© Italiano LinguaDue, n. 2. 2011. P. B. Viganò, *I corpora e il loro sfruttamento in didattica.*

I *CORPORA* E IL LORO SFRUTTAMENTO IN DIDATTICA

Paola Bruna Viganò[1]

INTRODUZIONE

Nell'ambito dell'insegnamento dell'italiano a stranieri, sembra che lo sfruttamento dei *corpora* sia ancora lontano dall'aver preso piede, sia a livello di impiego indiretto (progettazione di materiali di riferimento o produzione di materiali didattici) sia a livello di impiego diretto da parte degli insegnanti e degli studenti sotto la guida dell'insegnante o in autonomia. Chi scrive è invece una convinta sostenitrice dell'utilità del cosiddetto *data-driven learning* (DDL) di cui si tratterà a continuazione. Per giustificare quest'affermazione, si chiarirà il concetto di *corpus* trattandone le caratteristiche principali e facendo qualche accenno alla sua costruzione; successivamente si offrirà una panoramica sui tipi di *corpora* e sulle tecniche e possibilità di impiego, per poi focalizzarsi specificamente sul loro sfruttamento nell'ambito dell'insegnamento/apprendimento e infine fare qualche riflessione su vantaggi e problematiche connesse.

1. CORPUS: DEFINIZIONE E CARATTERISTICHE

Al di là del loro impiego, il concetto di *corpus* è ormai abbastanza diffuso, almeno intuitivamente. Si cercherà qui di fare chiarezza riportando e commentando due definizioni: quella data da Sinclair (1991a: 171) e quella data da McEnery e Wilson (2001: 197). La scelta di riportare due definizioni – e non una sola – è data dalla differenza tra le due, legata alla prospettiva fornita dai due autori: rispetto a quella presentata da Sinclair, McEnery e Wilson danno una definizione – per così dire – stratificata, come sarà evidente dalla definizione stessa, riportata e commentata poco sotto.

Si comincerà da quella di Sinclair (*ibid.*): «A corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language».

È importante notare che un *corpus* viene definito come una **raccolta di testi**, ma soprattutto che tali testi sono «naturally-occurring», ovvero si deve trattare di campioni **genuini** (scritti o orali che siano) e non di formulazioni linguistiche create *ad hoc* per fornire esempi. Ciò al fine di studiare la lingua nella sua forma spontanea e dunque naturale, senza le falsificazioni che possono derivare dalla creazione di un esempio. Inoltre, nella definizione data da Sinclair (*ibid.*) si dichiara l'obiettivo di tale raccolta: «to characterize a state or variety of a language», con riferimento alla **rappresentatività**, uno degli attributi imprescindibili di un *corpus* che sarà a breve definita in maggior dettaglio. Immediatamente dopo il testo citato, Sinclair (*ibid.*) prosegue sostenendo che

[1] Master Promoitals, Università degli Studi di Milano.

le **dimensioni** di un *corpus* devono essere tali da permettere di individuare le strutture tipiche della lingua, la cui creatività porta a un'enorme varietà; in concreto, Sinclair (*ibid.*) parla di molti milioni di parole, evidentemente facendo riferimento a un *corpus* di lingua generale. A proposito delle dimensioni dei *corpora*, Bowker e Pearson (2002: 45-48) dichiarano che non ci sono regole ferree né linee guida ideali. Inoltre, sostengono che non è sempre vero che più grandi sono le dimensioni meglio è; infatti, a loro giudizio, grandi quantità di parole possono rendere difficile il reperimento di dati importanti ma meno abbondanti. In sostanza, affermano che le dimensioni dipendono da tre fattori: lo scopo del *corpus*, la disponibilità di risorse e il tempo a disposizione per la compilazione. In concreto, riferendosi alla costruzione di un *corpus* specialistico, gli stessi autori dichiarano che dimensioni comprese tra 10.000 e centinaia di migliaia di parole sono proprie di *corpora* che nella loro esperienza si sono rivelati utili. È evidente che centinaia di migliaia di parole costituiranno un *corpus* specialistico, mentre le 10.000 parole citate potranno essere le dimensioni di un *corpus* costruito *ad hoc* per risolvere uno specifico problema (ad esempio, uno specifico problema linguistico). Per concludere le osservazioni sulla definizione di *corpus* fornita da Sinclair (*ibid.*), bisogna dire che l'autore non fa alcun accenno al **formato** in cui vengono raccolti i testi, che di norma è elettronico. Ciò porta una serie di vantaggi, tra cui, innanzitutto, la possibilità di consultarli per mezzo di strumenti informatici che velocizzano il processo di ricerca dell'oggetto di studio e restituiscono dati quantitativi in tempi brevi.

Passiamo ora alla definizione di *corpus* data da McEnery e Wilson (*op. cit.*: 197): «(i) (loosely) any body of text; (ii) (most commonly) a body of machine-readable text; (iii) (more strictly) a finite collection of machine-readable text, sampled to be maximally representative of a language or variety».

