CORESPI e CORITE, due nuovi strumenti per l'analisi dell'interlingua di lingue affini

Sonia Bailini, Aldo Frigerio[1]
Università Cattolica del S. Cuore di Milano

In this paper, two parallel longitudinal corpora of interlingua are presented: the CORpus del ESPañol de los Italianos (CORESPI, 474 texts, 124 648 words) and the CORpus del Italiano de los Españoles (CORITE, 385 texts, 103 147 words). The texts have been produced by 45 couples of Spanish and Italian as foreign language learners (A1- B2). The manner in which the texts have been collected, tagged and put on the Internet using the tool CATMA are illustrated. CATMA potentialities are briefly surveyed. The usefulness of this kind of corpora for the linguistic and parallel analysis and for language teaching theory is emphasized.

Keywords: CORITE, CORESPI, learner corpora, second language acquisition, interlanguage

1. CORESPI e CORITE: criteri di raccolta dei dati e caratteristiche

Il *CORpus del ESPañol de los Italianos* (CORESPI, 474 testi, 124 648 parole) e il *CORpus del ITaliano de los Españoles* (CORITE, 385 testi, 103 147 parole) sono due corpora di interlingua longitudinali paralleli costituiti dai testi di 45 coppie di apprendenti di spagnolo e italiano come lingua straniera (A1 – B2, QCERL 2001) raccolti attraverso il metodo tandem (Hehmann *et. al.* 1997; Hehmann *et. al.* 2003) nel corso di un anno accademico. Come confermano gli studi più recenti (Brezina y Flowerdew 2018; Granger *et al*. 2017; Castello *et al.* 2015), i *learner corpora* sono strumenti utili per l'analisi dell'interlingua perché permettono di osservarne caratteristiche specifiche frequenti nelle varie tappe evolutive su un campione elevato di testi. Tuttavia, come afferma Alonso Ramos (2016: 3), sono

[1] Sebbene questo lavoro sia frutto della collaborazione di entrambi gli autori, Bailini ha curato in particolare il par. 1 e Frigerio il par. 2.

CHIMERA. Romance Corpora and Linguistic Studies 5.2 (2018), 313-319. ISSN 2386-2629

ancora relativamente pochi i *learner corpora* che raccolgono dati provenienti da apprendenti di spagnolo come lingua straniera la cui L1 sia diversa dall'inglese)[2]. Inoltre, in molti casi si tratta di corpora che raccolgono l'interlingua di informanti con diverse lingue materne e, tra questi, sono pochi quelli che contengono anche l'interlingua dello spagnolo di italofoni (CORANE, Cesteros Mancera & Penadés Martínez 2009; Corpus ESPALEX, Bustos Gisbert & Sánchez Iglesias 2012). Lo stesso si può dire per i *learner corpora* che raccolgono l'interlingua di apprendenti ispanofoni di italiano (LIPS, Vedovelli *et. al.*; VALICO; ADIL2, Palermo 2009*)*. Ad oggi, gli unici *learner corpora* esclusivamente dedicati all'interlingua dello spagnolo di italofoni sono il *Corpus E.L.E.I. (Español Lengua Extranjera en Italia*) di Martín Sánchez e Solís García (2013) e il *Corpus de conversaciones en italiano y en español/LE* (*C.I.E.L.E*) di Pascual Escagedo (2013). Per quanto riguarda invece l'interlingua dell'italiano di ispanofoni ricordiamo il *Corpus A.MA.DIS*, elaborato dal *Grupo A.MA.DIS* della Universidad Complutense di Madrid. Ancora meno sono i *learner corpora* paralleli relativi a queste due interlingue: a *CORINÉI* (*Corpus orale di interlingua spagnolo e italiano*), frutto di un progetto congiunto tra l'Università di Salerno, la Suor Orsola Benincasa di Napoli e l'Università di Alicante, si aggiungono ora CORESPI e CORITE, che raccolgono l'interlingua scritta di questi due gruppi di apprendenti.

I grafici seguenti (Fig. 1) mostrano la distribuzione dei testi per livello in ciascuno dei due corpora:

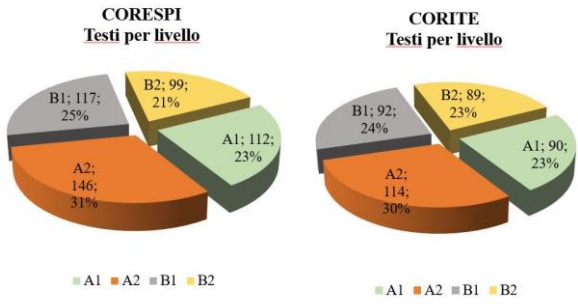

Figura 1. CORESPI e CORITE: distribuzione testi per livello

[2] Per un quadro completo sui learner corpora si veda Centre for English Corpus Linguistics (18/06/2018): Learner Corpora around the World. Louvain-la-Neuve: Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (Accesso 18 giugno 2018).

