E-Aesla n.º 5 (2019)

CORESPI Y *CORITE*: CRITERIOS DE CONSTRUCCIÓN E IMPLEMENTACIÓN EN LÍNEA DE DOS CORPUS DE INTERLENGUA PARALELOS

Sonia Bailini
Università Cattolica del Sacro Cuore
Aldo Frigerio[1]
Università Cattolica del Sacro Cuore

RESUMEN

En este artículo se presentan los criterios de construcción y las características de dos corpus de interlengua escrita paralelos y longitudinales: el CORpus del ESPañol de los Italianos (CORESPI) y el CORpus del ITaliano de los Españoles (CORITE). Asimismo, se muestran las funciones de la herramienta CATMA para su análisis con el objetivo de destacar su potencial para la investigación en adquisición de segundas lenguas.

Palabras clave: corpus de hablantes no nativos, interlengua, español e italiano como LE, lenguas afines

ABSTRACT

This article presents the criteria of building and the features of two parallel and longitudinal corpora of written interlanguage: the CORpus del ESPañol de los Italianos (CORESPI) and the CORpus del ITaliano de los Españoles (CORITE). Moreover, the functions of the CATMA tool for learner corpora analysis are shown in order to highlight its potential for second language acquisition research.

Keywords: learner corpora, interlanguage, Spanish and Italian as a foreign language, cognate languages

[1]Sonia Bailini recogió y clasificó los datos de ambos corpus. Aldo Frigerio se encargó de su implementación en línea a través de *CATMA*. Sonia Bailini es autora de los apartados 1 y 3 y Aldo Frigerio del apartado 2.

1. *CORESPI* Y *CORITE:* CRITERIOS DE CONSTRUCCIÓN Y CARACTERÍSTICAS

Si, por un lado, los corpus de hablantes no nativos son un ámbito de investigación relativamente nuevo (Granger et al., 2017; Castello et al., 2015; Brezina y Flowerdew, 2018), por otro, como afirman Granger (2015) y Alonso Ramos (2016), lo son aún más cuando recopilan datos procedentes de aprendientes cuya primera (L1) o segunda (L2) lengua son distintas del inglés. Los dos corpus que se presentan en este artículo forman parte de este segundo grupo y pretenden ser una herramienta más para la investigación en adquisición de segundas lenguas, especialmente cuando éstas son afines a la L1 de los aprendientes.

El marco teórico de referencia es el Análisis Contrastivo de la Interlengua (CIA, Granger, 1996, 2015), ya que la comparación entre ambos corpus permite determinar si las particularidades o déficits en una interlengua concreta (del español o del italiano) resultan de la influencia de la L1, si son consecuencia del *input* o de patrones de desarrollo universales.

El *CORpus del ESPañol de los Italianos* (*CORESPI*) y el *CORpus del Italiano de los Españoles* (*CORITE*) empiezan a recogerse en el año 2008/2009 a raíz de la necesidad de recopilar datos empíricos lo suficientemente representativos para el estudio de la interlengua de aprendientes de lenguas afines (Bailini, 2012).

El e-tándem por correo electrónico (Hehmann et. al., 1997 y 2003) se configuró desde el principio como el método más adecuado para recoger textos que reflejaran una comunicación bidireccional equilibrada, auténtica y monitorizable, al mismo tiempo que una modalidad de interacción que permitiera recoger *output* con características parecidas.

Los 90 informantes estaban aprendiendo sus respectivas LE en contextos de aprendizaje formales y fueron organizados en 45 parejas de hablante nativo/no nativo de español e italiano a partir de su nivel de competencia comunicativa: 11 parejas de nivel A1, 12 del A2, 14 del B1 y 8 del B2. También se tuvieron en cuenta sus estilos cognitivos[2], la edad (entre 18-40 años) y el conocimiento de otras lenguas extranjeras: con respecto a este último criterio cabe señalar que el español o el italiano representaban la L3 para todos los informantes.

[2]Se ha utilizado el cuestionario elaborado por Luciano Mariani, y que se encuentra disponible en http://www.learningpaths.org/Questionari/stilil2compatto.htm [30/04/2018].