Come si può notare, gli autori in questione danno una definizione che, poco sopra, abbiamo definito stratificata e procedono con una prospettiva che dal generale si focalizza sul particolare, fornendo una definizione sempre più dettagliata di *corpus* in ambito linguistico. Nel punto (i) lo si definisce come un **insieme di testi** e nel punto (ii) si specifica che tali testi devono essere processabili da un **software**. Il punto (iii) fornisce le caratteristiche fondamentali che deve possedere un *corpus*: la finitezza delle sue dimensioni («finite»), il derivare da campionatura («sampled») e la rappresentatività («to be maximally representative of a language or variety»). Per quanto riguarda la **finitezza** del *corpus*, McEnery e Wilson (*op. cit.*: 30) specificano che le dimensioni vengono stabilite nella fase di progettazione dello stesso[2]. Per quanto riguarda il tema della **campionatura dei testi** e della **rappresentatività** del *corpus*, questi due temi saranno qui di seguito trattati congiuntamente in quanto interdipendenti e si farà riferimento a quanto sostenuto da McEnery e Wilson (*op. cit.*: 77-81) e a quanto sintetizzato da Baker, Hardie e McEnery (2006). La campionatura dei testi va operata sulla base dello scopo che si prefigge il *corpus* e, di conseguenza, dei relativi criteri stabiliti in fase di progettazione. Svolgendo quest'operazione, però, bisogna considerare la questione fondamentale della

[2] Quest'affermazione vale solo parzialmente per i cosiddetti *monitor corpora*, che non hanno dimensioni finite per definizione. Infatti, questo tipo di *corpus* viene progettato e realizzato in quella che potremmo definire 'la sua prima versione', per poi essere periodicamente ampliato; l'ampliamento avviene in modo che la proporzione tra le tipologie testuali rimanga costante, spiega Hunston (2002:16). L'obiettivo è quello di monitorare, appunto, determinati cambiamenti linguistici nel tempo (spesso cambiamenti di tipo lessicale, ma non solo).

rappresentatività. Infatti la costruzione del *corpus* non può prescindere da una selezione dei testi, includendone alcuni ed escludendone altri, e ciò va fatto tenendo presente che il *corpus* dovrà andare a costituire un campione ragionevolmente completo e accurato della popolazione linguistica in esame, affinché le osservazioni che verranno poi fatte sul *corpus* siano generalizzabili alla popolazione nel suo complesso. Va da sé che, prima di procedere alla campionatura, è necessario definire il più chiaramente possibile i confini della popolazione che si vuole studiare. Per fare questo si può procedere all'individuazione di una bibliografia vasta e completa o selezionare i materiali all'interno di quelli messi a disposizione da una determinata biblioteca specializzata nell'area di interesse. Inoltre, identificare un'articolazione della popolazione linguistica in esame, ad esempio individuando i generi in cui si possono classificare i testi, può rappresentare una guida per una campionatura completa della popolazione in questione. Seguire una tale classificazione e rispettarne le proporzioni permette anche di **bilanciare** il *corpus* al suo interno, come appena ipotizzato in base ai generi testuali individuati in fase di campionatura, oppure in base ai domini[3] relativi, e pertanto, in definitiva, può aiutare a costruire un *corpus* sia rappresentativo della popolazione linguistica in esame sia bilanciato nella sua articolazione interna. Altro elemento da tenere in considerazione è l'**affidabilità** delle fonti da cui vengono estratti i testi selezionati; questo aspetto non va affatto sottovalutato se si considera che, per comodità, spesso i testi che confluiscono nel *corpus* vengono reperiti on-line e la fonte o la data dell'ultimo aggiornamento non sono sempre identificabili o precisati. Per concludere questo tema, è opportuno accennare agli **errori di campionatura**. Come specifica MacMullen (2003), questi possono imputarsi al caso o alla presenza di occorrenze rare all'interno del materiale selezionato. Ad ogni modo, lo stesso autore afferma che tali errori sono inversamente proporzionali alle dimensioni del *corpus* e, pertanto, un *corpus* adeguatamente dimensionato può garantire che eventuali errori di campionatura vengano neutralizzati.

2. TIPI DI *CORPORA*

I *corpora* si possono classificare a seconda degli scopi che l'investigazione degli stessi si propone. L'esposizione che segue vuole fornire una panoramica sui tipi fondamentali di *corpora* e si rifarà parzialmente alla classificazione esposta da Bowker e Pearson (*op. cit.*: 11-13) e Aston (1999).

Una prima distinzione può essere fatta tra *corpora* **generali** e *corpora* **specialistici**. I primi sono rappresentativi di una determinata lingua e da essi si possono derivare osservazioni che avranno validità su quella lingua nel suo complesso. I secondi differiscono dai primi perché in questo caso si seleziona un particolare aspetto della lingua in questione: area tematica, tipologia testuale, varietà linguistica, etc.

Una seconda distinzione può essere fatta tra *corpora* **monolingui** e **multilingui**. Come è facilmente intuibile, i primi comprendono testi in una sola lingua. I secondi comprendono testi in due o più lingue. A loro volta, i *corpora* bilingui possono essere

[3] L'impiego di «dominio» è qui mutuato dalla terminologia, dove ha l'accezione di settore principale; questo può essere oggetto di suddivisioni in ulteriori ambiti più ristretti, denominati «sottodomini». Ad esempio, il dominio medicina si può articolare in vari sottodomini, tra cui quello della patologia, a sua volta suddivisibile in patologia generale e speciale.

ulteriormente suddivisi in due categorie: *corpora* **comparabili** e **paralleli**. I primi comprendono testi in diverse lingue, in cui gli stessi sono stati scritti in originale e il criterio alla base del loro raggruppamento può essere ad esempio la tematica che affrontano, la tipologia testuale, etc. I secondi possono essere **unidirezionali** e contenere testi in una determinata lingua con le corrispondenti traduzioni in un'altra lingua, oppure **bi-direzionali** o **reciproci** e contenere testi originali in una determinata lingua (L1) con le corrispondenti traduzioni in un'altra lingua (L2), ma anche testi originali in L2 con le corrispondenti traduzioni in L1. Quest'ultimo tipo di *corpus* riunisce le caratteristiche di un *corpus* parallelo e di uno comparabile.

Una terza distinzione può essere fatta tra *corpora* **aperti** e **chiusi**. I primi sono i *monitor corpora* a cui abbiamo accennato poco sopra in nota 1. I secondi, invece, una volta realizzati, non vengono più ampliati.

Altra importante, anche se intuitiva, distinzione è quella tra *corpora* che riuniscono testi **scritti** e *corpora* composti da trascrizioni di testi **orali**.

