o di un fenomeno. Inoltre, è possibile estendere il corpus oggetto di studio, oltre i limiti materiali che si pongono nel caso di analisi tradizionali (e, di converso, diventa possibile trattare problemi che hanno bisogno, per essere studiati, di *corpora* molto vasti, difficilmente affrontabili con strumenti qualitativi tradizionali).

Tutto questo è tanto più indispensabile quanto più è ampio il numero di testi che devono, o vogliono, essere oggetto di confronto. Lo studio quantitativo di un'ampia raccolta di testi, anche letterari, analizzati rinunciando, almeno all'inizio, a una lettura particolareggiata, testo per testo, ma analizzati in maniera complessiva, per individuare tendenze, macroanalogie, similarità, si inserisce in quella prospettiva di studio che è stata chiamata *distant reading* (Moretti 2003).

2. Le ipotesi sull'identità di Elena Ferrante

Come è noto, Elena Ferrante è un fenomeno editoriale e giornalistico di particolare successo. A partire dal 1992 l'autrice ha scritto sette romanzi (quattro dei quali compongono una quadrilogia che va sotto il nome *L'amica geniale*) e un racconto per bambini. Inoltre ha pubblicato una raccolta di scritti metaletterari (prevalentemente sotto forma di interviste e di lettere) con il titolo *La frantumaglia*[1].

Elena Ferrante esiste solo come autrice. Si tratta, infatti, di uno pseudonimo; la vera identità della scrittrice (o dello scrittore) è tenuta strettamente segreta dagli editori, con una discrezione che regge ormai da più di un ventennio. È diventata famosa per le sue storie anche, e forse soprattutto, fuori dall'Italia, in particolare negli Stati Uniti; anche il suo successo in Italia si è rafforzato dopo il successo all'estero. Ma alla sua popolarità ha certamente contribuito il mistero

1 Precisamente, i primi tre romanzi sono: *L'amore molesto*, Roma, E/O, 1992; *I giorni dell'abbandono*, Roma, E/O, 2002; *La figlia oscura*, Roma, E/O, 2006; La quadrilogia è formata da *L'amica geniale*, Roma, E/O, 2011; *Storia del nuovo cognome. L'amica geniale volume secondo*, Roma, E/O, 2012; *Storia di chi fugge e di chi resta. L'amica geniale volume terzo*, Roma, E/O, 2013; *Storia della bambina perduta. L'amica geniale volume quarto*, Roma, E/O, 2014. Il racconto per bambini si intitola *La spiaggia di notte* (Roma, E/O, 2007), mentre la raccolta di frammenti metaletterari è *La Frantumaglia*, di cui sono uscite tre edizioni, via via aumentate (l'ultima, Roma, E/O, 2016).

della sua identità: anche da questo è nato il vero e proprio mito di Elena Ferrante, una vera e propria *Ferrante fever*. La febbre per Elena Ferrante ha portato a fare numerosi nomi di possibili veri autori delle sue opere: sono stati sospettati scrittori come Guido Ceronetti, Erri De Luca, Francesco Piccolo, Michele Prisco, Fabrizia Ramondino, Domenico Starnone, saggisti e studiosi come Goffredo Fofi e Marcella Marmo, sceneggiatori e registi come Mario Martone, traduttori come Anita Raja (e quasi certamente ci sono sfuggite ulteriori proposte).

Numerosi sono i giornalisti e gli studiosi che si sono occupati dell'identità di Elena Ferrante, con metodi diversi e risultati, almeno in parte, diversi. A tutt'oggi, però, il tema non è stato oggetto di trattazioni pubblicate in sedi autorevoli, secondo le procedure della ricerca scientifica, ormai consolidate anche in ambito umanistico e sociale. Le proposte, la documentazione, i dati, i confronti sono stati annunciati e discussi soprattutto in articoli giornalistici o in blog. Si tratta, tuttavia, di interventi che in molti casi hanno alle spalle analisi scrupolose e metodi accurati, che consentono di discuterne in questa sede.

La prima proposta ampiamente documentata è quella di Luigi Galella, che nel 2005, in un articolo pubblicato sulla «Stampa» del 16 gennaio (p. 27: *Ferrante-Starnone. Un amore molesto in via Gemito*), sulla base di precise ricorrenze tematiche, e in parte lessicali, tra *L'amore molesto* di Elena Ferrante e *Via Gemito* di Domenico Starnone, giungeva a identificare in quest'ultimo l'autore delle opere firmate Elena Ferrante. La prospettiva di Galella è stata rilanciata e rafforzata in due riprese da Simone Gatto, che nel blog «Lo specchio di carta» (http://www.lospecchiodicarta.it), espressione dell'Osservatorio sul romanzo italiano contemporaneo dell'Università di Palermo, ha pubblicato i contributi *Starnone-Ferrante: quando il senso di colpa genera doppi* (pubblicato il 28 ottobre 2006) e *Una biografia, due autofiction. Ferrante-Starnone: cancellare le tracce* (26 ottobre 2016).

Si deve alla sollecitazione di Luigi Galella il primo esperimento con l'utilizzo di metodi quantitativi, quello del fisico matematico Vittorio Loreto, di cui ha dato notizia sempre Galella nell'«Unità» del 23 novembre 2006 (*Ferrante è Starnone. Parola di computer*). Vittorio Loreto ha testato la similarità tra i romanzi di Elena Ferrante e quelli di Domenico Starnone, Goffredo Fofi, Fabrizia Ramondino, Michele Prisco, Erri De Luca, utilizzando gli algoritmi di compressione. Ne è risultata una marcata similarità dei testi di Ferrante e di Starnone, che si abbinano come se fossero opera di un solo autore, e risultano separati da quelli degli altri autori. Anche il confronto tra coppie di romanzi avvicina sempre le opere di Starnone a quelle di Ferrante. A risultati analoghi è giunta la più ridotta ricerca della società svizzera OrphAnalytics, sulla quale ha riferito Alessia Rastelli (*Elena Ferrante, lo studio statistico richiama in causa Starnone*, «Corriere della Sera» 12 ottobre 2016, p. 37).

Si basa, invece, sul confronto tra indizi cronologici e topografici presenti nell'opera e dati biografici la proposta di Marco Santagata (nell'articolo *Elena Ferrante è …*, apparso nella «Lettura», rivista letteraria del «Corriere della Sera», del 13 marzo 2016, pp. 2 e 5): analizzata nel dettaglio la parte del secondo volume della quadrilogia ambientata alla Normale di Pisa, Santagata è giunto alla con-

clusione che l'autore abbia verosimilmente frequentato la Scuola Normale negli anni Sessanta, ma prima del 1966, provenendo da Napoli. A rispondere a questo identikit è una studentessa del tempo, ora professoressa di storia contemporanea all'Università Federico II di Napoli, Marcella Marmo.

Infine, si è appoggiato a dati extratestuali il giornalista Claudio Gatti. Il 2 ottobre 2016 Gatti ha pubblicato nel «Sole 24 ore» (e in tre testate di altri Paesi) il risultato di un'indagine patrimoniale, dalla quale emerge che i compensi riconosciuti dalla casa editrice «e/o» ad Anita Raja, traduttrice dal tedesco per quella casa editrice e moglie di Domenico Starnone, si possono spiegare solo indentificandola con Elena Ferrante.

Dunque, dalle proposte meglio argomentate e documentate escono tre nomi come possibili vere identità di Elena Ferrante: Domenico Starnone (che emerge da tutte le ricerche che si sono basate sul confronto tra testi), Marcella Marmo (la cui identità proviene anch'essa da indizi presenti nei testi), Anita Raja (chiamata in causa in base a riscontri extratestuali).

L'interesse per individuare la vera identità di Elena Ferrante è stato criticato da molti, in articoli giornalistici italiani e stranieri: è stata considerata una curiosità morbosa, che non tiene conto della volontà apertamente espressa dall'autrice di stare nell'ombra e oscura il valore letterario dell'opera, entità autonoma e indipendente dall'identità dell'autrice. Noi crediamo, invece, che la ricerca delle similarità tra le opere di Elena Ferrante e altre opere della letteratura contemporanea (o anche altri scritti), con il risultato collaterale di portare argomenti sull'identificazione dell'autrice, si possa basare su motivazioni importanti e possa avere un rilevante valore scientifico.

La prima motivazione è che la ricerca è inevitabilmente attratta dai misteri e impegnata nella loro soluzione, soprattutto in quanto si tratta di situazioni nelle quali si mettono alla prova i metodi e gli strumenti messi a punto dalla ricerca di base: quello di Elena Ferrante è un mistero esibito, che ha attirato l'attenzione di molti, che, come si è detto, non è stato affrontato con i criteri di verificabilità propri dell'attività scientifica. Ma proprio perché, sia pure attraverso le pagine dei giornali e dei blog, sono state messe in campo numerose tecniche per l'attribuzione d'autore e sono emerse diverse ipotesi, è utile affrontare in modo sistematico il tema anche nell'ambito della ricerca scientifica. Quello di Elena Ferrante è un mistero complesso: sono stati sospettati saggisti, traduttori, studiosi, non solo scrittori. È, quindi, un enigma che obbliga a effettuare analisi in campi che non sono stati abitualmente affrontati nelle ricerche sull'attribuzione d'autore.

La presenza tra i sospettati di donne e uomini fa intravvedere anche un interesse critico dietro l'identificazione dell'autore: non sarebbe irrilevante giungere alla conclusione che la scrittura di romanzi, nei quali molti critici (ma anche molti lettori) individuano la rappresentazione di un punto di vista prettamente femminile nella narrazione dei rapporti familiari e amicali delle protagoniste (Ceccoli 2017; Chemotti 2009; Dow 2016; Lee 2016), sia opera, esclusiva o meno, di un autore di genere maschile.

Infine, il successo internazionale delle opere di Elena Ferrante, in misura incomparabilmente superiore a quello di altri prodotti della nostra letteratura odierna, pone alla critica la domanda di quali siano le ragioni di tale successo: si tratta solo del fascino esercitato dalla Napoli degli anni Cinquanta e dalle sue dinamiche sociali, o ci sono anche ragioni legate allo stile o alla scrittura? La "lettura distante" può portare elementi anche alla discussione di questo aspetto.

3. Metodo e corpus

Ci piace introdurre la descrizione della metodologia usata per sviluppare una ricerca sulla posizione delle opere di Elena Ferrante all'interno della produzione scrittoria dell'ultimo trentennio parafrasando una frase di Alessandro Baricco (da *Novecento. Un monologo*, Milano, Feltrinelli, 1994): dal nostro punto di vista possiamo dire che nelle ricerche sulla similarità tra testi non si è fregati veramente se si hanno da parte un buon corpus e un buon metodo con cui interrogarlo.

Tra i tanti metodi disponibili per la misurazione dell'affinità dei testi (Stamatatos 2009), abbiamo usato quelli che si basano sull'analisi delle corrispondenze (una tecnica statistica multivariata di tipo esplorativo, che utilizza le frequenze delle parole nei testi e trova il miglior compromesso per rappresentare su un piano cartesiano l'associazione tra testi, tra parole e tra testi e parole: Lebart et al. 1984; Greenacre 2007; Murtagh 2010) e sulla distanza intertestuale di Labbé (che misura la similarità dei profili lessicali di due o più testi confrontando la frequenza relativa con cui compaiono le parole nei testi sottoposti ad esame, valutando la differenza parola per parola e poi sintetizzando il risultato in un unico valore dato dalla somma di tutte le differenze: Labbé & Labbé 2001, con le modifiche alla procedura di Cortelazzo et al. 2013).

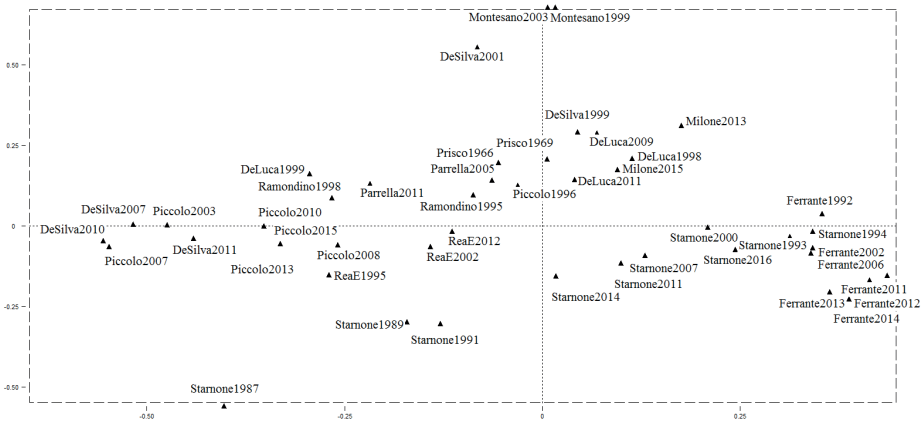
Come corpus abbiamo utilizzato una raccolta di 150 romanzi di 40 autori diversi, messa insieme appositamente per questa ricerca, in base a criteri che tengono conto delle osservazioni più rilevanti avanzate sull'opera di Elena Ferrante da parte della critica accademica e di quella militante e giornalistica. Il *corpus* comprende romanzi di successo (scelti tra quelli vincitori di premi letterari o che hanno conseguito rilevanti risultati di vendita) e romanzi di buon valore letterario (secondo il giudizio raccolto tra esperti), per creare uno sfondo il più possibile rappresentativo della letteratura italiana dell'ultimo trentennio; romanzi di autori sui quali è aleggiato il sospetto che possano aver scritto i romanzi attribuiti a Elena Ferrante; romanzi di autori campani, per verificare se le similarità con Domenico Starnone, accertate dalla critica, non fossero riconducibili semplicemente alla comune provenienza napoletana; romanzi scritti da autrici di genere femminile. A 143 romanzi appartenenti a queste categorie (anche contemporaneamente a più di una di esse) sono stati affiancati i 7 romanzi di Elena Ferrante. Per ognuno degli autori individuati, sono stati presi in considerazione due o più romanzi, risalenti all'ultimo trentennio. Si è derogato a

questo limite temporale solo nel caso di Michele Prisco, chiamato in causa come possibile identità di Elena Ferrante, e, per un'opera ciascuna, in quello di Dacia Maraini e Marta Morazzoni.

Gli autori presi in esame sono, oltre a Elena Ferrante: Eraldo Affinati, Niccolò Ammaniti, Andrea Bajani, Marco Balzano, Alessandro Baricco, Stefano Benni, Enrico Brizzi, Gianrico Carofiglio, Mauro Covacich, Erri De Luca, Diego De Silva, Giorgio Faletti, Marcello Fois, Paolo Giordano, Nicola Lagioia, Dacia Maraini, Margareth Mazzantini, Melania Mazzucco, Rossella Milone, Giuseppe Montesano, Marta Morazzoni, Michela Murgia, Edoardo Nesi, Paolo Nori, Valeria Parrella, Francesco Piccolo, Tommaso Pincio, Michele Prisco, Christian Raimo, Fabrizia Ramondino, Ermanno Rea, Tiziano Scarpa, Clara Sereni, Domenico Starnone, Susanna Tamaro, Chiara Valerio, Giorgio Vasta, Sandro Veronesi, Simona Vinci.

4. I primi risultati: Elena Ferrante e gli scrittori campani

In questa fase, presentiamo i dati relativi al subcorpus costituito da Elena Ferrante e dagli altri scrittori di origine campana (De Luca, De Silva, Milone, Montesano, Parrella, Piccolo, Prisco, Ramondino, Rea, Starnone). L'analisi delle corrispondenze applicata ai 46 romanzi scritti dagli 11 autori campani di questo corpus ci restituisce il seguente grafico che rappresenta il piano fattoriale dei primi due assi cartesiani:



Le posizioni reciproche assunte dai romanzi sul primo piano fattoriale dell'analisi delle corrispondenze, che si possono leggere in termini di affinità lessicali, sono molto interessanti. Innanzi tutto, le opere di Elena Ferrante si trovano tutte al margine destro del grafico. Questo indica che la sua è una scrittura molto caratterizzata e differenziata rispetto a quella degli altri autori campani, con la

sola eccezione di Domenico Starnone, ma solo per quel che riguarda le opere pubblicate dal 1993 in poi. Le opere pubblicate in precedenza (dal 1987 al 1991) si distaccano visibilmente dalle opere successive e si collocano in un quadrante diverso. L'immagine che emerge dalla rappresentazione fondata sull'analisi delle corrispondenze è caratterizzata, quindi, da almeno tre fatti:

1. particolarità della posizione di Elena Ferrante;
2. affinità tra le opere di Elena Ferrante e quelle di Domenico Starnone successive al 1994;
3. cesura tra le prime opere di Domenico Starnone e quelle posteriori. Lo spartiacque è dato proprio dal 1992, anno di pubblicazione dell'*Amore molesto*, primo romanzo firmato da Elena Ferrante.

L'affinità tra le opere di Elena Ferrante e quelle di Domenico Starnone è confermata dall'analisi della distanza testuale tra le opere del corpus. Presentiamo qui di seguito le graduatorie (*ranking*) di ogni opera di Elena Ferrante rispetto a quelle di tutti gli autori campani: per ogni opera, sono elencate le quindici opere del subcorpus che risultano più simili, in ordine decrescente del valore della distanza intertestuale. È opportuno sottolineare che la lista mostra la similarità tra le opere da un punto di vista lessicale, ma non dimostra una dipendenza diretta, come è indicato anche dal fatto che l'ordine di affinità spesso non rispetta l'ordine cronologico di pubblicazione delle opere.

Ferrante 1992	**Ferrante 2002**	**Ferrante 2006**	**Ferrante 2011**	**Ferrante 2012**	**Ferrante 2013**	**Ferrante 2014**
Starnone 1993	Ferrante 2006	Ferrante 2002	Ferrante 2012	Ferrante 2011	Ferrante 2014	Ferrante 2013
Ferrante 2006	Starnone 1993	Starnone 1993	Ferrante 2014	Ferrante 2014	Ferrante 2012	Ferrante 2012
Ferrante 2002	Ferrante 1992	Ferrante 2013	Ferrante 2013	Ferrante 2013	Ferrante 2011	Ferrante 2011
Starnone 1994	Starnone 2016	Ferrante 2014	Ferrante 2006	Ferrante 2006	Ferrante 2006	Ferrante 2006
Ferrante 2011	Starnone 1994	Ferrante 2012	Ferrante 1992	Starnone 1993	Starnone 2014	Starnone 2014
Ferrante 2012	Ferrante 2013	Ferrante 1992	Starnone 1993	Starnone 2014	Starnone 1993	Starnone 1993
Starnone 2000	Starnone 2007	Ferrante 2011	Starnone 2000	Starnone 2011	Starnone 2016	Starnone 2016
Ferrante 2014	Ferrante 2012	Starnone 1994	Starnone 2014	Starnone 2016	Starnone 2011	Ferrante 2002
Ferrante 2013	Ferrante 2014	Milone 2015	Milone 2015	Starnone 2000	Ferrante 2002	Starnone 2011
Milone 2015	Ferrante 2011	Starnone 2011	Starnone 2011	Ferrante 1992	Starnone 2000	Ferrante 1992
De Luca 1998	Milone 2015	Starnone 2007	De Luca 1998	Ferrante 2002	Starnone 2007	Starnone 2000
Starnone 2007	Starnone 2011	Starnone 2014	Ferrante 2002	Milone 2015	Milone 2015	Piccolo 2008
Starnone 2016	De Luca 1998	Starnone 2016	Starnone 2016	De Silva 1999	Ferrante 1992	Milone 2015
Starnone 2011	Starnone 2014	De Luca 1998	Piccolo 1996	De Luca 1998	Piccolo 2008	Starnone 2007
Starnone 2014	Starnone 2000	De Luca 2011	De Silva 1999	Piccolo 2008	De Silva 1999	De Luca 1992

In queste graduatorie, Elena Ferrante, con le sue 7 opere, compare 42 volte; Domenico Starnone, del quale abbiamo considerato 10 romanzi, è presente 43 volte; ad essi si aggiungono De Luca (7 volte), De Silva (3 volte), Milone (7 volte), Piccolo (3 volte)[2].

Per quel che riguarda le prime posizioni, si può osservare che tutte le opere di Elena Ferrante, tranne la prima, risultano simili innanzi tutto a un'altra opera della stessa autrice. In particolare, tutte le opere della quadrilogia risultano molto simili tra di loro. Subito dopo vengono gli altri romanzi di Elena Ferrante oppure, inframmezzati ad essi, quelli di Domenico Starnone: in tutte le colonne, corrispondenti ognuna a un'opera, le prime otto posizioni sono occupate dalle opere di Ferrante o da quelle di Starnone. La vicinanza con i testi di Starnone appare più forte per le prime opere: *Eccesso di zelo* (del 1993) appare la seconda opera più simile a *I giorni dell'abbandono* e alla *Figlia oscura* e addirittura l'opera più simile all'*Amore molesto* (ancor più degli altri romanzi di Elena Ferrante).

Risultati più sorprendenti emergono dalle graduatorie basate sulla misurazione della distanza testuale tra le opere di Domenico Starnone e quelle degli altri autori campani presenti nel corpus:

2 Per comodità del lettore, presentiamo l'elenco completo delle opere presenti nelle graduatorie, dove sono indicate con il nome dell'autore e l'anno di edizione: Erri De Luca, *Tu, mio* (Milano, Feltrinelli, 1998), *I pesci non chiudono gli occhi* (Milano, Feltrinelli, 2011); Diego De Silva, *La donna di scorta* (Ancona, PeQuod, 1999), *Non avevo capito niente* (Torino, Einaudi, 2007), *Mia suocera beve* (Torino, Einaudi, 2010), *Sono contrario alle emozioni* (Torino, Einaudi, 2011); Rossella Milone, *Il silenzio del lottatore* (Roma, Minimum fax, 2015); Valeria Parrella, *Behave* (Milano, RCS Quotidiani, 2011), *Per grazia ricevuta* (Roma, Minimum fax, 2005); Francesco Piccolo *Storie di primogeniti e figli unici* (Milano, Feltrinelli, 1996), *Allegro occidentale* (Milano, Feltrinelli, 2003), *L' Italia spensierata* (Roma-Bari, Laterza, 2007), *Separazione del maschio* (Torino, Einaudi, 2008), *Momenti di trascurabile felicità* (Torino, Einaudi, 2010), *Il desiderio di essere come tutti* (Torino, Einaudi, 2013), *Momenti di trascurabile infelicità* (Torino, Einaudi, 2015); Domenico Starnone *Ex cattedra* (Roma, Rossoscuola e Il manifesto, 1987), *Il salto con le aste* (Milano, Feltrinelli, 1989), *Fuori registro* (Milano, Feltrinelli, 1991), *Eccesso di zelo* (Milano, Feltrinelli, 1993), *Denti* (Milano, Feltrinelli, 1994), *Via Gemito* (Milano, Feltrinelli, 2001), *Prima esecuzione* (Milano, Feltrinelli, 2007), *Autobiografia erotica di Aristide Gambia* (Torino, Einaudi, 2011), *Lacci* (Torino, Einaudi, 2014), *Scherzetto* (Torino, Einaudi, 2016). Per i romanzi di Elena Ferrante rimandiamo all'elenco della nota precedente.

Starnone 1987	**Starnone 1989**	**Starnone 1991**	**Starnone 1993**	**Starnone 1994**	**Starnone 2000**	**Starnone 2007**	**Starnone 2011**	**Starnone 2014**	**Starnone 2016**
Starnone 1991	Starnone 1991	Starnone 1989	Ferrante 1992	Starnone 1993	Ferrante 2011	Ferrante 2006	Ferrante 2006	Ferrante 2014	Ferrante 2002
Starnone 1989	Starnone 2014	Starnone 2014	Ferrante 2002	Ferrante 1992	Ferrante 2012	Starnone 2011	Starnone 2007	Ferrante 2013	Starnone 1993
Piccolo 2010	Piccolo 2008	Piccolo 2008	Ferrante 2006	Ferrante 2006	Ferrante 1992	Ferrante 2002	Ferrante 2013	Piccolo 2008	Ferrante 2013
Piccolo 2015	Piccolo 2015	Starnone 2011	Starnone 1994	Ferrante 2002	Ferrante 2013	Starnone 1993	Ferrante 2012	Starnone 2016	Starnone 2014
Piccolo 2007	Ferrante 2013	Starnone 1993	Ferrante 2013	Ferrante 2011	Ferrante 2014	Ferrante 2013	Ferrante 2014	Ferrante 2012	Ferrante 2014
Parrella 2011	Starnone 2011	Ferrante 2011	Ferrante 2014	Ferrante 2012	Starnone 2007	Starnone 2016	Starnone 2014	Ferrante 2006	Ferrante 2012
Piccolo 2008	Ferrante 2012	Starnone 2007	Ferrante 2012	Milone 2015	Starnone 1993	Ferrante 2012	Ferrante 2011	Ferrante 2011	Ferrante 2006
Piccolo 2003	Ferrante 2011	Ferrante 2012	Starnone 2016	Ferrante 2013	Starnone 2011	Ferrante 1992	Starnone 1993	Starnone 2011	Ferrante 2011
Piccolo 2013	Ferrante 2014	Ferrante 2013	Ferrante 2011	Starnone 2007	Ferrante 2006	Starnone 2000	Piccolo 2008	Piccolo 2015	Starnone 2007
De Silva 2010	Starnone 1987	Ferrante 2006	Milone 2015	Starnone 2016	Ferrante 2002	Ferrante 2014	Ferrante 1992	Starnone 1989	Ferrante 1992
De Silva 2011	Starnone 2000	Piccolo 2015	Starnone 2007	Ferrante 2014	Starnone 2016	Starnone 2014	Starnone 2000	Starnone 1993	Starnone 2011
Parrella 2005	De Silva 2010	Starnone 1987	De Luca 1998	Starnone 2011	Starnone 2014	Starnone 1994	Ferrante 2002	Starnone 1991	Milone 2015
Starnone 2007	Piccolo 2010	Ferrante 2014	Starnone 2011	De Luca 1998	De Silva 1999	Ferrante 2011	Starnone 2016	Ferrante 2002	Starnone 1994
De Silva 2007	Parrella 2011	Starnone 2000	De Silva 1999	Starnone 2000	Starnone 1994	Milone 2015	De Silva 1999	Starnone 2007	De Silva 1999
Starnone 2014	Starnone 1993	Ferrante 1992	Starnone 2014	Parrella 2005	Piccolo 1996	De Silva 1999	Milone 2015	Ferrante 1992	Starnone 2000

Nelle graduatorie, Elena Ferrante compare 59 volte, Starnone 56, De Luca 2, De Silva 9, Milone 5, Piccolo 15, Parrella 4.

