¿QUÉ CORPUS EN LÍNEA UTILIZAR PARA QUÉ FINES EN LA CLASE DE ELE?

KRIS BUYSE
Katholieke Universiteit Leuven / Lessius Hogeschool

RESUMEN: En el boletín de ASELE n° 37 (2007), Elena Alonso Pérez-Ávila ya presentó las ventajas del uso del corpus lingüístico como instrumento en la clase de ELE. En este artículo completaremos estas informaciones con el objetivo de ofrecer a los profesores de español una herramienta que les permita elegir el corpus más adecuado para el propósito elegido. Para estos fines, completaremos el artículo mencionado de tres maneras, a saber: que añadiremos otros corpus, los compararemos todos a partir de quince parámetros, y finalmente formularemos sugerencias para el uso de cada uno de los corpus en los niveles del Marco Común Europeo de Referencia, para que se enseñen gradualmente sus posibilidades respectivas.

1. INTRODUCCIÓN

En este artículo nos proponemos ofrecer a los profesores de español una herramienta que les permita elegir el corpus más adecuado para el propósito elegido. Nuestras preguntas de investigación son las siguientes: ¿Qué corpus pueden servir para la clase de ELE? ¿Cuáles son las diferencias entre ellos? ¿Para qué pueden servir y en qué nivel de enseñanza? Después de definir el concepto de corpus (1), compararemos una selección de corpus del español a partir de quince parámetros (2). Finalmente, formularemos sugerencias para el uso de cada uno de los corpus y otras herramientas técnicas para cada uno de los niveles del Marco Común Europeo de Referencia, para que se enseñen gradualmente sus posibilidades respectivas (3).

2. ¿CORPUS? ¿QUÉ CORPUS?

Un corpus lingüístico 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. Para Pérez-Ávila (2007: 11) las ventajas para el alumno y el profesor son numerosas. Para el alumno: constituye (1) una base sólida para elegir las *estructuras lingüísticas más frecuentes en las producciones reales* de los hablantes nativos de una lengua; (2) una herramienta que les otorga a los estudiantes la *autonomía* de elegir por sí mismos qué aprender, cómo aprenderlo y en qué orden; (3) un instrumento para *encontrar respuestas* a una tipología muy variada de dudas concretas y de profundizar en ellas por medio del acceso a amplios contextos reales. Para el profesor (no nativo), el corpus permite, además de ello, (1) basarse no exclusivamente en su intuición y en ejemplos elaborados ad hoc, sino en una *fuente amplia y fiable de recursos lingüísticos*; y (2) seleccionar un *input suficiente y de calidad* al que enfrentar a sus alumnos de forma que tenga lugar el *intake*, es decir, la adquisición de cualquier tipo de contenido.

Existen diferentes tipos de corpus. Para nuestros fines, los más importantes son los siguientes:

- corpus propios
  - corpus de español general o para fines específicos (económicos, médicos, jurídicos, etc.; véase, para unos ejemplos, Buyse 2003, 2010)
  - caso especial: "corpus de aprendices" (*learner corpus*; véase, para un ejemplo, Buyse 2009)
- corpus que se pueden adquirir (p.ej. http://www.elda.org/)
- corpus de acceso libre disponibles en internet:
  - Google, www.google.es
  - Webcorp, http://www.webcorp.org.uk/
  - Wortschatz, http://wortschatz.uni-leipzig.de
  - Corpus de Referencia del Español Actual (CREA), http://www.rae.es > http://corpus.rae.es/creanet.html (incluye pequeña guía bajo "Ayuda")
  - Corpus del Español, www.corpusdelespanol.org (toma el "Tour de 5 minutos" bajo "Más información" a la derecha en la pantalla)
  - corpus paralelos (p. ej. http://www.euro parl.eu.int/).

3. ¿QUÉ CORPUS ELEGIR Y PARA QUÉ?

