Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli

Stefania Gandin

Introduzione

La metodologia dei corpora applicata all'analisi della traduzione rappresenta un fenomeno di ricerca ancora piuttosto recente, che può offrire tuttavia significativi contributi sia in termini di analisi linguistiche avanzate, sia come strumento didattico per l'insegnamento e l'acquisizione delle lingue straniere, in particolare per l'apprendimento dei linguaggi specialistici. L'utilizzo dei corpora nello studio della traduzione consente non solo di individuare importanti aspetti sintattici, grammaticali e terminologici di due o più lingue di riferimento attraverso la comparazione diretta di testi originali (source texts, o ST) e testi tradotti (target texts, o TT), ma rappresenta inoltre un valido strumento di supporto e di apprendimento per traduttori professionisti, studenti e altri soggetti (madrelingua e non) che desiderano acquisire un'adeguata conoscenza terminologica, stilistica e concettuale di determinati linguaggi e generi linguistici. Nelle successive sezioni verranno illustrate le possibili aree di applicazione della metodologia dei corpora per lo studio della traduzione, con particolare riferimento alla traduzione dei linguaggi specialistici: verranno fornite innanzitutto alcune definizioni per delineare il contesto teorico-pratico della linguistica computazionale applicata alla traduzione e,successivamente, verranno descritti gli aspetti più importanti relativi alla compilazione di corpora paralleli e all'analisi dei dati linguistici in essi contenuti. In conclusione verranno suggeriti ulteriori percorsi di sviluppo per l'applicazione della metodologia dei corpora in altri ambiti di ricerca.

1. *La metodologia dei corpora e la traduzione: alcune definizioni*

Come ben noto, un *corpus* rappresenta un insieme di testi in formato elettronico che possono essere 'letti', suddivisi e analizzati attraverso appositi software, al fine di individuare e classificare dati linguistici rilevanti per l'analisi delle caratteristiche specifiche di determinati generi linguistici.

Le prime applicazioni della linguistica dei corpora furono dedicate soprattutto ad analisi monolinguistiche o ad analisi comparative (fra due o più lingue) finalizzate all'identificazione delle somiglianze e/o divergenze fra testi originali di generi linguistici simili. L'applicazione della *corpus analysis* allo studio di testi tradotti per identificarne caratteristiche linguistiche e strategie traduttive è un fenomeno di ricerca recente della linguistica dei corpora (Baker 1996, Laviosa 1998 e 2002, Olohan e Baker 2000) e ancora poco esplorato, soprattutto per quanto riguarda i linguaggi specialistici. In molti progetti di ricerca sono stati costruiti diversi *corpora paralleli* e *comparabili* che contengono testi originali in due (o più) lingue e corrispettive traduzioni.

Per *corpus parallelo* si intende un corpus formato da una serie di testi originali in una determinata *lingua di origine* (definita tecnicamente anche *Source Language*, o SL) e dalle relative traduzioni in un altra *lingua* (o altre lingue) *di destinazione* (*Target Language*, o TL) (Olohan 2004: 24-25). Esistono varie combinazioni e modelli da utilizzare per la creazione di corpora paralleli, come ad esempio:

- il modello uni-direzionale, che contiene testi in una sola lingua d'origine (es. inglese) e relative traduzioni in una sola lingua di destinazione (es. italiano);
- il 'modello bi-direzionale' (Johansson 2003: 138), che contiene testi originali in due lingue (es. inglese e italiano) e corrispettive traduzioni nelle stesse due lingue (i.e. dall'inglese all'italiano e dall'italiano all'inglese);
- il 'modello a stella' (ibid.: 140), costituito da testi originali in una sola lingua (es. inglese) e traduzioni in due o più lingue (es. italiano, francese, tedesco, portoghese etc.);
- il 'modello a diamante' (ibid.: 139), che prevede l'inserimento di testi originali in tre (o più) lingue (es. testi originali in inglese, italiano e francese) e relative traduzioni combinate (es. traduzioni dall'inglese all'italiano e francese, dall'italiano all'inglese e francese e dal francese all'inglese e italiano). Si tratta di un modello molto complesso sia per la difficoltà materiale nel reperire combinazioni di testi e traduzioni in un numero così elevato di lingue, sia per il tipo di analisi linguistiche che tale modello permette di eseguire (analisi comparative di testi originali appartenenti allo stesso genere; analisi di testi originali e relative traduzioni in una o più lingue; analisi di testi originali e traduzioni nella stessa lingua; analisi comparative di traduzioni appartenenti allo stesso genere).

Un *corpus comparabile* invece è formato da una serie di soli testi originali o sole traduzioni appartenenti agli stessi generi testuali, scritti in una determinata lingua (corpora monolingui) o in due o più lingue (corpora bilingui o plurilingui) (Olohan 2004: 35).

I corpora paralleli bi-direzionali, a stella e a diamante rappresentano le risorse di ricerca più ricche e complete, in quanto essi possiedono già una dimensione di analisi comparabile che è possibile ricavare estraendo dal loro interno un sub-corpus di soli testi originali o sole traduzioni, permettendo di effettuare potenziali analisi linguistiche di tipo appunto comparativo sui generi testuali rappresentati nel corpus, oltre a quelle relative allo studio dei fenomeni linguistici inerenti le traduzioni.

