

Leńko-Szymańska, Agnieszka - Boulton, Alex (a cura di), "Multiple Affordances of Language Corpora for Data-driven Learning", John Benjamins, Amsterdam/Philadelphia, 2015. ISBN 978-90-272-0377-9

di **Luciana Forti**

Lo studio dei fenomeni linguistici ha visto, negli ultimi decenni, un ricorso sempre più frequente ai *corpora*, definibili come raccolte di testi autentici, dotate di una rappresentatività, e adeguate al trattamento automatico. A partire dagli anni Sessanta del secolo scorso, grazie ai *corpora* è stato possibile iniziare a osservare l'uso di un'unità linguistica sul piano della sua frequenza, della sua distribuzione e delle sue proprietà combinatorie, con il vantaggio di unire all'analisi qualitativa una dimensione quantitativa fondata sulle tecniche della statistica.

Nell'indicare i molteplici ambiti di applicazione dei *corpora*, si fa spesso riferimento all'opportunità di un loro sfruttamento pedagogico, dal momento che in molti casi tali raccolte sono in grado di offrire un'immagine dell'uso reale di una lingua che né il libro di testo, né il dizionario, né il modello linguistico rappresentato dall'insegnante riuscirebbero a fornire. Tuttavia, i tentativi sistematici volti a esplorare, nella dimensione pratica, questa possibilità applicativa sono molto scarsi, in particolar modo per quanto concerne l'apprendimento dell'italiano L2. Come si potrà immaginare, la situazione riguardante l'inglese L2 è nettamente migliore: le sperimentazioni in tal senso, infatti, si susseguono da almeno trent'anni, assumendo forme compiute a partire dai primi anni Ottanta del Novecento, grazie al lavoro di Tim Johns. È a questo docente, che ha svolto per molti anni la propria attività di studioso presso l'Università di Birmingham, che si deve la locuzione *Data-driven learning (DDL)*, ovvero *apprendimento guidato dai dati*, laddove per *dati* ci si riferisce alle informazioni estrapolabili dai *corpora*. Esito di una selezione di studi presentati al convegno *TaLC 10 (Teaching and Language Corpora)* tenutosi a Varsavia nel 2012, *Multiple Affordances of Language Corpora for Data-driven Learning* raccoglie una serie di studi volti a rispondere ad alcune delle domande più pressanti del settore: per quali obiettivi di apprendimento possono essere usati i *corpora*? L'uso di queste raccolte di testi autentici è adatto a ogni tipologia di apprendente e a ogni livello di competenza? Quali sono i vantaggi nell'utilizzo di un *corpus* rispetto a risorse tradizionali come il dizionario? Pos-

siamo considerare anche il web come un *corpus* per scopi legati all'apprendimento linguistico?

Dopo la parte introduttiva, due contributi preziosi precedono le varie sezioni del volume. Il primo, di Lynne Flowerdew intitolato *Data-driven learning and language learning theories*, discute le possibilità applicative del DDL all'interno di tre teorie sull'apprendimento di una lingua seconda: la *noticing hypothesis*, il costruttivismo e la teoria socioculturale di Vygotskij. La studiosa applica alcuni dei principi sottesi a ciascuna teoria in tre sperimentazioni didattiche con i *corpora*, riflettendo su come l'osservazione guidata delle occorrenze in un *corpus*, relative a specifiche strutture lessico-grammaticali, possa fornire la base per attività di autocorrezione o correzione collaborativa degli errori prodotti in esercizi di scrittura.

Il secondo contributo, di Chris Tribble, si intitola *Teaching and language corpora. Perspectives from a personal journey*, e propone una storia del DDL attraverso una cronologia dettagliata che va dal 1975 al 2014. Vengono inoltre riportati e discussi i risultati di un'indagine svolta nel 2012 sulla diffusione del DDL tra gli insegnanti di lingue, risultati che sono stati confrontati con quelli desunti da un'indagine analoga condotta nel 2008. In entrambi i casi, l'uso didattico dei *corpora* risulta maggioritario nelle università rispetto alle scuole, nei corsi di scrittura accademica rispetto ai corsi generali, e nei corsi di livello avanzato rispetto a quelli di livello iniziale. Tuttavia, l'indagine del 2012 indica una distribuzione meno polarizzata dei dati, riflettendo un lento, ma regolare incremento di sperimentazioni nelle scuole e con apprendenti di livello non necessariamente avanzato. A seguire, il volume si divide in tre sezioni. La prima, *Corpora for language learning*, riguarda l'uso dei *corpora* per l'apprendimento degli aspetti lessico-grammaticali della lingua.