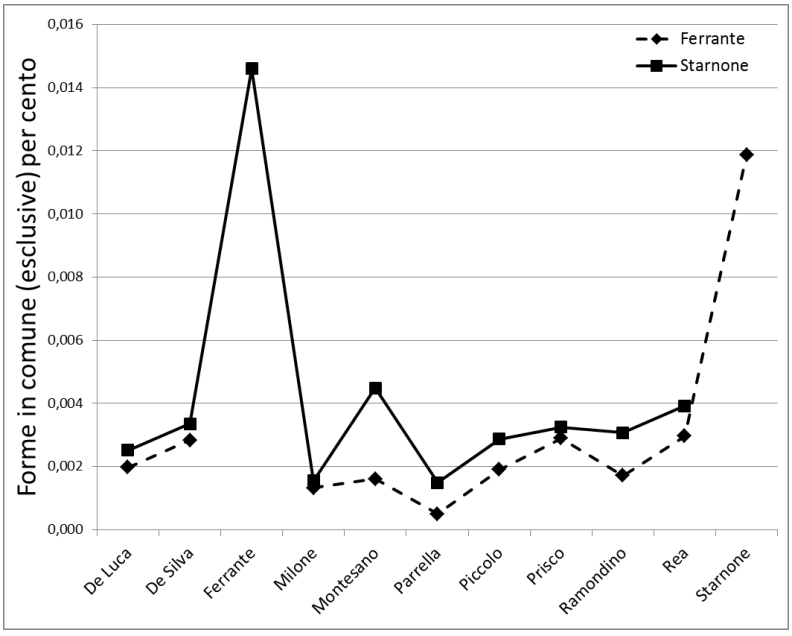
Su 10 opere di Starnone esaminate, 6 sono più vicine a uno dei romanzi di Elena Ferrante, 4 a uno degli altri libri di Starnone. In particolare, dopo il 2000 la prima opera più vicina a Starnone è sempre un'opera di Elena Ferrante. Quando si confrontano le opere dal 1993 in poi, nelle liste di similarità si alternano le opere di Ferrante e Starnone, con qualche sporadica apparizione di altri autori, in genere nelle posizioni più basse: le presenze di altri autori si fanno più frequenti a partire dalla tredicesima posizione; in posizioni più alte troviamo solo Francesco Piccolo autore della terza opera più simile a *Lacci* (e poi nella nona posizione ancora per *Lacci* e per *Autobiografia erotica di Aristide Gambia*), Rossella Milone, in

settima posizione per *Denti* e in decima per *Eccesso di zelo*, Erri De Luca, in dodicesima posizione ancora per *Eccesso di zelo*.

Un quadro del tutto diverso caratterizza le prime tre opere, che sono innanzi tutto simili ad altri lavori dello stesso Starnone (in prima posizione compare sempre un'altra opera della fase iniziale della produzione letteraria dell'autore), poi ai romanzi di Francesco Piccolo; poi, solo dalla quinta e dalla sesta posizione, rispettivamente per *Salto con le aste* e *Fuori registro*, appare Ferrante (che è, invece, totalmente assente dalle similarità relative al primo libro, *Ex cattedra*), e più avanti altri degli autori campani inseriti nel corpus. Solo in riferimento a queste opere iniziali presenta un certo grado di similarità Valeria Parrella, assente, invece, dalle liste che riguardano Elena Ferrante.

6. Uno sguardo sulle affinità lessicali

Ai risultati quantitativi presentati fino ad ora, ne possiamo aggiungere uno che ci introduce a una visione qualitativa della configurazione lessicale delle opere del corpus. Abbiamo calcolato il tasso di parole che ogni autore ha in comune solo, rispettivamente, con Elena Ferrante e con Domenico Starnone, l'autore che dai sondaggi appena presentati risulta quello decisamente più vicino a Elena Ferrante. I dati che emergono da questa verifica sono rappresentati nel seguente grafico:



Si vede chiaramente che il tasso di forme che Elena Ferrante ha in comune esclusivamente con Starnone (picco della linea continua) e il tasso di forme che Starnone ha in comune esclusivamente con Elena Ferrante (picco della linea tratteggiata) sono i più alti, in una misura che si stacca decisamente rispetto a quella che caratterizza le forme in comune con gli altri autori: il tasso di parole presenti nei romanzi di Elena Ferrante che ricorrono esclusivamente nei romanzi di Domenico Starnone è dello 0,0146% (l'autore che si trova al secondo posto è Giuseppe Montesano, con un tasso di forme in comune solo con le opere di Starnone dello 0,0045), lo 0,0119% delle parole presenti nei romanzi di Domenico Starnone sono presenti anche, e solo, nei romanzi di Elena Ferrante (il secondo autore di questa ipotetica graduatoria è Ermanno Rea che ha un tasso di parole in comune esclusivamente con Elena Ferrante dello 0,0030% , simile a quelli di De Silva e Prisco).

Se diamo uno sguardo alle parole che risultano più significative, compaiono tre parole che si caratterizzano diatopicamente come legate all'area napoletana: *sfottente*, che vanta 71 occorrenze nel subcorpus campano (43 in Ferrante, 28 in Starnone), *risatella* (30 occorrenze, di cui 20 in Ferrante e 10 in Starnone) e *malodore* (17 occorrenze, di cui 12 in Ferrante, 5 in Starnone). Per quest'ultimo lemma, si ritrova nel corpus la variante *maleodore*, usata 13 volte dal solo Francesco Piccolo. Sorprende che *sfottente* e *risatella*, così frequentemente presenti nelle opere di Elena Ferrante, non ricorrano in nessun altro autore di area campana se non in Domenico Starnone. Anche *malodore* presenta la stessa configurazione (presenza ricorrente in Elena Ferrante; e, in maniera numericamente più ridotta, solo in Starnone); inoltre, l'unico altro autore che presenta il tipo lessicale, lo usa sistematicamente secondo una variante diversa.

7. Prime conclusioni

La fase iniziale della ricerca delle similarità tra le opere di Elena Ferrante e un insieme di opere narrative italiane dell'ultimo trentennio ha portato a una serie di risultati interessanti. Innanzi tutto ha confermato l'ipotesi avanzata già a partire dal 2005 in studi quantitativi e qualitativi: è riconoscibile una marcata similarità tra i romanzi di Elena Ferrante e quelli di Domenico Starnone. Lo sguardo dall'alto utilizzato nella nostra ricerca dà ragione allo sguardo più ravvicinato con il quale Luigi Galella e Simone Gatto hanno individuato dettagliate affinità tematiche, contestuali e lessicali tra le produzioni dei due autori e allo sguardo distante di Vittorio Loreto, che è giunto alle medesime conclusioni. Da questo punto di vista la nostra ricerca risponde a un principio fondamentale della ricerca scientifica, quello della verifica e della validazione dei risultati della ricerca attraverso la replica degli esperimenti effettuati, utilizzando gli stessi metodi o, come in questo caso, metodi diversi.

La similarità tra i due autori risulta rafforzata dal fatto che Domenico Starnone mostra un deciso cambiamento della sua collocazione stilistica successiva-

mente all'avvio dell'attività scrittoria di Elena Ferrante. Solo in parte può influire sulla bipartizione della produzione letteraria di Starnone il fatto che le prime opere si configurino come raccolte di bozzetti di vita scolastica, molti dei quali hanno avuto come prima destinazione la pubblicazione su giornali e riviste: tra le prime opere sottoposte ad analisi c'è anche il primo romanzo di Starnone, *Il salto con le aste*, che appare del tutto coerente con le opere precedenti al 1992.

Le similarità tra Elena Ferrante e Domenico Starnone non si possono ridurre alla comune provenienza campana. L'argomento è stato più volte opposto da Domenico Starnone per spiegare le somiglianze evidenziate dalle analisi quantitative e qualitative. L'illustrazione più compiuta di questo punto di vista è affidato a un passo dell'ultima parte della *Autobiografia erotica di Aristide Gambia* (Starnone 2011: 432):

> - Ci sono le caratteristiche regionali, - argomentai, - e, diciamo, storicosociologiche. Io e Ferrante abbiamo in comune la Campania, Napoli, gli anni Cinquanta, l'ambiente piccolo borghese, gli stessi oggetti d'epoca, la stessa eco dialettale nella frase. Per forza che qualche somiglianza c'è.
> [...]
> - Certo. Galella ha mostrato che due scrittori molto diversi - uno di sesso maschile incline all'ironia e l'altra di sesso femminile incline ai sentimenti profondi - possono avere tratti in comune che dipendono dall'area dentro cui sono cresciuti. È interessante. Sono cose di cui la critica letteraria parla poco o niente, ormai. Ma non è sufficiente per costruirci un'intera pagina e segnalare addirittura il pezzo in prima.

La parte della ricerca di cui diamo conto in questo contributo dimostra l'inconsistenza di questa giustificazione: se davvero le ragioni delle somiglianze riscontrate fossero legate alla persistenza di caratteristiche regionali nella letteratura italiana contemporanea, o in parte di essa, le opere dei due autori sarebbero dovute risultare simili a quelle di altri autori campani. Questo, invece, non accade, se non per la vicinanza a Francesco Piccolo, e ad altri, del primo Starnone, ma non di quello successivo al 1992. In particolare, l'esame delle affinità lessicali tra Ferrante e Starnone, che stanno verosimilmente alla base della vicinanza individuata dagli strumenti automatici di analisi, mostra l'esistenza di un numero eccezionale di coincidenze lessicali che legano i due autori, e solo loro, anche all'interno del campo costituito dagli scrittori campani.

L'affinità tra Elena Ferrante e Domenico Starnone è un dato che emerge con indiscutibile nettezza da tutte le analisi testuali, da quelle qualitative, imputabili di soggettività, a quelle quantitative, più asettiche. Le conclusioni delle ricerche stilometriche contrastano con gli esiti delle indagini patrimoniali condotte dal giornalista Claudio Gatti. Le sue risultanze si basano sui dettami del giornalismo d'inchiesta e non sono verificabili secondo i criteri della ricerca scientifica (non possiamo verificare o falsificare i risultati, perché non conosciamo né i dati raccolti né le fonti), ma sono frutto di indagini rigorose secondo le metodologie in uso in ambito giornalistico e come tali da ritenere pienamente fondate. Non è però accettabile, dal punto di vista dell'analisi stilistica, la motivazione che Clau-

dio Gatti dà a questa disparità: la similarità tra i testi di Elena Ferrante e quelli di Domenico Starnone nascerebbe dal fascino che Christa Wolf ha esercitato congiuntamente su Anita Raja (per Gatti senz'altro identificabile con Elena Ferrante) e Domenico Starnone. Ma l'ammirazione, per quanto profonda e appassionata, per uno scrittore straniero, difficilmente può tradursi in concrete e minute similarità lessicali, quali sono quelle che vengono intercettate dai metodi di analisi quantitativa dei testi. La soluzione deve porsi, quindi, su un altro piano. È necessario immaginare che Elena Ferrante non nasconda un autore unico, ma la cooperazione (in forme difficilmente prefigurabili) di almeno due autori, uno dei quali può emergere dalle indagini patrimoniali, l'altro dalle ricerche stilistiche.

Ora si tratta di proseguire la ricerca, in almeno tre direzioni. La prima consiste nell'estendere all'intero corpus i confronti che, in questa prima fase, abbiamo circoscritto agli autori campani. La seconda comporta la presa in considerazione delle personalità sospettate di essere Elena Ferrante che non hanno al loro attivo una produzione letteraria: in primo luogo Anita Raja e Marcella Marmo. Per far questo, è in corso la costituzione di un secondo corpus, per forza di cose più ristretto, che comprenda i frammenti metaletterari di Elena Ferrante, raccolti nel volume *La frantumaglia* e opere saggistiche o divulgative degli altri candidati da prendere in esame. Infine, a partire dalle risultanze dell'analisi quantitativa, sarà opportuno tornare alle indagini qualitative, soprattutto sul piano della descrizione delle maggiori caratteristiche lessicali di Elena Ferrante e delle affinità con gli autori contemporanei che le risultano più vicini.

RIFERIMENTI BIBLIOGRAFICI

Ceccoli V.C. (2017) "On Being Bad and Good: My Brilliant Friend", *Studies in Gender and Sexuality*, 18(2), pp. 110-114.

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.

Chemotti S. (2009) *L'inchiostro bianco. Madri e figlie nella narrativa italiana contemporanea*, Padova, Il Poligrafo.

Dow G. (2016) "The 'biographical impulse' and pan-European women's writing", in *Women's Writing, 1660-1830: Feminisms and Futures*. Ed. by Batchelor J. & Dow G., London, Palgrave Macmillan, pp. 193-213.

Greenacre, M. J. (2007), *Correspondence Analysis in Practice*. London, Chapman & Hall.

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

Lebart L., Morineau A., Warwick K. M. (1984) *Multivariate Descriptive Statistical Analysis. Correspondence Analysis and Related Techniques for Large Matrices*, New York, Wiley.

Lee A. (2016) "Feminine Identity and Female Friendships in the 'Neapolitan' Novels of Elena Ferrante", *British Journal of Psychotherapy*, 32(4), pp. 491-501.

Moretti F. (2013) *Distant Reading*, London, Verso.

Murtagh F. (2010) "The Correspondence Analysis platform for uncovering deep structure in data and information", *Computer Journal*, 53(3), pp. 304-315.

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

Starnone D. (2011), *Autobiografia erotica di Aristide Gambia*, Torino, Einaudi.

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.

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

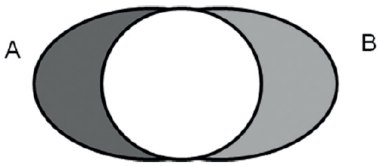
Lo studio condotto da Ondelli e Viale (2010: 59) sulle traduzioni di tipo giornalistico si concludeva con una prospettiva di ricerca incentrata sulla distanza intertestuale come strumento utile a "cercare di mettere ordine in un quadro oltremodo confuso a causa della pluralità di fattori concomitanti 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

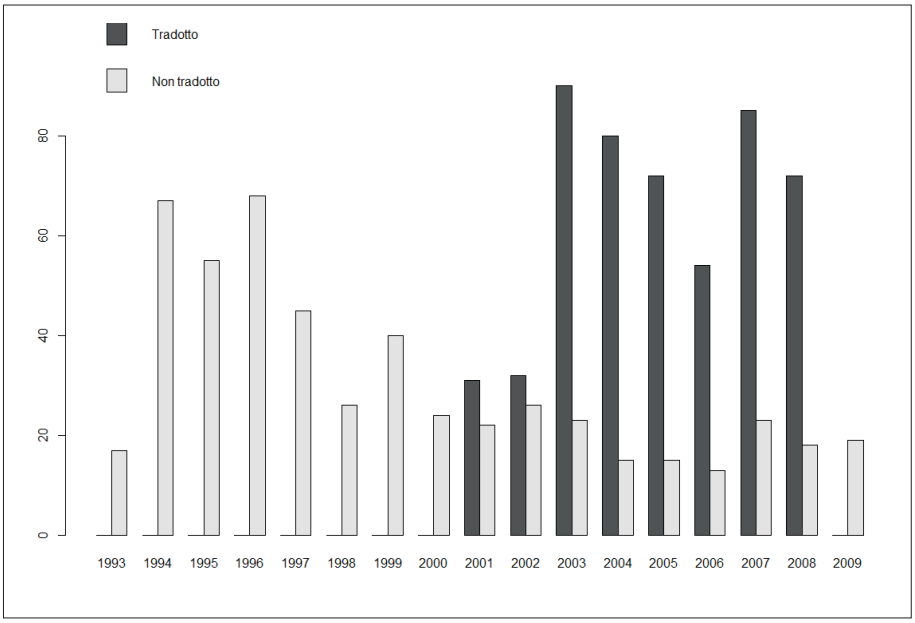
sulla rivista *Internazionale* tra il 1993 e il 2009. Al termine della selezione alcune lingue di partenza, come naturalmente l'inglese, risultavano comunque più presenti e ne abbiamo limitato l'incidenza a circa 50 articoli (Tabella 1):

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.

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) Poliremiatiche	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,

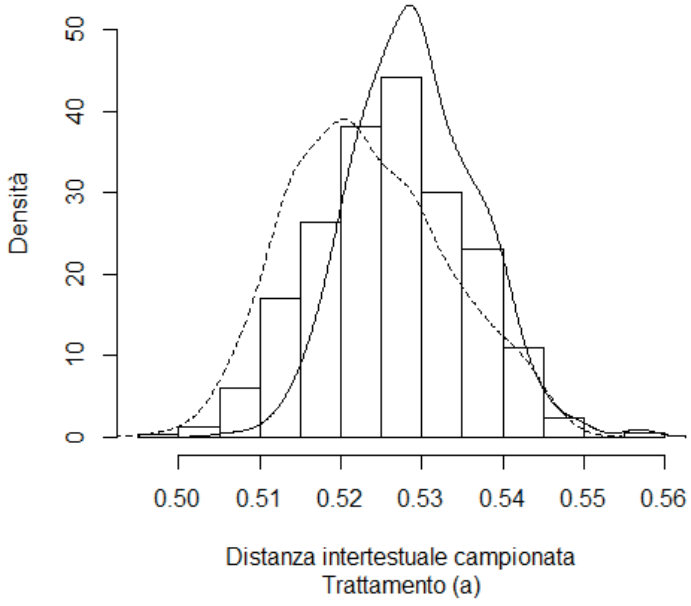
un risultato che parrebbe confermare l'universale traduttivo della convergenza o *levelling out*: "steering a middle course between any two extremes, converging towards the center, with the notion of center and periphery being defined from within the translation corpus itself" (Baker 1996: 184). In altre 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.

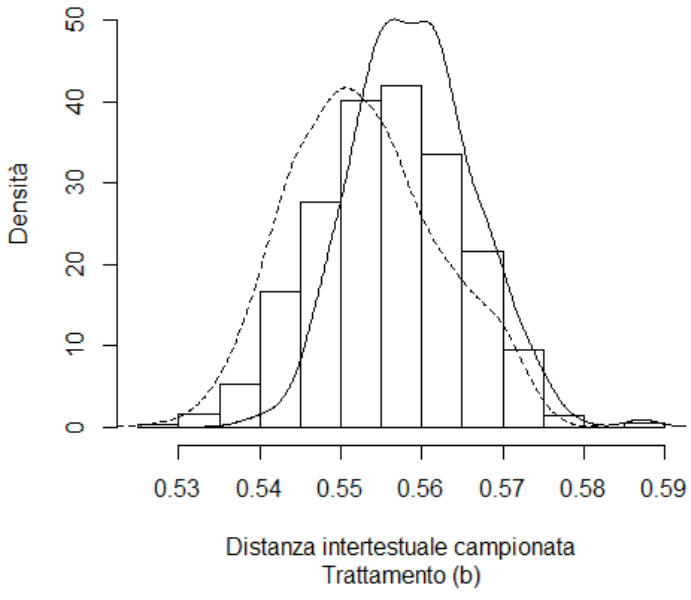
Queste conclusioni valgono per tutti i trattamenti a cui è stato sottoposto il corpus, con un leggero incremento della deviazione standard della distanza intertestuale calcolata considerando solo le parole grammaticali, ma anche con la notevole eccezione del trattamento *e*. In questo caso i valori all'incirca 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.

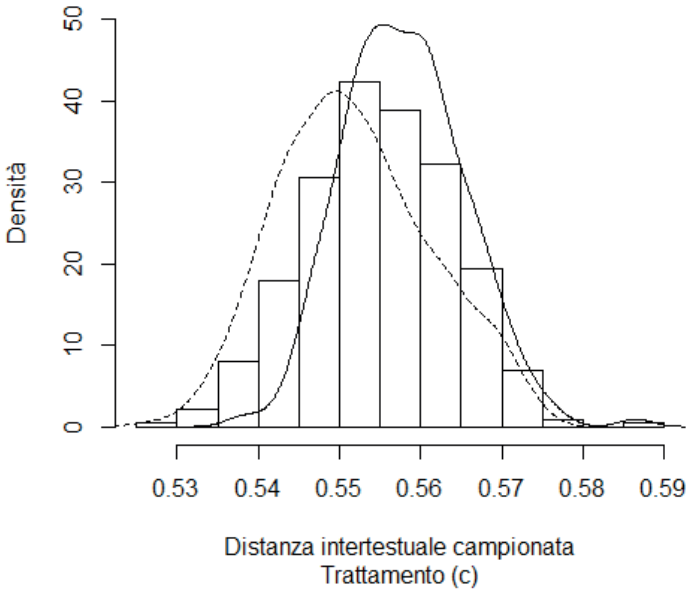
Finora abbiamo considerato la distanza intertestuale media e la sua variazione standard all'interno di ciascun subcorpus (tradotto e nativo), ma ciò che ci preme è, ovviamente, sapere se la distanza intertestuale tra le traduzioni risulti (sempre) minore rispetto alla distanza intertestuale tra una traduzione 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).

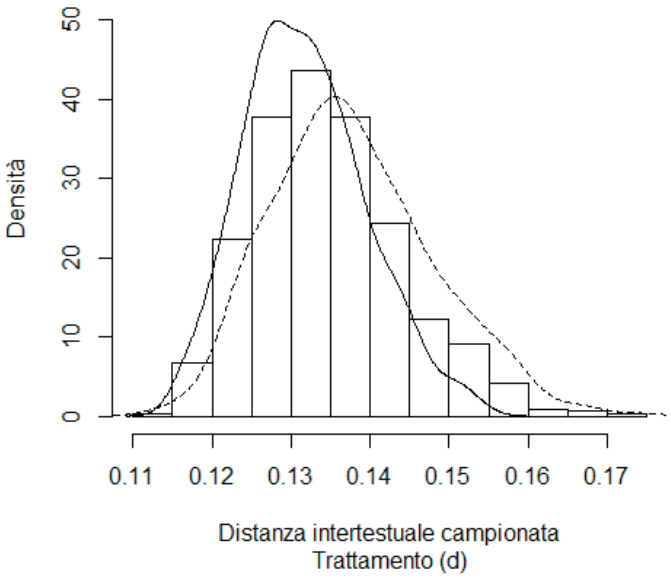
Rispetto alla distribuzione generale (i rettangoli al centro), la linea continua tende a formare curve più alte e caratterizzate da andamenti più "ripidi" e basi più "strette" della linea tratteggiata, che però si posiziona quasi sempre "a sinistra" delle linee continue (fa eccezione il trattamento *d*): 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.

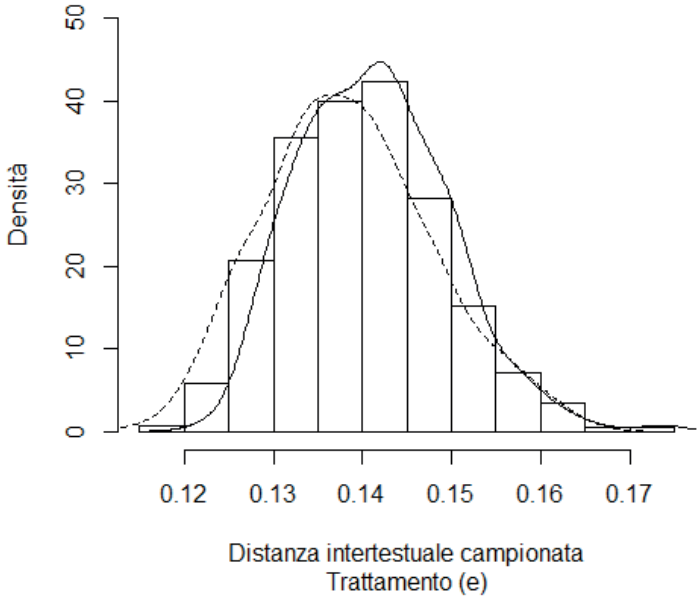
Figura 1. Distribuzione delle distanze intertestuali tra macrotesti nativi e tradotti.