La sistematización de los textos recogidos se realizó en tres etapas:

1) En un primer momento se reunieron todos los textos producidos por aquellos aprendientes italianos de español/LE que respetaran las condiciones mínimas para un seguimiento longitudinal de la evolución de su interlengua, es decir, aquellos que hubiesen escrito de un mínimo de 5 a un máximo de 25 textos a lo largo de los 7-8 meses en que se recogieron los datos. En esta primera fase se reunieron, bajo el nombre *Spanish Corpus of Italian Learners* (*SCIL*), 457 composiciones escritas (124 186 palabras) realizadas por 43 informantes (Bailini, 2013).

2) En un momento sucesivo se reunieron los textos de aproximadamente el 50% de los informantes (24 parejas) y se comprobó que los textos de ambos corpus fuesen efectivamente homogéneos y paralelos por extensión, géneros textuales y temas tratados, pues no hay que olvidar que se trata de textos de interlengua cuyos estímulos a la producción eran en parte semi-guiados y en parte libres. Se reunieron en esta fase la primera parte de *CORESPI* (293 textos, 82 203 palabras) y de *CORITE* (257 textos, 70 307 palabras) (Bailini, 2018).

3) En la tercera etapa, que es la que se presenta en este artículo, se añadieron los textos de las restantes 21 parejas de informantes, hasta sumar 474 textos (124 648 palabras) en *CORESPI* y 385 textos (103 147 palabras) en *CORITE*, por un total de 45 parejas. Además, ambos corpus están, a partir de ahora, disponibles en línea en acceso abierto desde la página web http://corespiycorite.altervista.org.

En los gráficos siguientes (véase la Figura 1) se muestra la repartición de textos por cada nivel en sendos corpus:

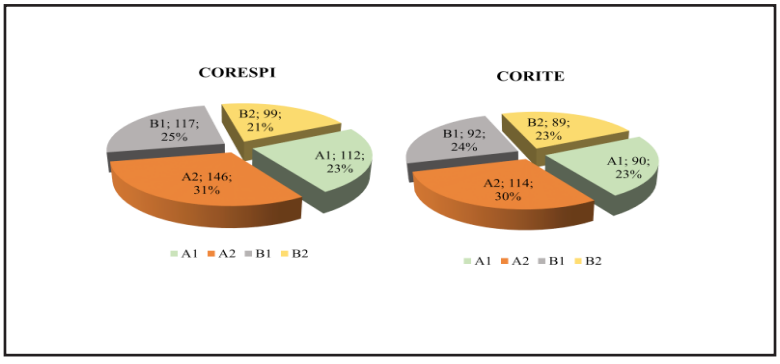

Figura 1. CORESPI y CORITE: repartición de los textos por nivel

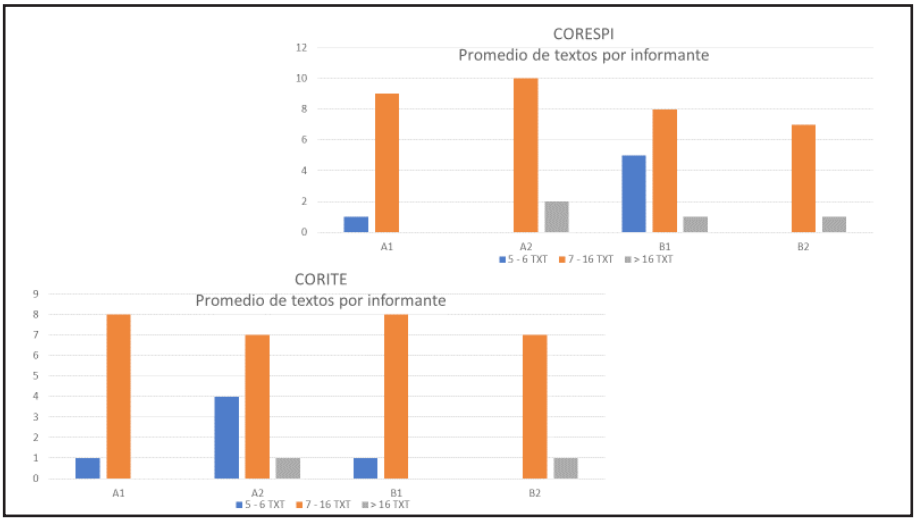

Figura 2. CORESPI y CORITE: promedio de textos por informante

En la Figura 2, además, se muestra el promedio de textos por informante, pues es un dato útil para el análisis longitudinal, es decir, el seguimiento de la evolución de la interlengua de un informante concreto. Se han incluido en este grupo solo aquellos informantes que hubiesen escrito al menos 5 textos en un período de 7-8 meses (43 en *CORESPI* y 38 en *CORITE*). Como se puede ver, la mayoría de ellos (34 informantes de *CORESPI* y 30 de *CORITE*, lo que equivale al 79%) produjo entre 7 y 16 textos, o sea entre uno y dos textos al mes, mientras que tan solo 6 informantes produjeron entre 5 y 6 textos (14% en *CORESPI* y 16% en *CORITE*), o sea menos de un texto al mes, y solo dos parejas escribieron más de 16 textos.

Por último, señalamos las características principales de los textos: en primer lugar, cabe decir que se trata de producciones escritas tanto semi-guiadas, a partir de temas propuestos por el profesor, como libres, para cuya realización los informantes podían utilizar cualquier tipo de recurso (diccionarios, manuales, páginas web, etc.), no tenían límites de extensión ni de tiempo y sabían que sus textos no iban a ser evaluados. Como iban dirigidos a un interlocutor real (el compañero tándem), se pueden incluir dentro del macro-género *carta informal*, aunque se pueden identificar en ellas al menos cinco estilos (narrativo, descriptivo, argumentativo, expositivo e instructivo) y catorce áreas temáticas (información personal, familia, gustos y aficiones, actividades cotidianas, geografía, música, cocina, viajes, cine, recuerdos, vivienda, lecturas, fiestas y

tradiciones y temas de actualidad). Estas características permiten seleccionar textos que contengan léxico o estructuras lingüísticas recurrentes relacionados con algún tema concreto: por ejemplo, si se quieren observar las dificultades de uso del imperfecto o su evolución en las respectivas interlenguas y en diferentes niveles de competencia, será más fácil encontrar muestras seleccionando textos que hablen de *recuerdos*.

A continuación se presenta la interfaz a través de la cual se pueden consultar los dos corpus.

2. IMPLEMENTACIÓN EN LÍNEA DE *CORESPI* Y *CORITE*, SISTEMA DE ETIQUETADO Y TIPOS DE CONSULTAS POSIBLES

Para la implementación en línea de *CORESPI* y *CORITE* se ha utilizado *CATMA* (*Computer Assisted Textual Markup and Analysis*: www.catma.de), un software de acceso abierto que ha sido desarrollado por la Universidad de Hamburgo. Los motivos de la elección de esta interfaz son al menos cuatro:

1) permite etiquetar un texto de manera sencilla utilizando un sistema visual;

2) permite modificar fácilmente las etiquetas (añadir, borrar, cambiar de nombre, etc.);

3) permite consultas avanzadas utilizando simultáneamente criterios distintos y los resultados de las búsquedas se caracterizan por un alto grado de legibilidad;

4) permite compartir los corpus y su etiquetado entre varios usuarios.

Los textos de *CORESPI* y *CORITE* se han anotado por nivel de competencia en la lengua extranjera, género, edad, otros idiomas conocidos y palabras clave relacionadas con los temas tratados en cada texto. Además, al tratarse de corpus paralelos longitudinales, se han agregado dos etiquetas más: una, que permite identificar las parejas de informantes para reconstruir sus intercambios, y otra que indica si los textos de cierto informante son suficientes para realizar un análisis longitudinal de su interlengua o si, en cambio, solo es posible un análisis transversal.