Fra i numerosi e più importanti corpora paralleli finora progettati, possiamo ricordare:

- il *CEXI*, un corpus parallelo che contiene testi originali in italiano e inglese e corrispondenti traduzioni. È stato realizzato presso la Scuola per interpreti e traduttori di Forlì e contiene una collezione di testi di fiction suddivisi in due sub-corpora: fiction per adulti e fiction per bambini (Zanettin 2000);
- l'*ENPC* (*English Norwegian Parallel Corpus*), un corpus parallelo bi-direzionale di testi originali (con generi di fiction e non) e delle loro rispettive traduzioni dall'inglese al norvegese e viceversa, che recentemente è stato integrato dall'*OMC* (*Oslo Multilingual Corpus*), una raccolta di corpora paralleli di testi originali e traduzioni (prevalentemente letteratura fiction) in diverse combinazioni linguistiche (norvegese, inglese, francese, tedesco, olandese e portoghese), che include inoltre due ulteriori corpora paralleli di testi e traduzioni (generi fiction e non) in inglese e svedese e viceversa (sito web dell'OMC);
- *COMPARA,* un corpus parallelo bi-direzionale aperto (in quanto il progetto di ricerca prevede un continuo apporto di testi per poter effettuare analisi linguistiche aggiornate anche da un punto di vista diacronico). COMPARA è costituito da una raccolta di testi originali di generi di fiction e relative traduzioni dal portoghese all'inglese e viceversa, e consente di analizzare, oltre alle tipiche possibilità di ricerca dei corpora, anche elementi quali le note del traduttore, i forestierismi, i titoli, le frasi enfatizzate etc. (sito web di COMPARA);

- il *TRANSEARCH*, un corpus parallelo aperto di frasi tratte dai dibattiti del parlamento canadese tradotte dall'inglese al francese e viceversa, suddiviso in quattro database: lo *House of Commons Hansard* (che contiene le trascrizioni dei dibattiti della camera dei deputati canadese dal 1986 e le corrispondenti traduzioni in francese); il *Senate Hansard* (che contiene le trascrizione dei dibattiti del senato canadese dal 1996 e le corrispondenti traduzioni dall'inglese al francese e vice versa); il *Canadian Courts rulings* (una raccolta di documenti relativi alle decisioni della corte suprema canadese dal 1986). Il TRANSEARCH contiene inoltre una sezione formata da documenti originali dell'*ILO* (*International Labour Organization*) e relative traduzioni dall'inglese allo spagnolo e viceversa;
- l'*MLCC*, un corpus multilingue composto da una sezione comparabile di articoli tratti da riviste finanziarie in sei lingue (francese, inglese, italiano, olandese, spagnolo e tedesco) e da un corpus parallelo a stella di testi in inglese e francese tradotti in nove lingue europee (danese, francese, inglese, greco, italiano, olandese, portoghese, spagnolo e tedesco) forniti dalla Commissione Europea (sito web dell'*ELRA - European Language Resources Association).*

Per quanto riguarda invece i corpora comparabili dedicati allo studio della traduzione, possiamo ricordarne uno dei più importanti, ovvero:

- il *TEC*, un corpus monolingue di traduzioni in inglese, creato presso il Centre for Translation and Intercultural Studies (CTIS) della University of Manchester a supporto di due grandi progetti di ricerca volti a determinare le caratteristiche universali della traduzione[1] e la ricerca sulle caratteristiche stilistiche dei traduttori imputabili proprio alla loro attività di traduzione (sito web del CTIS). Il TEC contiene traduzioni in inglese di testi scritti originalmente in numerose lingue europee e non-europee [arabo, cinese, ebraico, francese, gallese, italiano, polacco, portoghese (europeo e variante brasiliana), spagnolo (europeo e varianti dell'America centro-meridionale), tamil, tailandese e tedesco]. I generi testuali rappresentati includono testi di fiction (più dell'80%), biografie (circa il 15%), articoli di giornali e riviste aeree.

Nonostante esista un elevato numero di corpora dedicati allo studio delle traduzioni, come quelli appena descritti, i progetti di ricerca destinati allo

studio dei linguaggi specialistici in traduzione sono ancora poco numerosi. Infatti, i generi linguistici rappresentati in questi grandi progetti hanno incluso sino ad oggi solo, o principalmente, testi di fiction, come ad esempio nel caso del CEXI, del COMPARA e del TEC. I corpora paralleli dedicati ai linguaggi specialistici sono ancora pochi (es. il TRANSEARCH o l'MLCC) o di dimensioni molto limitate (es. l'ENPC), e non consentono perciò di effettuare ricerche approfondite per identificare le strategie traduttive più efficaci, quelle meno efficaci, gli aspetti della traduzione più difficili da tradurre e le possibili soluzioni.

Ma cosa si intende esattamente per linguaggi specialistici? Un *linguaggio specialistico*, o *LSP* (Language for Special Purposes) viene definito come una lingua utilizzata "to discuss specialized fields of knowledge" (Bowker e Pearson 2002: 25). Diversamente dalla *LGP* (Language for General Purposes) utilizzata "to talk about ordinary things in a variety of common situations" (ibid.), le LSP possono includere temi che riguardano qualsiasi attività professionale o persino gli hobby, ma sono caratterizzate da un elevato grado di specificità determinato da:

- presenza di un vocabolario specializzato e usato solo (o prevalentemente) in tale settore;
- collocazioni e aspetti stilistici e grammaticali usati solo in tale contesto specialistico.