Infine, una ulteriore distinzione tra *corpora* di testi elaborati da madrelingua e i cosiddetti ***learner*** *corpora*. Questi ultimi sono composti da materiali prodotti da studenti non nativi e possono essere particolarmente utili, se taggati[4], ad esempio per indagare il tipo di errori commessi dagli alunni. Tale indagine può essere utile sia per l'elaborazione di materiali specifici, sia per l'intervento didattico mirato.

Ulteriori distinzioni potrebbero essere fatte dal momento che ci possono essere tanti *corpora* quante possono essere le loro applicazioni. Infatti – come si approfondirà nel par. 4., concentrandosi sull'insegnamento nel par. 4.1. – la costruzione di *corpora* può essere abbinata a varie discipline per indagare una vasta gamma di aspetti linguistici: lessicografici, socio-linguistici, storici, apprendimento del linguaggio, traduzione, redazione tecnica, etc.

3. TECNICHE D'INDAGINE DEI *CORPORA*

Per quanto riguarda gli strumenti di indagine dei *corpora*, allo stato attuale sono disponibili diversi **software** che recuperano occorrenze ed eseguono indagini in modo automatico, come ad esempio *AntConc* oppure il pacchetto *WordSmith Tools*.

È opportuno osservare che le operazioni svolte in modo automatico dai software sono di tipo essenzialmente **quantitativo**, come ad esempio generazione di *wordlist* e *keyword*, che descriveremo a breve; l'estrazione di dati in base alla loro frequenza può essere al limite incrociata con la ricerca di elementi taggati, ad esempio 'etichettati' con la relativa categoria grammaticale o, nei *learner corpora*, in base al tipo di errore, come accennato in nota 3. La possibilità di consultazione per mezzo di strumenti informatici che velocizzano il processo di ricerca e restituiscono dati quantitativi in tempi brevi è un vantaggio noto e il risparmio di tempo che l'utente ottiene rispetto a ricerche e conteggi

[4] Come anche in altri ambiti, applicare dei *tag* a determinati elementi, significa 'etichettarli' assegnandogli una parola in qualche modo utile per descriverli. In linguistica dei *corpora*, i *tag* servono a catalogare dati contenuti nei *corpora* e facilitarne il reperimento in base alle categorie definite dalle 'etichette'. Queste possono seguire diverse logiche, anche specifiche e dettate da chi ha progettato i *corpora* in questione; si può trattare di categorie grammaticali, informazioni lessicali, socio-linguistiche, etc. Nel caso specifico, si tratta di tipologie di errori contenuti in *learner corpora*.

manuali può essere investito nell'interpretazione dei dati quantitativi ottenuti e nella cosiddetta **analisi qualitativa**. Infatti, come sostengono McEnery e Wilson (*op. cit.*: 76-77), le due metodologie di analisi sono diverse ma complementari e, in particolare, l'analisi quantitativa è in grado di fornire dati statistici affidabili e generalizzabili, mentre l'analisi qualitativa è in grado di fornire maggior ricchezza interpretativa a partire da quei dati affidabili.

Senza entrare nel dettaglio delle possibilità offerte dal singolo software, pare utile fare una rassegna delle **tecniche di base** che si possono generalmente impiegare nel sondare i *corpora*:

– *Wordlist*
  elenco in ordine di frequenza delle unità linguistiche[5] presenti nel *corpus* in esame. Possono essere utili, ad esempio, per identificare il lessico fondamentale di una lingua e definire, per confronto, ad alunni di che livello somministrare determinati testi.

– *Keyword*
  lista di parole presenti nel *corpus* che si vuole analizzare, che risultano insolitamente frequenti in rapporto a quelle di un altro *corpus*. Tipicamente, confrontando una lista di frequenza di un *corpus* specialistico con la lista di frequenza di un *corpus* generale, si potrà ottenere una *keyword list* in cui figureranno i termini propri dell'ambito specialistico in esame.

– *Cluster*
  sono gruppi di parole (due o più) che si ripetono affiancate con una frequenza minima specificata nel momento dell'interrogazione; la ricerca dei cluster può essere utile sia dal punto di vista lessicale, ovvero nella ricerca di termini complessi, sia nell'indagine di comportamenti lessico-grammaticali, come ad esempio la reggenza preposizionale di un sostantivo.

– *Concordanze*
  si tratta del risultato della ricerca di una parola (o sintagma o espressione), presentato in una serie di righe (normalmente si può scegliere un numero massimo) visualizzate nel formato KWIC (*Key-Word-In-Context*); con questo formato, al centro dello schermo compare l'elemento ricercato (o nodo) e, alla sua destra e alla sua sinistra, una porzione di testo equivalente agli orizzonti specificati (ovvero quantità di parole o caratteri); ciò permette di verificare le occorrenze dell'oggetto della ricerca e di andare a indagare il suo comportamento linguistico all'interno del contesto in cui compare. In alcuni casi, al momento della ricerca, c'è la possibilità di specificare parole o elementi contestuali e impiegare anche i cosiddetti caratteri jolly[6]. Normalmente è anche possibile ordinare le concordanze specificando quale posizione a destra o a sinistra del nodo deve essere presa in considerazione per

[5] Si impiegano volutamente i termini "unità linguistiche" e "parole", per non introdurre in questa trattazione i termini *tokens* e *types*.

[6] Si tratta di caratteri che non rappresentano sé stessi, bensì un singolo carattere sconosciuto (tipicamente si impiega il carattere «?») o un insieme di caratteri ignoti (tipicamente si impiega «*»); sono dunque elementi utili quando la ricerca non può o non vuole essere precisa; ad esempio, possiamo inserire la stringa di ricerca «tavol?» per recuperare tutte le occorrenze di «tavolo, tavola, tavoli, tavole».

ordinarne alfabeticamente gli elementi; questo al fine di poter effettuare analisi manuali e qualitative per individuare ad esempio collocazioni, fraseologia, strutture grammatico-lessicali, etc.