CATMA permite realizar distintos tipos de consultas: en primer lugar, se puede visualizar la lista completa de las palabras que constituyen sendos corpus y la frecuencia con que aparece cada una de ellas. Esta función es especialmente útil en los corpus de interlengua porque permite visualizar también los ítems incorrectos. Además, las listas de palabras se pueden poner en orden alfabético

o de frecuencia. En segundo lugar, es posible buscar ítems parecidos al que se está buscando, o los que empiecen, contengan o terminen con una forma determinada. En tercer lugar, *CATMA* permite consultas complejas a través de la búsqueda de ítems que no aparezcan en secuencia uno inmediatamente detrás de otro en el texto, sino distanciados por una o más palabras. Esta función permite extraer combinaciones de palabras, ver su frecuencia de aparición y detectar algunos errores idiosincrásicos recurrentes.

Marcando la casilla *visible in KWIC* al lado de una de las palabras de la lista, se pueden visualizar todos los contextos en que aparece y restringir o ampliar su visualización de un mínimo de 2 a un máximo de 60 palabras de co-texto. Asimismo, se pueden visualizar simultáneamente tanto el contexto en que aparecen palabras distintas como el texto completo en que cierta palabra aparece.

También se pueden realizar consultas sencillas o avanzadas a partir de los criterios de etiquetado de los textos. En el primer caso se pueden visualizar los ítems que cumplan con cierta etiqueta (por ej. *hay* en los informantes de nivel B1) o, en el segundo caso, se pueden combinar criterios diferentes (por ej. *hay* en los informantes de nivel B1 en los textos que hablen de *vivienda*). También se pueden juntar resultados que se hayan extraído con criterios distintos (por ej. visualizar los textos en que aparece *hay* en los informantes de B1 *y* de B2) o refinar los resultados que se hayan obtenido con un criterio añadiendo otro. También es posible juntar los resultados de la búsqueda de más de un ítem (por ej. las partículas *ci* y *ne* del italiano en textos de informantes españoles que conozcan el catalán).

Las consultas se pueden realizar utilizando tanto el lenguaje específico de *CATMA* para las búsquedas como, más sencillamente, la herramienta *query builder*, que guía paso a paso al usuario en la construcción de una consulta según varios niveles de complejidad.

El sistema de etiquetado de *CORESPI* y *CORITE* también permite realizar consultas longitudinales visualizando, por ejemplo, la frecuencia y el contexto en que aparece cierto ítem en los textos de un informante concreto, ya que los textos también están anotados según un criterio cronológico. Esto permite observar la evolución en el uso de cierto ítem en la interlengua, o bien de un informante concreto, o bien de un grupo de informantes con características comunes. De esta manera se puede detectar qué errores persisten, cuáles desaparecen o cómo se modifican a lo largo del tiempo.

CATMA también tiene una función denominada *KWIC as a double tree* que permite visualizar un árbol doble en cuyo centro aparece el ítem que se ha buscado.