3.1. Google

El corpus más conocido es el de Google (www.google.es). A pesar de su uso frecuente, constatamos que no suelen agotarse las posibilidades del instrumento, tales como las comillas (“”) para buscar una combinación de palabras (hasta colocaciones lingüísticas), la añadidura de “site” seguida del dominio de un país para limitar la búsqueda en términos geográficos (p.ej. *corpus site:es*), el uso de operadores como AND (para buscar contextos donde aparezcan al mismo tiempo las palabras a la izquierda y a la derecha del operador) y NO (para excluir contextos donde aparezcan las palabras a la derecha del operador); en particular las “búsquedas avanzadas” de Google son poco utilizadas, de modo que no siempre se aprovechan los parámetros de “idioma”, “región”, “dominio” (que equivale a la añadidura de *site*), “formato”, “fecha”, “con todas las palabras” (equivale a *AND*), “sin las palabras” (equivale a *NO*), “con la frase exacta” (equivale a las comillas), etc. Por ejemplo, al leer en una redacción una frase como *Renunciado a eso, fue uno de los mejores conciertos que he visto* (donde “renunciado” concuerda con el sujeto de “he”, lo cual es posible en neerlandés), se puede recurrir a una “búsqueda avanzada” de Google como se aprecia en la fig. 1: se limita al idioma español, se busca la frase exacta “renunciado a eso” y se excluye la coocurrencia de palabras como “he”, “has”, “ha”. El resultado es que Google nos devuelve sólo cuatro casos, donde “renunciado a eso” coocurre con “había(s/n)”, así que se puede concluir que tal estructura sin auxiliar probablemente no es posible en español.

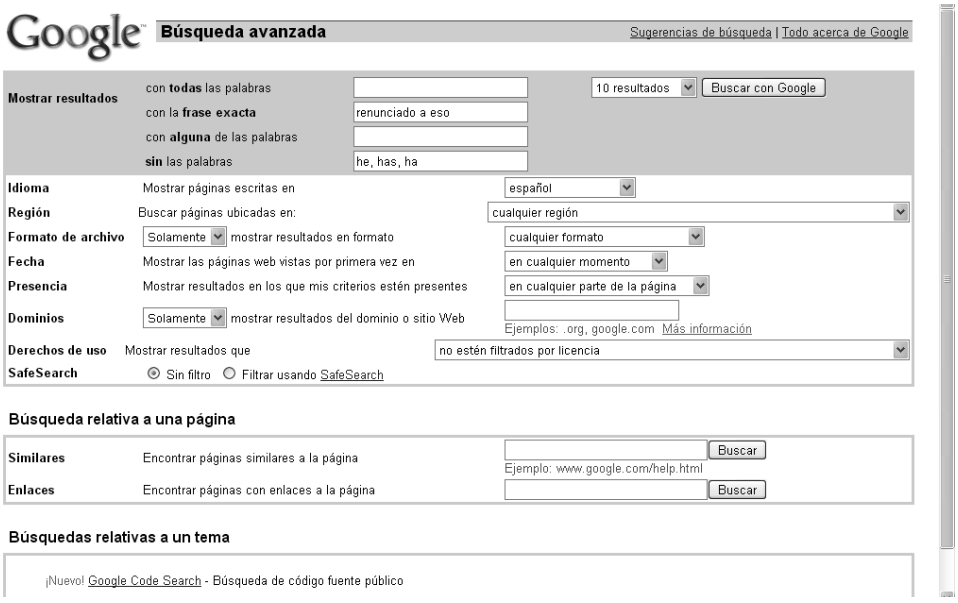

Fig. 1. Búsqueda “avanzada” en Google.

He aquí unas características de Google:

1. es fácil de usar
2. es multilingüe
3. se mezclan textos diacrónicos y sincrónicos (-)
4. no hay selección: se busca en todo lo que se encuentra en Internet, inclusive "textos" de no nativos, textos menos "cuidados" de nativos, etc.; la cantidad de textos es alta, la calidad mediocre, la actualización rápida
5. se trata de un corpus NO lematizado (si buscas "trabajar", el sistema sólo devolverá ocurrencias de "trabajar", pero no de todas las demás formas del verbo, como "trabajo", "trabajé", etc.)
6. NO hay POS tagging ("Parts Of Speech", es decir: categorías gramaticales): e.g. el sistema no establece ninguna diferencia entre la forma verbal y la forma sustantiva *duda*)
7. NO hay etiquetado semántico: e.g. no se pueden buscar los sinónimos de "limpio"
8. NO se toman en cuenta las mayúsculas, los acentos ("está" frente a "esta") y la puntuación (e.g. no puedes comprobar si "Para seguir," se puede emplear como conector aditivo, porque no se toma en cuenta la coma)
9. se puede buscar con comodines y operadores (AND, NO...)
10. hay una herramienta básica para buscar combinaciones de palabras ("frase exacta" o las comillas)
11. no se puede limitar la búsqueda a ciertas áreas temáticas (como "medicina"), tampoco a ciertos tipos de textos (como "orales", "académicos", etc.), pero sí a ciertas áreas geográficas (países)
12. no se pueden comparar distintos subcorpus (e.g. "sin papeles" en textos académicos y orales)
13. no se pueden sacar concordancias (para el concepto de "concordancias", véase el apartado siguiente)
14. por lo tanto, tampoco se pueden reordenar las concordancias
15. las informaciones sobre frecuencias y distribución de lo que se busca son muy limitadas

3.2. Webcorp

Una herramienta que permite paliar algunas de las desventajas de Google (y que por lo tanto se puede enseñar a los alumnos como alternativa recomendable) es Webcorp (http://www.webcorp.org.uk/), donde es posible *sacar concordancias*, limitar la búsqueda a ciertas *áreas temáticas* y tomar en cuenta las mayúsculas, los acentos y la puntuación.