– *Collocazioni*
alcuni programmi permettono anche un'estrazione automatica dei collocati relativi all'oggetto dell'interrogazione.

A questo proposito, prima di proseguire, pare opportuna una breve ma specifica trattazione del significato di **collocazione**, per cercare di far luce sul concetto.

3.1 *Collocazioni*

Il concetto di collocazione è stato trattato da vari autori e con approcci spesso distanti oltreché differenti. Nesselhauf (2005) passa in rassegna e cerca di sistematizzare i principali approcci; McKeown e Radev (2000) ne tratteggiano una visione linguistica e una lessicografica, per poi trattarne l'estrazione dai *corpora* e i possibili usi; Manning e Schütze (1999) espongono le varie metodologie statistiche che ne permettono l'individuazione. È poi scontato ma non superfluo ricordare che, tra gli altri, Firth ha aperto la strada a una trattazione dell'argomento e Sinclair lo ha più volte affrontato.

Dandone una definizione del tutto generica, possiamo affermare che il termine collocazione si riferisce a **un qualche tipo di relazione sintagmatica**[7] **tra parole che co-occorrono**. La definizione fornita è volutamente vaga e ha solo la funzione di introdurre il tema, mentre si procederà ora a illustrare gli approcci principali.

Nesselhauf (*op. cit.*: 11-24) identifica due principali approcci al tema delle collocazioni: il primo è stato definito «statistically oriented» o «**frequency-based approach**» e il secondo è stato chiamato «significance oriented» o «**phraseological approach**». Nel primo caso le collocazioni sono identificate come frequenti co-occorrenze di parole a una certa distanza, dove spesso la frequenza viene ulteriormente precisata nei seguenti termini (*ibid.*): «more frequent than could be expected if words were combined randomly in language».

Nel secondo caso le collocazioni sono identificate come un tipo di combinazione tra parole, dove spesso il tipo di combinazione viene precisato come segue (*ibid.*): «[a combination] that is fixed to some degree but not completely».

Il primo approccio è adottato ad esempio da Sinclair e ciò è chiaro dalla sua definizione di collocazione (1991: 170): «the occurrence of two or more words within a short space of each other in a text».

Nella citazione si parla di un breve **spazio**, ma questo non viene meglio precisato. In effetti, la quantità di co-testo presa in considerazione per la ricerca di collocazioni è una delle varianti che ciascun sostenitore dell'approccio *frequency-based* stabilisce. Da parte sua, Sinclair parla di circa quattro parole a destra e a sinistra della parola di cui si vogliano ricercare le collocazioni. Come si può notare dalla definizione data da Sinclair, un'altra variante è costituita dal **numero di parole coinvolte**, oltre al fatto che queste

[7] Ježek (2005: 168) precisa che la dimensione sintagmatica (o orizzontale) della lingua può realizzarsi in sintagmi, ma anche in unità superiori come le frasi e i testi.

debbano o meno essere consecutive. Anche la **frequenza** è oggetto di discussione e di scelta da parte dei singoli fautori di questo approccio; per citare gli estremi, alcuni sostengono che la frequenza debba essere superiore a uno, altri considerano collocazioni co-occorrenze con una frequenza qualsiasi. Infine altro punto di discussione e di scelta è se si debbano prendere in considerazione le **unità lessicali** o le singole **forme declinate**; nel primo caso, "sistema di controllo" e "sistemi di controllo" sarebbero due realizzazioni di una medesima collocazione, mentre nel secondo caso sarebbero due collocazioni distinte.

Il secondo approccio è adottato ad esempio da Ježek: l'autrice (*op. cit.*: 178) identifica le collocazioni come «una combinazione di parole soggetta a **restrizione lessicale»**, ovvero un tipo di combinazione ristretta, in opposizione alle combinazioni libere e alle locuzioni (o espressioni idiomatiche).

È dunque opportuno fare un breve *excursus* per definire cosa s'intende per restrizione: il fatto che le parole non possano essere combinate del tutto liberamente, ma debbano rispettare regole sintattiche e limitazioni semantiche (o congruenze concettuali). Ježek (*op. cit.*: 168-173) distingue tre tipi di restrizioni di tipo semantico, che vincolano la combinazione delle parole: restrizioni concettuali, restrizioni lessicali basate su una solidarietà semantica e restrizioni lessicali basate su una solidarietà consolidata dall'uso. Nel primo caso tali restrizioni derivano «dalle proprietà intrinseche del referente della parola» e sono esemplificate col fatto che un oggetto inanimato come una sedia non può svolgere un'azione come parlare. Nel secondo caso l'autrice si riferisce a quelle che chiama «solidarietà semantiche», che esemplifica col fatto che il verbo "calzare" può essere riferito a scarpe e guanti, ma non a una cravatta, ad esempio. Nel terzo e ultimo caso si tratta di restrizioni legate a combinazioni «caratterizzate da un elemento di convenzionalità», esemplificate dal fatto che si dica "prendere una decisione" ma non "*prendere una scelta".

Come detto poco sopra, secondo Ježek (*op. cit.*: 178) le collocazioni sono combinazioni soggette a restrizioni di tipo lessicale e possono fondarsi su restrizioni **semantiche lessicali** (il secondo dei tre tipi illustrati) o su restrizioni **consolidate dall'uso** (il terzo tipo). In realtà l'autrice considera collocazioni propriamente dette solo quelle fondate su restrizioni consolidate dall'uso.

Ad ogni modo (*ibid.*): «la scelta di una specifica parola (il collocato) per esprimere un determinato significato, è condizionata da una seconda parola (la base) alla quale il significato è riferito». Chi scrive utilizza collocazione proprio secondo questa accezione.