Nel primo contributo di questa sezione, ovvero nel quarto capitolo del volume, troviamo il lavoro di Guy Aston dal titolo *Learning phraseology from speech corpora*. Lo studioso descrive l'uso didattico di un *corpus* di parlato composto a partire dalle trascrizioni di 500 conferenze TED. L'obiettivo delle attività legate al *corpus* consiste nello sviluppo della fluidità verbale, attraverso, per esempio, l'osservazione guidata delle caratteristiche prosodiche di determinati insiemi fraseologici, grazie a programmi come *WinPitch*.

Nel quinto capitolo, intitolato *Stealing a march on collocation: Deriving extended collocations from full text for student analysis and synthesis*, James Thomas propone una serie di procedure guidate finalizzate a stimolare l'osservazione di collocazioni nei loro intorni lessicali e grammaticali, dando agli studenti la possibilità di operare generalizzazioni a partire da

quanto osservato.

La questione relativa allo sfruttamento dei *corpora* per i livelli di competenza iniziali è invece al centro del capitolo 6, intitolato *A corpus and grammatical browsing system for remedial EFL learners*, ad opera di Kiyomi Chujo, Kathryn Oghigian e Shiro Ahasagawa. Il gruppo di ricerca ha adattato le concordanze estratte da un *corpus* di riferimento, trasformando il cotesto di varie unità lessicali in frasi sintatticamente e lessicalmente semplificate. Nasce così il *corpus* SCoRE (*Sentence Corpus of Remedial English*) che, integrato allo strumento GPPS (*Grammatical Pattern Profiling System*), è in grado di far osservare profili collocazionali e colligazionali in cotesti brevi e semplici.

Con il capitolo 7, intitolato *Same task, different corpora. The role of personal corpora in EAP classes*, Maggie Charles apre la seconda sezione del volume, riguardante lo sviluppo di competenze linguistiche legate alla scrittura e alla lettura, dal titolo *Corpora for skills development*. Nell'ambito di un corso di lingua inglese per scopi accademici, Charles chiede a ciascun apprendente di costruire un *corpus* di articoli scientifici relativi al proprio ambito di studio e di svolgere una serie di compiti, basati sull'osservazione di aspetti caratteristici del lessico accademico. Ciascun apprendente ha quindi a disposizione un proprio *corpus* personale su cui lavorare, ma i compiti da svolgere sono gli stessi per tutti, e consistono, per esempio, nell'osservare il modo in cui vengono usati i pronomi personali e i connettivi, usi che possono in qualche misura variare da una disciplina all'altra, e dai quali è possibile operare importanti generalizzazioni d'uso. Nel capitolo 8, Svitlana Babych propone un contributo dal titolo *Textual cohesion pattern for developing reading skills. A corpus-based multilingual learning environment*. Per lo sviluppo di strategie di lettura, la studiosa propone una classificazione dei connettivi per tre lingue: ucraino, russo e inglese. Viene inoltre costruito un *corpus* parallelo di testi giornalistici, reso disponibile per gli apprendenti attraverso una piattaforma online. Tali risorse vengono sperimentate nell'ambito di un corso universitario di lingua ucraina, in cui si fa leva sulla conoscenza del russo e dell'inglese.

Alejandro Curado Fuentes fornisce, nel capitolo 9 intitolato *Exploiting keywords in a DDL approach to the comprehension of news texts by lower-level students*, il secondo contributo basato sull'applicazione del DDL in contesti di competenza iniziale. Nel raro, e dunque estremamente apprezzabile, tentativo di valutare l'efficacia dell'approccio rispetto a metodi tradizionali, lo studioso crea una serie di attività sperimentali per migliorare le abilità di comprensione attraverso l'osservazione delle parole chiave in un *corpus* di testi giornalistici e in un *corpus* di riferimento. Il capitolo

10 apre la sezione del volume dedicata all'uso dei *corpora* nella formazione di traduttori: *Corpora for translation training*, e si intitola *Webquests in translator training. Introducing corpus-based tasks*. Gli studiosi, Teresa Molés-Cases e Ulrike Oster, descrivono una sperimentazione orientata al miglioramento delle competenze traduttive dall'inglese e dal tedesco verso il catalano. Gli studenti sono guidati a

definire contrastivamente e induttivamente i tratti d'uso di un insieme di morfemi e di modali a partire dal *Corpus Valencià de literatura Traduïda (COVALT)*, e ricorrendo a strumenti online quali *DWDS*, *Linguee*, *Linguatools* e *Pons*.