Fra le LSP rientrano, ad esempio, il linguaggio giuridico, il linguaggio medico-scientifico, il linguaggio turistico etc.

La necessità di utilizzare correttamente una LSP può interessare diversi soggetti e situazioni, come persone semi-esperte o non-esperte che intendono comunicare con esperti di un determinato settore, studenti (madrelingua e non) che hanno necessità di acquisire le necessarie conoscenze per comunicare attraverso linguaggi specialistici e traduttori, la cui professione richiede una competenza linguistica tale da riconoscere le caratteristiche specifiche di una LSP nella lingua d'origine e di conoscere e riportare caratteristiche equivalenti nella lingua di destinazione. Rispetto ad altre risorse linguistiche quali dizionari (specializzati e non), testi stampati o la consultazione diretta di professionisti di un determinato settore, la linguistica dei corpora offre notevoli vantaggi per l'apprendimento, la comprensione e l'analisi dei linguaggi specialistici. Innanzitutto il formato elettronico dei corpora consente di:

- avere a disposizione dei dati di analisi più numerosi;

- un aggiornamento più rapido e *up-to date* rispetto alle risorse cartacee;
- una ricerca dei dati in maniera più facile e veloce.

Inoltre, i testi inseriti nei corpora rappresentano dei testi "autentici" che forniscono uno spaccato reale della lingua in uso corrente di un determinato settore e consentono, perciò, di utilizzare i corpora come valide "guide di stile" per la compilazione di testi e traduzioni specialistici in una determinata lingua di riferimento, nonché di effettuare analisi linguistiche con dati di ricerca sempre aggiornati e aggiornabili.

Per illustrare le possibili implicazioni didattiche e di ricerca dei corpora paralleli, verranno illustrati qui di seguito alcuni esempi tratti da un corpus parallelo appositamente creato ai fini di questa analisi. Il corpus, che chiameremo DSMPE, è stato compilato selezionando una serie di Dichiarazioni Scritte dei membri del Parlamento Europeo riferite al periodo 2004-2009 in inglese (ST) e in italiano (TT). Il corpus è stato creato utilizzando il programma Multiconcord®: questo software è in grado di analizzare dati linguistici in undici lingue europee (danese, finlandese, francese, greco, inglese, italiano, olandese, portoghese, spagnolo, svedese e tedesco); consente inoltre di scegliere la coppia di lingue e i testi sui cui effettuare la ricerca di singole parole, frasi intere etc. nella SL di riferimento e di mostrare, parallelamente, i risultati delle corrispondenti traduzioni nella TL selezionata.

Le seguenti figure illustrano i risultati di una breve e casuale selezione di termini ed espressioni tipici del linguaggio giuridico inglese e delle corrispondenti traduzioni in italiano,quali:

- la preposizione '*whereas*', tradotta nei testi della TL, tramite un processo di trasposizione[2], con il verbo al gerundio '*considerando*';

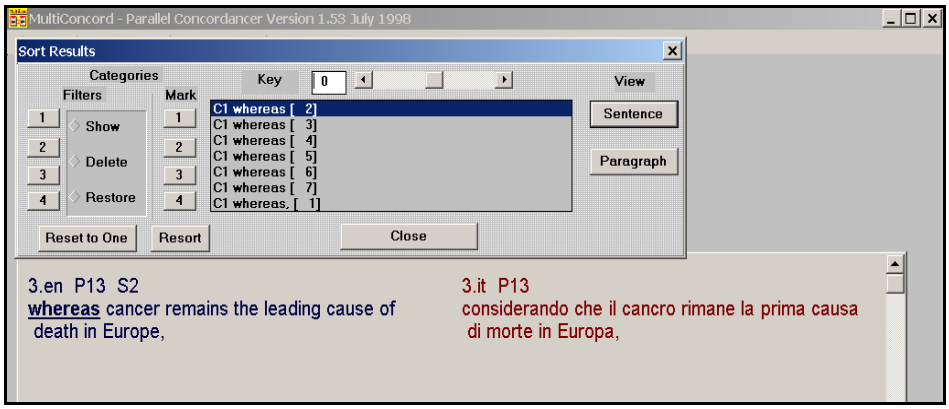

Figura A: 1° risultato di ricerca della preposizione '*whereas*' nel corpus DSMPE;

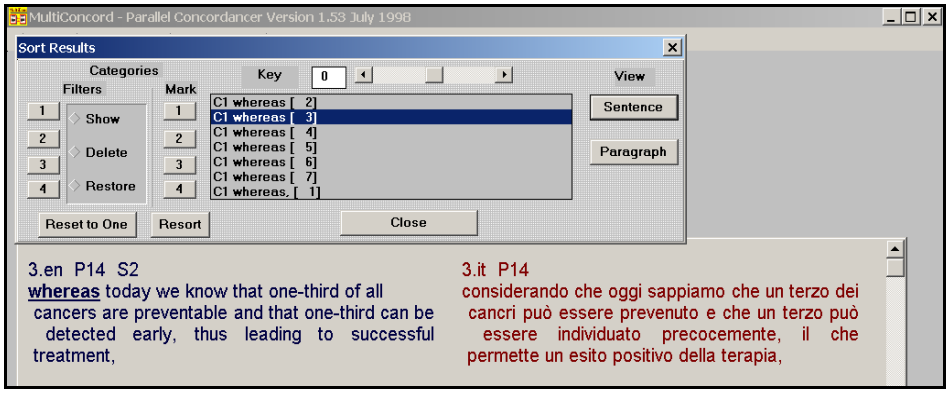

Figura B: 2° risultato di ricerca della preposizione '*whereas*' nel corpus DSMPE;