Ježek (*op. cit.*: 179) dichiara che il fenomeno della collocazione è interessante soprattutto dal punto di vista **interlinguistico** e dunque ad esempio nella traduzione e nell'insegnamento. Infatti i componenti di una collocazione non sono liberamente sostituibili e diverse lingue possono presentare differenze anche rilevanti nella scelta dei collocati. McKeown e Radev (*op.cit.*: 12-14) sono dello stesso parere e specificano che le collocazioni sono spesso «language-specific» e aggiungono che, nella maggior parte dei casi, non possono tradursi composizionalmente. Allo stesso modo avvertono che il fatto che in una data lingua un concetto sia espresso mediante una collocazione non significa che questo valga anche in un'altra lingua. Queste osservazioni, oltre alla descrizione del concetto di collocazioni illustrato poco sopra nel presente paragrafo, rivelano l'importanza delle collocazioni per creare testi, scritti o orali che siano, che risultino

adeguati in lingua d'arrivo. Barnbrook (2007: 193), riferendosi all'*idiom principle*[8] di Sinclair, afferma espressamente: «there are clear implications for the modes of production of texts […].».

4. TIPI DI SFRUTTAMENTO DEI *CORPORA*

Come sottolineano O'Keefe, Carter e McCarthy (2006), non c'è un unico *corpus* che soddisfi tutte le esigenze, ma diversi tipi di *corpora* con diverse specifiche (innanzitutto progettuali) possono essere funzionali a diversi tipi di sfruttamento.

I *corpora* possono essere impiegati sia in ambito di ricerca linguistica, sia in contesto di insegnamento/apprendimento, sia nella pratica di traduzione. Nel paragrafo immediatamente successivo ci addentreremo specificamente nell'impiego in ambito di insegnamento/apprendimento, ma prima di ciò è utile dare almeno una panoramica di **alcune delle principali aree d'uso** dei *corpora*, anche in base a Partington (1998c) e O'Keefe, Carter e McCarthy (*ibid.*). Le applicazioni dei *corpora* possono riguardare:

– *Lessico*
  indagini di tipo semantico o lessico-grammaticali mediante analisi di frequenza, concordanze e collocazioni, anche in relazione a diverse tipologie testuali e varietà linguistiche.

– *Grammatica*
  investigazione delle strutture grammaticali e della relativa frequenza, nonché loro diversità di uso in differenti varietà di una stessa lingua.

– *Sintassi*
  esplorazione della combinazione di parole (o persino dei singoli significati di un lemma) in unità di maggiore dimensione legate da «qualche tipo di relazione sintagmatica» (Ježek, *ibid.*).

– *Testualità*
  studio dei fenomeni linguistici a livello del testo.

– *Lingua parlata*
  studio, tra l'altro, di fenomeni legati al parlato come pause, ripetizioni ed espressioni di attenuazione.

– *Registri linguistici*
  analisi con l'obiettivo di confrontare diverse varietà di una stessa lingua o studiare lingue speciali.

– *Lessicografia*
  indagine della frequenza e dell'uso (tra l'altro anche sociolinguistico) di un lemma nelle sue accezioni, con anche estrazione di esempi autentici soprattutto ai fini della creazione di materiali di consultazione.

[8] Sinclair (*op.cit.*: 173) lo ritiene uno dei principi fondamentali nell'organizzazione del linguaggio e consiste nel fatto che la scelta di una parola condiziona la scelta delle parole che compaiono nelle sue vicinanze nel co-testo.

– *Studi stilistici e autorialità*
  analisi che mirano a identificare le caratteristiche peculiari di scrittura di un autore o cercano di identificare l'autore di uno scritto, con particolare riferimento alla linguistica forense (testamenti, richieste di riscatto, casi di plagio, etc.).

– *Studi di tipo storico*
  studi su testi relativi a un dato periodo storico o studi di linguistica diacronica per ricavare informazioni sul cambiamento linguistico.

– *Traduzione*
  impiego di *corpora* (monolingui, comparabili o paralleli) durante il processo di traduzione per estrazione di informazioni linguistiche, semantiche, pragmatiche o impiego per lo studio del processo di traduzione.

– *Sociolinguistica*
  indagine delle variazioni linguistiche legate a diversi fattori sociolinguistici, come ad esempio età, genere, educazione scolastica, ambito socio-economico, etc.

4.1. *Sfruttamento per l'insegnamento*

I diversi approcci nell'impiego di *corpora* per l'insegnamento sono stati trattati a livello teorico da diversi autori, tra cui quelli a cui si farà riferimento nel presente paragrafo, Aston (2000) e McEnery e Xiao (2010), per citarne alcuni piuttosto esaustivi. Molti di loro si rifanno alla classificazione di Leech (1997) che, sulla scia di Fligelstone (1993), suddivide l'utilizzo di *corpora* in ambito di insegnamento/apprendimento in base a tre usi: insegnamento **riguardo** ai *corpora*, **sfruttamento di *corpora*** per l'insegnamento e **insegnamento dello sfruttamento** dei *corpora*. Nel primo caso il riferimento è all'insegnamento della linguistica, in particolare computazionale e applicata; nel secondo caso i *corpora* passano da fine a strumento per l'insegnamento di lingua, linguistica, lingue speciali e CLIL; nel terzo e ultimo caso agli apprendenti viene illustrato come servirsi direttamente dei *corpora* per le loro diverse necessità.

Riprendendo tale classificazione, Römer (2008) si focalizza sulla suddivisione tra impiego di tipo **indiretto** o **diretto**. Nel primo caso i soggetti coinvolti possono essere ricercatori e compilatori di materiali, perciò le 'ricadute' si possono avere a livello di elaborazione di sillabi, opere di riferimento e materiali d'insegnamento, ovvero su che cosa insegnare e quando. Nel secondo caso, quello del cosiddetto *Data-Driven Learning* (DDL), l'interazione può essere tra insegnanti e *corpora* o tra discenti e *corpora*, con 'ricadute' su come si insegna e su come si apprende.