Nella sperimentazione di Josep Marco e Heike van Lawick, riportata nel capitolo 11 e intitolata *Enhancing translator trainees' awareness of source text interference through use of comparable corpora*, gli studenti lavorano su testi tradotti dall'inglese o dal tedesco verso il catalano, prodotti da studenti degli anni precedenti. La valutazione dell'efficacia dell'attività *data-driven* viene effettuata misurando il grado di influenza della L1 nelle traduzioni prodotte dal gruppo di apprendenti prima e dopo la sperimentazione.

Nel capitolo 12, Patricia Sotelo propone un contributo intitolato *Using a multimedia corpus of subtitles in translation training. Design and applications of the Veiga corpus*, e presenta una sperimentazione basata su un *corpus* multimediale in un corso di traduzione e sottotitolazione audiovisiva. Alcuni vantaggi del *corpus* consistono nella semplicità d'uso dell'interfaccia abbinata alla possibilità di svolgere attività di annotazione, e dunque di personalizzazione, del *corpus* stesso, che è in grado di restituire a una determinata chiave di ricerca una concordanza nella lingua sorgente, nella lingua di arrivo e, per ciascuna stringa, il segmento video corrispondente.

Applying data-driven learning to the web è il contributo che conclude il volume, al capitolo 13, e reca la firma di Alex Boulton. L'autore illustra il modo in cui alcune funzionalità di Google possono permettere agli apprendenti di usare il web come *corpus*, tenendo conto di tutte le limitazioni del caso, con il vantaggio di avere a disposizione un'interfaccia più semplice e familiare rispetto a quella di un *corpus* tradizionale. Boulton discute le opzioni *GoogleFight* e *autofill*, anche in relazione ai lavori svolti da altri studiosi che hanno coniato le sigle GALL (*Google assisted language learning*) e GDLL (*Google driven language learning*).

Le ultime pagine del volume contengono una lista delle risorse online citate nei vari contributi, suddivisa in *corpora*, programmi per l'estrazione di dati, *tagger*, programmi per la trascrizione e l'allineamento audio nei *corpora* di parlato, e un elenco di siti web contenenti materiali basati, ancora una volta, su *cor-*

pora costruiti per l'apprendimento e l'insegnamento. Complessivamente, il volume fornisce una panoramica decisamente ampia su alcune delle più recenti sperimentazioni pedagogiche basate sull'utilizzo degli stessi *corpora* in classe. Tuttavia, come spesso indicato dagli stessi autori, le limitazioni riguardanti le sperimentazioni contenute in questo volume sono diverse e generalmente note, sebbene discusse e approfondite solo sommariamente nella sede di questa raccolta. Innanzitutto, emergono almeno tre criticità riguardanti la natura del campione di apprendenti. In primo luogo, la dimensione: una sperimentazione fondata su un campione ridotto, generalmente costituito da poche decine di studenti, difficilmente permette di superare le variabili legate alle differenze individuali che caratterizzano l'apprendimento linguistico. In questi casi, operare delle generalizzazioni scientifiche risulta alquanto arduo, dal momento che le ridotte dimensioni di un campione contribuiscono a ridurre anche la rappresentatività. Certamente, la direzione dei risultati ottenuti da studi pilota svolti con pochi studenti può tracciare il percorso per sperimentazioni su scala più vasta. In secondo luogo, con l'eccezione dei pochi casi menzionati, i campioni sono generalmente soltanto sperimentali: risultano spesso assenti, infatti, campioni di controllo opportunamente organizzati al fine di misurare il raggiungimento di un dato obiettivo di apprendimento sulla base di due percorsi differenti, tipicamente distinti nei termini di una presenza o di una assenza dell'uso dei *corpora*. Inoltre, nei casi in cui il gruppo di controllo è presente, non vengono indicati con chiarezza i metodi con cui viene costruito il percorso di apprendimento da porre in fase con l'approccio DDL: molto spesso, infatti, il termine di paragone per una sperimentazione DDL risulta essere il normale svolgimento del corso, una scelta che, però, pone una serie di problemi, e, non da ultimo, il fatto che l'attenzione specifica su determinate aree della competenza linguistica potrebbe essere molto differente non solo in termini di qualità, ma anche in termini di quantità, rendendo la comparazione poco bilanciata e dunque potenzialmente fuorviante.