I 90 informanti sono stati suddivisi in 45 coppie (11 di livello A1, 12 di A2, 14 di B1 e 8 di B2) formate da madrelingua/non madrelingua in base al loro livello di competenza nelle rispettive LS e allo stile di apprendimento dominante, rilevato attraverso un questionario specificamente predisposto.[3] Va segnalato che, per tutti gli informanti, lo spagnolo o l'italiano rappresentavano la terza lingua straniera (dopo l'inglese per il 100% degli informanti italiani e l'84% degli spagnoli e dopo il francese per il 30% degli italiani e il 39% degli spagnoli) e che buona parte degli ispanofoni (75%) conosceva anche il catalano.

CORESPI e CORITE permettono sia un'analisi trasversale che longitudinale di entrambe le interlingue, poiché per la maggior parte degli informanti sono stati raccolti da un minimo di 5 a un massimo di 25 testi. Infatti, il 79% di essi (34 di CORESPI e 30 di CORITE) ha scritto tra i 7 e i 16 testi, che, su un periodo di 7-8 mesi di raccolta dati, corrisponde a uno o due testi al mese, mentre solo 6 informanti hanno scritto 5 o 6 testi nello stesso arco di tempo, cioè meno di un testo al mese, e solo due coppie hanno elaborato più di 16 testi.

Per l'elaborazione dei testi gli informanti potevano utilizzare qualsiasi tipo di supporto linguistico (dizionari, manuali, grammatiche, etc.), senza alcun limite di estensione né di tempo, sapevano che le loro produzioni non sarebbero state oggetto di valutazione e non erano a conoscenza del fatto che sarebbero state utilizzate per fini di ricerca. Poiché erano testi indirizzati ad un interlocutore reale, possono essere classificati come "lettere informali", anche se è possibile riconoscere cinque macro-generi testuali (narrativo, descrittivo, argomentativo, espositivo e istruttivo). Inoltre, in base al loro contenuto principale, i testi sono stati classificati in quattordici aree tematiche. Queste caratteristiche permettono di restringere la ricerca, oltre che per livello e caratteristiche dell'informante, a testi in cui ci sono maggiori probabilità di trovare un certo tipo di strutture linguistiche o lessico relativo a un determinato campo semantico. Nel prossimo paragrafo verranno presentate le caratteristiche dell'interfaccia attraverso la quale è possibile consultare i due corpora.

2. CATMA, l'interfaccia per l'analisi di CORESPI e CORITE

Per l'etichettatura e la messa on line dei due corpora si è scelto di usare il software CATMA (*Computer Assisted Textual Markup and Analysis*: www.catma.de), sviluppato dall'Università di Amburgo. La scelta è ricaduta su questo strumento per almeno quattro ragioni:

[3] Luciano Mariani, http://www.learningpaths.org/Questionari/stilil2compatto.htm (Accesso 30 aprile 2018).

a. L'etichettura viene effettuata in maniera visuale ed è quindi una operazione relativamente semplice;
b. Le etichette sono facilmente modificabili anche dopo essere state applicate;
c. Permette di effettuare ricerche complesse, utilizzando molteplici criteri;
d. Permette di condividere i corpora fra più utenti.

I due corpora sono stati etichettati per informante, per livello di conoscenza della lingua straniera, per genere, per età, per altre lingue conosciute e per parole tematiche, relative agli argomenti di cui un testo tratta. CATMA permette di realizzare vari tipi di ricerche:

- È possibile visualizzare la lista delle parole di ciascun corpus e il numero di occorrenze di ciascuna di esse. Ciò è importante in un *learner corpus*, che contiene inevitabilmente molte parole scritte in modo scorretto.
- È possibile ricercare parole simili a una parola data o parole che iniziano o finiscono o che contengono una certa stringa. Anche questo tipo di ricerche può essere utile in un *learner corpus*.
- È possibile ricercare stringhe costituite da due o più parole intervallate da un numero *n* di parole.
- È possibile visualizzare, per ognuna delle parole di un corpus, i contesti in cui quella parola occorre. Inoltre, è possibile allargare o restringere la finestra del contesto in cui la parola compare e anche visualizzare l'intero testo.
- È possibile cercare parole o stringhe di parole che corrispondano a certi criteri, cioè che siano state taggate in un certo modo. È possibile, inoltre, incrociare più criteri. Per esempio, è possibile ricercare tutte le occorrenze di una certa parola contenute in testi di informanti di livello B2 oppure le occorrenze di una certa parola contenute in testi di informanti di livello B2 che conoscano anche una determinata lingua straniera.