A questo proposito, Partington (*ibid.*) specifica ulteriormente. L'insegnante può sondare la raccolta di testi da sé e per sé, per rispondere a diverse esigenze: 'semplicemente' l'auto-formazione o l'approfondimento di alcuni temi (magari in risposta a qualche quesito posto dagli studenti), verificare i dati fornitigli dalla sua intuizione, ricercare esempi da impiegare in classe, produrre materiali *ad hoc* da utilizzare in momenti di spiegazione o di esercitazione e rinforzo. Nel caso dell'interazione tra discenti e *corpora*, lo sfruttamento può avvenire dentro o fuori il contesto classe e può essere di tipo **mediato** o **non mediato**. Infatti gli apprendenti possono 'metter mano' ai *corpora*, sia sulla scia di una ricerca guidata o proposta dal docente, sia con la libertà di

seguire una traccia dettata da interesse personale o per risolvere difficoltà da loro incontrate in precedenza.

Da un'altra prospettiva, Aston (1997) suggerisce una ulteriore discriminazione: impiego di *corpora* come **riferimento** e **navigazione** di *corpora*. Come riferimento, i *corpora* si rivelano uno strumento di consultazione complementare a più tradizionali dizionari, grammatiche ed enciclopedie, per attingere a esempi e chiarire dubbi specifici. Parlando di strumento di navigazione, è evidente la metafora informatica, infatti Aston (*ibid.*) fa riferimento al fatto che l'esplorazione di *corpora* può rappresentare di per sé un'attività e, come con l'ipertesto, può condurre a esplorazioni successive.

4.1.1. *Osservazioni sullo sfruttamento per l'insegnamento*

Lo sfruttamento dei *corpora* nell'insegnamento/apprendimento comporta dei cambiamenti in tale contesto: sia per quanto riguarda il ruolo dell'insegnante sia per quanto riguarda il ruolo dell'apprendente, sia nel processo di apprendimento sia nella metodologia di insegnamento.

Per quanto riguarda **l'insegnante**, questi assume un ruolo di guida, che Laviosa (1999) definisce anche «facilitatore» e Ramamoorthy (2009) «coordinatore». Ciò avviene soprattutto in una situazione di esplorazione di *corpora* da parte degli studenti, mediata dall'insegnante; tale ruolo si addice anche a quanto poco sopra definito insegnamento dello sfruttamento, momento in cui il docente, per avviare i discenti allo sfruttamento autonomo o para-autonomo dei *corpora*, li introduce agli strumenti e alle tecniche necessarie; in particolare, il docente dovrà illustrare l'utilizzo di determinati software, *concordancer*, che possono essere a disposizione in ambito scolastico o eventualmente reperibili on-line. Non bisogna inoltre dimenticare che l'insegnante dovrà anche essere in grado di creare i presupposti per invogliare i discenti all'indagine dei *corpora*, creando le condizioni affinché ciò risulti per loro significativo. Bernardini (2000) annota come questa nuova prospettiva possa avere anche una lettura più generale, ovvero l'insegnante, in realtà, guidando gli studenti allo sfruttamento dei *corpora*, li indirizza a rendere le loro strategie di apprendimento più efficaci. La figura dell'insegnante si rivela sempre più poliedrica.

L'utilizzo di corpora nella didattica influisce in profondità anche sul ruolo del **discente**: questi è stato paragonato a un esploratore e a un viaggiatore (Bernardini, 2004), nonché a una sorta di ricercatore (Partington, *ibid.*). Condividendo questa prospettiva, Laviosa (*ibid.*) lo dichiara responsabile in prima persona delle investigazioni relative a dati che gli interessino e Ramamoorthy (*ibid.*) sostiene che il discente impara ad imparare esercitando le sue capacità di osservazione e interpretazione. Portato dunque a partecipare attivamente al processo di apprendimento, sarà avviato a una gestione autonoma dello stesso, come sostiene, tra gli altri, Römer (*ibid.*); raggiunta sufficiente autonomia potrà inoltre modellare il proprio apprendimento in base alle sue preferenze cognitive, ai suoi interessi e stili di apprendimento, come specifica Bernardini (2000). Bisogna sicuramente notare che, ad esempio, anche una ricostruzione di conversazione o la scoperta di una regola ricercando verbi in un testo sono attività che insegnano al discente a imparare, mediante l'osservazione e l'interpretazione. L'impiego di *corpora* offre però qualcosa in più e qualcosa di diverso: mette a disposizione dati autentici e in grande quantità (sempre relativamente alle dimensioni del *corpus* in esame) e permette

allo studente di acquisire maggiore familiarità con le possibilità di ricerca e con i suoi specifici interessi, nonché maggiore autonomia nella ricerca stessa, man mano che il ruolo di guida dell'insegnante viene meno e il discente accede direttamente a dati non più pre-selezionati, fino a diventare, appunto, una sorta di ricercatore.

Per quanto riguarda la **metodologia di insegnamento**, Carter e McCarthy (1995) sottolineano il passaggio dall'approccio detto delle tre P (*presentation*, *practice*, *production*) a quello delle tre I (*illustration*, *interaction*, *induction*). Il primo, di stampo tradizionale, si compone di presentazione (da parte dell'insegnante) seguita da pratica e produzione (da parte degli studenti) ed è ancora necessario soprattutto nell'ambito di esercizi di rinforzo, ma va complementato con il secondo, di tipo esplorativo, scandito da illustrazione, interazione e induzione, spiegano gli autori. Con «illustrazione» ci riferisce al fatto di prendere in considerazione testi o concordanze estratte da *corpora* per stimolare i discenti a notare di volta in volta determinate strutture e comportamenti linguistici; con «interazione» ci riferisce alla condivisione di osservazioni e opinioni (tra i discenti e con l'insegnante); con «induzione» ci si riferisce al fatto di ricavare generalizzazioni che saranno via via perfezionate man mano che si incontreranno maggiori dati che le confermeranno; in alternativa, potranno essere modificate e riviste nel caso nuovi dati le smentiscano in tutto o parzialmente. In base a questa visione, il **processo di apprendimento** procede dunque mediante successive e parziali generalizzazioni, sulla via verso l'ottenimento di una regola pienamente soddisfacente, almeno fino a prova contraria.