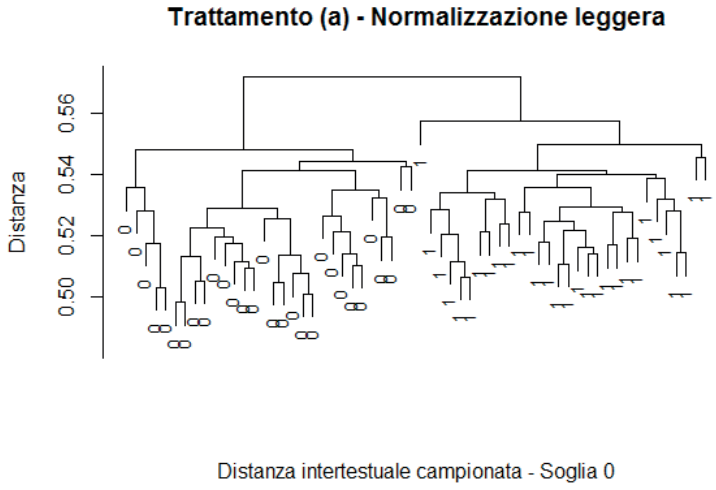

Densità
0
10
20
30
40
50
0.53
0.54
0.55
0.56
0.57
0.58
0.59
Distanza intertestuale campionata
Trattamento (c)

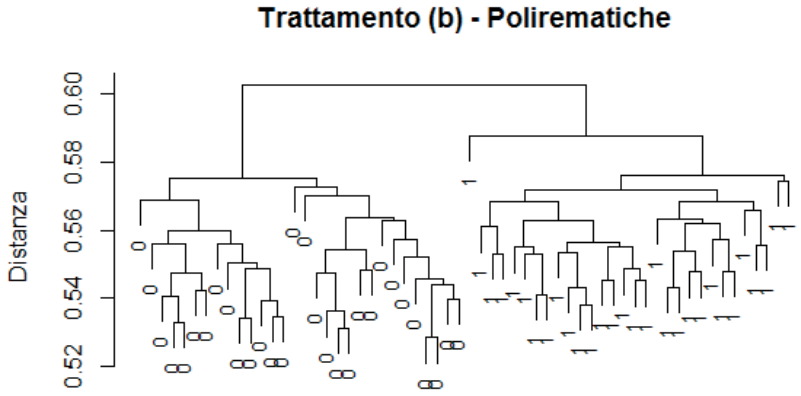

Densità
0
10
20
30
40
50
0.11
0.12
0.13
0.14
0.15
0.16
0.17
Distanza intertestuale campionata
Trattamento (d)

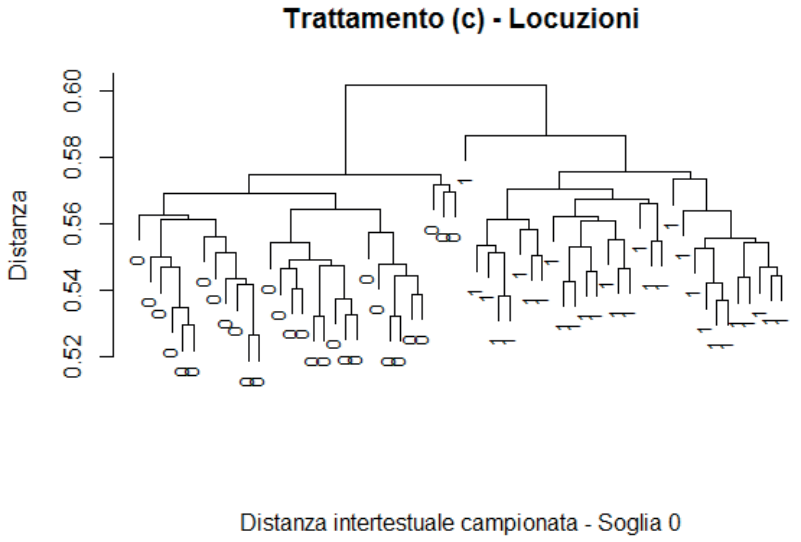


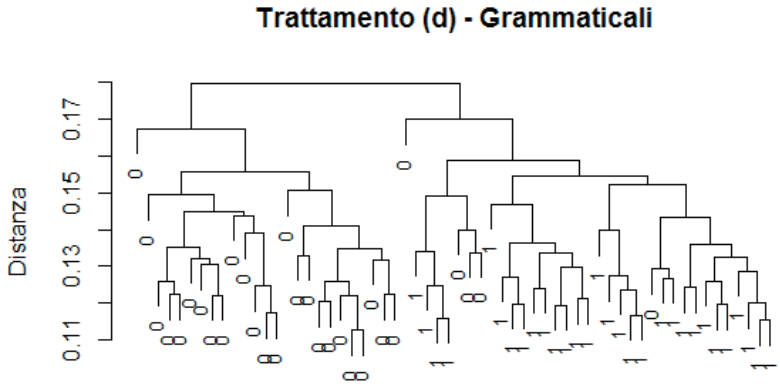
Per illustrare più efficacemente i risultati dei nostri calcoli uno strumento molto utile è il dendrogramma. Il dendrogramma consente di visualizzare in maniera intuitiva gruppi diversi di elementi (nel nostro caso: i macrotesti) e come questi siano collegati tra loro in base a una misura di distanza 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.

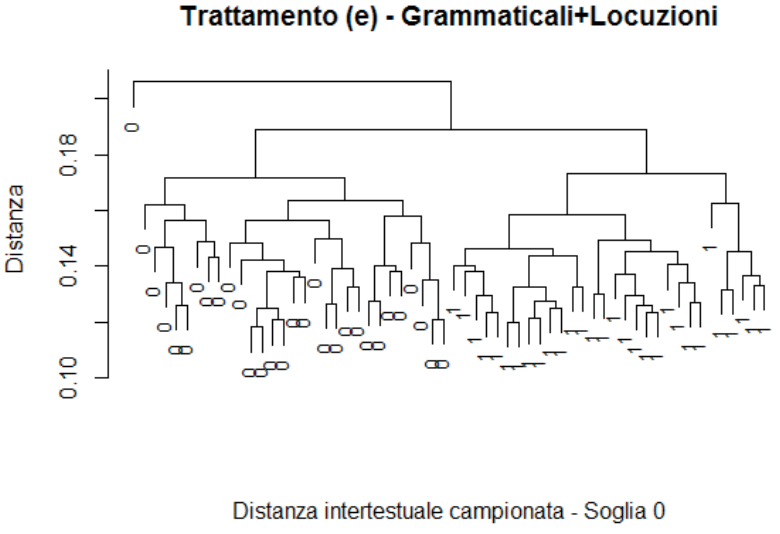
Figura 2. Dendrogrammi per i 60 macrotesti (0 = nativi, 1 = tradotti).




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 poliremiche (che dovrebbe essere più efficaci nel cogliere i contenuti).

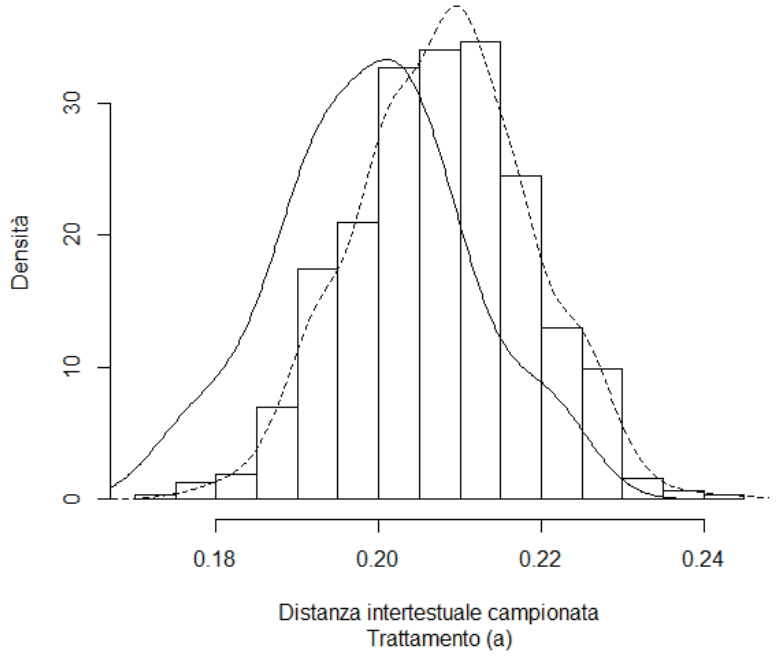
Tabella 2. Media della distanza intertestuale campionata e deviazione standard per macrotesti tradotti dalla stessa lingua e macrotesti tradotti da lingue diverse

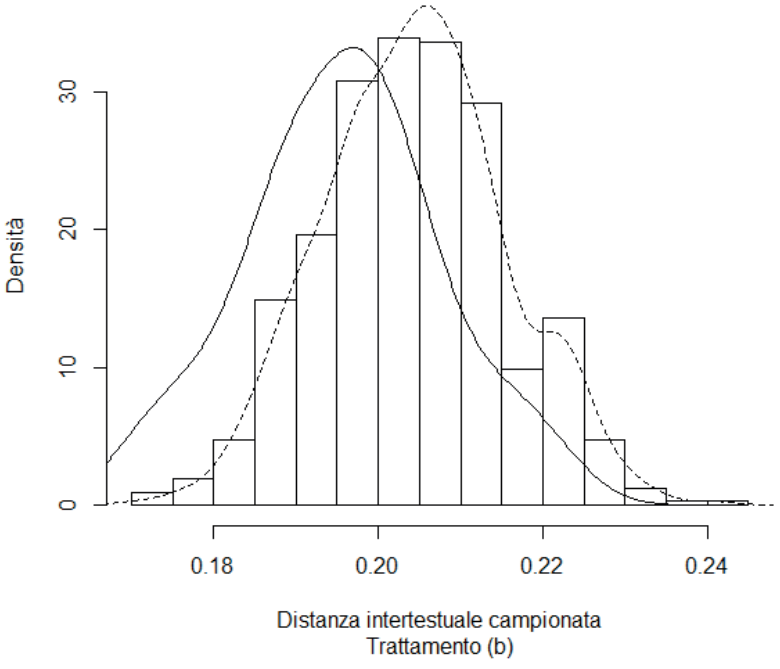
Trattamento	*Distanza campionata media infra-lingua*	*Deviazione standard infra-lingua*	*Distanza campionata media extra-lingua*	*Deviazione standard extra-lingua*
a) Normalizzazione leggera	0,199	0,012	0,208	0,011
b) Locuzioni	0,196	0,012	0,204	0,011
c) Polirematiche	0,196	0,012	0,204	0,011
d) Grammaticali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

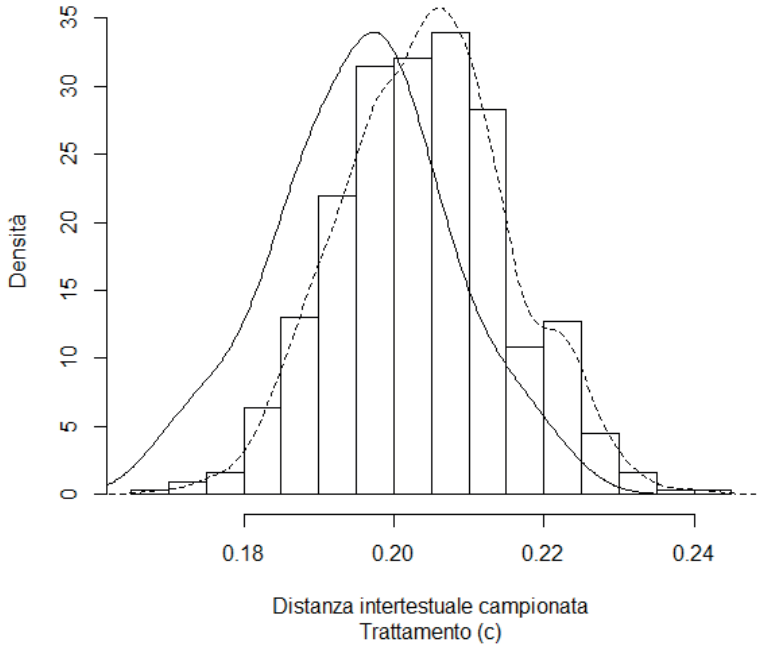
Come abbiamo già fatto nel caso del confronto tra testi nativi e testi tradotti (Figura 1), grazie ai grafici possiamo visualizzare nel dettaglio la distribuzione delle distanze intertestuali per ogni trattamento del corpus (Figura 3). Stavolta le linee continue rappresentano le distanze tra traduzioni 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.

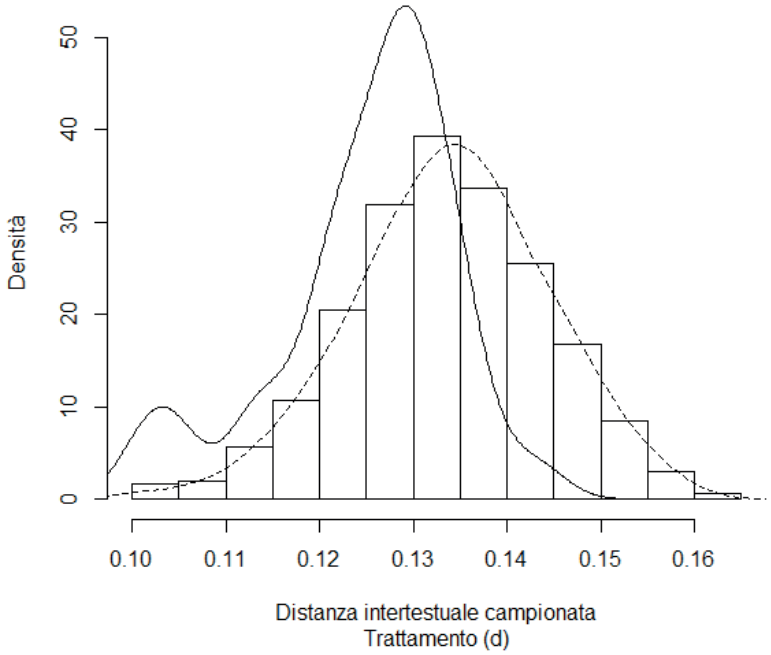
Può essere interessante notare come per i diversi trattamenti siano presenti alcune irregolarità nelle distribuzioni, che non sempre assumono la forma delle classiche "campane" normali. Soprattutto possiamo fare riferimento ai trattamenti *b*, *c* e *d*. Nei primi due è presente una "gobba" alla destra 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.

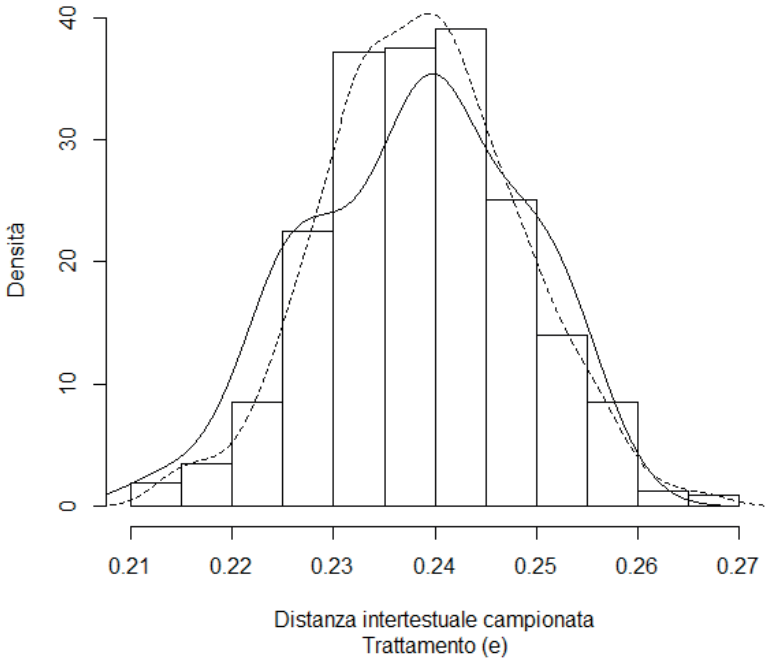
Figura 3. Distribuzione delle distanze intertestuali tra macrotesti tradotti dalla stessa lingua e da lingue diverse.





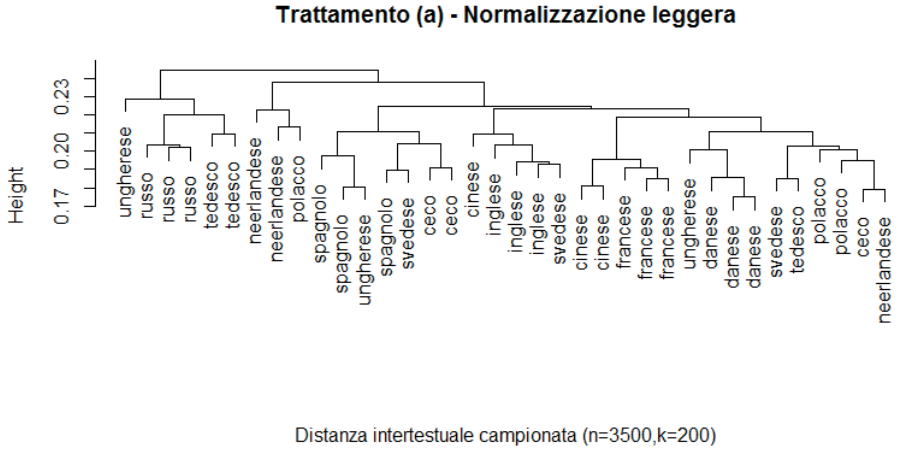

Densità
0
5
10
15
20
25
30
35
0.18
0.20
0.22
0.24
Distanza intertestuale campionata
Trattamento (c)

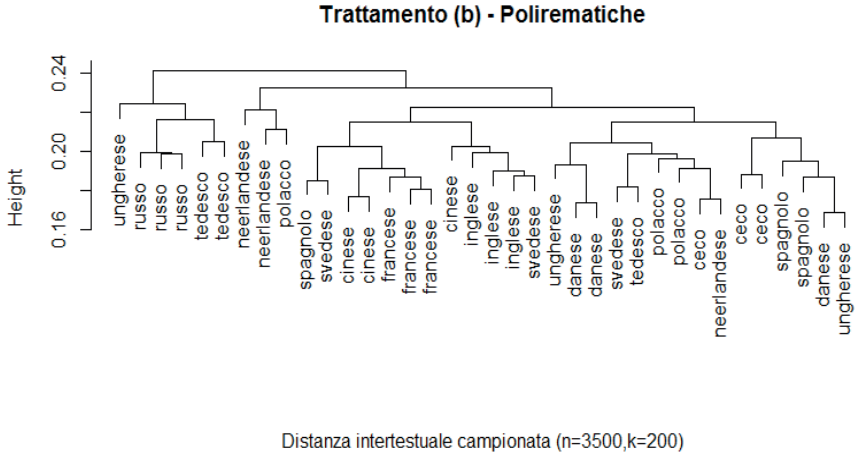

Densità
0
10
20
30
40
50
0.10
0.11
0.12
0.13
0.14
0.15
0.16
Distanza intertestuale campionata
Trattamento (d)

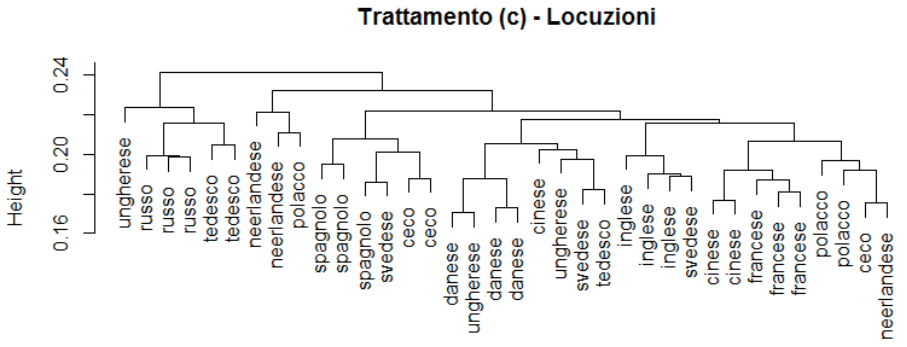


Passando alla rappresentazione dei dati sotto forma di dendrogrammi, la Figura 4 riporta gli accoppiamenti dei macrotesti in base alle lingue fonte secondo i diversi trattamenti a cui abbiamo sottoposto il subcorpus. È possibile notare una certa tendenza dei macrotesti ad aggregarsi per lingua di partenza. 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).

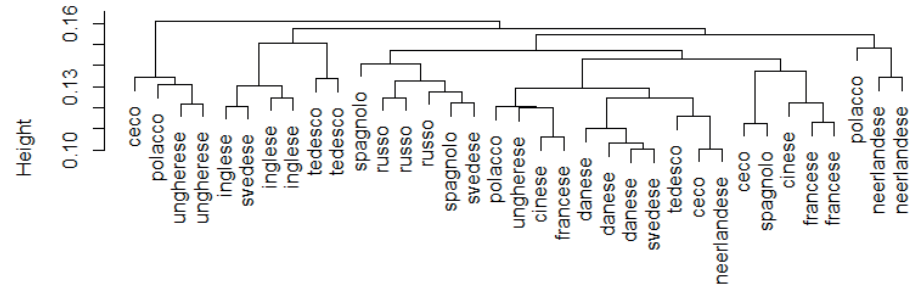
Figura 4. Dendrogrammi dei raggruppamenti di traduzioni dalla stessa lingua.



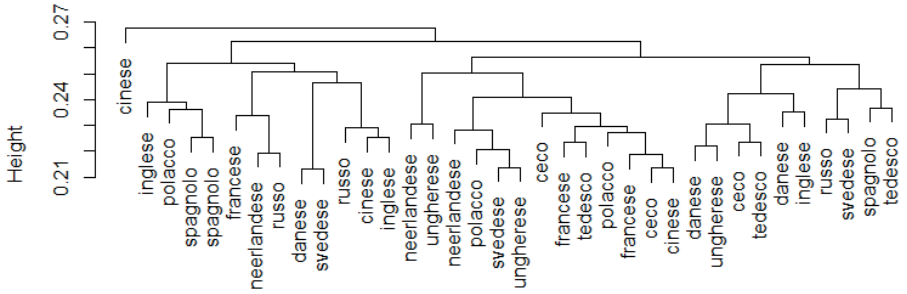




Distanza intertestuale campionata (n=3500,k=200)



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.

DOI: 10.13137/978-88-8303-913-3/18481

4. A cross-cultural contrastive analysis of interpersonal markers in promotional discourse in travel agency websites

ERSILIA INCELLI
Università di Roma Sapienza

Abstract

The present study is an investigation into the use of interpersonal markers in English and Italian tourism texts obtained from three successful online travel agency websites: one American, one British and one Italian. The aim is to explore cross-linguistic, cross-cultural pragmatic perspectives through a comparative analysis of the discoursal and pragmatic features related to the interpersonal use of language, so as to better understand how the discursive patterns of tourism texts from different cultures might affect the communicative function, or the interactional metadiscourse strategies, in promotional discourse. Results show how a comparative analysis of interpersonal devices in tourism websites offers insights not only into the way in which culture is conveyed and transmitted via tourism discourse, but more specifically into how authorial stance is constructed in the websites and how the audience is engaged in the discourse. The premise is that an awareness of the cross-linguistic and cross-cultural differences involved in interpersonal discourse can contribute to improving cross-cultural understanding, and ultimately contribute to the fields of intercultural and translation studies.

Keywords

Engagement, interpersonal markers, metadiscourse, stance markers, tourism discourse.

1. Introduction

The present article focuses on the use of interpersonal markers in English and Italian tourism texts from a cross-linguistic, cross-cultural pragmatic perspective and presents results of an ongoing research project concerning corpus-based studies in the language of tourism. This part of the research employs a specifically designed corpus of tourism texts from three successful online travel agency websites from the USA, Britain and Italy. By applying a comparative analysis of English and Italian data, the investigation focuses on the discoursal and pragmatic features related to the interpersonal use of language (Halliday 1994), the aim being to understand to what extent the discursive patterns of tourism texts from different cultures might affect the communicative function, and more specifically how they affect interactional metadiscourse strategies (Hyland 2005a) in essentially promotional discourse. In light of this, the study takes a model of interpersonality originally devised by Vande Kopple (1985) and Crismore et al. (1993), later developed by Hyland (2005a; 2005b) and Hyland and Tse (2004). This model establishes the categories of stance and engagement (author and reader) as key elements in social interaction.

The general hypothesis is that an analysis of interpersonal devices in tourism websites can offer insights not only into the way in which culture is conveyed and transmitted via tourism discourse, but more specifically into how authorial stance, i.e. conventions which shape the writing of the website texts, is constructed and how the audience (the reader) is engaged in the discourse. Given the premise that interpersonal discourse and its related pragmatic features can vary cross-linguistically and cross-culturally, an awareness of the differences can contribute to improving cross-cultural understanding, ultimately contributing to the fields of intercultural studies and translation studies. However, this work is not in itself about translation, but rather a study of the cultural interpretation of words and lexical items which function in a specialized way.