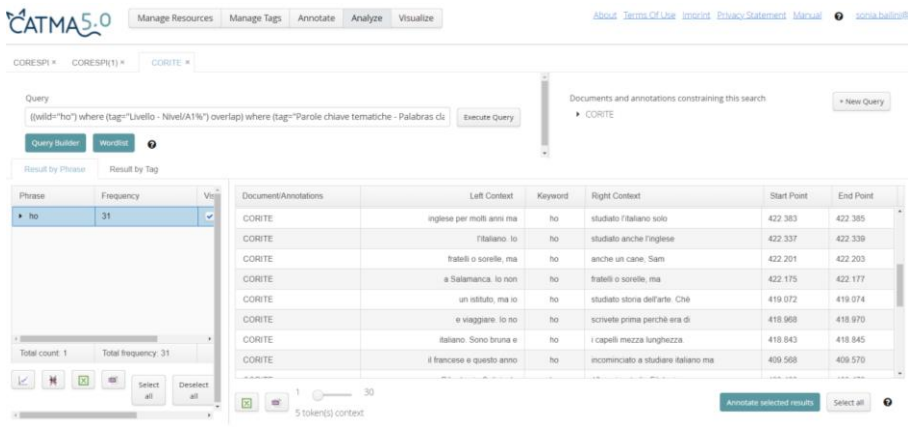

Figura 2. Un esempio di ricerca in CATMA (occorrenze di "ho" in CORITE, utilizzate da informanti di livello A1 in testi che parlano di "informazioni personali".

- CATMA ha una funzione chiamata "KWIC as double tree". Essa permette all'utente di visualizzare un doppio albero al centro del quale compare l'item cercato. Ognuna delle foglie dell'albero è costituita da una delle parole che segue o precede nel corpus quell'item. Quanto maggiore è la dimensione della parola tanto più è alto il numero di occorrenze che compaiono prima o dopo di essa. Cliccando su una delle parole che la seguono o la precedono viene visualizzata l'intera stringa, in cui vengono evidenziate in rosso le parole che nel corpus seguono o precedono la parola che è al centro dell'albero. Le figure 3 e 4 mostrano i risultati di una ricerca realizzata con gli stessi parametri in CORESPI e CORITE così come viene visualizzata con la funzione "KWIC as a double tree". Nell'esempio sono stati selezionati rispettivamente "ho" in CORITE e "tengo" in CORESPI in informanti di livello A1 in testi taggati come "informazioni personali":