- l'espressione '*having regard*', tradotta in italiano tramite gerundio (*considerando*) o tramite trasposizione verso un participio passato (*visto*);

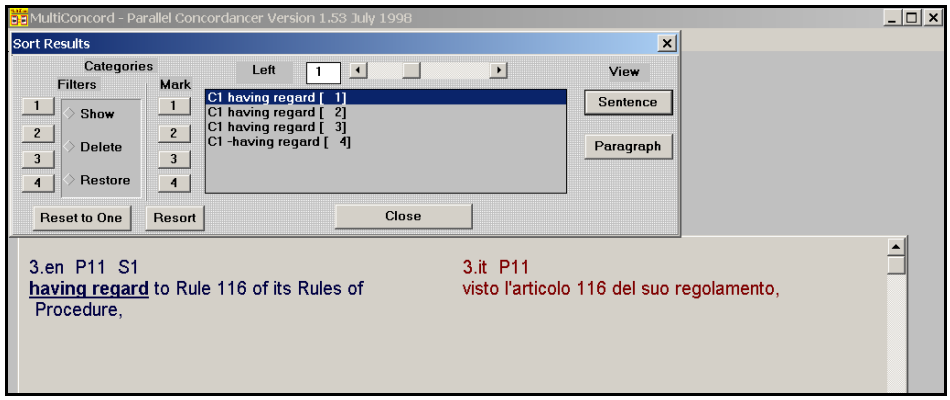

Figura C: 1° risultato di ricerca dell'espressione '*having* regard' nel corpus DSMPE;

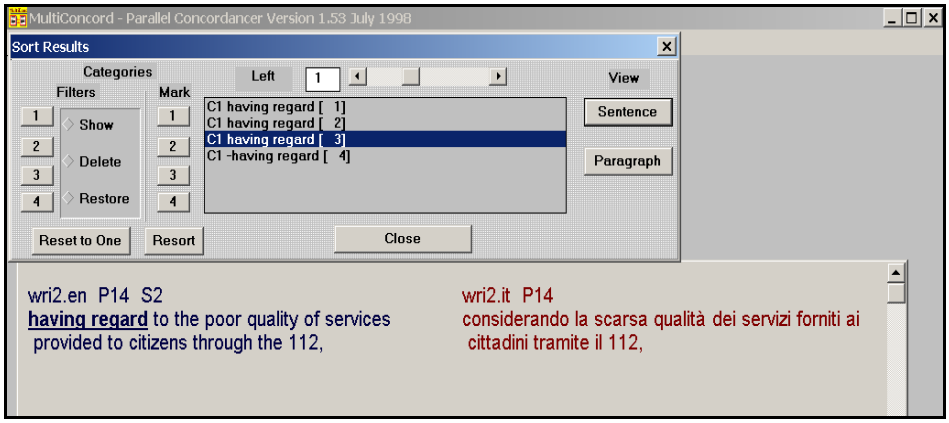

Figura D: 3° risultato di ricerca dell'espressione '*having regard*' nel corpus DSMPE;

- i phrasal verbs to '*to call on*' e '*to call upon*', sempre tradotti nei testi della TL attraverso il verbo '*invitare*';

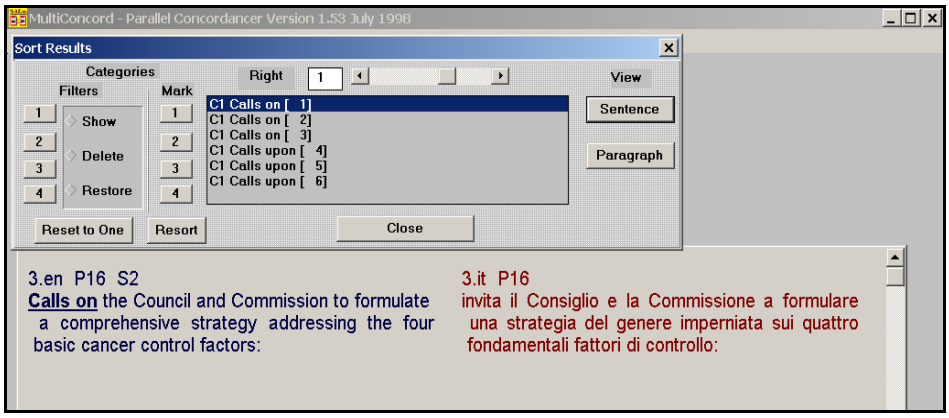

Figura E: 1° risultato di ricerca del verbo
'*to call on*' nel corpus DSMPE;

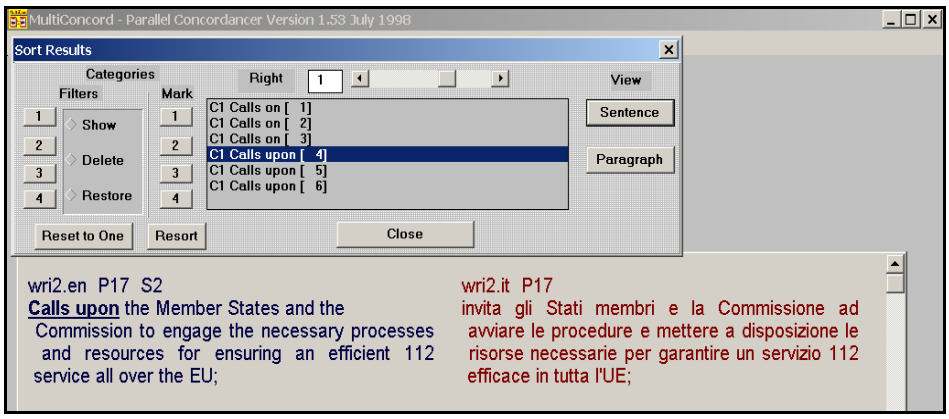

Figura F: 1° risultato di ricerca del verbo
'*to call upon*' nel corpus DSMPE.