Para ello hace falta abrir las “advanced search options”, cambiar el “output format” a “plain text (KWIC)” y limitar si necesario a la extensión de un país (“country”) y un área temática (“textual domain”). Por ejemplo, si buscamos el verbo o los verbos que entran en una colocación con “anamnesis”, rellenamos el formulario como en la fig. 2 y aparecerán gradualmente una serie de concordancias (es decir, líneas de texto con la palabra buscada en el medio, como en la fig. 3), que nos permiten encontrar inmediatamente verbos como “realizar”, “ejecutar” y “hacer”.

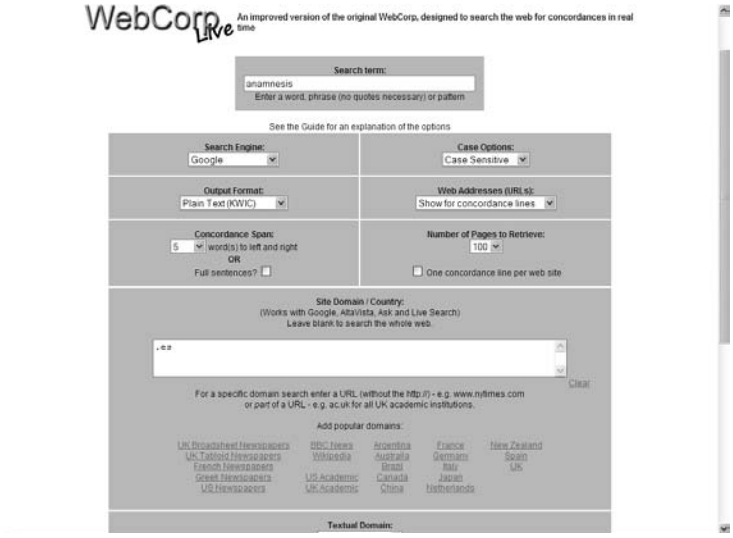

Fig. 2. Búsqueda con Webcorp.



Fig. 3. Parte de unas concordancias sacadas de Webcorp.

De ahí que Webcorp ofrezca unas ventajas con respecto a Google (las diferencias van en itálicas):

1. es fácil de usar
2. es multilingüe
3. se mezclan textos diacrónicos y sincrónicos
4. no hay selección, la cantidad de textos es alta, la calidad mediocre, la actualización rápida
5. se trata de un corpus NO lematizado
6. NO hay POS tagging
7. NO hay etiquetado semántico
8. *se toman en cuenta las mayúsculas, los acentos y la puntuación*
9. se puede buscar con comodines y operadores (AND, NO...)
10. *hay varias herramientas para buscar combinaciones de palabras ("phrase"* y *"collocations")*
11. *se puede limitar la búsqueda a ciertas áreas temáticas* y a ciertas áreas geográficas (países)*, pero no* a ciertos tipos de textos
12. no se pueden comparar distintos subcorpus (e.g. "sin papeles" en textos académicos y orales)
13. *se pueden sacar concordancias*
14. no se pueden reordenar las concordancias
15. las informaciones sobre frecuencias y distribución de lo que se busca son muy limitadas

3.3. Wortschatz

Otras desventajas de Google se pueden evitar recurriendo a pequeños corpus fácilmente manejables como los de Wortschatz de la Universidad de Leipzig (http://wortschatz.uni-leipzig.de: selecciona "corpora", luego "Spanish" y teclea la forma en la ventana a la derecha de "Word", por ejemplo "matrimonio", con el objetivo de encontrar el verbo con el que se suele emplear). A continuación el sistema devuelve una página como en la fig. 4, con el número de ocurrencias, un *grado de frecuencia, unos ejemplos contextualizados, las formas coocurrentes a la izquierda y a la derecha*, ambas en orden de frecuencia decreciente, y una visualización gráfica (*mapa semántico*). En este caso se ve entre los "vecinos significativos" a la izquierda de *matrimonio* que el verbo que buscamos es "contraer". Por otro lado, nos permite constatar que se trata de un corpus no lematizado, puesto que se listan aparte las formas "contraer", "contrajo", "Contrajo", "contraen", etc. Tampoco se puede limitar a ni comparar entre varias áreas geográficas, ámbitos temáticos o géneros textuales.