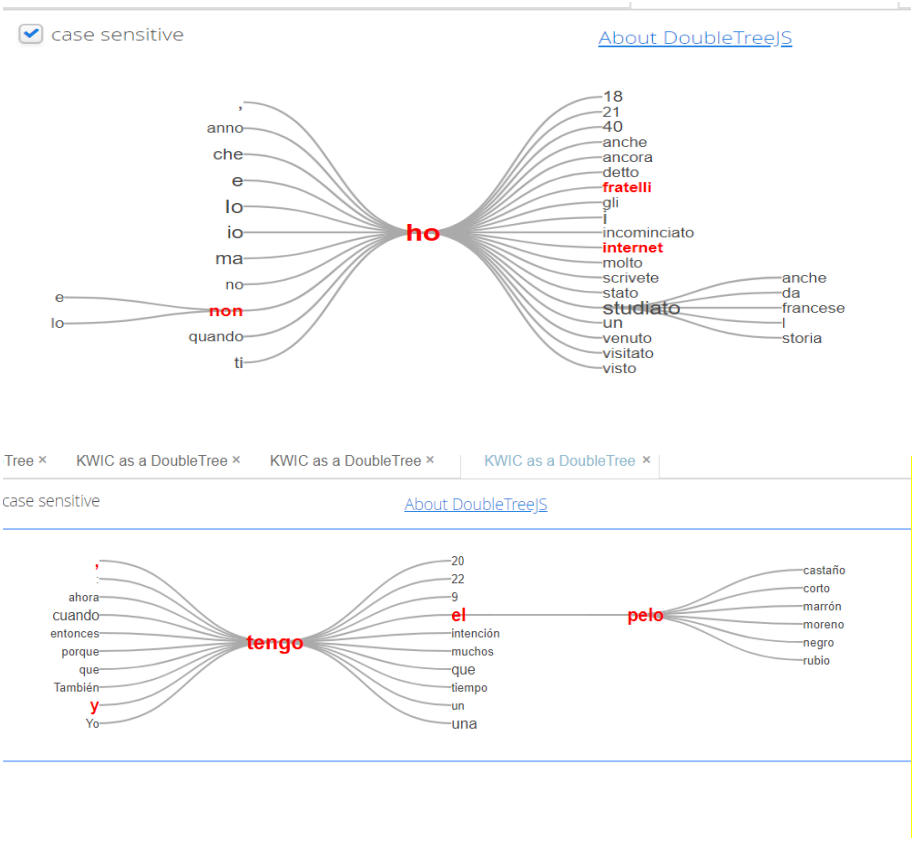

Figure 3 e 4. KWIC di “ho” (CORITE) e “tengo” (CORESPI) in livello A1, tag “info pers”.

3. Conclusioni

In questo contributo sono stati illustrati brevemente i criteri di costruzione di CORESPI e CORITE e le caratteristiche del software CATMA per la loro consultazione on line con l’obiettivo di mostrarne le potenzialità per l’analisi delle varie tappe evolutive delle interlingue di lingue affini. L’osservazione empirica dei dati, infatti, è la base di partenza per lo studio delle interferenze -sia positive che negative- che inevitabilmente si manifestano nel processo di acquisizione di lingue tipologicamente poco distanti dalla L1 dell’apprendente.

Bibliografia

Alonso Ramos, M. (ed.) 2016. *Spanish Learner Corpus Research. Current trends and future perspectives*. Amsterdam/Philadelphia: John Benjamins.

A.Ma.Dis, *Adquisición de Marcadores Discursivos,* http://www.marcadores-discursivos.es/ (Accesso 18 giugno 2018).

Brezina V. & Flowerdew L. (eds) 2018. *Learner Corpus Research. New Perspectives and Applications*. London/New York: Bloomsbury.

Bustos Gisbert, J. & Sánchez Iglesias, J.J. 2012. ESPALEX: un corpus para la adquisición del español como lengua extranjera. In C. Hernández González, A. Carrasco Santana & E. Álvarez Ramos (eds), *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera. Actas del XXII Congreso ASELE*. Valladolid: ASELE/Universidad de Valladolid, 149-159.

Castello E., Ackerley K. & Coccetta, F. (eds) 2015. *Studies in Learner Corpus Linguistics*. Bern: Peter Lang.

Cesteros Mancera, A. & Penadés Martínez, I. 2009. *Corpus de textos escritos para el análisis de errores de aprendices de E/LE (CORANE)*. Alcalá: Universidad de Alcalá.

Granger S., Gilquin G. & Meunier, F. (eds) 2017. *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

Hehmann G. *et al*. (eds) 1997. *Guida per imparare le lingue in Tandem via Internet.* Torino: Trauben.

Hehmann G. & Ponti, D. 2003. *Apprendimento autonomo delle lingue in tandem*. Torino: Trauben.

Martín Sánchez, T. & Solís García, I. 2013. Un corpus de interlengua de español en Italia. In M. V. Calvi, A. Cancellier, & E. Liverani (eds), *Frontiere: soglie e interazioni. I linguaggi* ispanici *nella tradizione e nella contemporaneitá. Atti del XXVI Convegno AISPI.* Trento: Università degli Studi di Trento, 255-265.

Palermo, M. (2009). *Percorsi e strategie di apprendimento dell'italiano lingua seconda: sondaggi su ADIL2*. Perugia: Guerra.

Pascual Escagedo, C. 2013. El corpus oral E.L.E.I. In M. V. Calvi, A. Cancellier & E. Liverani (eds), *Frontiere: soglie e interazioni. I linguaggi ispanici nella tradizione e nella contemporaneitá. Atti del XXVI Convegno AISPI*. Trento: Università degli Studi di Trento, 313-328.

VALICO (*Varietà di Apprendimento della Lingua Italiana*) www.valico.org (Accesso 18 giugno 2018).

Vedovelli, M. *et al. Corpus LIPS*, http://www.parlaritaliano.it/index.php/it/corpora-di-parlato/653-corpus-lips (Accesso 18 giugno 2018).