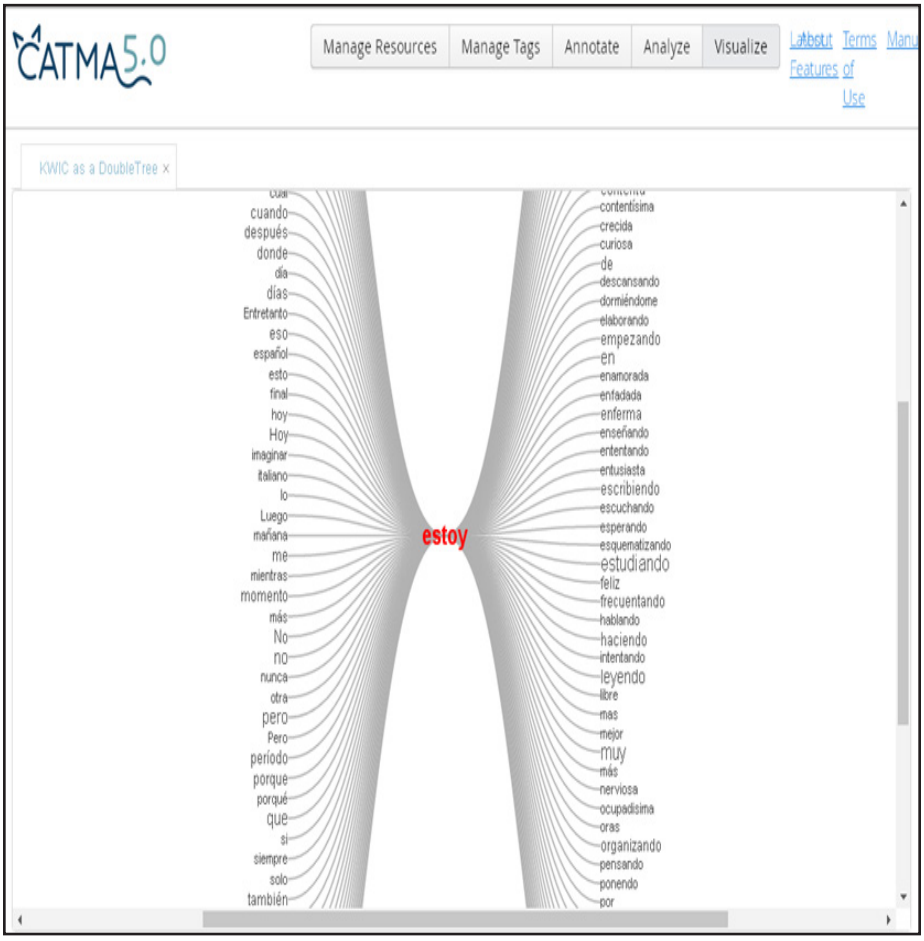

Figura 3. CORESPI: búsqueda del ítem *estoy* con la función *KWIC as a double tree*

Cada una de las hojas del árbol está formada por las palabras que la siguen o preceden. Cuanto más grande es el tamaño de la palabra, mayor es el número de veces que aparece en esa posición. Por ejemplo, en la Figura 3 se puede ver una parte del árbol que se ha generado buscando *estoy* en *CORESPI*. Si se clica en una de las hojas, aparecen las palabras que, en el corpus, siguen o preceden a la que está en el centro. Clicando en la palabra *muy* (Fig. 4), aparecen, en rojo, todas las palabras que la siguen cuando está después de *estoy*. También en este caso, el tamaño de una palabra destaca su índice de frecuencia. Esta operación se puede repetir clicando, por ejemplo, en una de las palabras que siguen a *muy*: de esta manera se pueden ver secuencias enteras de ítems.