In fact, issues regarding linguistic and cultural representations in all types of genre arising from the interplay of systemic or pragmatic differences across languages and cultures are gaining significance with increasing internationalization in all spheres, especially in tourism. This in turn leads to changes in terms of marketing strategies. The Web has transformed the way marketers and customers interact. Today's travel agencies are increasingly faced with the big challenge of capturing and retaining the attention of potential travelers aware of their greater control over information. The traditional brochure has now evolved into

the dynamic website, joined by other digital resources such as smartphone apps, blogs, and travel forums and wikis. This has led to a new type of "hyper interactive travel consumer", forcing a convergence of all marketing and distribution channels into a single channel, i.e. the "customer engagement channel" (Eye for Travel 2011). These new emerging global contexts make it necessary to give intercultural and cross-cultural issues more attention, especially in terms of engagement, which is still an under-researched aspect in the field of discourse analysis (Suau-Jiménez 2017).

One important element of travel and tourism websites is multimodality, which is paramount for attracting the customer (Francesconi 2014). Despite acknowledging the fundamental role of multimodality, this paper focuses on the verbal word in the belief that the written text performs a key role in the decision-making process which draws people to a tourist destination. Hence, the focus of the analysis is on language as a commercial tool in promotional discourse, and how it conveys meaning in specialized semiotic spaces.

2. Research question

It is generally agreed that the language of tourism represents a particular type of specialised language made up of a wider range of stylistic, pragmatic and lexical features intertwined with and influenced by different registers (Dann 1996; Gotti 2006). Its characteristics have been studied both at the linguistic and social level by a variety of scholars, in comparative studies and in various types of tourism genres (Diani 2017). However, the study of this language as specialized discourse is incomplete if we do not take into account its metadiscourse, used to help the reader organize, interpret and evaluate given information (Hyland 2005a). Much has been done on metadiscourse in academic discourse (Bortoluzzi 2000; Bondi 2006), but far fewer studies have been carried out on interpersonal strategies in non-academic genres in different cultures. It is the cultural load present in tourism genres which makes the field particularly suitable for cross-cultural analysis. Hence, the main research question can be formulated as follows: to what extent do the American, British and Italian cultures differ in their interpersonal strategies in order to attract tourists and to promote tourist destinations?

It is hypothesized that the different values held by the three cultures will each foster a different interpersonal and interactional stance and approach to the persuasive strategies of the websites, and potentially different cultures will use different forms and ways of promoting destinations according to differing linguistic and textual systems affected by cultural filters. Before continuing, certain assumptions need to be made explicit, first regarding language systems in general and second regarding the relationship between language, cognition (knowledge) and culture. This chapter does not go into the intricacies of contrastive or language-typological differences but points out, if only in brief, that English and Italian

are first and foremost two distinct languages, the former Germanic and the latter a Romance language, each with long, rich histories, and with two very different ways of expressing thoughts. The languages offer not merely two parallel ways of saying the same things, but rather different ways of thinking about them – two distinct lenses through which to see the world (Crystal 2004). It follows that writing conventions, rhetoric and style are culturally shaped, and that linguistic and cognitive elements are interwoven with the cultural frame (Stubbs 2001).

The paper is structured as follows: section 3 presents the theoretical premises to the study, section 4 describes the corpora, data and methodology. Section 5 discusses the key findings, focusing on high frequency lexis and phraseological units emerging from the data. The final section takes into account the results and draws conclusions.

3. Theoretical frameworks

The concept of metadiscourse is defined here as "the linguistic resource used to organize a discourse or the writer's stance towards either the content or the reader", dealing with "the ways writers project themselves into their discourse" (Hyland and Tse 2004:156-157). These authors developed a metadiscursive taxonomy of markers which distinguished between textual and interpersonal or interactional strategies, involving two broad categories: a) interactive and b) interactional. The former helps to organize the discourse by indicating topic shifts and deals mainly with propositional content; the latter modifies and highlights aspects of the text and reveals the writer's attitude through particular features that draw the reader into the discourse and gives them an opportunity to interpret and react to it (Hyland 2005a: 52). This paper focuses on the second function of metadiscourse, i.e. the interactional strategies binding the writer and reader relationship, involving the voices of stance and engagement. *Stance* is represented by markers which express "textual" or "community" voice (Hyland 2005b: 176) and includes features referring to how writers present themselves, convey opinions, judgments and commitments, imprint their personal authority onto the text or, as in academic discourse, how they purposely step back and disguise their involvement in the text. Stance markers generally comprise four main elements: *self-mentions, hedges, boosters, attitude markers*. *Engagement* on the other hand is the way writers relate and align their readers with respect to the propositions in the text, involving readers in the discourse and connecting with them by using direct ways of address, e.g. the pronoun *you*.

Since culture, like other aspects of reality, is actively construed through language and reveals itself in language, a useful approach to the culture of a society is to focus on its lexis or capture patterns in its language (Bednarek and Bublitz 2007). For this reason, this paper also draws on insights from studies in phraseology (Sinclair 1991; Tognini-Bonelli and Manca 2002) and intercultural commu-

nication studies (Hall 1989; Hofstede 2001; Katan 2004) only briefly mentioned here for space constraints. Following Sinclair's (1991) ideas on extended lexical units of meaning, Tognini-Bonelli and Manca (2002) developed a methodology for identifying "functionally equivalent units of meaning" when comparing two languages. Their approach proposes translation equivalence by considering the collocational profile of a given node word in the source language, rather than a one-to-one correspondence between words, which can then be interpreted within the framework of linguistic representations of conceptual and cultural schemata.

Cultural orientations are a particular way of perceiving reality (Katan 2004), which inevitably has implications on language. Cross-cultural communication studies are often linked to the theories of High Context Cultures (HCC) and Low Context Cultures (LCC) elaborated by Hall (1989) and Hofstede (2001) who added more cultural dimensions to the model, e.g. the individualism/collectivism distinction. The present article takes these intercultural models into account, but acknowledges the need to redefine cultural boundaries which are becoming fuzzy as a result of globalization and the increasing use of social media.

4. Corpora

To verify the assumptions made above and address the research question, the study is empirically based on data from three comparable and relatively small corpora consisting of texts downloaded in the year 2016 from the websites of three large, successful adventure holiday travel agencies, namely the US travel agency *Grand American Adventures, Exodus Adventure Holidays* in the UK, and the Italian travel agency *Viaggi - Avventure nel mondo*. I say comparable because the websites contain similar texts with the same communicative purpose, selling destinations as *adventure* (not extreme adventure). The texts can thus be said to contain lexically homogeneous data, especially in relation to ideological categories such as *adventure, nature, environment, freedom*. To ensure further comparability and equivalence, only texts describing the main destinations and the travel agency's history, purpose and mission were sourced. Texts describing hotels, accommodation, legal services and contracts were not included in the corpus.

Source – Travel Agency	**Total words: token/type ratio**
Grand American Adventures - USA http://www.grandamericanadventures.com/	80,352/5,749
Exodus Adventure Holidays - UK https://www.exodus.co.uk/	130,405/10,733
Viaggi Nel Mondo - IT www.viaggiavventurenelmondo.it	91,151/13,027

Table 1. Corpora used for the study.

The number of running words totaled 80,352 in the American tourism corpus (hereafter UST), 130,405 words in the British tourism corpus (UKT) and 91,151 in the Italian tourism corpus (ITT). Although comparable, the UK corpus is notably larger, so relative frequencies were considered in the analysis.

5. Methodology

The research methodology adopted for this study follows frameworks which integrate quantitative and qualitative techniques for analysis, such as corpus-assisted discourse analysis (CADS; Partington et al. (2013)), and refers to studies which have adopted corpus linguistic retrieval techniques in cross-cultural communication studies and translation, especially in the fields of tourism (Tognini-Bonelli and Manca 2002).

As far as the methodological procedure is concerned, after reading the texts for first impressions, the first step was to create frequency lists for each corpus (not shown here for space constraints). I used *Wmatrix* (Rayson 2003) for word lists and part-of-speech (POS) tagging; *ConcApp* and *ConcGram* (Greaves 2009) for concordance lines, collocation patterns and concgram configurations; *Sketch Engine* (Kilgarriff et al. 2014) for Italian word frequency lists and POS tagging.

In the English corpora the items of most relevance to this investigation ranked in the top ten/twenty of the word frequency lists, i.e. the pronoun *you*, the possessive adjectives *our/your* and subject pronoun *we*. I then looked for the equivalent Italian interpersonal markers, i.e. the 1st person plural subject and object pronouns *noi* and *ci* ('we' and 'us'), the possessive adjectives *nostro**[1] and *vostro** ('our' and 'your'), the 2nd person plural subject and object pronouns v*oi* and *vi* ('you'), the 1st person plural of verb forms, e.g. **iamo, *emo*, and the 2nd person plural of verb forms, e.g. **ate, *ete, *ite*. However, only *si* (impersonal 3rd person singular 'one') and *ci* (1st person plural object pronoun 'us') as candidate pronouns were high frequency items in the ITT corpus. I then checked their occurrences to exclude any pragmatic use as an adverb, e.g. *ci siamo andati* (we went *there*). The next stage was to analyse the items within their concordance lines followed by expanded text and context, and the collocational profiles of the word, which can tell us a great deal about the linguistic environment of the interpersonal marker and the noun phrase it supports, including evaluative and affective attributes.

1 The * symbol means the word or verb can be inflected or conjugated according to gender of person or object, and according to singular or plural.

6. Analysis

A detailed contrastive analysis of the texts in each corpus shows that specific communicative functions such as the description of destinations, their historic, geographical or cultural aspects, exemplification and explanation, have basic similarities. On the contrary, what appears to be different is the interpersonal metadiscourse.

6.1 Stance

For the purpose of this paper, given the input of the quantitative results and to efficiently manage the data, I quantify and investigate the high frequency features of s*elf-mentions* and *hedges* and only mention the use of *boosters* and *attitude markers* when they occur within the examples, acknowledging that these latter two elements are also incisive on the persuasive intensity of promotional discourse but are not the focus of this research.

6.1.1 Self – mentions

Self-mention as a category is displayed in the corpora principally through: the name of the company, i.e. *Exodus, Grand Real Adventures, Viaggi - Avventure nel mondo*; the 1st person plural subject and object pronouns *we* and *us* in the UST and UKT corpora; the equivalent subject and object pronouns *noi* and *ci,* and the 1st person plural verb form in the ITT corpus; and the possessive pronouns *our* and the Italian equivalent *nostro**.

Self-mention presents a "discoursal self" (Ivanič1998), which can produce a powerful rhetorical strategy for constructing authorial identity. In corporate discourse and in particular in promotional discourse, *self-mention* becomes a key strategy for implementing competitive marketing through positive identity construction. In actual fact, differences were found not only between the Italian and English corpora, but there are also notable differences between the UST and UKT corpora. For example, the subject pronoun *we* in the UST corpus is almost always an "exclusive authorial *we*" (Quirk et al., 1985) used to refer to the travel agency itself, and rarely includes the customers. On the other hand, *we* in the UKT corpus and the *we* equivalents in the ITT corpus had a variety of contextual and situational 1st person plural constructions involving both inclusive and exclusive *we*.

Table 2 quantifies the high frequency self-mentions (excluding company names) in the three sub-corpora, which are then explored within their textual environment for comparative analysis.

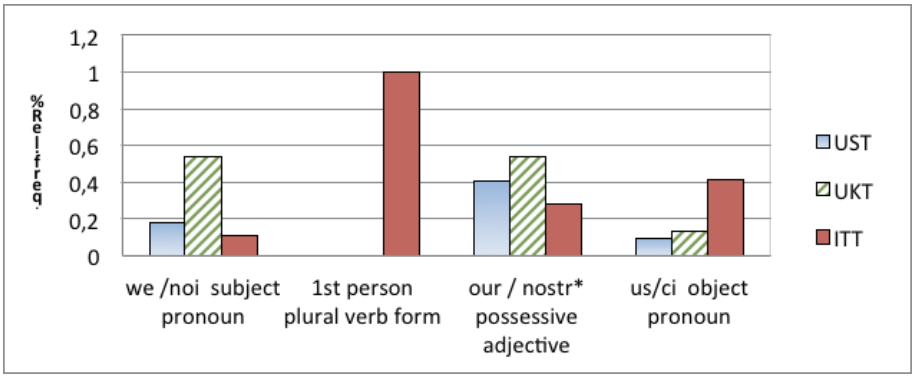


Table 2. Self-mention stance markers in the corpora.

We can see how the Italian corpus, although smaller than the UKT corpora, is heavily represented by self-mentions in the 1st person plural of the verb, with a total of 966 instances (1% relative frequency) compared to the total *we* mentions in the UST corpus (167 instances, 0.18 %) and the UKT corpus (797 instances, 0.54%). The subject pronoun *noi* ('we') is less frequent in the ITT corpus (90 instances, 0,11%) due to the prominent use of the inflected verb forms, and probably also due to the fact the Italian, unlike English, can use verbs without expressing the subject. The object pronoun *ci* (*us*) is more frequent in the Italian corpus (0.4%) than in the English sub-corpora, each with less than 0.1%. *Ci* functions both as a reflexive pronoun, e.g. *ci rilassiamo* (*we* relax) or as a direct or indirect pronoun, e.g. *ci mostrano* ('they show us'), sometimes in an enclitic position, e.g. *fateci sapere/guidandoci* ('let us know/driving us'). In the ITT corpus, the 1st person plural is represented mainly by the present indicative **iamo* (847 instances), and to a lesser extent the future tense **emo* (99 instances) and the conditional **emmo* (17 instances), e.g. *abbiamo, proseguiremo, potremmo.* The examples below illustrate how *self-mentions* are constructed in each sub-corpus.

(1) **Grand American Adventures** specialises in small group holidays…, with **unrivalled** knowledge and experience, **we are committed to** bringing **you** the **finest** small group adventures. UST

(2) **Our tours** are aimed at people of all ages, although most of **our** travellers are aged between 25 and 55, **we believe** that age is most definitely just a number. UST

(3) We love finding new ways of discovering the world to share with **our** customers. UST

Frequent reference is made to the travel company's name (example 1) in all three sub-corpora. In the UST corpus close investigation of *we* concordance lines showed how exclusive *we* and *our* often combined with lexis referring to the company's *credo, belief, commitment, style*, as in examples (1) - (3) *we are committed to, we believe, we love*, also creating a level of informality in the register. Besides, the examples show how self-mention patterns, for example the retrieved concgram *Grand American Adventure/we/ our* juxtaposed near positive evaluative adjectives and adverbs increase the (persuasive) illocutionary force of the statement, e.g. *Grand American Adventure has unrivalled knowledge and the finest adventures.* Example 2 highlights their strategy of marketing flexibility and inclusion open "to all ages". Let's now turn to examples from the UKT corpus.

(4) **We** are **very proud** to have **won the Best** Overall Special Interest Tour Operator accolade at this year's **British Travel Awards.** UKT

(5) That is what **Exodus** was founded upon … exploring this amazing planet we all live on. UKT

(6) W**e** always remember that **we** are only guests. So **we** travel courteously and respectfully, in smaller groups to minimise **our** impact. UKT

(7) **We believe** in small environmentally aware ships … to ensure **we travel responsibly. Our** vessels burn Marine Gas Oil … clean fuel with a low emission factor. UKT

Example 4 is a very corporate-like way of constructing identity. The mention of the award received by the company is a key strategy in competitive marketing to convince customers of a company's expertise and excellence.

In example 5 inclusive *we* links the company to the reader and the whole of the human race (*this amazing planet we live on*). Examples 6 and 7 construct the company's image and mission as an environmentally aware and socially responsible travel company. It is interesting to see how *we* alternates between vaguer and more precise references, as in 6: *we are guests* can ambiguously include both the company and the participants or even humanity itself. In fact, inclusive *we* strategically embraces the customer/reader in socially responsible actions, e.g. *We believe in small environmentally aware ships* (7), thus co-creating values of membership and purpose. The marketing strategy is based on well-defined messages reinforced by evaluative boosters and adverbials.

On the surface, the use of self-mentions as a marketing strategy in the ITT corpus appears similar to that in the English corpora. However, the analysis revealed some interesting differences in the collocates and co-text of person markers.

(8) **Noi** di **Avventure nel Mondo** siamo stati creatori e protagonisti di **questa formula** di affrontare l'esperienza del viaggio.

(9) **Noi forniamo** una puntuale documentazione, garantendo la scrupolosa preparazione dell'itinerario. Consulta la **nostra** libreria e scegli il tuo libro per viaggiare.

(10) **I nostri** sono **viaggi** disorganizzati riservati a viaggiatori **culturalmente motivati**.

The subject pronoun *noi* ('we') is used in examples 8 and 9 to emphatically strengthen the authorial stance of the travel company, e.g. *Noi di Avventure*. In examples 9 and 10 the 1st person plural possessive pronoun *nostri* ('our') and the 1st person plural *forniamo* ('we provide') co-create values based on a 'cultural motivation' (see example 10) for travel, highlighting the fact that *Avventure nel Mondo* (henceforth *AM*) is a well-read travel agency through lexis such as *documentazione, libreria, preparazione*. The aim is to appeal to an 'educated', 'cultured' public.

(11) La **nostra formula …, le nostre avventure** sono viaggi scomodi.

(12) Per chi fa gia' parte della **nostra grande famiglia**, …, riconosce in pieno la **nostra** competenza, **la nostra** professionalita' e **la nostra** preziosa originalita'.

The *we/our* patterns in the ITT corpus construct a strong identity with characteristics which are quite different from those highlighted by the American and British travel agencies. The Italian travel agency seems to use rhetorical repetition, based on the possessive adjective *nostr* + noun/noun phrase*, with the aim of constructing a niche for itself in the travel market. The most frequent collocates of exclusive *nostr**, are *formula* (16 hits) and *avventura* (11 hits), other collocates include *viaggi* (trips/tours), *libreria* (bookshop), *famiglia* (family). The emphasis on *nostra formula* ('our formula') constructs 'uniqueness' and binds customers by appealing to their own 'uniqueness'. Examples 10 and 11 imply exclusion and self-selection, e.g. *I nostri viaggi sono scomodi/disorganizzati* ('our trips are uncomfortable/disorganized'). Here *AM* has a clear idea of its target audience and prefers to 'exclude' customers who may not be suitable for their adventure holidays. This makes their marketing strategy very different to the Anglo-American model, which makes every effort to 'include' a reader/customer. In effect, *AM*'s marketing strategy may be a defence measure against tourists who have expectations *AM* may not be able to meet. At the same time, they strongly believe in this *formula* as the key to their *success*, their *originality* and *professionality* (see example 12). In other words, they seem to propose a whole philosophy or way of life and an alternative to the mass market ideology. The identity construction is particularly evident in the use of emphatic reflexive

noi stessi ('ourselves'), e.g. *Viaggiare per conoscere noi stessi; l'Avventura è in noi stessi* ('we travel to know ourselves'; 'the adventure is inside us').

Example 13, from the UKT corpus, shows how sometimes authorial stance overlaps with engagement, as illustrated in the *we /you* proposition.

(13) At **Exodus**, **we know** what makes **you** tick when it comes to holidays. UKT

Stance and engagement are in fact "two sides of the same coin' (Hyland, 2005b:176), as both contribute to the interpersonal dimension of discourse with overlap in the functions of the two voices. The result is a powerful interpersonal engagement strategy, the key concept being *We have exactly what you want for a perfect holiday.* Overlapping strategies of voice are frequent in the UST and UKT sub-corpora. The retrieved concgram *we/you* in the UST (63 instances) and UKT (192 instances) corpora is representative of this strategy. The pattern *we* (the travel agency) + *verb* + *you* (customer) underlines the strategy of assuring the client (*you*) that (*we*) the company can take care of them, e.g. *we recommend/suggest/advise you to*.

In contrast, there are fewer instances of the 1st and 2nd person markers within the same proposition in the ITT corpus. This gives the impression that the overlapping interpersonal strategy is less represented in the Italian travel agency website. On the whole, we can conclude that self-mention interpersonal markers are particularly effective in constructing identity, uncovering different marketing priorities.

6.1.2 Hedges

Although there are various ways of hedging both in English and in Italian, for example through attitude markers (e.g. *it is interesting to*), high frequency hedging markers in the tourism corpora are represented by the modals *can, may*, and the Italian equivalent *potere* (with all its inflections) and *if* conditionals (*se* in Italian; Table 3). *Potere* and *volere* are called 'verbi servili/modali' so they have the same pragmatic function as modal verbs in English. Hedging devices generally indicate "the writer's decision to withhold complete commitment to a proposition" (Hyland, 2005: 178): this makes modal verbs good candidates for hedging as they allow information to be expressed indirectly or 'hedged' as an opinion rather than as a fact.

On closer manual inspection of concordance lines and expanded co-text, the pragmatic use of *can* or *may* in the UST and UKT sub-corpora and *potere** in the ITT corpora are not always used to express hedging: for example, *you may apply for an East African Visa* (UKT) expresses permission. After checking meanings in expanded text, the total numbers were lower than expected, e.g. *can* had 360 overall instances but, on verification, 227 instances as a hedging marker.

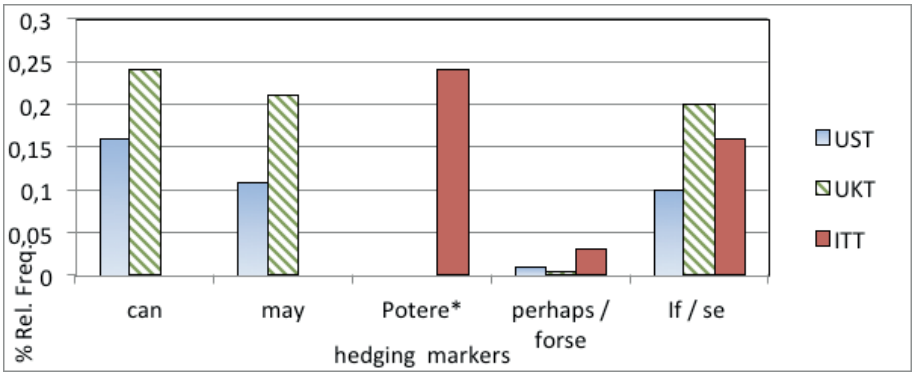


Table 3. Sample hedging markers in the three corpora.

We can see in Table 3 above that cumulatively, *can* and *may* in the UKT corpus have more instances than *potere** in Italian, (227 instances, approximately 0.2%). The UST has lower frequencies (e.g. *can*: 110 instances, or 0.15%). *Potere** (182 instances, 0,2%) comprises various conjugated forms, the most frequent of which are the 2nd person plural future tense *potrete* ('you will be able to'; 27 instances), and *potete* ('you can/may'; 20 instances), followed by a few other inflected forms, e.g. *può, potresti. possiamo, potremmo.* Adverbs such as *perhaps* have lower frequencies in the English sub-corpora than in the Italian corpus (*forse* has 32 hits in ITT; *perhaps* has 12 hits in UST and 3 in UKT).

(14) The facilities **can** be quite simple but cosy with all huts heated by a wood-burning stove. UST

(15) There is a stunning view from baboon cliffs…, and here we **may** see the small furry Rock hyrax. UKT

(16) From June to September it is monsoon season, it will be hot and humid in Nepal and you **may** well get rain. UKT

In the examples above we can see how claim-making is hedged so as not to raise too many expectations, but guarantee a certain amount of customer satisfaction. As claiming certainties is risky business, travel agencies prefer to invoke potential barriers in the way of their (future or past) actions which can help them disclaim responsibility for the absoluteness of their propositions. This defensive device, enacted through the pragmatic use of modals and *if* clauses, is convenient when talking about accommodation (14), wildlife (15), and the weather (16), widely 'hedged' topics in the UST and UKT corpora.

Compare the following examples of *potere** ('can/may') in the ITT corpus.

(17) Quest'oggi **potrete** dedicarvi allo shopping, al relax, **oppure** ad un "pub crawl" nella città di Galway, **probabilmente** la città più graziosa e vivace d'Irlanda.

(18) **Oppure, potreste** dividere le Vostre energie tra Galway e le isole Aran.

(19) **Tempo e clima permettendo**, **potrete anche** provare la pesca sul lago ghiacciato.