Questo procedimento di tipo esplorativo e induttivo, procede dal basso verso l'alto, ovvero dai dati alle generalizzazioni, come rilevano McEnery e Xiao (*ibid.*). La considerazione che sembra maggiormente rilevante e che può dare adito a qualche perplessità, come rilevato anche da O'Keefe, Carter e McCarthy (*ibid.*), è il venir meno di una rigida contrapposizione tra giusto e sbagliato, in favore di un'ottica che mette in luce il probabilismo (di occorrenza di unità linguistiche, di loro associazioni, etc.) e porta ad operare scelte in base ai dati analizzati e alle astrazioni ricavate dall'analisi, o meglio dalla combinazione di *illustration* e *interaction*.

4.1.2. *Pro e contro l'impiego diretto*

Sulla scia delle perplessità sollevate dal fatto che gli studenti abbiano a che fare con probabilità e scelte, è utile passare in rassegna quali possono essere vantaggi e problematiche legate all'impiego diretto dei *corpora* in contesto di insegnamento/apprendimento.

L'impiego dei *corpora* legato a un'ottica probabilistica e al fatto di non voler o poter dare sempre solidi riferimenti in termini di giusto o sbagliato ha spesso portato a confinarne l'impiego **a livelli avanzati**, ritenendo che a livelli più alti fosse in qualche modo più facile procedere induttivamente mentre a livelli base ci fosse il rischio di creare confusione (McEnery e Xiao, *ibid.*). In realtà, un maggiore impiego anche a livelli base sarebbe auspicabile.

L'importanza di impiegare l'analisi di *corpora*, concordanze estratte dagli stessi e materiali basati su di essi è data da diversi vantaggi: affidabilità e rappresentatività della lingua nei *corpora*; risveglio della motivazione e spinta all'autonomia degli studenti; efficacia del lavoro sui *corpora*. Per quanto riguarda **affidabilità** e **rappresentatività**,

queste devono evidentemente essere garantite 'a monte', ovvero dalla progettazione e realizzazione dei *corpora* stessi, secondo i parametri specificati nei paragrafi iniziali e secondo specifici parametri dettati dalle finalità previste per le singole raccolte di testi. Per quanto riguarda l'**autonomia** degli studenti è già stato accennato come certi tipi di sfruttamento diretto, più o meno mediato dall'insegnante, siano volti a massimizzarla; per quanto riguarda la **motivazione**, questa è strettamente legata alla spinta verso l'autonomia. Infatti, la ricerca autonoma o pseudo-autonoma legata a materiali autentici e magari anche a compiti autentici, ad esempio diretti alla risoluzione di problematiche sperimentate dagli stessi discenti, sono fattori chiave per la motivazione degli apprendenti, come rileva anche Aston (2000).

Eventualmente, a livelli base, c'è la necessità di una **guida** più salda da parte dell'insegnante. Per non creare infatti situazioni di frustrazione (cfr Bernardini, *ibid.*: 230) e per non fare affidamento esclusivamente su procedimenti induttivi, è necessario che i compiti siano attentamente graduati e che i materiali di lavoro siano stati accuratamente pre-selezionati dall'insegnante, in particolare le concordanze.

Osborne (2000) predilige poi un tipo di **approccio misto**, che associ compiti induttivi a spiegazioni vere e proprie e si rimanda al suo intervento per quanto riguarda la realizzazione di un software che mira proprio a questo, ovvero una sorta di *concordancer* in cui sono presenti una serie di link che forniscono spiegazioni di tipo metalinguistico e mirano a guidare l'interpretazione dei dati.

Bernardini (*ibid.*) porta un altro interessante esempio di impiego non mediato nell'apprendimento dell'inglese, testimoniando l'utilizzo di *corpora* da parte degli studenti per ricerche libere nella modalità di **navigazione** cui si è accennato nel paragrafo 4.1. Bernardini (*ibid.*) rileva l'**efficacia** delle attività, con effetti di maggiore durata legati al coinvolgimento diretto e attivo dei discenti. Contemporaneamente, rileva problemi legati all'ottenimento di troppi o troppo pochi risultati e precisa la necessità di intervento dell'insegnante per indicare come specificare o generalizzare le interrogazioni.

In realtà, bisogna rilevare che un lavoro sui *corpora* dipende fortemente dalla **preparazione dei docenti** e dalla capacità degli stessi di formare gli alunni nell'uso di pacchetti informatici. Ovviamente, è necessario che docenti e studenti abbiano a loro disposizione le **risorse necessarie**: computer, pacchetti software e *corpora*. Oggigiorno sono disponibili anche software gratuiti e *corpora* scaricabili o consultabili online, ma la formazione dei docenti e, evidentemente, l'accessibilità di postazioni a PC restano essenziali.

5. CONCLUSIONI

Questo intervento è nato dalla volontà di approfondire e sostenere l'impiego, in particolare diretto, dei *corpora* in ambito di insegnamento/apprendimento.

Ci si auspica che questa rassegna su cosa sono i *corpora* e come possono essere impiegati possa essere utile per introdurre al concetto di *corpora* e alle relative possibilità di sfruttamento chi ancora non conosca questo ambito e le possibilità che offre e, in particolare, per sollecitarne l'impiego in ambito di insegnamento/apprendimento.