Questi esempi dimostrano chiaramente l'utilità dei corpora come strumenti di ricerca e di apprendimento, attraverso i quali è possibile acquisire la necessaria competenza linguistica e concettuale dei linguaggi specialistici, sia in lingua straniera che nella propria lingua madre. La loro tecnologia consente infatti di:

- individuare una terminologia specialistica accurata e completa attraverso software che generano le liste di parole presenti nel corpus e la relativa frequenza, permettendo di identificare quelle più ricorrenti, più specifiche etc.;

- identificare le collocazioni terminologiche tipiche del genere analizzato[3];
- analizzare le strutture grammaticali, sintattiche e stilistiche caratteristiche di una LSP;
- risalire al significato concettuale di un termine attraverso la visualizzazione immediata del suo contesto di utilizzo.

Risulta quindi evidente come la linguistica dei corpora meriti di essere ulteriormente applicata allo studio delle LSP ed estesa, in particolare, all'analisi delle relative traduzioni, visto che è in grado di offrire strumenti efficaci per migliorare sia le didattiche di apprendimento e insegnamento delle lingue straniere, sia l'attività e le tecniche di traduzione.

Nella prossima sezione verranno descritti i più importati criteri per la compilazione di corpora paralleli da utilizzare sia per la ricerca linguistica generale e dei linguaggi specialistici in particolare, sia per eventuali attività didattiche.

2. *Criteri di compilazione e implicazioni di ricerca dei corpora paralleli*

La tipologia di corpora più utile per lo studio delle traduzioni (di LSP ma anche di LGP) è quella dei corpora paralleli che, come precedentemente spiegato, contengono testi originali in una o più lingue di origine e relative traduzioni in una o più lingue di destinazione. I corpora paralleli, oltre a permettere di investigare gli aspetti specifici relativi alle traduzioni, possono essere utilizzati anche come strumenti per il confronto delle similitudini e delle divergenze fra i vari generi testuali inseriti nel corpus e, di conseguenza, possono essere potenzialmente impiegati anche come strumento didattico per insegnare e fare apprendere efficacemente le caratteristiche di stile, registro, forma etc. dei generi linguistici in lingua straniera (o nelle lingue straniere) contenuti nel corpus di riferimento. Come affermato da Pinna (2004/2007: 19) "corpora can be exploited to study given language phenomena in order to show the connection between text and context in both cultural and functional-rhetorical dimensions". Infatti il materiale linguistico dei corpora può essere agevolmente impiegato nella programmazione didattica di un corso di lingua per illustrare agli studenti esempi pratici di applicazione della lingua che rivelino "[…] the cultural

connotations or rhetorical functions associated with language choices both in general language use and in specialized contexts" (ibid.).

Oltre alle applicazioni didattiche, la costruzione di corpora paralleli è naturalmente finalizzata alla ricerca linguistica sulla traduzione. Come già spiegato nella precedente sezione, essi forniscono un valido supporto per effettuare analisi linguistiche approfondite sui più importanti aspetti lessico-grammaticali e stilistici che contraddistinguono le metodologie traduttive di determinati generi linguistici, per poterne così individuare le caratteristiche di utilizzo, i punti di forza, i limiti e gli eventuali aspetti da migliorare. Per questi motivi, la compilazione di corpora paralleli destinati alla ricerca o all'applicazione didattica deve tenere conto di diversi elementi quali il design del corpus, la codifica dei testi e gli obbiettivi del progetto di ricerca.

2.1 Design

Il design di un corpus si basa sulla determinazione di numerosi aspetti (Bowker e Pearson 2002: 45-53 , Olohan 2004: 45-61) che comprendono:

- l'area tematica e le tipologie testuali di analisi;
- le dimensioni del corpus;
- la scelta fra la creazione di un corpus aperto o di un corpus chiuso;
- la selezione di testi interi o estratti;
- la quantità e la qualità di testi e relativi autori.

Anche può sembrare un aspetto ovvio, la definizione dell'area di analisi e delle tipologie testuali da inserire in un corpus rappresenta una fase fondamentale nella compilazione dei corpora. Delineare chiaramente l'area tematica di ricerca consente di stabilire i criteri di selezione dei testi, le modalità di ricerca e gli obbiettivi del progetto, determinando di conseguenza tutte le scelte relative alla creazione del corpus.

Per stabilire le dimensioni di un corpus è necessario considerare innanzitutto le esigenze del progetto di ricerca, nonché i dati e il tempo effettivamente disponibili per effettuare l'analisi linguistica. Infatti, a seconda degli obbiettivi della ricerca potrebbe non essere necessario costruire dei corpora dalle dimensioni vastissime, anche se è sempre consigliabile cercare di garantire un elevato grado di rappresentatività dei

testi inseriti nel corpus per evitare di ottenere dati linguistici fuorvianti, non completi e facilmente criticabili.