When we investigate close collocates of *potere** what is interesting in the Italian corpus is how the modal is juxtaposed to other modifiers and conditions creating a double hedging effect, e.g. the hedging patterns *potrete* ('you can') + *oppure/probabilmente* and *tempo e clima permettendo* + *potrete*. The examples also show how the 2nd person plural future tense *potrete* is more frequent than the other tenses. Although the prima-facie translation in English is *you will be able to*, its pragmatic function is to 'politely' make a suggestion to the reader. The effect is 'we're telling you what you *can* do' offering options, rather than 'we're telling you what to do'. Compare: *Take a tour to Burrow Hill* with *Se volete, potete godere di un tour a Burrow Hill*. ('If you want you can take a tour to Burrow Hill'). The hedging in Italian has two pragmatic functions: first it has a politeness and mitigating effect; second, the proposition gives the reader/customer the impression they are in control of their holiday. This is all part of the agency's stance. The concgram *se /volere*/ potere** is a recurrent pattern (8 instances) in the ITT corpus (Figure 1), whereas the *if /want/ can* concgram does not occur in the two English corpora.

```
1          Se non vi basta, alla fine del tour, potete se volete,  prolungare il vostro soggiorno a
2           la più antica università dell'Irlanda dove se volete potete ammirare nella Old Library (Vecchia
3       di una contea famosa per le sue mele da sidro. Se volete, potete godere di un tour del sidro a
4       e le maestose ed imponenti Scogliere di Moher. Se vorrete, dopo il tour potete prolungare il vostro
5            Che ci crediate o no, accade anche oggi! Se proprio volete rivivere quelle immagini, potete
```

Figure 1. *Se + volere/potere* concgram pattern in the Italian corpus.

To conclude, the main difference between the English corpora and the Italian corpus is in the pragmatic use of the hedging devices. In the former , the modals *can/may* and *if* sentences create a defence strategy in marketing the product, whereas hedging in the Italian corpus focuses on modifying the illocutionary force of the proposition for politeness and mitigation.

6.2 Engagement features

Engagement is the means by which writers bring readers into the discourse by anticipating their possible expectations and interpretations. Two main engagement strategies can be identified in the tourism corpora, confirming previous studies (Hyland, 2005b). The first strategy uses linguistic devices aimed at meeting the readers' expectations of inclusion and solidarity. Readers are addressed as participants by means of the personal subject or object pronoun *you*, the Italian equivalents being the subject pronoun *voi*, the object pronoun *vi* and the *2nd person plural* verb forms, **ate, *ete, *ite*, e.g. *visitate, proseguite*, (the 2nd person singular pronoun *tu* and the 2nd person singular verb forms are not frequent in the ITT corpus), and the possessive adjectives *your* and *vostr**. The second strategy consists in rhetorically aligning and positioning the audience, guiding the reader to interpret or carry out particular actions. This process is achieved through *directives* and *questions* (see section 4.2.2; Hyland (2005b) also identifies other engagement features common to academic discourse, such as *references to shared knowledge* and *personal asides*, which are not discussed here).

6.2.1 2nd person engagement markers

Table 4 quantifies 2nd person reader engagement markers across the corpora.

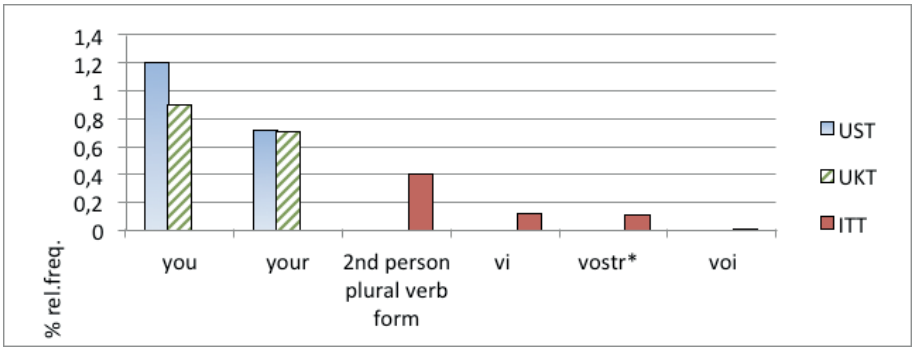


Table 4. Person reader markers in the tourism corpora.

As we can see in Table 4 *you* and *your* are high frequency pronouns in the UST and UKT corpora. The 2nd person plural subject pronoun *voi* (17 instances) is not common, (also related to the fact that Italian does not need to express the subject); it is used to address the audience or as a polite form of reference. The context of the situation usually reveals who *voi* is referring to, but here in the tourism texts the item can be ambiguous and we are left to interpretation. There is a clear preference for the direct and indirect pronoun *vi* (108 instances, 0.11%), sometimes acting as

a reflexive particle or as an enclitic pronoun, e.g. *vi potete lavare, vi suggeriamo, fatevi*. Overall, *you* in the ITT corpus is most frequently expressed in the 2nd person plural verb form, for a total of 365 hits (0.40%), e.g. *avete, continuate, siete.* Nevertheless, propositions using these subject verb forms in the ITT corpus are far less frequent than *you* propositions in the UST (885 instances, 1.2%) and UKT (1120, 0.90%) sub-corpora. The possessive adjective *vostr** (0.11%) is also less frequent in the ITT corpus than in the UST (0.71%) and the UKT (0.72%) corpora. It is worth pointing out here that the low frequencies may also be due to the fact that Italian commonly uses the impersonal third person singular to address an audience, e.g. *si può visitare*, a polite but distant way of addressing the customer. *Si* refers to the impersonal 'one', e.g. *one can visit*, but it can also be translated by the vague 'you'. Of the 600 instances of *si*, approximately 200 are for impersonal address.

The examples below illustrate how engagement strategies in each corpus are qualitatively constructed in relation to specific cultural and institutional contexts.

(20) Have **you** ever fancied a bit of time travel? These routes are guaranteed to transport **you** back to some of the world's oldest continually inhabited settlements. UKT

(21) Daily **you will** see ancient sites and experience the countryside and the gastronomic delights. UKT

(22) The moment **you** fly over the rim [Grand Canyon]is an experience **you'll never forget.UST**

(23) The snowy peaks and mountain lakes of this superb national park **will leave you falling head over heels** for this incredible area. UST

(24) **You'll** soon realise just what an enchanted world **you've** had the **privilege** to enter into. UKT

In examples 21, 22 and 23 customers are engaged in the co-creation of values like history and culture (*ancient sites, the world's oldest, gastronomic delights*), nature and environment (*countryside, canyon*), through the persuasive strategy of linking 'you' the customer to exhilarating experiences *you'll never forget, you'll feel head over heels*. Notice also the engagement question in example 20 (*have you ever…?*). The informal personal question creates a pseudo-dialogue which takes the form of an intimate, private conversation with 'you', appealing directly to the customer. Likewise, the company constructs a privileged *you* which is taken 'special' care of; the reader/customer is made to feel the only person in the world to live such an extraordinary travel experience.

(25) Whatever trip **you** decide upon, there's always an **inclusive** atmosphere, with all activities taking place as a **group**. UST

(26) Exodus offers **FREE** airport arrival and departure transfers …so no matter which flight **you** choose to arrive on **you** will be met and transferred to **your** hotel. UKT

(27) All of **our** tours are designed to be flexible …, helping **you tailor your** tour to **your own** interests, budget and level of adventure. UKT

Example 25 refers to values related to group inclusion, a clear signal of membership binding writer and reader. Examples 26 and 27 are meant to co-create the values of care, ease and efficiency in the services offered. The items *your* and *you* directly align the readers in the co-creation of 'economic' values such as a *free* service and *budget* holiday. The uniqueness of the holiday is emphasized: it is *personally tailored* to suit *your own interests* and needs. Here the engagement strategy is constructed around the lexical unit *your own*, which enhances 'personal freedom'. The pattern *your own + noun phrase* is recurrent in both the UST and UKT corpora (59 instances, 0.07% v. 237, 0.17% respectively), but with different collocates reflecting different cultural orientations. Figure 2 shows sample concordance lines of the recurrent collocation pattern in the UST corpora.

```
1     the whole experience of setting up camp, doing your own cooking, getting supplies. It was very satisfying
2   land of rock spires that leave you searching for your own description of this bizarre landscape. Day 5
3    Hire 52 If you want to take your adventure into your own hands, our range of campervans will give you every
4    more leisurely feel, allowing you to explore at your own pace. With plenty of free time and optional
5    national parks where you are encouraged to find your own path off the trails. Alternatively, take to the
6     itinerary Optional Solo Sleeper available, for your own room and tent Meals: 10 breakfasts, 12 lunches,
7       is compulsory). However, if you would prefer your own space then we can offer an optional Solo Sleeper
```

Figure 2. Sample concordance lines for *your own + noun phrase* in the UST corpus

In the UST corpus the closest collocates of *your own* are *room* (27), *pace* (12), *space* (6), *interests* (4). This phraseological pattern focuses on 'doing/having something yourself', conveying the idea that the participant is the key player in the holiday, actively taking the adventure into their 'own hands' (figure 2, line 3).

(28) On the park's extensive system of trails, **you'll** encounter **a kaleidoscope of colour** and **a fantasyland** of rock spires that leave **you** searching for **your own** description of this bizarre landscape. UST

In example 28 powerful evaluative noun phrases (*a kaleidoscope of colour, fantasyland* etc.) construct the company as an agent who can make dreams come true. Interestingly, in the UKT corpus *your own* appears to collocate more with items related to tourist services, providing practical information and advice, e.g. *dates* (190), *flights* (13), *visa* (8), *expenses* (5), as in *If you miss the transfer you must make your own way back at your own expense*.

Turning now to engagement strategies in the ITT corpus, as 2nd person singular/plural markers are less frequent, on the surface it would appear almost like the readers/customers are barely taken into account in the construction of shared values. Nevertheless, where the Italian equivalents of *you* and *your* engagement markers do occur (i.e. *voi/vi*, 2nd person plural verb forms, and *vostr**), they have similar pragmatic functions.

(29) È possibile scegliere tra tantissime attività per **personalizzare** il **vostro** viaggio.

(30) Benvenuti nel **mondo dei sogni** realizzabili…, il mondo intero è ai **vostri** piedi.

(31) Se la fortuna è con **voi**, **assisterete** ad un fenomeno indimenticabile.

The 2nd person engagement readers personalize and tailor the holiday to the customers' tastes (examples 29-30), and aspire to bring the world to 'your feet' (*ai vostri piedi*). What is different is the fact that Italian prefers the collective plural, i.e. the 2nd person pronoun *voi/vi* and the 2nd person plural verb form, with verbal actions often expressed in the future tense, e.g. *assisterete*. This could be interpreted as a polite way of suggesting a tourism activity, where perhaps the English would use a directive.

6.2.2 Directives

Directives represent an important engagement device in the tourism corpora. Their pragmatic function is to create the shortest distance with readers. As a basic speech act they are hortative because they encourage action (Searle 1976) and they can be conveyed in different ways. In the English corpora they are mainly signaled by the imperative mood, e.g. *click, discover*; similarly, in Italian: *clicca, scoprite*. There are also occurrences of "indirect directives" (Quirk et al. 1985) often with verbs of recommendation which suggest or give advice, e.g. *we recommend/suggest you, vi consigliamo/suggeriamo* (IT), as well as instances of modals of obligation, such as *must* and *should*, e.g. *You must have adequate travel insurance.*

In the tourism corpora, directives are used as metadiscourse mostly to guide readers around the website for information, recommending the use of services, or enhancing the enjoyment of specific values. On the webpages, the travel companies constantly point to basic booking online actions, e.g. *request your free brochure, look for the blue flag.* Hyland (2005b:184) calls these types of directives "textual acts". He also identifies two other types of directives, those involving "physical acts", e.g. *Drive along Route 66,* and those involving "cognitive acts", which have the function of guiding readers through a line of reasoning, e.g. *Think about it, paradise!* In fact, directives involving cognitive acts can be highly persua-

sive, because of the power they have in raising the tourist/traveler's imagination. Similar acts and metadiscourse functions are found in the ITT corpus, e.g. *fate un giro* ('take a tour'). However, directives occur to a much lesser extent in the ITT corpus (Table 5). Once again, this may be due to the commonly used impersonal pronoun *si* in Italian.

I calculated the number of (direct) directives first by generating a verb list (with POS tagging) for the simple present tense of verbs (base form) in each corpus. I then examined all the concordance lines to check for the imperative mood, and investigated the linguistic environment of the verb to observe the directive acts in context. This type of analysis revealed differences not only between the English and Italian corpora, but also between the American and English data, reflecting different cultural concepts and norms.

Category	UST Total /Rel. Freq.	UKT Total / Rel. Freq.	ITT Total /Rel. Freq.
Directives	432 /0.5% e.g. *Enjoy* *Book* *Discover Drive* *Explore* *Find* *Relax* *Take* *Walk* *Hike*	268/ 0.2% e.g. *Choose* *Enjoy* *Request* *Feel* *Book* *Talk* *Drive* *Fly* *Visit*	78/ 0.08% e.g. *Godetevi* *Continuate* *Guidate* *Fate* *Clicca* *Prenota* *Prendi* *Provate* *Proseguite*

Table 5. (Direct) directives in the three sub-corpora

As we can see from the results in Table 5, directives in the imperative form are highly represented in the English corpus, above all in the UST corpus (432, or 0.5%). The directives differ in type and frequency across all three corpora. The UST corpus employs a lot of physical action verbs in the imperative, e.g. *drive, explore, hike, walk*. Cognitive directives appear to be more common in the UKT corpus, perhaps to strategically create an interactive dialogue to increase the reader's desires, e.g. *Think crystal clear waters, pristine beaches!; learn more about how to make bibimbap!* The following example is representative of the English corpora.

(32) **Make sure you try** the dish with Walnut and Pommegranate sauce! **To die for**. **Go for a walk** along the bridges in Esfahan and **mingle** with the locals. **Don't be afraid** to chat and have tea. UKT

The directives engage readers/customers in creating their perfect holiday: *make sure you…; go for a walk; don't be afraid*. Figure 3 presents sample directives in the imperative in the ITT corpus.

Rwanda , Tanzania , Eritrea , Gibuti , Madagascar)	chiedi	l' amicizia a questo gruppo aperto Centro Africa :
Se vuoi condividere questo viaggio con i tuoi amici	clicca	sull' icona Facebook a lato CLASSIFICAZIONE ,
il tutto al riparo dentro una calda tenda Sami (lavvo) .	Fateci	sapere in anticipo se desiderate guidare la slitta
lasciatevi tranquillamente trasportare,	fidatevi	dei vostri compagni di viaggio . Rispettiamo
lasciandoli per poter guidare liberamente.	Godetevi	la costa artica con le sue fantastiche vedute su Tromsø
esplorare le strette strade con piccoli negozi e pub .	Godetevi	una visita alla Galleria d' Arte Waterside che presenta
vedono l ' ora di raccontarvi il perché .	Visitate	uno dei tanti bars o cafés , vi prometto che non riuscirete
sicurezza e di praticabilità di ogni servizio scelto .	Siate	pure critici ma esprimete le vostre perplessità

Figure 3. Sample imperatives in the ITT corpus

Notice that most of the directives in Italian are in the 2nd person plural; some are reflexive or enclitic, e.g. *fidatevi*, *fateci sapere* ('trust us/let us know'). On the whole, they have the same type of textual or physical function as their equivalents in the English corpora, e.g. *Visitate uno dei tanti bars o cafés* ('Visit one of the many bars and cafes'). However, we can conclude from the quantitative data retrieved that direct directives are not a common engagement marker in the ITT corpus. Italian appears to have alternatives: the use of the impersonal *si*, or a softer approach through hedging devices. For example, compare the following extracts. They are not direct translations in that they present different destinations and situations, but they are representative of the preferred rhetorical style peculiar to the language, culture and genre.

(33) Il cielo è uno spettacolo…, **si può godere** di una stellata meravigliosa.

(34) **Enjoy** sunset views of the inner circle and Colorado River from Plateau Point.UST

The generic impersonal structure (*si può godere*) creates a vague 'you', a polite distancing effect, conveying something which is programmed and routinely done, rather than drawing the reader into some novel experience. The English directive *enjoy* is more concise and follows a rhetoric of explicitness, with information often appearing in snippets alongside multimodal devices.

It is interesting to note that the most frequent directive in the UST corpus is *enjoy!* (101 hits), whereas of the 60 instances in the UKT corpus only 10 act as a directive. The explanation for this perhaps lies in the American culture, in how it views enjoyment and the cultural connotations of *enjoy* (c.f. Bednarek and Bublitz, 2007). Moreover, in the ITT corpus the prima-facie translation of *enjoy - godere** appears only twice as a directive (in the 2nd person plural) *godetevi*, e.g. *Godetevi una visita alla Galleria d'Arte Waterside* ('Enjoy a visit to Waterside Gallery'). One presumes that the ITT corpus uses other ways of conveying pleasure. Other verbs referring in Italian to the concept of *enjoy* are *divertire, gustare, piacere,* but these rarely occur in the imperative in this case study; in some cases, they are used with an inclusive 'we', e.g. *ci gustiamo una meritata birra* ('we enjoy a well-deserved beer').

7. Conclusions

The results of this case study of travel and tourism texts show that English and Italian seem to favour certain interpersonal categories independently of the genre. Despite some similarities, the stance and engagement markers in the tourism promotion websites are different both quantitatively and qualitatively, with idiosyncratic peculiarities in each corpus implying that the three cultures operate differently from one another. This in turn reflects the adoption of different marketing strategies. Quantitatively speaking it can be noted that engagement is more highly represented in English than in Italian, where stance seems to bear the propositional force of the communication, priority is given to the writer's identity and authority, and pre-conceived values seem to be imposed on readers, e.g. in the way *AM* presents their holidays as a philosophy. This attempt to find a niche reflects an 'exclusive' marketing strategy, very different from the Anglo-American one which makes a great effort to accommodate and 'include' a wider audience.

In the American and English corpora, the customers are well aligned and engaged through the use of personal pronouns and directives. This happens less in the Italian corpus, which does not mean that they have a weaker engagement strategy, but rather that persuasion is realized in a different way, that is, either through politeness or through evaluative language. Hedging, as we have seen, generally creates a softer approach, proposing services and activities rather than imposing them on the customer. Italian also shows a preference for traditional writing conventions and rhetorical style, through the use of the impersonal 3rd person structure *si*, despite evidence of a progressive convergence towards a global homogenization of web-marketing.

Differences also came to light between American and British cultural norms. For example, the UST corpus employs more directives, used (informally) to engage the customer and promote services. In addition, differences emerged in the cultural concept of ideologies like enjoyment and freedom, potential areas for future research.

All in all, the findings presented here support the claim that interpersonality in language (regardless of genre and discipline) is conditioned by cultural elements. In this particular case study, the interpersonal strategies are reflected in the marketing style of the culture and language. The issue is whether the travel agency websites would be able to attract someone from another country or culture. This is where the importance of understanding cross-cultural features lies. Features of metadiscourse need to be translated appropriately so as to avoid the loss of nuances. Writing conventions are cultural and linguistically rooted, and when language and culture come into contact, variation at different discoursal and pragmatic levels can be an area of potential difficulty and cross-linguistic misunderstanding. The results presented here may be a valuable source of information for travel agency companies and new marketing styles. Finally, this type of comparative analysis has highlighted the need to redefine intercultural paradigms due to shifting cultural identities in an era of fast globalization.

REFERENCES

Bednarek M. & Bublitz W. (2007) "*Enjoy!*: The (Phraseological) Culture of Having Fun", in *Phraseology and Culture in English*. Ed. by P. Skandera, Berlin/New York, Mouton de Gruyter, pp. 109-135.

Bondi M. (2005) "*Metadiscursive Practices in Academic Discourse: Variation across Genres and Disciplines*", in *Dialogue within Discourse Communities: Metadiscursive perspectives on academic genres*. Ed. by J. Bamford and M. Bondi, Tubingen, *Niemeyer*. pp. 3-30.

Bortoluzzi M. (2000) "Person Markers in Italian and English: A Comparative Discourse Perspective of Undergraduate Writing", *Lingua e Stile*, a. XXXV, 2, pp. 273-302.

Crismore A., Markkanen R. & Steffensen. M. (1993) "Metadiscourse in Persuasive Writing: A study of texts written by American and Finnish students", *Written Communication*, 10. pp. 39-70.

Crystal D. (2004) *The Stories of English*, London, Penguin.

Dann G. (1996) *The Language of Tourism: A sociolinguistic perspective*, Wallingford, U.K., Cab International.

Diani G. (2017) "*The Appeal of Travel Blogs: The image of Italy through American eyes*", in *The Discursive Construal of Trust in the Dynamics of Knowledge Diffusion*. Ed. by R. Salvi and J. Turnbull, Newcastle upon Tyne, Cambridge Scholars Publishing, pp. 46- 61.

Eye for Travel (2011) *Engaging the New Hyper-Interactive Travel Consumer*, http://www.eyefortravel.com/mobile-and-technology/engaging-new-hyper-interactive-travel-consumer (last visited 30 July 2017).

Francesconi S. (2014) *Reading Tourism Texts: A Multimodal Analysis*, Bristol, Channel View Publication.

Gotti M. (2006) "The Language of Tourism as a Specialized Discourse", in *Translating Tourism: Linguistic/Cultural Representations*. Ed. by O. Palusci and S. Francesconi, Trento, Università degli Studi di Trento, pp.15-34.

Greaves C. (2009) *ConcGram 1.0. A Phraseological Search Engine*, Amsterdam/Philadelphia, John Benjamins.

Greaves, C. (2005) *ConcApp*, http://concapp1.software.informer.com/ (last visited 30 July 2017).

Hall E. T. (1989) *Beyond Culture*, New York, Doubleday.

Halliday M.A.K. (1994) *An Introduction to Functional Grammar*, London, Edward.

Hofstede G. (2001) *Culture's Consequences: Comparing Values, Behaviours, Institutions, and Organizations across Nations*, Thousand Oaks, CA, Sage Publications.

Hyland K. (2005a) *Metadiscourse: Exploring Interaction in Writing*, London/New York, Continuum.

Hyland K. (2005b) "Stance and Engagement: A Model of Interaction in Academic Discourse", *Discourse Studies*, 7(2), pp. 172-193.

Hyland K. & Tse, P. (2004) "Metadiscourse in Academic Writing: A Reappraisal", *Applied* Linguistics, 25(2), pp. 156-77.

Ivanič R. (1998) *Writing and Identity: The discoursal construction of identity in academic writing*, Amsterdam/Philadelphia, John Benjamins.

Katan D. (2004) *Translating Cultures*, Manchester, St. Jerome.

Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P. & Suchomel V. (2014) "The Sketch Engine: Ten years on", *Lexicography*, 1(1), pp. 7-36.

Quirk R., Greenbaum S., Leech G. & Svartvik, J. (1985) *A Comprehensive Grammar of the English Language*, London, Longman.

Partington A., Duguid A. & Taylor C. (2013) *Patterns and Meanings in Discourse: Theory and practice in corpus-assisted discourse studies (CADS)*, Amsterdam/Philadelphia, John Benjamins.

Rayson P. (2003) *Wmatrix*, Lancaster University, http://ucrel.lancs.ac.uk/wmatrix3.html (last visited 30 July 2017).

Searle J. (1969) *Speech Acts*, Cambridge, Cambridge University Press.

Sinclair J. (1991) *Corpus Concordance Collocation*, Oxford, Oxford University Press.

Stubbs M. (2001) *Words and Phrases: Corpus Studies of Lexical Semantics*, Oxford, Blackwell.

Suau-Jiménez F. (2017) "Engagement of readers/customers in the discourse of e-tourism promotional genres", unpublished paper, https://www.uv.es/suau/Engagementinetourism.pdf, (last visited 30 September 2017).

Tognini-Bonelli E. & Manca E. (2002) "'Welcoming Children, Pets and Guests: A Problem of Non-equivalence in the Languages of 'Agriturismi' and 'Farmhouse Holidays'", *Textus*, 15(2), pp. 317-334.

Vande Kopple W. J. (1985) "Some Exploratory Discourse on Metadiscourse", *College Composition and Communication*, 36, 1, pp. 82-93.

DOI: 10.13137/978-88-8303-913-3/18482

5. Finding traces of transnational legal communication: cross-referencing in international case law

KATIA PERUZZO
Università di Trieste

Abstract

National and international judicial systems are today in constant interaction. This interaction has been mainly studied by legal scholars, who termed it "transjudicial communication". This communication manifests itself in case law and may also be of interest for linguists. The focus in this chapter is on one of the possible linguistic manifestations of transjudicial communication in the judgments delivered by the European Court of Human Rights (ECtHR). They consist in "external cross-references", i.e. references that point to sources of legislative or judicial law other than the Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and its Protocols or the ECtHR case law. The chapter presents a feasibility study conducted on a small corpus of three English judgments delivered by the Grand Chamber of the ECtHR. The aim of the study is first to develop a language-specific methodology for the semi-automatic extraction of cross-references and then to conduct a qualitative analysis of the extracted cross-references that fall within the category of "external cross-references". These cross-references are analysed to identify both the type of sources they point to and the function they perform. As regards the sources, they are both judicial (case law) and normative (legislation); therefore, an alternative term to refer to the communication that emerges is proposed, namely "transnational legal com-

munication". The three functions cross-references are seen to perform are: (i) description of the factual background and the legal history of the case, (ii) recall of relevant domestic law or other legal provisions, and (iii) provision of a backbone for the legal reasoning and argumentation of the ruling.

Keywords

Transnational legal communication, cross-references, intertextuality, ECtHR judgments, international case law.