Questo inquadramento teorico ha cercato di fornire un quadro il più possibile esaustivo, mettendo le basi per un futuro approfondimento del tema da declinare nella

pratica dell'impiego diretto di *corpora* sia in aula sia al di fuori di questa per un lavoro, pseudo-autonomo prima e autonomo poi, degli studenti.

RIFERIMENTI BIBLIOGRAFICI

Aston G. (1997), "Enriching the learning environment: Corpora in ELT", in Wichmann A., Fligelstone S. e McEnery T. (1997), *Teaching and Language Corpora*, Longman, London, pp. 51-64.

Aston G. (1999), "Corpus use and learning to translate" in *Textus*, 12, pp. 289-314.

Aston G. (2000), "Corpora and language teaching", in Burnard L. e McEnery T. (a cura di), *Rethinking language pedagogy from a corpus perspective: papers from the 3rd international conference on Teaching and language corpora*, Lang, Francoforte.

Baker P. Hardie A. e McEnery T. (2006), *A Glossary of Corpus Linguistics*, Edinburgh University Press, Edinburgh.

Barnbrook G. (2007), "Sinclair on collocation", in *International Journal of Corpus Linguistics*, 12, 2, pp. 183-199.

Bernardini S. (2000), "Systematising serendipity: Proposals for concordancing large corpora with language learners", in Burnard L. e McEnery T. (a cura di), *Rethinking language pedagogy from a corpus perspective: papers from the 3rd international conference on Teaching and language corpora*, Lang, Francoforte.

Bernardini S. (2004), "Corpora in the classroom: An overview and some reflections on future development", in Sinclair J. (a cura di), *How to use Corpora in Language Teaching*, John Benjamins, Amsterdam e Philadelphia.

Biber D. Conrad S. e Reppen R. (1998), *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge.

Bowker L. e Pearson J. (2002), *Working with Specialized Language: a practical guide to using corpora* , Routledge, London e New York.

Burnard L. e McEnery T. (a cura di) (2000), *Rethinking language pedagogy from a corpus perspective: papers from the 3rd international conference on Teaching and language corpora*, Lang, Francoforte.

Carter R. e McCarthy M. (1995), "Grammar and spoken language", in *Applied Linguistics*, 16, 2, pp. 141 -158.

Corino E. e Marello C. (2009), "Didattica con i corpora di italiano per stranieri", in *Italiano LinguaDue*, 1, pp. 279-285.

Daskalovska N. (2009), "Using corpora to teach collocations", *Online coverage of the 43rd Annual International IATEFL Conference in Cardiff.*

Fligelstone S. (1993), "Some reflections on the question of teaching, from a corpus linguistics perspective", in *ICAME Journal*, 17, pp. 97-110.

Hunston S. (2002), *Corpora in Applied Linguistics*, Cambridge University Press Cambridge.

Ježek E. (2005), *Lessico: classi di parole, strutture, combinazioni*, Il Mulino, Bologna.

Laviosa S. (1999), "Come studiare e insegnare l'italiano attraverso i corpora", in *Italica*, 76, 4, pp. 443-453.

Leech G. (1997), "Teaching and language corpora: a convergence", in Wichmann A. Fligelstone S. McEnery T. e Knowles G. (a cura di), *Teaching and Language Corpora*, Longman, London, pp. 1-23.

Lewis M. (1997), *Implementing the Lexical Approach: Putting Theory into Practice*, Language Teaching Publications, Hove.

MacMullen J. (2003), "Requirements definition and design criteria for test corpora in information science", on-line all'indirizzo: http://sils.unc.edu/sites/default/files/general/research/TR-2003-03.pdf (visitato 02/01/2010).

Manning C. e Schütze H. (1999), "Collocations" in Manning C. e Schütze H. (a cura di), *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge (Massachusetts), pp. 141-177.

McEnery T. e Wilson A. (2001), *Corpus Linguistics: an Introduction*, Edinburgh University Press, Edinburgh.

McEnery T. e Xiao R. (2010), "What corpora can offer in language teaching and learning", in Hinkel E. (a cura di), *Handbook of Research in Second Language Teaching and Learning*, 2, Routledge, London e New York.

McKeown K. R. e Radev D. R. (2000), "Collocations", on-line all'indirizzo: http://tangra.si.umich.edu/~radev/papers/handbook00.ps (visitato 29/12/2010)

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

O'Keefe A. Carter R. e McCarthy M. (2006), *From Corpus to Classroom. Language Use and Language Teaching*, Cambridge University Press, Cambridge.

Osborne J. (2000), "What can students learn from a corpus?: Building bridges between data and explanation", in Burnard L. e McEnery T. (a cura di), *Rethinking language pedagogy from a corpus perspective: papers from the 3rd international conference on Teaching and language corpora*, Lang, Francoforte.

Partington A. (1998c), *Patterns and Meanings: Using Corpora for English Language Research and Teaching*, Benjamins, Amsterdam.

Prada M. (2010), "LIPSI. Il lessico di frequenza dell'italiano parlato in Svizzera", in *Italiano LinguaDue*, 1, pp. 182-205.

Ramamoorthy L. (2009), "Language teaching through corpora", in *Language in India*, 9, 11.

Römer U. (2008), "Corpora and language teaching", in Lüdeling A. e Kytö M. (a cura di), *Corpus Linguistics. An International Handbook, Vol. 1*, Mouton de Gruyter, Berlin.

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

Viganò P. B. (2009), "Collocazioni in un data base per traduttori", in Nedelcu I., Nicolae A, Toma A. e Zafiu R. (a cura di), *Studii de lingvistică. Omagiu doamnei profesoare Angela Bidu-Vrănceanu*, Editura Universităţii din Bucureşti, Bucureşti, pp. 349-371.