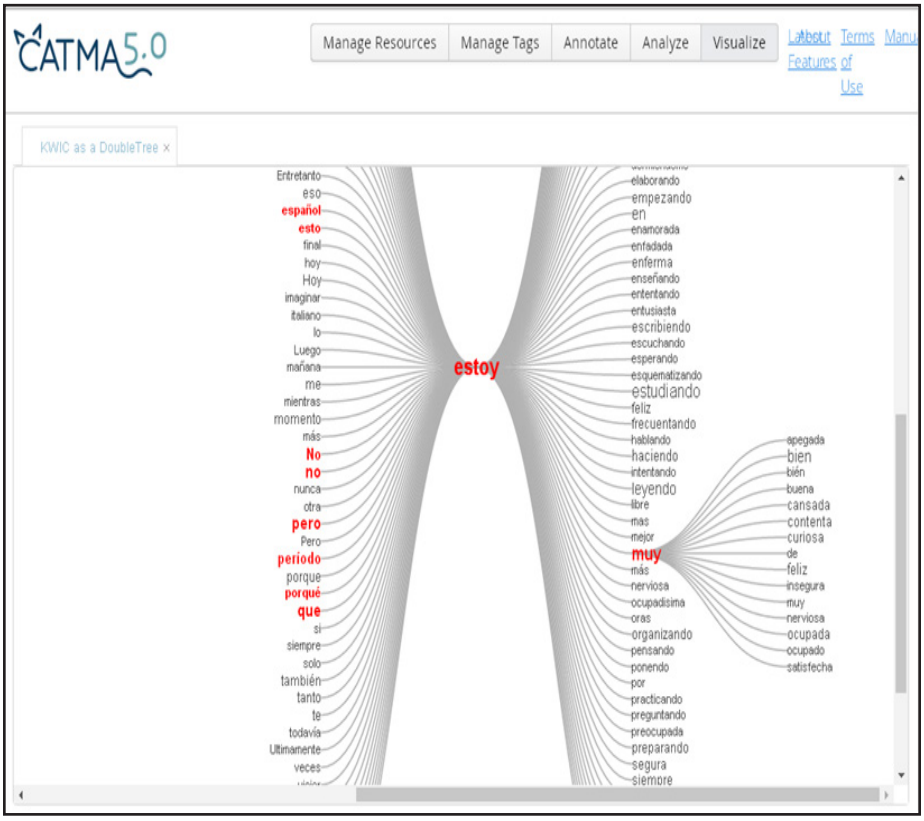

Figura 4. CORESPI: secuencias de palabras con la función *KWIC as a double tree*

Los resultados de las búsquedas se pueden exportar a un archivo de Excel para poder trabajar fácilmente con los datos.

CATMA permite compartir los corpus y sus anotaciones entre los usuarios registrados. La inscripción es gratuita y solo se precisa rellenar el siguiente formulario: https://docs.google.com/forms/d/1G79DAsRfAM4tBEGX-4h77RRLSQKZSbn4A0qbKIexIKM8/edit, que se encuentra accesible también desde la página web de los corpus: http://corespiycorite.altervista.org

El usuario registrado tendrá acceso libre a ambos corpus y a sus anotaciones y podrá realizar todos los tipos de consulta mencionados arriba, pero no podrá modificar ni los textos ni los criterios de etiquetado.

3. CONCLUSIONES

En este artículo se han presentado los criterios de construcción, las características de los informantes y de los textos que constituyen el *CORpus del ESPañol de los Italianos* y el *CORpus del ITaliano de los Españoles*. Asimismo, se han ilustrado las funcionalidades de la interfaz de consulta *CATMA* y el sistema de etiquetado de ambos corpus con un muestrario de las consultas posibles.

Aunque los corpus de hablantes no nativos son cada vez más numerosos (https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html), en la actualidad todavía no existen corpus de la(s) interlengua(s) del español y del italiano paralelos y longitudinales consultables en línea. *CORESPI* y *CORITE* aspiran a colmar este vacío y, sobre todo, pretenden poner al alcance de la comunidad académica una herramienta más para el estudio de la interlengua de lenguas afines.

REFERENCIAS BIBLIOGRÁFICAS

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

Bailini, S. 2012. *Interacción y desarrollo de la interlengua: e-tándem español-italiano*. Tesis doctoral. Salamanca: Universidad de Salamanca.

Bailini, S. 2013. "SCIL: a Spanish Corpus of Italian Learners", *Procedia – Social and Behavioural Sciences, Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th International Conference on Corpus Linguistics (CILC 2013)*, 95: 542-549. URL: http://www.sciencedirect.com/science/journal/18770428/95

Bailini, S. 2018. "CORESPI y CORITE: criterios de construcción, características y potencialidades". En S. Chiapello, T. Martín Sánchez, C. González Royo y N. Puigdeval Balafuy (eds.), *Telecolaboración y corpus para el estudio de lengua y cultura. Telecollaborazione e corpus per lo studio di lingua e cultura*. Alicante: Publicaciones de la Universidad de Alicante, 131-146.

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

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

Granger, S. 1996. "From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora". En K. Aijmer, B. Altenberg y M. Johansson (eds.), *Languages in Contrast. Text-based cross-linguistic studies*. Lund: Lund University Press, 37-51.

Granger, S. 2015. "Contrastive Interlanguage Analysis. A reappraisal" *International Journal of Learner Corpus Research* 1: 7-24.

Granger, S., Gilquin, G. y Meunier, F. (eds.) 2017. *The Cambridge H-anbook of Learner Corpus Research*. Cambridge: Cambridge University Press.

Hehmann, G., Ponti Dompé, D., Brammerts, H. y Little, D. (eds.) 1997. *Guida per imparare le lingue in Tandem via Internet.* Torino: Trauben.

Hehmann, G., Ponti Dompé, D. y Brammerts, H. (eds.) 2003. *Apprendimento autonomo delle lingue in Tandem*. Torino: Trauben.