1. Introduction

The second half of the 20th century witnessed the emergence of a universal process of globalisation that is today still shaping the economic, cultural, and social landscape. This process has resulted in continuous interaction and intensified co-operation across and above national legal systems and international organisations, with legislative and judicial law-making increasingly taking place on various levels, i.e. in globalized, regional, national or even private legal frameworks. The focus in this chapter is on the visible traces of the interaction between different legal systems in judicial law-making, with particular attention on the judgments delivered by the European Court of Human Rights (ECtHR). Such traces are to be found in what is here termed "external cross-references", i.e. cross-references to sources of law, either legislative or judicial, that do not belong to the supranational legal system stemming from either the Convention for the Protection of Human Rights and Fundamental Freedoms (ECHR) and its Protocols or the case law developed by the ECtHR.

In order to investigate this form of "intertextuality of judgments" (Mattila 2011: 96), a case study on the English version of a corpus of three judgments passed by the Grand Chamber of the ECtHR is presented. The aim is two-fold: first, to propose a methodology for the semi-automatic extraction of cross-references from legal text corpora in English, and second, to analyse the cross-references thus extracted. A legal and a linguistic framework are first provided to situate cross-references within the broad phenomenon of transjudicial communication (Sections 2 and 3). The methodology is then illustrated in Section 4; the main idea behind its development is the possibility to replicate it on a larger scale for future studies. Note that since the methodology illustrated here is based on the analysis of the corpus described in Section 4.1, the list of lexical patterns used for the case study cannot be considered exhaustive and would need to be fine-tuned every time the methodology was applied to a new corpus. Section 5 contains a classification of the cross-references extracted from the corpus into intratextual and intertextual cross-references and further divides the latter type into internal and

external cross-references. Given the emphasis of the study on transjudicial communication, a more detailed discussion is devoted to external cross-references (Section 5.1), in an attempt to provide greater insight into the functions that they perform in Grand Chamber judgments.

2. Transjudicial communication: A legal perspective

Since the second half of the 20th century, the globalisation process has led to a situation in which national legal systems are encouraged, expected and even required to constantly interact and co-operate with international organisations and vice versa. This means that law-making, both legislative and judicial, is no longer confined within rigid national borders. The boundaries within which law-making takes place are increasingly fluid and permeable. In this scenario, the interaction and co-operation between national and supranational/international legal systems are made possible in different ways: directly, when hierarchical relations are established between, for instance, national and supranational legislation, such as in the European Union; or indirectly, as in the development of trade customs or the exchange of good practices. However, both direct and indirect interaction and co-operation call for communication that goes beyond national boundaries and jurisdictions. It follows that this form of legal communication may occur in legislative and judicial law-making. In the latter case, its natural setting is transnational litigation, i.e. litigation that "encompasses domestic and international tribunals" and "includes cases between states (with individuals typically in the wings), between individuals and states, and between individuals across borders" (Slaughter 2003: 192).

Interaction between courts, especially at national level, is not a new phenomenon, as is well recognized by Wagner (2011: 439):

> dialogue between domestic courts has been in practice for a considerable period of time. The shared heritage of common law countries has facilitated and necessitated communication between courts of this tradition since their inception. In the civil law world, domestic courts engaged in dialogue even before the period of codification, which started with the creation of the French Civil Code in the late 18th century and its promulgation in 1804.

Switching the focus on modern times, Wagner (2011: 439) recognizes that

> many of Europe's highest courts have been communicating with one another since the end of World War II. This communication sometimes takes place through hierarchical and somewhat predetermined mechanisms, such as the European Union's Court of Justice and its predecessors or the Council of Europe's European Court of Human Rights. On the other hand, courts also communicate directly with one another. This phenomenon is not confined to these spheres; rather, it has taken on global dimensions.

Several authors have approached the communication between courts – whether national, international or supranational – from different perspectives and named it differently. The first author to use the term "international judicial dialogue" was Andrew L. Strauss, who applied it to "the discussion that could take place between members of different courts and even different judiciaries" (Strauss 1995: 378). Other expressions used to refer to the interaction between different courts and justices include "transnational judicial dialogue" (Burley 1992: 1923; Waters 2005), "transjudicial communication" (Slaughter 1994), "transjudicial dialogue" (Wiener & Liste 2014: 267), "global judicial dialogue" (a term derived from L'Heureux Dube 1998), "transjudicialism" (Bahdi 2002) or "constitutional comparativism" (Alford 2005). These labels may refer to a variety of forms of interaction, ranging from the explicit invocation of either foreign or supranational/international case law to the comparative analysis of foreign legislation found in judicial deliberations and even to direct interactions among judges and networking events for the members of the judiciary.

The background for the present study is Slaughter's typology of "transjudicial communication", broadly defined as the "communication among courts – whether national or supranational – across borders" (Slaughter 1994: 101). This typology is based on a variety of factors. The first factor has to do with the participants in the communication. Slaughter distinguishes between horizontal communications between courts of the same status, vertical communication between national and international/supranational courts, and mixed vertical-horizontal communication, in which international or supranational judicial bodies work as the medium allowing communication among different jurisdictions. The second factor is the "degree of reciprocal engagement manifested by the courts involved" (Slaughter 1994: 112). Slaughter distinguishes between "direct dialogue", in which different courts are actively involved, "monologue", where there is no ongoing conversation but rather a one-sided borrowing of ideas and principles by foreign courts, and "intermediated dialogue", where international/supranational organisations broker communication between national courts. The third factor in Slaughter's classification concerns the functions of transjudicial communication. As regards the interaction between international and supranational law and domestic law, Slaughter (1994: 117) identifies two functions concerning "the role of transjudicial communication in strengthening international regimes", namely the enhancement of the effectiveness of supranational tribunals and the assurance and promotion of the acceptance of reciprocal international obligations. However, the function that has most attracted both interest and criticism from other scholars is "cross-fertilization", i.e. the process by which ideas are disseminated from one legal system to another. According to Slaughter (1994: 118) herself, the process "is likely to be very difficult to track", since it may occur without the recipient court needing to acknowledge the source. A fourth function attributed by Slaughter to transjudicial communication is the enhancement of the persuasiveness, authority or legitimacy of individual judicial deci-

sions, while the final function is what Slaughter (1994: 119) terms "collective deliberation", i.e. the process whereby different actors cooperate in the deliberation on common problems.

Slaughter's seminal work on transjudicial communication has been followed by further studies on cross-fertilization, the increasingly direct interactions between judges, and the active construction of a global judicial community (see Slaughter 2003). However, her standpoint, together with that of advocates of transnational judicial dialogue, has not been immune from criticism. Law and Chang (2001: 523), for instance, mounted a powerful attack on the use of the word "dialogue", considering it "conceptually and factually inaccurate to characterize the manner in which constitutional courts cite and analyze foreign jurisprudence", since "[a]s a conceptual matter, constitutional courts do not cite one another for the purpose of communicating with another, while as an empirical matter, there is little evidence to suggest that one-sided citation of a handful of highly prestigious courts has given way to genuine two-way dialogue". Wiener and Liste, in turn, question one of the functions identified by Slaughter in her typology, i.e. cross-fertilization. Although not denying the existence of "transjudicial dialogue", which they define as "direct interactions between involved legal practitioners including judges, lawyers, or prosecutors", they see cross-fertilization as "a process of legal systems mutually affecting one another", i.e. "an ongoing effect of inter-judicial practice such as cross-referencing" (Wiener and Liste 2014: 267). Based on the results of an empirical test carried out by means of semi-structured interviews with legal practitioners, the two authors argue that the cross-jurisdictional referencing of legal norms and decisions has not yet led to a "global community of courts", as envisaged by Slaughter (2003). Despite such objections, however, it can be confidently stated that "the [international judicial discourse] theory's advocates far outnumber its critics" (Krotoszynski 2006: 1329).

In an increasingly globalized world it is impossible to think of judicial systems working in complete isolation. The existence of some sort of interaction between judges and courts, be it in the form of a monologue or dialogue – to use Slaughter's own words – is undeniable (see, for instance, Waters 2005: 555 ff.). The perspective adopted in this study is, however, linguistic rather than legal. "Transjudicial communication" is here ascribed a narrower meaning than in some of the definitions provided above, since it is conceived as a form of intertextuality, i.e. the linguistic manifestation of the interaction between courts in judicial deliberations. All other forms of interaction, such as face-to-face meetings and networking among legal practitioners, are excluded from the study. Even so, the typology devised by Slaughter, is still judged to be a useful framework for the data to be analysed.

3. Transjudicial communication: A linguistic perspective

Transjudicial communication has so far been analysed exclusively through the lenses of legal studies. However, given the indissoluble connection between law and language, it would be naïve to assume that the language in which transjudicial communication as a legal phenomenon is carried out remains unaffected by the ongoing interaction between different courts. In this study transjudicial communication is intended as the actual linguistic manifestation of the interaction between different courts. Interestingly, Wiener and Liste exemplify interjudicial practice by mentioning a linguistic phenomenon, i.e. cross-referencing (Wiener and Liste 2014: 267), and it is precisely cross-referencing that is taken here as one possible linguistic manifestation of transjudicial communication. The involvement of courts in transjudicial communication may be more or less conscious or intentional, in the sense that there may be interaction between courts without them being aware of it or without them being compelled to give credit to the source. To put it simply, some courts may draw inspiration from other courts but refrain from quoting the source. The focus of this study, however, is on cross-references that signal some sort of interaction between different courts: for this reason, the covert type of transjudicial communication has been ignored in favour of an analysis of the overt type. In particular, the analysis of cross-references is conducted on a corpus of judgments issued by the ECtHR (described in Section 4.1).

In order to place this corpus in the framework of transjudicial communication from a legal perspective, ECtHR judgments have been considered against Slaughter's typology described above. Within this typology, the ECtHR is the natural forum for vertical communication, since it is in charge of resolving disputes originally brought before national courts. Taking the degree of reciprocal engagement into account, the ECtHR is used by Slaughter herself to exemplify "intermediated dialogue", since in her view the ECtHR "effectively brokers communication among national courts" (Slaughter 1994: 113). As for the function, Slaughter sees the ECtHR as an example of "collective deliberation", since it "surveys the common consitutional provisions and practices of the states under its jurisdiction and disseminates national norms from one country to another through the medium of national courts" and, by doing so, "it is effectively collaborating with these national courts on the development of a common European law of human rights" (Slaughter 1994: 120-121).

4. Methodology

As stated above, in this study cross-references are seen as a possible linguistic manifestation, or "trace", of transjudicial communication. Transjudicial communication is also assumed to occur in the international case law produced by

the ECtHR, which constitutes the subject of this case study. Such communication was further assumed to leave traces in ECtHR judgments in the form of cross-references. This assumption was confirmed by skimming a random sample of ECtHR judgments. Once evidence of the presence of cross-references in these judgments was found, the need for their systematic corpus-based analysis emerged. To this end, a small corpus of ECtHR judgments was compiled, which is illustrated in more detail in subsection 4.1, with a brief digression on the structure of ECtHR judgments. The first challenge for the investigation of cross-references was how to identify them in the corpus. The methodology described in subsection 4.2 was developed to semi-automatize their extraction. However, given the peculiarities of the structure of ECtHR judgments, the methodology was only applied to certain sections ("The facts" and "The law") of the texts included in the corpus.

4.1 ECtHR judgments and the corpus

The study presented in this paper was intended as a pilot study to verify the feasibility of a three-step method for the semi-automatic extraction of cross-references from international case law. In Table 1 the main features of the analysed corpus are summarised, attesting to the homogeneity of the corpus components in terms of text type and content.

	Corpus Total word types: 3613 Total word tokens: 50515		
Title	CASE OF HERMI v. ITALY (Application no. 18114/02)	CASE OF MARKOVIC AND OTHERS v. ITALY (Application no. 1398/03)	CASE OF SEJDOVIC v. ITALY (Application no. 56581/00)
Word types:	1822	2297	1794
Word tokens:	14434	21485	14596
Parties to the case:	a Tunisian national; the Italian Republic	ten nationals of Serbia and Montenegro; the Italian Republic	a national of the then Federal Republic of Yugoslavia; the Italian Republic
Year of delivery of Grand Chamber judgment:	2006	2006	2006
Article of the ECHR allegedly violated:	Article 6	Article 6	Article 6

Table 1. Details of the three ECtHR Grand Chamber judgments used for compiling the corpus.

The three judgments were delivered by the Grand Chamber in 2006. In all three cases the applicants were nationals of States that were not contracting States of the European Convention on Human Rights (ECHR); the respondent State was always Italy. Moreover, in all the three cases the Article of the ECHR that was allegedly violated by the Italian Government was Article 6 on the right to a fair trial. As for the language of the judgments, it should be borne in mind that, under Rule 34(1) of the Rules of Court, "[t]he official languages of the Court shall be English and French" and, under Rule 76(1), "[u]nless the Court decides that a judgment shall be given in both official languages, all judgments shall be given either in English or in French". Due to historical and legal reasons, when Italy is the respondent State, the language of proceedings and of the resulting judgment is usually French. However, "there is an assumption that only difficult and complicated cases are submitted to it [the Grand Chamber]" (Garlicki 2009: 394) and, due to their importance, the judgments of this Chamber are published in the official reports of the Court. For these reasons, they are required to be available in both official languages (see Rule 76(2)). For this study the English version of the three judgments was used and the data relating to the number of word tokens and types given in Table 1 refer to that version. The choice of English is also relevant for the method presented in subsection 4.2, since the extraction criteria that were used are language-specific.

The feasibility study presented here was only conducted on a part of the corpus. The reason for this has to do with the particular way the content of the analysed judgments is organized. In the words of Senden (2011: 21),

> [i]n the more than 50 years since its inception the ECHR has developed its own style of reasoning and its own style of judgment. While other international courts, like the CJEU [Court of Justice of the European Union], have had a clear inspirational model, the ECtHR and its judgments have not been modelled on a particular national legal tradition.

The ECtHR has thus developed an elaborate judicial style (see, for instance, Garlicki 2009: 391) which can be observed in both its decisions and judgments. Despite their peculiar style, however, ECtHR judgments can be compared to judgments rendered by any other court. Therefore, when considered from a general perspective, judgments can be seen either as a judicial activity or as documents (see Iacoviello 2001: 'motivazione'). If considered as an activity, they consist in legal reasoning and argumentation on the one hand, and decision-making on the other. When considered as documents, their complex structure comes to the fore: it includes an operative part, which contains the decision taken on the basis of a set of premises, and an argumentative part, which consists in the elicitation of such set of premises. ECtHR judgments are no exception, containing both an operative and an argumentative part. The analysis of the structure of the Grand Chamber judgments included in the corpus reveal a high degree of regularity, with their content being subdivided into the sections described in Table 2.

Activity	Section	Content
INTRODUCTON	Opening section	Information about the Chamber delivering the judgment (here, the Grand Chamber), the parties involved, the application number and year, the date of delivery, the composition of the judging body.
	"Procedure"	Information about the identity of the parties involved, as well as of their representatives and advisers.
ARGUMENTATION; LEGAL REASONING	"The facts"	Divided into two or three subsections identified by a heading: (i) "Circumstances of the case", giving background information on the applicants and reporting, diachronically, the circumstances of the case and the proceedings before domestic courts; (ii) "Relevant domestic law", providing an overview of the relevant national legislation; (iii) "Other relevant provisions", making references to other provisions (e.g. international conventions and recommendations of the Committee of Ministers).
	"The law"	Analysis of the circumstances of the case in the light of the ECHR, divided into subsections, all of which are not always present: (i) The Government's preliminary objection; (ii) Admissibility of the application; (iii) Alleged violation of Article no. of the Convention (iv) Application of Articles 46 and 41 of the Convention.
DECISION-MAKING	Operative part	Ruling of the Court (here, the Grand Chamber).
CLOSING	Closing section	Information about the language(s), mode, place and date of the delivery of the judgment. Signatures of the President and the Registrar.

Table 2. Content organisation in ECtHR Grand Chamber judgments.

Since the purpose of this paper is to shed light on cross-references as indicators of transjudicial communication in ECtHR judgments, only the sections mentioned in Table 2 in which explicit traces of interaction between legal systems can be found were selected and analysed. Therefore, of the three judgments in Table 1, only "The facts" and "The law" sections were subject to the semi-automatic extraction process described below and the subsequent analysis.

4.2 Three-step extraction procedure

In order to extract cross-references from the ECtHR judgments listed in Table 1, a semi-automatic corpus-analysis approach was adopted. For the codification of text segments in the relevant sections ("The facts" and "The law") in each judgment, a qualitative corpus-analysis software programme[1] was used. In particular, the corpus sections were analysed against a list of potential extraction criteria to identify segments containing cross-references as pointers to instances of transjudicial communication and, once those segments were found, they were coded as instances of transjudicial communication. Further possible lexical patterns or other extraction criteria pointing at such communication were recorded. The methodology adopted was thus conceived as a sequential process in which, once the first extraction of cross-references from the corpus was completed, the obtained results could either become keywords or provide suitable hints for a second round of extraction. The second round could then be followed by a further round and the methodology could be repeated until all the cross-references were extracted.

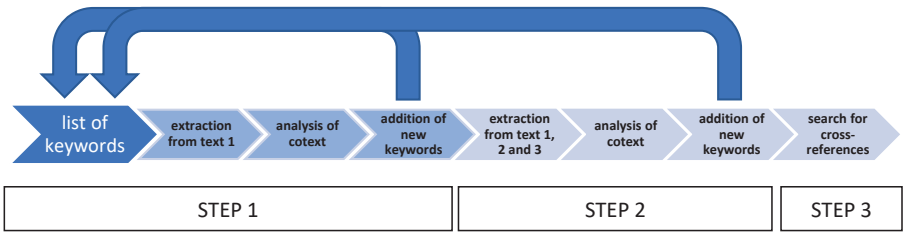


Figure 1. Representation of the three-step extraction procedure.

For the purposes of this feasibility study, the methodology was divided into three steps, each of them entailing a round of extraction of candidate cross-references. The individual steps are outlined below.

1 QDA Miner Lite, version 2.0.

Step 1

The first step in the methodology consisted in the extraction of candidate cross-references based on a list of keywords. The first extraction was performed from only one text in the corpus, i.e. *Markovic and others v. Italy*[2], rather than from the entire corpus. Taking into consideration that the judgments under analysis had been delivered against Italy, the expectation was that cross-references would mainly point to Italian legislation or case law. Based on prior knowledge on ECtHR judgments, such cross-references were supposed to co-occur with a series of words that constitute prototypical cross-referencing patterns in legal language and may thus collocate with possible cross-references. Therefore, an initial list of keywords (see Appendix 1) was compiled to be used for the extraction of cross-references. This list contained three types of keywords. First, verbs were included for which it seemed plausible to expect proximity to cross-references, such as "to provide" and "to state". Second, prepositions and prepositional phrases were added, such as "under" and "in accordance with". Finally, given the assumption that ECtHR judgments contain cross-references to other legislative and judicial sources and that the corpus to be analysed consists of judgments where the Italian Government is a party to the case, terms referring to legal instruments and case law or to their respective sections were included, such as "law", "decree", "judgment", "convention", "article" and "paragraph". In order to illustrate this step, the results of the use of two keywords, i.e. "provides" and "Convention", are discussed below.

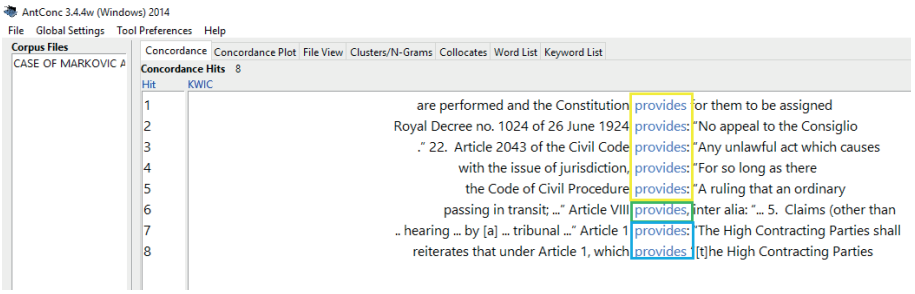


Figure 2. Concordances of "provides" in *Markovic and others v. Italy.*

In Figure 2, all the concordances of "provides" in *Markovic and others v. Italy* are reported. Since the co-text provided by the concordancer[3] is not enough to understand the full cross-citation, the following colour scheme is used in the Figure: yellow for references to Italian legislation, green for international legislation (in this case, the London Convention of 19 June 1951), and blue for the ECHR and its Protocols. The same colour scheme was applied in relation to the first twenty occurrences of "Convention" in the same text (Figure 2). However, since conven-

2 The case was selected due to its higher number of word tokens and types.

3 AntConc, version 3.4.4w.

tions are international rather than national legal instruments, in this case no reference to a national piece of legislation was found.

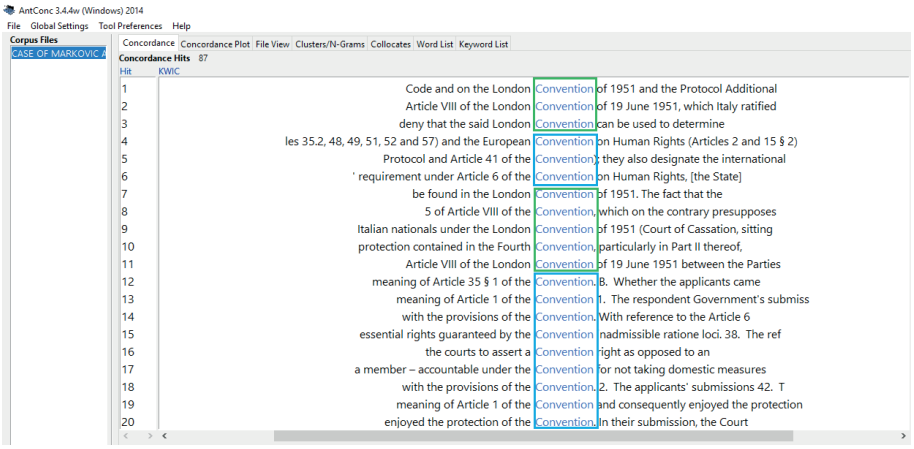


Figure 3. Concordances of "Convention" in *Markovic and others v. Italy.*

The analysis of the concordances obtained by using the keywords in the initial list, such as those in Figures 2 and 3, and of their broader co-text allowed for the identification of further keywords to be added to the initial list. Indeed, in Figure 2, possible keywords such as "Civil Code" and "Code of Civil Procedure" can be noticed, as well as the use of punctuation (inverted commas and semi-colons), which can indicate the presence of a cross-reference nearby in the text. Therefore, the existing list was enriched not only with keywords pertaining to the categories already present in the original list (i.e. verbs, in the active and passive form, prepositions and prepositional phrases, and terms referring to legal instruments and case law), but also with new categories of keywords. The categories added to the list of extraction criteria are nouns (e.g. "rule", "provision"), collocations of 'noun+preposition' (e.g. "scope of", "applicability of") and punctuation (e.g. inverted commas, brackets). The list of extraction criteria obtained at the end of step 1 is available in Appendix 2.

Step 2

In step 2, the list of extraction criteria obtained at the end of step 1 was applied to the entire corpus. After the second round of extraction, other keywords (mainly verbs and 'noun+preposition' collocations) and punctuation marks were added to the list of extraction criteria, which is available in Appendix 3.

Step 3

The final step of the methodology consisted in a third round of semi-automatic extraction of cross-references, carried out by applying the list of extraction cri-

teria from Step 2 to the corpus. The cross-references extracted from the corpus were then subjected to a qualitative analysis to provide a tentative classification based on the sources referred to, as discussed in Section 5.