Anche la scelta fra la creazione di un corpus aperto o chiuso è determinata dagli obbiettivi finali dello specifico progetto di ricerca: se il corpus deve essere uno strumento per studiare le caratteristiche di un determinato linguaggio e relativa traduzione in uno specifico periodo di tempo, allora il modello da preferire sarà quello chiuso, che serve appunto ad offrire uno "snapshot of the state of a language at a given time" (Bowker e Pearson 2002: 48). Se invece l'obbiettivo di un corpus è quello di analizzare l'evoluzione di una lingua nel tempo, il modello da seguire sarà quello aperto[4] dato che questa tipologia permette di aggiungere o rimuovere testi dal corpus per riflettere "the changing state of a language" (ibid.). Questo modello è particolarmente consigliabile per la creazione di corpora paralleli dedicati allo studio di linguaggi specialistici, considerata la 'natura dinamica' (ibid.) di questi linguaggi in continua evoluzione.

Le dimensioni dei testi inseriti in un corpus sono ugualmente determinate dal tipo di ricerca che si intende effettuare. L'analisi delle LGP può essere effettuata anche attraverso sezioni limitate di un testo senza che ciò comprometta la natura dei dati linguistici del corpus. Per le LSP, invece, è preferibile l'utilizzo di testi interi, in quanto numerosi significati derivano spesso dalla struttura del testo, che risulterebbe inevitabilmente compromessa se il corpus venisse formato da estratti di testo selezionati casualmente.

Infine, anche la quantità e la qualità di testi e relativi autori rappresentano ulteriori aspetti da tenere in considerazione nel design di un corpus. Per ottenere dei dati di ricerca che riflettano ampiamente le caratteristiche linguistiche di una LSP o di una LGP è preferibile che i testi selezionati siano numerosi e che siano stati scritti da diversi autori. Bisognerà inoltre considerare la data di pubblicazione dei testi sempre a seconda degli obbiettivi della ricerca e selezionare quelli pubblicati più recentemente se, ad esempio, lo scopo dell'analisi è l'individuazione delle caratteristiche linguistiche più attuali di un determinato linguaggio. Sarà necessario inoltre bilanciare l'effettiva disponibilità dei testi, valutandone di conseguenza il mezzo di trasmissione (testi scritti o orali) e la possibilità di ottenere l'autorizzazione al loro utilizzo per scopi di ricerca da parte degli autori.

2.2 Codifica dei testi

Ulteriori fattori da considerare nella compilazione dei corpora sono gli aspetti puramente tecnici inerenti la codifica dei testi. Per poter inserire dati linguistici nei corpora è infatti necessario che i testi siano convertiti in formato elettronico[5]. Al fine di ottimizzare i tempi di ricerca, è consigliabile utilizzare testi che siano già in formato elettronico, ricercandoli ad esempio da risorse internet affidabili[6], da CD-rom etc. Nel caso non si disponga di materiale in formato elettronico, sarà necessario trascrivere i testi da analizzare attraverso un processo di scannerizzazione in OCR[7] o con dei software di trascrizione vocale, anche se queste operazioni richiedono tempi piuttosto lunghi per essere effettuate e l'accuratezza della trascrizione non è sempre garantita considerato che alcuni dati possono essere facilmente distorti o persi durante la fase di conversione, richiedendo di conseguenza un processo di verifica aggiuntivo e causando potenzialmente un'ulteriore dilatazione dei tempi.

I testi in formato elettronico devono essere poi ulteriormente "preparati" per poter essere interpretati correttamente dai programmi di analisi dei corpora. Questa fase di preparazione viene effettuata attraverso il processo di *markup* e, nel caso di corpora paralleli, anche attraverso il processo di *allineamento*.

Il *markup* consiste nel determinare (attraverso una 'etichettatura' dei dati linguistici, più propriamente indicata come 'sistema di *tags*') la struttura e l'apparenza di un testo affinché esso possa essere letto e interpretato da un determinato programma di analisi. Per facilitare il riconoscimento dei documenti sottoposti ad un processo di markup da parte di software differenti, negli anni '60 venne sviluppato il linguaggio *SGLM* (*Standard Generalized Markup Language*). Nel campo della linguistica dei corpora l'SGLM ha trovato la sua esplicitazione nel linguaggio *CES* (*Corpus Encoding Standard*), un insieme di parametri che permette di classificare i dati linguistici di un corpus in maniera tale che essi possano essere riutilizzati e inseriti in altri corpora per successivi progetti di ricerca (Bowker 2002: 43-75). Il CES include la categorizzazione di aspetti quali:

- la documentazione (informazioni bibliografiche, lingua etc.);
- dati primari inerenti la struttura del testo (titoli, paragrafi etc.);
- l'annotazione linguistica (un processo di inserimento di *tags* per rendere esplicite caratteristiche linguistiche quali parti del discorso, caratteristiche sintattiche e/o semantiche).

Il processo di markup viene generalmente effettuato in maniera automatica dai programmi contenuti nel software che si utilizza per costruire il corpus.

