*El corpus de aprendices **Aprescrilov** y su utilidad para la didáctica de ELE en la Bélgica multilingüe*

Kris Buyse
KU Leuven, Bélgica

Eva González Melón
Thomas More, Amberes, Bélgica

Resumen

En este artículo nos proponemos presentar la arquitectura y el uso del corpus de aprendices *Aprescrilov*, por un lado, y demostrar, por otro, su utilidad para la didáctica de ELE en la Bélgica multilingüe a partir de los resultados de un estudio acabado (el uso de *ser* y *estar* por neerlandófonos) y dos proyectos de tesis (el uso de las preposiciones *a, de, con, por* y *para*, y de los verbos de cambio). Para ello, describiremos algunos corpus lingüísticos a disposición del público, con el objetivo de caracterizar nuestro corpus *Aprescrilov*; luego presentaremos los resultados del estudio de *ser* y *estar* basado en este corpus; por último, esbozaremos brevemente la metodología de los dos estudios en desarrollo basados igualmente en datos extraídos de *Aprescrilov*.

1. El corpus lingüístico

Un corpus es "un conjunto de textos informatizados producidos en situaciones reales, que se han seleccionado siguiendo una serie de criterios lingüísticos explícitos que garantizan que dicho corpus pueda ser usado como muestra representativa de la lengua" (Alonso Pérez-Ávila 2007).

De esta definición destacamos varias ideas centrales:

- primero, que se trata de textos informatizados, lo que permite homogeneizar mejor la forma;
- segundo, que los textos se producen en situaciones reales, lo que añade un valor de autenticidad;
- tercero, que los criterios que se eligen para la selección de textos permiten diferenciar los corpus entre sí.

2. Corpus lingüísticos a nuestra disposición

Gracias al creciente apogeo de la lingüística de corpus y el desarrollo que está experimentando en todas las vertientes de la lingüística actual, tenemos a nuestra disposición varios corpus que nos pueden ser de utilidad con variadas aplicaciones a la

enseñanza de español como lengua extranjera (ELE), y cuya descripción sucinta nos permite presentar los criterios necesarios para caracterizar el corpus Aprescrilov. Para una descripción más elaborada, véase Buyse (2011).

2.1. Google

Si se habla de un corpus "en línea", "libre", "mixto", (es decir, con textos de todo tipo: históricos y sincrónicos, de todo tipo de temas, géneros, etc.), que sólo permite "búsquedas limitadas", y "no ha sido controlado ni anotado" (morfológica, sintáctica o semánticamente), estamos hablando de Google. Está al alcance de todos, pero es limitado, si por ejemplo, buscamos en Google "para seguir,", con el objetivo de averiguar si podemos utilizarlo como sinónimo de "a continuación": Google parece dar la impresión de que sí, pero esto ocurre porque no toma en cuenta la coma.

2.2. Webcorp

Otro corpus en línea, libre, mixto, no controlado ni anotado pero con mayores posibilidades de búsqueda que permiten sacar concordancias es Webcorp[1] (Figura 1), "puesto que" nos permite buscar palabras o frases, teniendo en cuenta las mayúsculas, los acentos y la puntuación, eligiendo el motor de búsqueda o el idioma, permitiendo restringir la búsqueda a una página en particular o a un dominio específico, así como a un área temática. Además, la aplicación devuelve una serie de "concordancias" que se pueden "reordenar" alfabéticamente según las palabras que preceden o siguen. Sin embargo, tal como en el caso de Google, el corpus mismo se compone de todos los textos no controlados de internet.



Fig. 1. Búsqueda en Webcorp.

2.3. Wortschatz

Otro corpus en línea, libre, no anotado pero controlado es Wortschatz[2] (Figura 2).

1. http://www.webcorp.org.uk/
2. http://wortschatz.uni-leipzig.de/

Se trata de un corpus de textos nativos seleccionados por un equipo de la Universidad de Leipzig, que también nos permite sacar concordancias.

Si queremos buscar, por ejemplo, qué verbo se utiliza con la palabra "matrimonio", introducimos la palabra y el sistema nos devuelve una página con el número de ocurrencias, grado de frecuencia, unos ejemplos contextualizados, las formas coocurrentes a la izquierda ya la derecha en orden de frecuencia decreciente, una visualización gráfica (mapa semántico), lo que nos permite concluir que el verbo que buscamos es "contraer". Entre las formas que aparecen encontramos "contraer", "contrajo", "contraen", etc. con lo que concluimos que se trata de un corpus "no lematizado", pero también "sin concordancias".