5. Classification of cross-references extracted from the corpus

The first insight resulting from the analysis of the extracted cross-references is that they refer to a variety of sources. Cross-references in ECtHR judgments can first be distinguished into two broad categories, i.e. intratextual and intertextual references. The former are references that point at information that can be found within the same judgment, such as in the following two examples[4]:

(1) The applicant had been arrested *in flagrante delicto* (see paragraph 12 above) and had at no stage in the proceedings attempted to deny the factual basis of the charges against him. (*Hermi v. Italy*)

(2) The Court therefore considers it unnecessary to indicate any general measures at national level that could be called for in the execution of this judgment. (*Sejdovic v. Italy*)

Intratextual references will not be discussed further, as this study focuses on *trans*judicial communication. The main focus is here on intertextual references, which direct readers to texts other than the one they are reading, as illustrated in these examples:

(3) In view of the circumstances outlined above, and on the basis of the Court's case-law in *Kamasinski v. Austria* (19 December 1989, Series A no. 168) and, conversely, in *Kremzow v. Austria* (21 September 1993, Series A no. 268-B), the Government concluded that the presence of the defendant at the appeal hearing was not required under the Convention. (*Hermi v. Italy*)

(4) In so far as the Government have cited the first of these provisions, the Court reiterates that under Rule 55 of the Rules of Court, any plea of inadmissibility must be raised by the respondent Contracting Party in its written or oral observations on the admissibility of the application (see *K. and T. v. Finland* [GC], no. 25702/94, § 145, ECHR 2001-VII, and *N.C. v. Italy* [GC], no. 24952/94, § 44, ECHR 2002-X). (*Sejdovic v. Italy*)

4 In all the examples, the emphasis (signalled by underlining) is added.

(5) The Court accordingly considers that, where, as in the instant case, an individual has been convicted following proceedings that have entailed breaches of the requirements of Article 6 of the Convention, a retrial or the reopening of the case, if requested, represents in principle an appropriate way of redressing the violation (see the principles set forth in Recommendation No. R (2000) 2 of the Committee of Ministers, as outlined in paragraph 28 above). (*Sejdovic v. Italy*)

(6) In its judgment of 25 November 2004 (no. 48738) in *Soldati*, the Court of Cassation (First Section) observed that leave to appeal out of time could be granted on two conditions: if the accused had not had any knowledge of the proceedings and if he or she had not deliberately avoided taking cognisance of the procedural steps. (*Sejdovic v. Italy*)

(7) They [the applicants] also relied in support of their claim on Article 174 of the Wartime Military Criminal Code and on the London Convention of 1951 and the Protocol Additional to the Geneva Conventions. (*Markovic and others v. Italy*)

While examples 1-7 above contain either intra- or intertextual references, the corpus also reveals the presence of intratextual and intertextual references within the same paragraph, such as in the following example:

(8) Article 599 § 2 of the CCP states that the proceedings are to be adjourned if "a defendant who has expressed a wish to appear has a legitimate reason for not attending" (see paragraph 31 above). (*Hermi v. Italy*)

Intertextual cross-references can be further divided into two subcategories or types, based on the source of the texts being referred to. The first type can be termed 'internal cross-references'. These references point at: (i) texts produced by the ECtHR (ECtHR judgments and decisions, as well as the Rules of Court) (see examples 3 and 4); (ii) texts produced by the Committee of Ministers of the Council of Europe (e.g. Recommendations and Resolutions) (see example 5); and (iii) the European Convention for the Protection of Human Rights and Fundamental Freedoms and its Protocols (see example 5). The second type of intertextual cross-references, instances of which can be found in examples 6 and 7 above, can be termed 'external cross-references'. This subcategory includes a greater variety of sources, such as international legal provisions (e.g. the Geneva Conventions of 12 August 1949 and the London Convention of 19 June 1951), national legal provisions (e.g. the Italian Code of Civil Procedure, Code of Criminal Procedure, Laws, Legislative Decrees and Royal Decrees), national case law (i.e. judgments), and other sources (e.g. notes from public prosecutor's offices). Internal cross-references were not taken into consideration for the discussion of the results presented in Section 5.1 because they lack a transnational character.

5.1 External cross-references: from transjudicial to transnational legal communication

In Section 2.2, transjudicial communication has been described as the linguistic manifestation of the interaction between different courts, and cross-references have been suggested as indicators of transjudicial communication. However, once cross-references are classified using the source as a criterion, it becomes clear that not all cross-references mark the presence of such communication – only external cross-references do so. Focusing on this subtype only, a closer look at examples 4, 5 and 7 further reveals that cross-references in ECtHR judgments are not limited to judicial deliberations, but also concern other – mainly legislative – sources. This means that, when considering external cross-references against the background of ongoing legal interaction in ECtHR judgments, the communication that emerges assumes a variety of forms. When reference is made to judicial sources, the implied interaction is between different courts and the communication can indeed be 'transjudicial'. However, if the invoked source is of a different origin, it would not be fully appropriate to call the interaction 'transjudicial'. Therefore, since the cross-references described in the remaining part of the paper fall within both types, we propose the use of an alternative expression to refer to the underlying interaction, i.e. "transnational legal communication"[5]. This term seems broad enough to include transjudicial cross-references, but also references to other texts that share these common features: they go beyond national boundaries and are legal in their nature, but are not the result of judicial decision-making.

Having established that external cross-references can be considered as indicators of transnational legal communication, the next issue to be addressed is the possible functions such references perform in the argumentative part of ECtHR judgments. The qualitative analysis of the cross-references extracted following the methodology illustrated in Section 4.2 indicates that they are used for three different functions[6]. The first function, which can be identified in both the "The facts" and "The law" sections, is to provide a description of the factual background and legal history of the case, as can be seen in the examples below. Given that in the three cases under analysis the respondent State was Italy and that all the cases to be decided by the ECtHR were already discussed extensively in domestic courts, it should come as no surprise that the references performing this function point at Italian judicial and legislative sources.

5 The term "translational legal communication" is not new and has been used by Wagner (2011). In his view, "transjudicial communication" or "transnational judicial dialogue" refers to "modes of communication between judges that can be traced through the official pronouncements in the publications of the courts", while "transnational legal communication" has a broader meaning, including forms of "open dialogue between domestic judicial institutions" (Wagner 2011: 441). In this study, however, the alternative term is used to go beyond the boundaries of judicial bodies and "legal" is used to refer to both law enforcement and law-making institutions.

6 To illustrate these functions, all the examples provided below are from *Markovic and others v. Italy*.

(9) (from "The facts") On 31 May 2000 the first four applicants brought an action in damages in the Rome District Court under Article 2043 of the Italian Civil Code. The other six applicants applied to be joined to the proceedings on 3 November 2000.

(10) (from "The facts") In a ruling (no. 8157) of 8 February 2002, which was deposited with the registry on 5 June 2002 and conveyed to the applicants on 11 June 2002, the Court of Cassation, sitting as a full court (*Sezioni Unite*), found that the Italian courts had no jurisdiction. It reasoned as follows:

> "...
>
> 2. The claim seeks to impute liability to the Italian State on the basis of an act of war, in particular the conduct of hostilities through aerial warfare. The choice of the means that will be used to conduct hostilities is an act of government. These are acts through which political functions are performed and the Constitution provides for them to be assigned to a constitutional body. The nature of such functions precludes any claim to a protected interest in relation thereto, so that the acts by which they are carried out may or may not have a specific content – see the judgments of the full court of 12 July 1968 (no. 2452), 17 October 1980 (no. 5583) and 8 January 1993 (no. 124). With respect to acts of this type, no court has the power to review the manner in which the function was performed.
> [...]"

(11) (from "The law") The applicants pointed out that the question whether their claim was well-founded or ill-founded under the domestic legal system should have been determined by a court. However, the Court of Cassation's decision had prevented them from asserting in the Italian courts a right recognised by Article 2043 of the Civil Code. Moreover, it was at variance with that court's existing case-law and subsequent decisions. In the applicants' submission, the Court of Cassation's judgment no. 5044 of 11 March 2004 (see paragraph 28 above) showed, firstly, that immunity from jurisdiction could never extend to the criminal law so that civil liability for criminal acts could not, therefore, ever be excluded and, secondly, that rules of international origin protecting fundamental human rights were an integral part of the Italian system and could therefore be relied on in support of a claim in respect of damage caused by criminal acts or by negligence. It followed that anyone alleging a violation of a right guaranteed by such rules was always entitled to the protection of the courts.

The second function is that of recalling relevant domestic law or other legal provisions and was observed in the "The facts" section only. The reason for this may be related to the rigid structure of ECtHR judgments described in Section 4.1: in "The facts", two subsections are specifically devoted to this function, i.e. "Relevant domestic law" and "Other relevant provisions". The following examples

relate to Italian domestic law, but while in example 12 the cross-reference is judicial, in example 13 it is legislative (constitutional).

(12) In a judgment of 10 July 1992 (no. 124/1993), the Court of Cassation, sitting as a full court, established the rule that the courts had no jurisdiction to hear cases against the authorities relating to political acts.

(13) The relevant provisions of the Italian Constitution are as follows:

Article 10 § 1

"The Italian legal system shall comply with the generally recognised rules of international law.
..."
[...]

As regards "Other relevant provisions", the example provided below may be useful to highlight the fact that the position of a cross-reference in a specific section of an ECtHR judgment does not necessarily determine its function. In fact, although the paragraph in Example 14 is found in "Other relevant provisions", a combination of the aforementioned functions can be observed: here, the Protocol Additional to the Geneva Conventions is used to refer to the international provisions invoked by the applicants (first function) and then a specific provision of the same Protocol is explicitly reported (second function).

(14) The applicants relied in the domestic courts on the Protocol Additional of 8 June 1977 to the Geneva Conventions of 12 August 1949, relating to the Protection of Victims of International Armed Conflicts (Protocol I). The Protocol, which Italy ratified through Law no. 672 of 11 December 1985, contains, *inter alia*, the following provisions:

Article 35 – Basic rules

"1. In any armed conflict, the right of the Parties to the conflict to choose methods or means of warfare is not unlimited.

2. It is prohibited to employ weapons, projectiles and material and methods of warfare of a nature to cause superfluous injury or unnecessary suffering.

3. It is prohibited to employ methods or means of warfare which are intended, or may be expected, to cause widespread, long-term and severe damage to the natural environment.

..."

The third function of external cross-references was identified in the "The law" section. Here, cross-references are used to provide either a backbone for the legal reasoning and argumentation of the ruling that is to be found in the operative part of the judgment or the object of such reasoning.

(15) Although it is not its role to express any view on the applicability of the Protocol Additional to the Geneva Conventions (Protocol I) or the London Convention, the Court notes that the Court of Cassation's comments on the international conventions do not appear to contain any errors of interpretation. There are two reasons for this: firstly, the statement that Protocol I regulates relations between States is true; secondly, the applicants relied on paragraph 5 of Article VIII of the London Convention, which concerns acts "... causing damage *in the territory of the receiving State* to third parties ..." (see paragraph 31 above), whereas the applicants' damage was sustained in Serbia, not Italy.

In example 15, the reference to the Court of Cassation's comments is used to refer to the object of discussion and the quotation of the London Convention is used to support the ECtHR's opinion on the correct interpretation by the Court of Cassation. However, at a closer examination, it can be said that in "the applicants relied on paragraph 5 of Article VIII of the London Convention" there is a slight overlap between the first and third function, since this segment is used both to reconstruct the legal history of the case and to support the ECtHR's standpoint.

6. Conclusion and future work

The feasibility study illustrated here represents the first step in a broader study on the interaction between different sources of law, both judicial and legislative, in ECtHR judgments. Drawing from the discussion on "transjudicial communication" as developed by legal scholars, it is argued here that this notion can also be examined from a linguistic perspective. Taking cross-references as possible linguistic indicators of interaction between different sources of law in case law in general and ECtHR judgments in particular, a corpus-driven study was conducted on the English version of three judgments delivered by the Grand Chamber of the ECtHR. To do so, a three-step language-specific methodology was developed for the semi-automatic extraction of cross-references and it was then applied to a small corpus so as to test its feasibility.

Based on the qualitative analysis of the extracted cross-references, a tentative classification was proposed in order to separate the cross-references that point to transjudicial communication from those that cannot say to perform this role. A first distinction was made between intratextual and intertextual cross-references, and the former category was excluded from further analysis because it lacks a

transnational character. Based on the invoked source, intertextual cross-references were subsequently divided into internal and external cross-references. The former were again excluded from the analysis, since they point at texts belonging to the same legal system as the analysed judgments or, in other words, texts produced by the organs of the Council of Europe, which should 'speak the same language' as the ECtHR. External cross-references, on the other hand, were further analysed with the aim of identifying the type of sources they link to. Given that these sources are not limited to judicial textual material, but also include legislative texts of various origins (e.g. national, international), it was concluded that what emerges from the analysis of external cross-references is probably not best described as "transjudicial communication". The term "transnational legal communication" was introduced to depict the interaction between different judicial and legislative sources of law in ECtHR judgments. The qualitative analysis was then extended to determine the functions performed by external cross-references. In the analysed judgments, three functions were identified: (i) description of the factual background and the legal history of the case, (ii) recall of relevant domestic law or other legal provisions, and (iii) provision of a backbone for the legal reasoning and argumentation of the ruling to be found in the operative part of the judgment or of the object of such reasoning.

The methodology developed for this case study was conceived so as to allow its future application to a larger corpus. The main idea for the near future is to build a larger corpus of ECtHR judgments in English and to continue investigating the field of transnational legal communication with the aim of shedding light on the dynamic relationship between different sources of law. Taking external cross-references as a starting point, it would be very interesting to observe the co-text of the cross-references rather than the cross-references only, so as to see what processes the invoked texts undergo when they are included in an ECtHR judgment. For instance, in some cases the text invoked is cited verbatim (as in example 15); in other cases it is quoted in inverted commas, but it is actually the result of a translation process (as in 10 and 13); and in other cases still the co-text is the result of reformulation of a text in another language (as in 12, where the legal principle established by the Italian Court of Cassation is summarised immediately after the reference to the relevant judgment).

REFERENCES

Alford R. P. (2005) "In Search of a Theory for Constitutional Comparativism", *UCLA Law Review*, 52, pp. 639-714.

Bahdi R. (2002) "Globalization of judgment: transjudicialism and the five faces of international law in domestic courts", *The George Washington International Law Review*, 34(3), pp. 555-603.

Burley A.-M. (1992) "Law among Liberal States: Liberal Internationalism and the Act of State Doctrine", *Columbia Law Review*, 92(8), pp. 1907-1996.

Garlicki L. (2009) "Judicial deliberations: the Strasbourg perspective", in *The Legitimacy of Highest Courts' Rulings: Judicial Deliberations and Beyond*. Ed. by N. Huls, M. Adams & J. Bomhoff, The Hague, TMC Asser Press, pp. 389-397.

Krotoszynski R. J. J. (2006) "'I'd like to Teach the World to Sing (In Perfect Harmony)': International Judicial Dialogue and the Muse: Reflections on the Perils and the Promise of International Judicial Dialogue", *Michigan Law Review*, 104(6), pp. 1321-1359.

L'Heureux-Dubé C. (1998) "The Importance of Dialogue: Globalization and International Impact of the Rehnquist Court", *Tulsa Law Journal*, 34(1), pp. 15-26.

Mattila H. E. S. (2011) "Cross-references in court decisions: a study in comparative legal linguistics", *Lapland Law Review*, 1, pp. 96-121.

Slaughter A.-M. (1994) "A Typology of Transjudicial Communication", *University of Richmond Law Review*, 29, pp. 99-137.

Slaughter A.-M. (2003) "A global community of courts", *Harvard International Law Journal*, 44(1), pp. 191-219.

Strauss A. L. (1995) "Beyond national law: the neglected role of the international law of personal jurisdiction in domestic courts", *Harvard International Law Journal*, 36(2), pp. 373-424.

Wagner M. (2011) "Transnational Legal Communication: A Partial Legacy of Supreme Court President Aharon Barak", *Tulsa Law Review*, 47(2), pp. 437-463.

Waters M. A. (2005) "Mediating norms and identity: The role of transnational judicial dialogue in creating and enforcing International law", *Georgetown Law Journal*, 93(2), pp. 487-574.

Wiener A., & Liste, P. (2014) "Lost Without Translation? Cross-Referencing and a New Global Community of Courts", *Indiana Journal of Global Legal Studies*, 21(1), pp. 263-296.

Appendix 1
Initial list of keywords for cross-reference extraction for step 1*

Verbs	Prepositions	Nouns referring to legislation or case law or parts thereof
to allow to provide to state	under in compliance with in accordance with	Article Constitution Convention Decree Judgment Law Paragraph Section Treaty

* The verbs are provided in the infinitive, but the extraction was carried out using inflected forms. For instance, for the verb "to provide", the forms "provides" and provided" were used as keywords.

Appendix 2
List of keywords and other criteria added to the initial list for step 2

Verbs	Phraseology	Nouns referring to legislation or case law or parts thereof	Nouns	Punctuation
to use (can be used) to find (can be found) to be laid down in to *** the rule laid down in to presuppose to be incompatible with (the provisions of) to be irreconcilable with to preclude to be established by to be recognised by/through to cover to preclude to rely on to be guaranteed by to be (not) applicable to find applicable to be set out in to afford to be required by to be defined in to justify to violate to be protected by to be compatible with to (not) apply to be provided in to be afforded by to be provided for in to entitle	within the meaning of with reference to as follows following in the light of for the purposes of in conjunction with on the basis of according to in accordance with as follows inter alia within	Code (Wartime Military Criminal Code, Criminal Code, Code of Civil Procedure, Civil Code) Constitutional Law Protocol	effect(s) of legislation rule provision (no) violation of explanation for scope of applicability of civil or criminal law international law case-law compliance with	“citation” : (…)”

Appendix 3
List of keywords and other criteria added to the initial list for step 3

Verbs	Phraseology	Nouns referring to legislation or case law or parts thereof	Nouns	Punctuation
to have regard to to hold to read (as follows) to be detailed in to be referred to to be indicated in to be contained in to be set forth in to be permitted by to require to be mentioned in to prevent to confer to be required by to be construed to be covered by to be irreconcilable with to be confirmed by to amend to disregard to introduce to enable to be infringed to afford (a right) to (not) specify to interpret to be derived from to be enshrined in to enunciate in to secure to be entitled to apply under to amend to be worded (as follows) to be in issue in	in breach of under the terms of pursuant to in connection with within the scope of by way of interpretation of	(relevant/integral) part(s) of	reference in/to requirement(s) of/ under application of spirit of the wording of a claim on the principle underlying right conferred by an analysis of	(Article XXX) [Article XXX] (see XXX)

DOI: 10.13137/978-88-8303-913-3/20551

6. Notes on investigating the native vs non-native distinction in written academic English

GIUSEPPE PALUMBO
Università di Trieste

Abstract

Texts written in English by non-native speakers can be considered instances of mediated language, where the mediation takes place between a writer's native language and English, seen, respectively, as the "source" and "target" poles. In investigating such texts, the methods of analysis can thus draw on some assumptions and approaches used in translation studies, starting from the idea that in mediated communication the target product always shows traces of interference from features and traits associated with the source material. This chapter reports on an investigation of written academic language in English. The investigation is corpus-based and the texts included in the corpus include research papers in two different academic disciplines written by either native speakers or non-native speakers of English. Initial findings of the investigation are discussed in relation to two specific aspects: part-of-speech distribution and preference for pre- or post-modification in noun groups.

Key words

Academic English, nativeness, language mediation, interference, noun groups.

1. Introduction

Written academic English has been and is still being extensively studied from many different perspectives. The reasons for this continued interest by scholars are manifold. English is the language of choice (some would say of necessity) of today's scientific and scholarly communities, having now established a global dominance that sowed its seeds in the 1950s or perhaps even earlier (Gordin 2015). This global dominance is now being cemented by the decision of many universities worldwide to use English as the medium of instruction in all or some of their degree programmes at both undergraduate and postgraduate level.

As the global language of academia, English is the object of a vast and diverse array of initiatives concerned with language research itself, pedagogy, publishing and language "brokering" (Lillis and Curry 2010). As regards pedagogy, the acquisition of English, whether it be for purposes of comprehension, production or both, is promoted through many instruction programmes, often informed by one or the other approaches or paradigms favoured by scholarly research. In terms of publishing, academic English has become the object of reference works (e.g. OUP 2014) and dedicated publications – some overtly instructional (e.g. Swales and Feak 2013), others with a more popularizing character (e.g. Sword 2012). Such works may be aimed not only at speakers of other languages but also at native English speakers who are still trying to master the specificities of academic expert communication. As regards language brokering (please note that Lillis and Curry (2010) originally used the term "literacy brokering"), this may be seen to include all the activities and practices of those who help authors to shape a manuscript into its final published form, thus including well-established and recognized practices such as copy-editing, translation and proof-reading but also the occasional and more diversified interventions of peers and other actors within an individual researcher's network of contacts.

In this chapter, I start with an overview of recent studies of written academic English, noticing how the focus has over the years increasingly moved on the production of scholarly writers who use English as an additional language to their native language (which in some academic disciplines amounts to say that they *only* use English in their academic publications). I then propose an approach to investigating written academic English using corpus-based methods and employing perspectives that are normally associated with studies in which the dimension of language "contact" or "mediation" is explicitly taken into consideration, in particular translation studies. Finally, I present an exploratory analysis of a corpus comprising research papers in English written by scholars with diverse language backgrounds (including native speakers) and conclude by pointing to ways of refining the analysis and extending it to other relevant aspects.

2. A brief overview of existing research on written academic English

Scholarly research on academic English has seen the emergence of different approaches, most of them sharing a pedagogical preoccupation. These approaches can be grouped according to the research traditions they draw from, the questions they investigate and the methods they employ. A recent critical overview is provided in Tribble (2017), who is essentially concerned with the impact of scholarly research on the effectiveness of academic writing instruction programmes. Tribble (2017: 30-33) distinguishes between genre-informed approaches, approaches based on writing and composition studies, "academic literacies" or "critical" approaches, and approaches based on the notion of English as a Lingua Franca (ELF).[1]

Genre-informed approaches emerge from the essentially UK-based tradition of studies on register and context. Representative, and highly influential, studies following a genre-based approach include Swales (1990) and Hyland (2000). More recently, studies in this tradition have started to apply corpus-based methods, extending their scope to comparisons between spoken and written academic registers (e.g. Biber 2006). The approaches based on composition and writing studies are mainly identified with US scholars and draw on rhetoric-based teaching. An influential study in this tradition (curiously not mentioned in Tribble's overview) is Bazerman (1988). The approaches based on the notion of "academic literacies" (Lea and Street 1998) tend to employ ethnographic methods (as in Lillis and Curry 2010) and are mainly preoccupied with the empowerment of students as writers, adopting at times openly critical stances against established academic conventions (as in Benesch 2001). The last, and more recent, group of approaches identified by Tribble (2017) in his overview is the one incorporating the notion of ELF and referred to by both Tribble and its proponents (e.g. Jenkins 2014) as ELFA, or English as a Lingua Franca Academic. These approaches emphasize the international character of scientific and scholarly communication and focus on how communicative practices and their outcomes are shaped by processes of revision and negotiation between participants.

Tribble's overview of the approaches to research on academic English is not neutral, in the sense that his main aim was to show how ELFA approaches are not as "challenging" or even "paradigm-changing" as some of their proponents (especially Jenkins 2014) would have them be. In particular, Tribble (2017: 33-35) notes the very scarce contribution that ELFA approaches have so far made to pedagogy and warns of the risks of exaggerating the importance and significance of the *native* vs *non-native* dichotomy. In his view, the effectiveness of writing instruction

1 For a definition of English as a Lingua Franca, see Jenkins (2014: 2): "English when it is used as a contact language between people from different first languages (including native English speakers)". Note not all scholars would agree to include native English speakers in this definition, reserving ELF for situations in which communication in English takes place exclusively between non-native speakers of the language.

is enhanced if the focus remains primarily on disciplinary expertise, as already happens in genre-informed approaches.

The relative merits of the approaches identified in Tribble's overview of the research on academic English would deserve a more in-depth discussion than is possible here. As far as the ELFA perspective is concerned, this may well not be – as noted by Tribble – the paradigm-changing approach that some its proponents consider it to be, but it certainly has the merit of having drawn the attention of language scholars and language users alike to the background of linguistic and cultural diversity against which a significant share of today's communication employing English takes place. Higher education in particular can easily be seen as a "prototypical ELF scenario" (Smit 2018: 387; cf. also Palumbo 2015), at least as much as all the other scenarios in which English is the contact language of choice for native speakers of other languages, one example being international institutions such as the UN and the EU.

3. Academic English from a "language mediation" perspective

In consideration of its multilingual background, communication based on ELF could be seen as one instance of "mediated communication"[2] or as a situation of "language contact" (Matras 2009) and thus considered amenable to the research methods employed in language studies that explicitly consider dimensions of mediation or contact. Areas in which these dimensions are explicitly taken into consideration include second language acquisition, pragmatics, linguistic typology, sociolinguistics, contrastive linguistics and translation studies.

On this basis, I would like to propose that texts written in English by non-native speakers can be considered as instances of mediated language, where the mediation takes place between a writer's native language and English, seen, respectively, as the "source" and "target" poles. In investigating such texts, the methods of analysis can thus draw on assumptions and approaches used in translation studies (as also suggested by Cook 2012), starting from the idea that in mediated communication the target product always shows traces of interference from features and traits associated with the source material. This interference, in turn, may also be seen to render translated texts in a given language somewhat different with respect to comparable non-translated texts, i.e. texts in that language that are not the result of translation and belong to the same genre or register.

2 I am using the notion of "mediation" in a more general sense than that sometimes used in translation studies, where it is linked (e.g. in Hatim and Mason 1997: 147) to the ideological significance of the translator's intervention in the transfer process. My idea of translation as "mediated communication" comes very close to Chesterman's (2004: 10-11) characterization of translation as one particular instance of "constrained communication", to be placed alongside reported speech and – especially significant in the present context – communication in a non-native language.

Investigations of translated texts along these lines have tended to focus, alternatively, on the difference between translations and non-translations or on the effects of interference from the source language. Not all distinctive features of translated language are due to interference from the source pole. Some features may be seen to emerge as an effect of the translation process, irrespective of the languages involved (for instance, "explicitation", whose status as a universal trait of translations has however been contested; see Becher 2010). Even interference itself can be observed in relation to two planes: one is that of the individual *texts* to be translated; the other is that of the *systemic* features of the source language. In order to assess the relative influence of these two planes, studies comparing translations and originals should be designed carefully and use appropriate methods and perspectives: textual and corpus-based analysis, in other words, should be complemented with investigations of cognitive processes and social-historical analyses. However, evidence from various studies (see for instance Mauranen 2005 and the other studies mentioned there) suggest that "that it is the source language system that influences the translator, not directly or only elements which the translator is faced with in the text" (Mauranen 2005: 78). Even more importantly for the purposes of the present chapter, Mauranen (2005: 77) warns that "it is unwise to neglect the ubiquity of transfer in bi- or multi-lingual situations".