La costruzione dei corpora paralleli prevede anche la fase di *allineamento*, un processo che collega i paragrafi, le frasi e i termini di un ST con i corrispettivi elementi presenti nel TT . Nel caso in cui non vi sia una assoluta corrispondenza tra gli elementi del ST e del TT (es. nel caso in cui alcuni paragrafi all'interno del TT risultino omessi, uniti, sdoppiati etc.), l'allineamento viene effettuato inserendo simboli o *tag* speciali in corrispondenza di tali discordanze. Anche l'allineamento viene effettuato direttamente dai programmi presenti nei software impiegati per la costruzione del corpus. Questi programmi riescono a generare i collegamenti fra ST e TT attraverso sistemi di calcolo probabilistico integrati a dizionari o a memorie di traduzioni con un livello di accuratezza elevato, anche se è inevitabile che vi possa essere qualche imprecisione, modificabile comunque manualmente.

2.3 Obiettivi della ricerca

Come accennato precedentemente, la compilazione di un corpus deve tenere sempre in considerazione gli obbiettivi della ricerca, e deve garantire che il design del corpus e la codifica dei testi consentano di raggiungere appunto tali obbiettivi. Sia per l'applicazione didattica che nel campo della ricerca linguistica avanzata, gli obbiettivi che ispirano la compilazione di un corpus sono generalmente rappresentati da:

- lo studio dei fenomeni linguistici più significativi tipici di un genere linguistico, o di un linguaggio specialistico o di un determinato autore o personaggio, attraverso l'analisi dei dati statistici forniti dai software del corpus che consento di individuare:
- il numero di *tokens*, cioè il numero totale di parole presenti in un corpus (Bowker 2002: 155);
- il numero di *types*, che indica il numero effettivo di parole diverse fra loro presenti in un corpus e rappresenta un indice per stabilire il tasso di *variabilità linguistica* del genere o del linguaggio analizzato (ibid);
- le liste indicanti le parole chiave (*key words*) caratterizzanti il corpus;
- la frequenza di ogni parola;

- le collocazioni presenti nel corpus, per poter individuare e analizzare il contesto d'uso effettivo di singoli termini o espressioni tipiche del genere o del linguaggio analizzato nel corpus.

Per i corpora paralleli in particolare, gli obbietti della costruzione possono essere rappresentati da:

- l'individuazione e l'analisi della terminologia caratterizzante i linguaggi oggetto di ricerca (LGP o LSP), incluso lo studio di neologismi;
- la realizzazione di glossari bilingui (o multilingui) approfonditi e facili da aggiornare e consultare (poiché in formato elettronico);
- lo studio delle combinazioni sintattiche e semantiche caratterizzanti i linguaggi analizzati, per individuare i modelli sintattici e stilistici più appropriati di un determinato linguaggio e le relative metodologie di traduzione più efficaci, allo scopo di creare delle importanti risorse di riferimento per traduttori, studenti etc.

Conclusioni

L'obiettivo di questa analisi è stato quello di dimostrare che la linguistica dei corpora rappresenta una metodologia che merita di essere ulteriormente applicata nel campo della ricerca linguistica e della didattica, soprattutto in riferimento allo studio della traduzione e delle LSP. I corpora rappresentano degli strumenti indispensabili per la ricerca linguistica e dei mezzi di supporto efficaci anche nell'attività di traduzione, nell' insegnamento e nell'apprendimento delle lingue straniere.

Le potenziali espansioni della linguistica dei corpora sono infinite, ma per concludere, ci limiteremo a suggerire quelle che, a nostro avviso, rappresentano al momento le applicazioni di ricerca più utili e rilevanti, quali:

- ulteriori approfondimenti sugli studi relativi alla ricerca degli *universali della traduzione*;
- la creazione di corpora di LSP di dimensioni più grandi rispetto a quelli creati finora, e con combinazioni linguistiche che coinvolgano una maggiore varietà e quantità di lingue;
- un maggiore impiego della linguistica dei corpora per analisi approfondite sulle lingue non-europee, che ad oggi risulta essere

ancora ostacolato dai problemi di codifica dei sistemi di scrittura non-occidentali[8], ma che meriterebbe un maggiore interesse da parte della ricerca in considerazione della necessità di ampliare le conoscenze linguistiche della popolazione in un contesto di relazioni internazionali e comunicazioni globali come quello attuale.

Note

[1] Gli studi sugli *universali della traduzione* effettuati dal CTIS dell'Università di Manchester hanno portato finora all'identificazione di quattro fenomeni linguistici tipici della traduzione:

- il fenomeno dell'"esplicitazione', rappresentato da tutti quegli accorgimenti linguistici che i traduttori utilizzano per 'svelare le cose in traduzione piuttosto che lasciarle implicite' (Baker 1996:180);
- il fenomeno della 'semplificazione', che si manifesta attraverso l'impiego di un linguaggio molto semplificato in traduzione a livello terminologico, sintattico etc. (ibid: 181-183);
- il fenomeno della 'normalizzazione o conservatismo' rappresentato da 'la tendenza ad esagerare le caratteristiche della lingua target e di conformarle ai suoi tipici modelli linguistici' (ibid: 183);
- il fenomeno dell'"appiattimento', che si esprime attraverso 'la tendenza dei testi tradotti a gravitare verso il centro di un continuum […] affinché il testo si distanzi da ogni estremo di marcatezza orale e letterale proveniente dalla SL o dalla TL' (ibid.: 184).