Fig. 2 Ejemplo de búsqueda en Wortschatz.

2.4. Corpus del Español

El corpus que más posibilidades ofrece, permitiendo la búsqueda en más de 20.000 textos del español, es el Corpus del Español (Figura 3). Según la propia descripción del corpus, "la interfaz permite realizar una búsqueda de diferentes maneras: palabras exactas o frases, comodines, etiquetas, "lemas", "categoría gramatical" o cualquier combinación de estos. También puede buscar colocaciones con un máximo de diez palabras."



Fig. 3. Ejemplo de búsqueda en el Corpus del Español.

3. El corpus de aprendices "aprescrilov"

Esta descripción limitada de algunos corpus en línea nos permite caracterizar el "corpus de aprendices" Aprescrilov[3] (Aprender a Escribir en Lovaina). Se trata de un corpus "en línea y libre" (para la investigación) que recoge más de 2700 textos escritos por estudiantes de lengua española de la Facultad de Letras de la KU Leuven y de la Lessius Hogeschool desde 2004-2005 hasta 2010-2011. En otras palabras, no se trata de textos escritos (mayoritariamente) por nativos, como en los casos anteriores, sino por "aprendices".

El corpus "no es mixto", puesto que todas las redacciones provienen de los tres primeros cursos de tres instituciones, recogidas entre 2005 y 2011, y fueron "controladas y anotadas" digitalmente con una misma versión personalizada del programa Markin,[4] cuya "barra de botones" permite "anotar de manera sistemática" los problemas o 'errores' contenidos en los textos. Las anotaciones cubren todos los componentes de la redacción –de la ortografía al texto, pasando por la puntuación, la morfología, la morfosintaxis, la pragmática, el léxico, etc.–.

Cuando accedemos a la página se nos presentan unos criterios de búsqueda que permiten buscar ejemplos y su contexto a partir de criterios como tipo de problema, palabra, categoría, carrera, curso, año académico, institución o tarea (Figura 4). El motor de búsqueda "no devuelve concordancias", sino contextos más amplios pero igualmente reducidos; además, un simple clic en un párrafo lleva al usuario a todo el texto correspondiente.



Fig. 4. Criterios de búsqueda en Aprescrilov

3. https://ilt.kuleuven.be/aprescrilov/
4. Markin, elaborado por Creative Technology, permite anotar redacciones de manera digital. "It is a Windows program which runs on the teacher's computer. It can import a student's text for marking by pasting from the clipboard, or directly from an RTF or text file. Once the text has been imported, Markin provides all the tools a teacher needs to mark and annotate the text. When marking is complete, the teacher can export the marked text as an RTF file for loading into a word-processor, or as a web page so that students can view the marked text in a web browser. Marked work can even be emailed directly back to the student, all from within the Markin program." (http://www.cict.co.uk/software/markin/).

Finalmente, el corpus va acompañado de una descripción cualitativa y cuantitativa de cada componente del mismo (número de redacciones, número de palabras por texto), lo cual permite calcular fácilmente la frecuencia absoluta y relativa de cada fenómeno.

4. Ejemplo de un estudio de análisis de errores en corpus de aprendices: sobre el uso de 'ser' y 'estar'

El uso de "ser" y "estar" se considera en el ámbito del español para neerlandófonos, tanto como en otros ámbitos de ELE, uno de los problemas típicos de la adquisición de la gramática española. Sin embargo, por falta de datos sobre la adquisición de "ser" y "estar" por los alumnos neerlandófonos, los profesores suelen basarse en su propia intuición, que suele corresponder con los datos sacados de estudios con alumnos ingleses como VanPatten (1987) y Ryan y Lafford (1992), según los cuales los alumnos no nativos ingleses en cursos de inmersión abusan de *ser* en una primera fase.

Por su parte, De Wolf (2012) concluye, a partir de la comparación de dos subcorpus del conjunto de textos de Aprescrilov, que en las redacciones de los alumnos de la KU Leuven el verbo *estar* es el que más problemas plantea en una primera fase, pero que el número de problemas (o "errores") baja rápidamente en fases posteriores; en las redacciones de la Lessius Hogeschool, al contrario, se abusa de *ser* en la primera fase y el número de problemas baja mucho más lentamente. Además, un tipo de error específico llama la atención por su frecuencia más alta que en las redacciones del otro subcorpus, a saber: **es claro que*...