Interference from the source language is normally considered one of the factors leading to "translationese", or the particular style described as being typical of translated texts. More specifically, interference – as in Toury's (1995) "law of interference" – manifests itself not so much through the presence of errors as through the recurrence of forms and structures that deviate from the norms of the target language or occur with greater than usual frequency in the target language.

Several studies have investigated the particular features of translationese, some of them even questioning its empirical basis. Tirkkonen-Condit (2002), for instance, showed that translationese is not readily recognizable by human subjects on the basis of linguistic features. Other studies have shown that when texts are subject to automatic analysis, there are cues for the distinction between translations and non-translations. Borin and Prütz (2001) investigated the distributional differences of part-of-speech n-grams in translated and non-translated texts and concluded that they "may turn out to be indicative of translation effects in the syntactic domain". Along the same lines, Baroni and Bernardini (2006: 259) found that one of the cues for discriminating between originals and translations was "the distribution of function words and morphosyntactic categories in general".

In the rest of the chapter I'll report on an attempt at framing the investigation of written academic texts in terms of the "translational" perspective briefly sketched above. The key interrelated notions in this perspective are those of transfer and interference. In particular, I will try to establish whether the language background of the writers may be seen to exert any influence on the texts at the level of morpho-syntax, taking a suggestion from some of the studies of

translated language mentioned above (in particular Borin and Prütz 2001) and employing an analytical procedure that was already used in a previous study (Palumbo 2015) with promising results. The following are some initial questions the investigation is intended to provide an answer to:

- Can any differences be observed in the morpho-syntactic profiles of the texts, as observed for instance in terms of part-of-speech (POS) distribution?
- How can the observed differences between texts be related to the author's native languages, or families of languages, in spite of the editorial process the texts have gone through?
- How do the linguistic differences attributable to national identities interact with the writers' disciplinary identities?

In the exploratory case study presented here I will not be able to provide conclusive answers to these questions. The case study is based on a small corpus and only proposes a crude quantitative analysis of some morpho-syntactic and structural features of the texts, based on simple descriptive statistics. The study is essentially intended to set the scene for subsequent, more rigorous analyses, which in turn might require a revision of both the criteria followed for compiling the corpus and the methods of analysis.

4. The corpus

The corpus used for this exploratory study was constructed on the basis of one general aim and according to a set of more detailed criteria following from this aim. The general aim was that of reflecting output in English as produced by two different groups of writers working as academic researchers: native speakers (NSs) of English on the one hand and non-native speakers (NNSs) of English on the other. Defining a "native speaker" for any given language may not be as straightforward as it seems (Davies 2003) and may be next to impossible when specialist writing is concerned (Tribble 2017: 34-35). The general operational criterion taken into account to guarantee that the corpus would reflect the "nativeness vs non-nativeness" distinction was an empirical one, related to a writer's publishing history: writers based in an English-speaking country, affiliated to a research institution there and only publishing articles written in English were taken to be representative of "native" output. Writers based in a non-English speaking country, affiliated to a research institution there and having a history of publishing both in English and in another language (presumably, their "native" language) were taken to be representative of "non-native" writing in English. (A bilingual writer might have ended up being included in this second group, but as defining "bilingualism" may be fraught with as many problems as defining "(non)-nativeness", no attempts were made at identifying possible bilingual authors.)

Not all academic disciplines would easily lend themselves to a search for "non-native" writers as defined above. In several disciplines most, if not all, publishing for research purposes is in English, and therefore authors generally have no history of publishing texts belonging to the same genre (i.e. academic articles) in another language, even when they are based in a non-English speaking country. After some research, two specific disciplines were identified in which authors could be observed to fall within the "native" or the "non-native" categories as described above: linguistics and agricultural economics.

A corpus with two components (or sub-corpora) was then constructed: one for linguistics (LING) and one for agricultural economics (AGRO). Each corpus component includes a total of 120 texts, all of them research articles, distributed as follows (see also Fig. 1):

- 20 texts written by native speakers of English;
- another 100 texts written by non-native speakers of English representing 5 different native-language, or L1, backgrounds, namely Croatian, German, Italian, Polish and Spanish; each of these native language backgrounds is represented with 20 texts.

Each corpus component can thus be seen to be composed of six different L1 sets, English being one of these L1s.

LING	AGRO
120 research articles from linguistics journals, representing the following native languages: English, German, Croatian, Polish, Italian, Spanish (20 texts each). Total size: 581,100 words	120 research articles from agricultural economics, representing the following native languages: English, German, Croatian, Polish, Italian, Spanish (20 texts each). Total size: 571,700 words

Table 1. Corpus composition

All the texts included in the corpus were taken from journals, with a preference for those adopting an open-access policy. All texts were considered in their final, published form. An attempt was also made at selecting both NS and NNS articles from the same journals, but this was not always possible. All texts were cleaned by removing (most) para-textual material. The corpus was compiled and POS-tagged using the Sketch Engine (Kilgarriff et al. 2014).

The six different native-language backgrounds for text authors are the same for the two corpus components. For the non-native English speaking authors, the native languages were chosen for opportunistic reasons (i.e. because it was easier for those languages to find texts written by authors meeting the requirements described above) but also with the aim of representing a variety of language families, so that results from the analysis could also be discussed in language-typological terms. Counting in English itself as one of the native language back-

grounds, both corpus components can thus to be said to represent writers with a Germanic language background (English, German), writers with a Romance language background (Italian, Spanish) and finally writers with a Slavic language background (Croatian, Polish). In the analysis, all the texts written by authors sharing a given native language were treated as separate sets.

5. An exploratory analysis of the corpus

The analysis presented here is a first exploratory attempt at answering the research questions illustrated above. The approach is similar to that already adopted in Palumbo (2015), in which a different corpus was analyzed. The analysis aims to explore differences and similarities among the sets of texts constituting each corpus component, where a set is understood as comprising all the texts written by authors sharing the same native-language, or L1, background. The level of description chosen for the analysis is morpho-syntax, based on the idea that traces of transfer from the writers' native language might emerge at the structural level. Two aspects are investigated in particular: 1) POS distribution in the L1 sets; and 2) the distribution of noun group structures in the L1 sets. The corpus was analyzed using the part-of-speech (POS) tagger available in the Sketch Engine (Kilgarriff et al. 2004). The results given here are no more than basic descriptive statistics; for an informed interpretation of the data, significance tests should be performed, which I reserve for later analyses.

5. 1 Part-of-speech distribution: Native vs non-native speakers

An initial general comparison can be made between POS distribution in the native and non-native texts, i.e. between the set of English L1 writers on the one hand and the other five L1 sets on the other. Figures 1 and 2 present the data for each corpus component (the data for the cumulative NNS set were obtained by averaging out the relative frequencies observed in the individual L1 sets for each POS class).

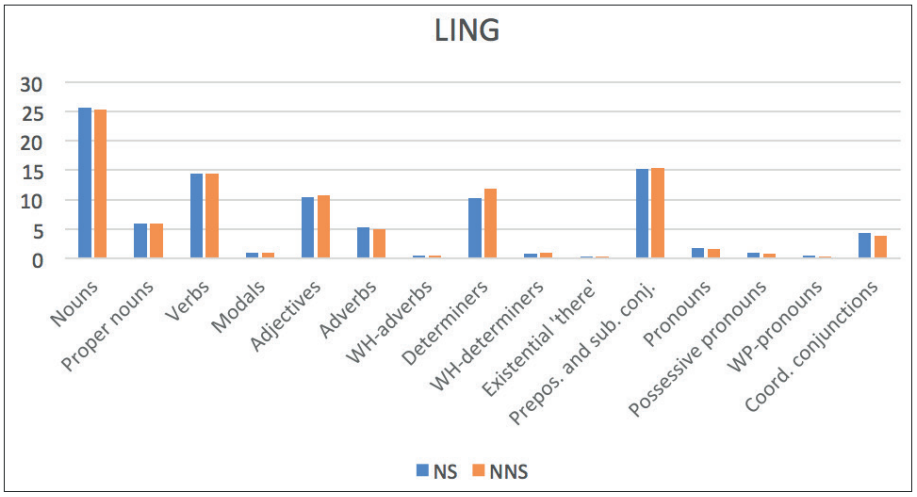


Figure 1. Part-of-speech distribution (in percentage terms) in the LING corpus: NS texts vs NNS texts.

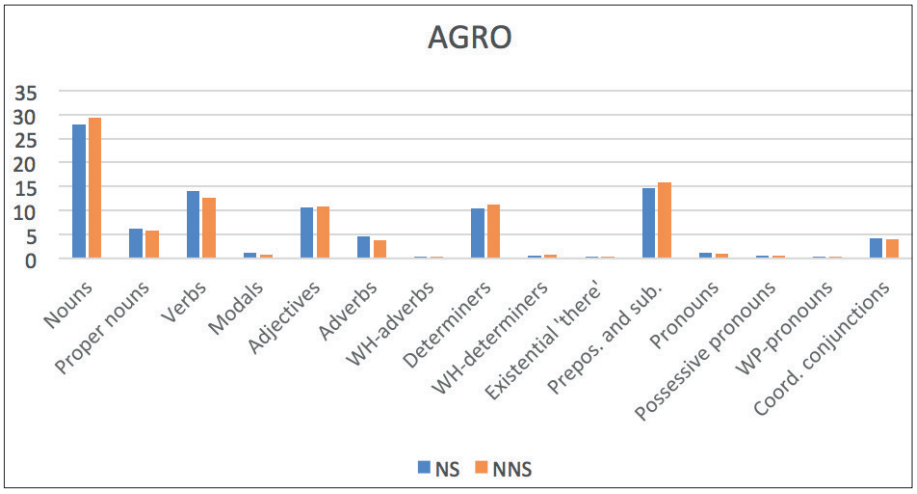


Figure 2. Part-of-speech distribution (in percentage terms) in the AGRO corpus: NS texts vs NNS texts.

What is certainly striking at first glance is the similarity of the distribution across the two disciplines. To take just two examples: the proportion of nouns is roughly double the proportion of verbs in both disciplines; prepositions/subordinating conjunctions and coordinating conjunctions are also distributed very similarly, with the latter amounting to roughly one third of the former in both cases. Consistent adherence to generic conventions of academic writing is likely to be at play here. When the data are considered across language backgrounds,

some differences emerge: determiners, for instance, are used consistently more by NNSs in both disciplines; verbs are used more by the NSs in the AGRO component, whereas in the LING component the percentage for verbs is almost exactly the same (14.45% for NSs and 14.48% for NNSs); finally, adverbs are used slightly more by NSs in both disciplines. Such small differences might of course be due to chance, but it could be worth investigating whether they reflect influences from the L1 of the non-native writers and whether such influences can be distinguished in terms of language families or language-typological differences.

5.2 Noun group structures per L1 sets

Nouns are by far the larger POS category represented in the two corpus components and across all language backgrounds. This marked preference for nominalization seems to be in line with the generic conventions of academic language. The figures so far, however, are only really telling us that all writers in the corpus use far more nouns than any other POS category. It could be worth checking how the nouns are organized, syntactically, into phrases or groups. This could make it possible to observe how pre- and post-modification (e.g. *article selection* vs *the selection of articles*; for a discussion, see Biber et al. 2001: 578-602) are distributed – the assumption being that in some NNS sets the preference for post-modified structures is more marked than in the native sets because of L1 interference. Rough measures in this respect can be obtained by counting occurrences of noun sequences (including proper names) with and without intervening prepositions or apostrophes (as in genitive pre-modified structures). For the purposes of this analysis, the following sequences of elements were searched for and counted in the corpus: 'NOUN + *of/in* (*the*) NOUN'; 'NOUN + NOUN; 'NOUN + NOUN + NOUN'; 'NOUN + *'s* + NOUN'; 'PROPER NOUN + *'s* + NOUN'. To provide comparative measures, the frequency of occurrence of items in each category was turned into a percentage ratio against all nouns present in the set under consideration. The resulting figures are given in Figures 3 and 4, where sequences of two ('NOUN + NOUN') and three nouns ('NOUN + NOUN + NOUN') are conflated into one category.

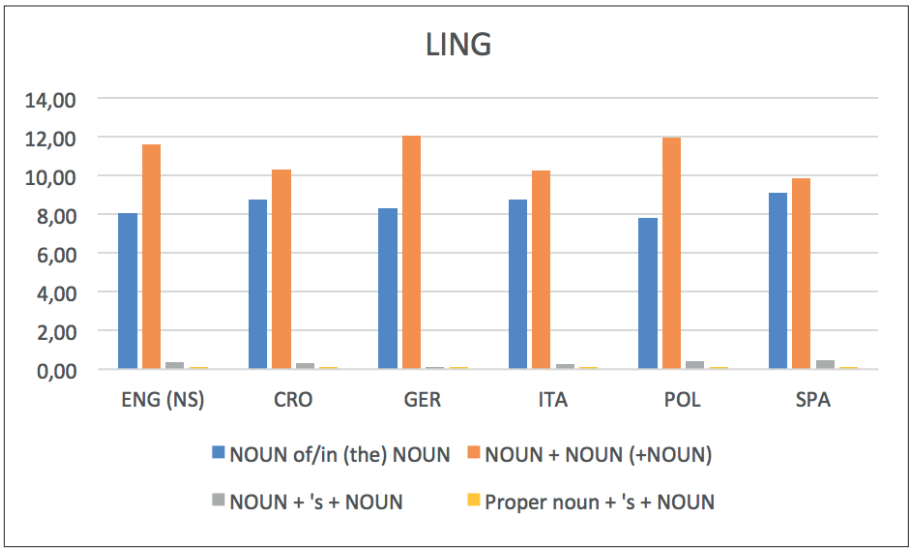


Figure 3. Noun group structures in the LING corpus (per language background of authors).

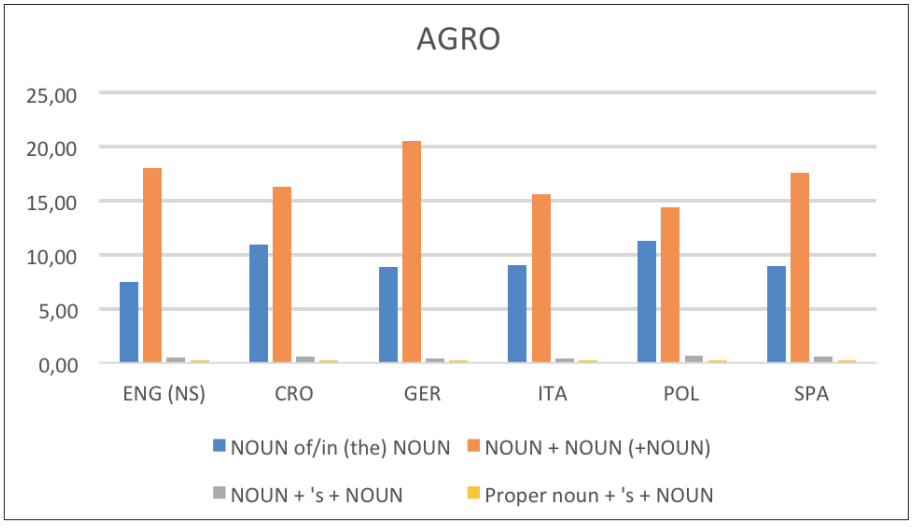


Figure 4. Noun group structures in the AGRO corpus (per language background of authors).

An intuitive expectation might be that writers coming from Romance languages would tend to use pre-modified structures to a lesser degree than writers from other language backgrounds, given that in Romance languages pre-modification is generally not available as a grammatical option for the modification of noun group heads. Conversely, German writers could be expected to make ample use of this syntactic option, which is one that they also have available in their own language. Data from the two corpus components show indeed that German writ-

ers are the ones that use pre-modification the most – even more than native English writers. The data for writers from Romance languages also seem to confirm expectations, at least with respect to writers from Germanic languages: in both AGRO and LING, Italian and Spanish writers use pre-modification less than both English and German writers (in LING Italian and Spanish writers are those that use pre-modification *the least*). For writers with a Slavic language background the data do not exhibit consistent trends. Particularly notable is the diverging behavior of Polish writers in the two corpus components. In the linguistics sub-corpus, Polish academics are among the writers that tend to use pre-modified structures the most, whereas in the agricultural economics sub-corpus Polish writers are the ones that use pre-modification *the least*. Not discounting chance as one possible cause for this difference, other explanations might have to do with the English language competence of writers. As part of their specific disciplinary expertise, Polish linguists might be more alert to the need to consider pre-modification as an option for the noun phrases in their English language writing. At this stage, however, this is no more than speculation.

6. Conclusions

The approach to investigating academic language in English proposed in this chapter tries to build on its nature as a locus of language contact. To reuse a metaphor that has been proposed (by Gellerstam 2006) in relation to translation, the approach is based on the idea that the native language of academics writing in English could leave "fingerprints" in their texts. These fingerprints, in turn, can be identified by comparing non-native writing to native writing. To use still another notion from translation studies, non-native academic writing can be observed in terms of its "textual fit" (Biel 2014) with respect to native writing. The idea of comparing and contrasting native and non-native writing in English in the context of academic language is not new, but to date it has mostly been used for investigating the language of learners (Nesselhauf 2005) or specific aspects of language, such as formulaic sequences (O'Donnel et al. 2013) and lexical bundles (Salazar 2014; Esfandiari and Barbary 2017).

The exploratory investigation proposed in this chapter is mainly intended as a pointer to aspects that appear to be amenable to further study and could take cues from analyses of translated language. Part-of-speech distribution, for instance, would benefit from further analysis based on an approach similar to Borin and Prütz's (2001), where the parts of speech are treated as sequences of n-grams and observed in terms of their positional differences. The texts in the corpus could also be analyzed using authorship attribution methods, such as the various measures of "intertextual distance" that are used to assess the similarity (and dissimilarity) between texts. (For an application of these methods to the analysis of translated texts, see the two chapters by Ondelli and Nadalutti in this volume.)

As far as specific language and textual traits are concerned, future analyses of the corpus might include all the aspects that are traditionally considered in comparisons and contrasts between translated and non-translated language: the distribution verb tenses and passive constructions; the use of encapsulating anaphoric references; the use of the definite article; the use of pronouns, as opposed to equivalent impersonal constructions; the frequency and distribution of connectives and other cohesive devices; the presence and function of questions; the use of lexical repetition, as opposed to the preference for synonyms; variation in sentence length; lexical preferences according to word etymology ('Latinate vs Anglo-Saxon vocabulary'). Any finding on such aspects could also be interpreted in terms of the general attitude displayed by writers with respect to readers, characterized, for instance, in terms of Hinds' (1987) distinction between "writer responsibility" and "reader responsibility" (see also MacKenzie 2015).

Some scholars (e.g., Römer 2009; Tribble 2017) have argued that the distinction between nativeness and non-nativeness in English academic writing is not as relevant as it would appear to be and that disciplinary expertise may play a much more significant role than native language background. Others have pointed out that the "native/non-native distinction remains a useful heuristic" (Gnutzmann and Rabe 2014: 39), while recognizing that the language demands made on researchers may vary significantly across disciplines. Adopting approaches normally associated with the study of translations does not imply, per se, that prominence should be given to the nativeness factor; however; it could contribute a fresh perspective on the study of how, in academic English, language and communicative norms are re-negotiated in an international scenario.

REFERENCES

Baroni M. and S. Bernardini (2006) "A new approach to the study of translationese: Machine-learning the difference between original and translated text", *Literary and Linguistic Computing*, 21(3), pp. 259-274.

Bazerman C. (1988) *Shaping Written Knowledge. The Genre and Activity of the Experimental Article in Science*, Madison, The University of Wisconsin Press.

Becher V. (2010) "Abandoning the notion of 'translation-inherent' explicitation: Against a dogma of translation studies", *Across Languages and Cultures*, 11(1), pp. 1-28.

Benesch S. (2001) *Critical English for Academic Purposes. Theory, Politics and Practice*, London/New York: Routledge.

Biber D. (2006) *University Language. A Corpus-Based Study of Spoken and Written Registers*, Amsterdam/ Philadelphia, John Benjamins.

Biel Ł. (2014) *Lost in the Eurofog: The Textual Fit of Translated Law*, Frankfurt am Main, Peter Lang.

Borin L. and K. Prütz (2001) "Through a glass darkly: part of speech distribution in original and translated text". In *Computational Linguistics in the Netherlands 2000*. Ed. by W. Daelemans, K. Sima'an, J. Veenstra and J. Zavrel, Amsterdam, Rodopi, pp. 30-44.

Chesterman A. (2004) "Hypotheses about translation universals". In *Claims, Changes and Challenges in Translation Studies*. Ed. by G. Hansen, K. Malmkjær and D. Gile, Amsterdam/Philadelphia, John Benjamins, pp. 1–13.

Cook G. (2012) "ELF and translation and interpreting: Common ground, common interest, common cause", in *Journal of English as a Lingua Franca*, 1(2), pp. 241-262.

Davies A. (2003) *The Native Speaker. Myth and Reality*, 2nd ed., Clevedon, Multilingual Matters.

Esfandiari R. and Barbary F. (2017) "A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles", *Journal of English for Academic Purposes*, 29, pp. 21–42.

Gellerstam M. (2005) "Fingerprints in Translation", in *In and Out of English: For Better, For Worse?* Ed. by G. Anderman and M. Rogers, Clevedon, Multilingual Matters, pp. 201-2013.

Gordin M.D. (2015) *Scientific Babel. The Language of Science from the Fall of Latin to the Rise of English*, Chicago, The University of Chicago Press.

Gnutzmann C. and Rabe F. (2014) "'Theoretical subtleties' or 'text modules'? German researchers' language demands and attitudes across disciplinary cultures", *Journal of English for Academic Purposes*, 13, pp. 31-40.

Hatim B. and Mason I. (1997) *The Translator as Communicator*, London/New York: Routledge.

Hinds J. (1987) "Reader versus writer responsibility: A new typology". In *Writing across Languages: Analysis of L2 text*. Ed. by U. Connor and R. B. Kaplan, Reading, Mass., Addison-Wesley, pp. 141-152.

Hyland K. (2000) *Disciplinary Discourses. Social Interactions in Academic Writing*, Harlow, Longman.

Lea M. R. and Street B. V. (1998) "Student writing in higher education: An academic literacies approach", *Studies in Higher Education*, 23(2), pp. 157-172.

Lillis T. and Curry M.J. (2010) *Academic Writing in a Global Context. The Politics and Practices of Publishing in English*, London/New York, Routledge.

Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P. & Suchomel V. (2014) "The Sketch Engine: Ten years on", *Lexicography*, 1(1), pp. 7-36.

Matras Y. (2009) *Language Contact*, Cambridge, Cambridge University Press.

Mauranen A. (2005) "Contrasting languages and varieties with translational corpora", *Languages in Contrast*, 5(1), pp. 73-92.

MacKenzie I. (2015) "Rethinking Reader and Writer Responsibility in Academic English". *Applied Linguistics Review*, 6(1), pp. 1-21.

Nesselhauf N. (2005) *Collocations in a Learner Corpus*, Amsterdam/Philadelphia, John Benjamins.

Palumbo G. (2015) "Studying ELF Institutional Web-based Communication by Universities: Comparison and Contrast with English Native Texts". In *English for Academic Purposes: Approaches and Implications*. Ed. by P. Thompson and G. Diani, Newcastle upon Tyne, Cambridge Scholars Publishing, pp. 245-264.

O'Donnell M. B., Römer U. and Ellis N. C. (2013) "The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm", *International Journal of Corpus Linguistics*, 18(1), pp. 83-108.

OUP (2014) *Oxford Learner's Dictionary of Academic English*, Oxford, Oxford University Press.

Römer U. (2009) "English in Academia: Does Nativeness Matter?", *Anglistik: International Journal of English Studies*, 20(2), pp. 89-100.

Salazar D. (2014) *Lexical Bundles in Native and Non-Native Scientific Writing: Applying A Corpus-Based Study to Language Teaching*, Amsterdam/Philadelphia, John Benjamins.

Smit U. (2018) "Beyond monolingualism in higher education. A language policy account". In *The Routledge Handbook of English as a Lingua Franca*. Ed. by J. Jenkins, W. Baker and M. Dewey, London/New York, Routledge, pp. 387-399.

Swales J. M. (1990) *Genre Analysis*, Cambridge, Cambridge University Press.

Swales J. M. and Feak. C.B. (2013) *Academic Writing for Graduate Students. Essential Tasks and Skills*, 3rd ed., Ann Arbor, Mich., Michigan University Press.

Sword H. (2012) *Stylish Academic Writing*, Cambridge, Mass., Harvard University Press.

Tirkkonen-Condit S. (2002) "Translationese - a myth or an empirical fact? A study into the linguistic identifiability of translated language", *Target*, 14(2), pp. 207-220.

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

Tribble C. (2017) "ELFA vs. Genre: A new paradigm war in EAP writing instruction?", *Journal of English for Academic Purposes*, 25, pp. 30-44.