Infine, i campioni vengono creati secondo modalità non probabilistiche, nel senso che non

tutte le unità della popolazione presentano la stessa probabilità di essere selezionati per far parte del campione. Ciò accade perché è la stessa definizione di popolazione a essere sfuggente, dal momento che consisterebbe in tutti gli apprendenti di una seconda lingua: un dato, questo, non semplice da ottenere. Si tende dunque a utilizzare quel che si ha a disposizione, per esempio gli studenti del corso universitario nel quale si insegna. Di conseguenza, generalizzare gli

esiti di una sperimentazione risulta difficile, visto che la selezione del campione non presenta i caratteri di una scelta casuale e rappresentativa in relazione alla popolazione.

Un'ulteriore limitazione riguarda il fatto che in molti casi il ricercatore coincide con lo sperimentatore, ovvero con l'insegnante che crea e mette in pratica le attività basate sui *corpora*, somministrando spesso anche questionari volti a cogliere le impressioni degli studenti. Ciò tende a creare vari *bias* scientifici, che nascono dall'influenza che le aspettative del ricercatore (o anche dei soggetti facenti parte del campione) esercitano sul risultato della sperimentazione stessa. Innanzitutto, il ricercatore ha in genere un interesse diretto nella buona riuscita della sperimentazione, dal momento che all'investimento di energie e di tempo vorrebbe che corrispondesse il raggiungimento di un risultato positivo, evitando così il cosiddetto *file drawer problem*. In secondo luogo, essere a conoscenza di far parte di una sperimentazione può determinare negli apprendenti un entusiasmo e una curiosità nei confronti della novità in grado di creare condizioni di apprendimento e insegnamento differenti rispetto a quelle autentiche. Inoltre, risulta compromessa la verifica di una delle potenzialità primarie del DDL, ovvero l'opportunità di diffonderne l'utilizzo presso gli insegnanti di lingua, che di fatto costituirebbero il principale canale di diffusione di questo tipo di approccio. Preziosi, in tal senso, sono le indagini di Chris Tribble e gli esperimenti di formazione degli insegnanti di Agnieszka Leńko-Szymańska.

Un altro punto sollevato frequentemente dai diversi contributi contenuti nel volume riguarda la necessità di sperimentare l'approccio per periodi di tempo superiori alle poche settimane. Mentre l'obiettivo a breve termine dell'uso dei *corpora* in classe è orientato al raggiungimento di una determinata competenza o conoscenza, quello a lungo termine vorrebbe condurre gli apprendenti a un uso autonomo degli strumenti legati a questo tipo di raccolte di testi autentici. La breve durata di una sperimentazione non permette la valutazione di quel che è forse il principale obiettivo del DDL, ovvero stimolare le capacità per diventare apprendenti migliori, con più risorse utilizzabili per scopi diversi, ed è per questo motivo che l'invito a condurre studi longitudinali più numerosi è fortemente sostenuto.

A ciò si può aggiungere anche la necessità di condurre sperimentazioni che riguardino lingue di apprendimento differenti dall'inglese, e combinazioni linguistiche che prescindano dall'inglese, o anche da lingue appartenenti alle medesime famiglie linguistiche, per avere una base più ampia e più solida sulla quale operare delle generalizzazioni.

Nel suo insieme, la lettura del volume delinea un quadro particolarmente variegato sulle modalità applicative del DDL dal punto di vista degli obiettivi di apprendimento, dei *corpora* utilizzabili, e delle procedure attuabili in classe. La varietà degli studi presentati è tuttavia armonizzata dalla presenza di contributi storici e teorici sul DDL, che incrementano l'accessibilità alla lettura anche da parte di chi avesse poca familiarità con il tema. Per tale motivo, il volume costituisce senz'altro un'importante e ricca fonte di stimolo per chi si occupa di riflessi applicativi dei *corpora*, così come di aggiornamento nel campo della didattica delle lingue moderne e della traduzione.