Estos resultados invitan a investigar las causas de estas diferencias, lo que lleva a la siguiente hipótesis con respecto a la importancia de la instrucción. En el corpus de la Lessius Hogeschool se confirma la conclusión de Van Pattten, porque el contexto de enseñanza se caracteriza por un período de instrucción corta y superficial, de modo que se parece a la situación de inmersión de los alumnos ingleses (poca instrucción; las primeras frases se construyen con "ser": *soy X, soy de X*, etc.); además, la instrucción no entra en los detalles, de modo que llega a explicar concretamente el caso de la aparente excepción a la regla general c **es claro que.* En la KU Leuven el input es detallado desde el principio y se prolonga durante tres años: hay más input sobre "estar" (el verbo "nuevo", puesto que "ser" se parece al único verbo neerlandés *zijn*), por lo que un abuso de *estar* parece lógico; dada la instrucción gradual, es normal que desaparezcan los problemas más rápidamente.

El estudio demuestra, junto con otros a partir del mismo corpus Aprescrilov, la importancia e incluso la necesidad de realizar un análisis de errores con el objetivo de optimizar la enseñanza, adaptándola a las necesidades del público. Sin embargo, dado que con respecto a la adquisición de otros aspectos del español por neerlandófonos no está claro en qué medida en la Flandes multilingüe las interferencias se deben al francés (L2) o al neerlandés (L1), nos proponemos ampliar la perspectiva e investigar en el caso de dos proyectos de tesis el impacto de L1 y L2, respectivamente.

5. Estudios en desarrollo de análisis de errores en Aprescrilov. Dos casos particulares: verbos de cambio y preposiciones

Actualmente se encuentran en desarrollo dos estudios de análisis de errores de los textos de Aprescrilov: uno centrado en el uso que los estudiantes flamencos de español hacen de los llamados verbos de cambio y el otro centrado en el uso de algunas preposiciones (*a, de, en, por, para*), dos aspectos del español que, o bien no tienen correspondencia plena en el francés ni en el neerlandés (es el caso de los verbos de cambio), o bien su uso se ve interferido por el francés, L2 de nuestros estudiantes flamencos.

Ambos estudios parten de las hipótesis siguientes:

(1) el análisis de errores del corpus de aprendices *Aprescrilov* sabrá determinar de manera objetiva qué errores se deben a la interferencia con el francés y en qué medida, por un lado, y con el neerlandés, por el otro;
(2) si la presentación de los temas se hace de manera contrastiva L1-L2-LE y si está basada en los problemas reales y más frecuentes que nos haya revelado el análisis de errores, los alumnos sabrán servirse mejor de las estructuras en cuestión y evitar en mayor medida los escollos de las interferencias típicas español-francés-neerlandés.

De ahí que una segunda fase de estos proyectos contemple el diseño de un instrumental didáctico que pretende ayudar a profesores y alumnos a la hora de explicar y aprender ambos temas considerados de cierta dificultad. Una vez que se haya aplicado este material didáctico y se haya comprobado su utilidad a través de grupos de control y un sistema de comprobación pre- y postest, veremos si se puede mejorar el input en temas conflictivos basándonos en los problemas comunes/ frecuentes y errores cometidos anteriormente por estudiantes de unas características similares (véase el artículo de Fernández y González en este mismo volumen).

Bibliografía

ALONSO PÉREZ-ÁVILA, E. (2007): "El corpus lingüístico en la didáctica del léxico del español como LE", *Boletín de la Asociación para la Enseñanza del Español como Lengua Extranjera*, 37, 11-27.

BUYSE, K. (2011): "¿Qué corpus en línea utilizar para qué fínes en la clase de ELE?" en J. de Santiago Cuervós, H. Bongaerts, J. J. Sánchez Iglesias y M. Seseña Gómez, *Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE*, Salamanca: ASELE, 277-289.

DE WOLF L. (2012): *Análisis de errores de alumnos neerlandófonos de español. El caso de* ser *y* estar. Tesis doctoral. Lovaina.

RYAN, J. y B. LAFFORD (1992): "Acquisition of lexical meaning in a study abroad environment: Ser and estar and the Granada experience", *Hispania*, 75, 714-722.

VANPATTEN, B. (1987): "Classroom learners' acquisition of ser and estar: Accounting for developmental patterns", *Foreign language learning: a research perspective*, Cambridge: Newbury House, 61-75.