[2] La trasposizione rappresenta una metodologia di traduzione che consiste nella 'sostituzione di una categoria grammaticale con un'altra, in base all'assunto che entrambe posseggano lo stesso peso semantico o un'equivalente densità semantica' (Alcaraz Varó e Hughes 2002: 181).

[3] Come negli esempi del corpus DSMPE nei quali, attraverso i *concordacing tools*, è stato possibile osservare i termini selezionati per la ricerca nel loro contesto d'uso e nelle loro collocazioni più frequenti.

[4] Il COMPARA rappresenta un esempio di corpus aperto: come precedentemente descritto, questo corpus viene costantemente aggiornato attraverso l'apporto di nuovi testi, in quanto l'obbiettivo di tale ricerca è quello di effettuare analisi linguistiche aggiornate anche da un punto di vista diacronico.

[5] In genere, i testi da inserire in un corpus devono essere convertiti nel formato *.txt* che rappresenta la formattazione informatica testuale più adatta ai programmi di analisi elettronica. Tuttavia, occorre tenere presente che la modalità di conversione in *.txt* non supporta la formattazione degli stili di scrittura e dei caratteri, le immagini o i caratteri di scrittura non-occidentali. Di conseguenza tali limitazioni possono rappresentare a volte dei grossi limiti alle applicazioni della ricerca e alterarne addirittura i risultati.

[6] Una valida risorsa on-line per la ricerca di testi da inserire in un corpus è il portale dell'Unione Europea (www.europa.eu) che, nella sezione dedicata ai documenti prodotti dalle Istituzioni Europee, contiene testi giuridico - politici e relative traduzioni in tutte le lingue dell'Unione.

[7] La sigla *OCR* (*Optical Character Recognition*) indica quel tipo di software in grado di esaminare un'immagine scannerizzata e convertirla in formato testo, attraverso un processo che paragona l'immagine ai caratteri testuali memorizzati nel database del software e, ogni qualvolta trova una coincidenza, converte tale immagine nel carattere di testo corrispondente (Bowker 2002: 26).

[8] I problemi di trascrizione dei caratteri non-occidentali possono essere attualmente superati attraverso il linguaggio di codifica UNICODE, capace di supportare caratteri di

scrittura come quelli delle lingue asiatiche che, a differenza dei caratteri occidentali ad 1 byte, richiedono l'impiego e la combinazione di 2 byte per poter essere trascritti e processati in formato elettronico (sito web del consorzio UNICODE).

Bibliografia

Alcaraz Varó E. e B. Hughes, 2002, *Legal Translation Explained*, St Jerome Publishing, Manchester;

Baker, M., 1996, "Corpus-based Translation Studies: the Challenges that Lie Ahead", in Harold Somers (ed) *Terminology, LSP and Translation*, John Benjamins, Amsterdam & Philadelphia: 175-186;

Bowker, L., 2002, *Computer - Aided Translation Technology: A Practical Introduction*, University of Ottawa Press, Canada;

Bowker, L. e J. Pearson, 2002, *Working with Specialized Languages: A practical guide to using corpora*, Routledge, London and New York;

Johansson, S., 2003, "Reflection on Corpora and their Uses in Cross-linguistic Research" in F. Zanettin, S. Bernardini and D. Stewart (eds) *Corpora in Translator*, St. Jerome, Manchster: 135-144;

Laviosa, S., 1998, "The English Comparable Corpus: A Resource and a Methodology", in L. Bowker, M. Cronin, D. Kenny and J. Pearson (eds) *Unity in Diversity: Current Trends in Translation Studies*, St. Jerome Publishing, Manchester: 101-112;

Laviosa, S., 2002, *Corpus-based Translation Studies: Theory, Findings, Applications*,: Rodopi, Amsterdam and New York;

Olohan, M., 2004, *Introducing Corpora in Translation Studies*, Routledge, London & New York;

Olohan, M. e M. Baker, 2000, "Reporting *that* in Translated English: Evidence for Subconscious Processes of Explicitation", *Across Languages & Cultures* 1(2): 141-158;

Pinna, A., 2004/2007, "Corpus linguistics: resources and activities for EFL" in *Annali della Facoltà di Lingue e Letterature straniere col.4*: 19-38;

Zanettin, F., 2000, "Parallel Corpora in Translation Studies: Issues in Corpus Design and Analysis" in M. Olohan, (ed.) *Intercultural Faultlines: Research Models in Translation Studies* I, St. Jerome, Manchester: 105–118.

Risorse on-line

Centre for Translation and Intercultural Studies:
http://www.llc.manchester.ac.uk/Research/Centres/CentreforTranslationandInterculturalStudies/ResearchProgrammesPhDMPhil/TranslationEnglishCorpus/ [ultimo accesso 20 gennaio 2008];

Compara:
http://www.linguateca.pt/COMPARA/ [ultimo accesso 20 gennaio 2008];

European Language Resources Association (ELRA):
http://www.elra.info/index.html [ultimo accesso 20 gennaio 2008];
Oslo Multilingual Corpus (OMC): http://www.hf.uio.no/ilos/OMC/
[ultimo accesso 20 gennaio 2008];
Portale dell' Unione Europea:
http://www.europa.eu/languages/it/home [ultimo accesso 20 gennaio 2008];
Transearch:
http://www.tsrali.com/ [ultimo accesso 20 gennaio 2008];
Unicoedw:
http://www.unicode.org/ [ultimo accesso 20 gennaio 2008].