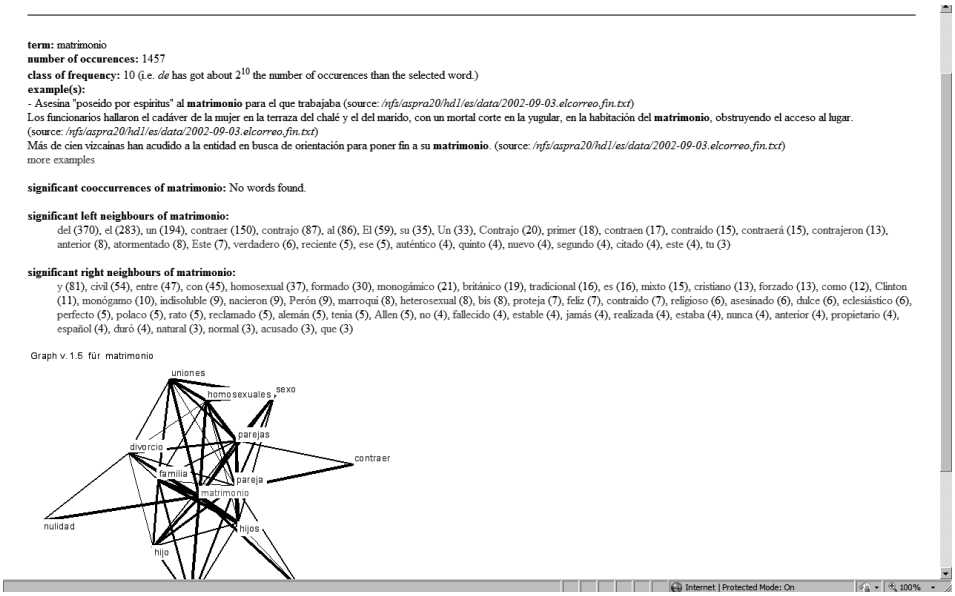

Fig 4. Ejemplo de una búsqueda en el corpus español de Wortschatz.

De ahí que Wortschatz ofrezca las ventajas siguientes con respecto a Google (las diferencias van en itálicas):

1. es fácil de usar
2. es multilingüe
3. *se trata de textos sincrónicos*
4. *se trata de una selección (más bien limitada) de textos, la cantidad de textos es limitada* (de modo que esta herramienta es menos útil para buscar palabras menos frecuentes), *la calidad más alta que la de Google*, *la actualización menos rápida*
5. se trata de un corpus NO lematizado
6. NO hay POS tagging
7. NO hay etiquetado semántico
8. *se toman en cuenta las mayúsculas, los acentos y la puntuación*
9. *NO se puede buscar con comodines y operadores*
10. *hay varias herramientas para buscar combinaciones de palabras ("significant left neighbour"* y *"significant right neighbour")*
11. no se puede limitar la búsqueda a ciertas áreas temáticas, tampoco a ciertos tipos de textos *o a ciertas áreas geográficas*
12. no se pueden comparar distintos subcorpus

13. no se pueden sacar concordancias
14. por lo tanto tampoco se pueden reordenar las concordancias
15. las informaciones sobre frecuencias y distribución de lo que se busca son más bien limitadas

3.4. CREA

Para paliar ciertas deficiencias como la cantidad limitada de textos y la falta de comparación de géneros, ámbitos y países, hace falta buscar en corpus más grandes como el CREA de la Real Academia Española (http://corpus.rae.es/creanet.html; también hay una colección de textos diacrónicos, denominado *CORDE*). En la primera pantalla de búsqueda se pueden elegir géneros ("medio"), países ("geográfico") y ámbitos ("tema"), hasta el año de publicación (véase la fig. 5):



Fig. 5. Búsqueda en el CREA

A continuación, el sistema permite *sacar concordancias* (pulse "recuperar concordancias"), pero muchos usuarios parecen desconocer la posibilidad de *reordenar las concordancias* alfabéticamente según la primera hasta la quinta palabra a la izquierda o a la derecha, aparte de la posibilidad de "agrupar" según el autor, el año, el país, el tema y el título. Por ejemplo, para encontrar la diferencia entre "(dar) una clase de" y "(dar) una clase sobre", reordenamos alfabéticamente según la primera palabra a la derecha, de modo que se listen juntos los casos con "de" y los de "sobre" (fig. 6):

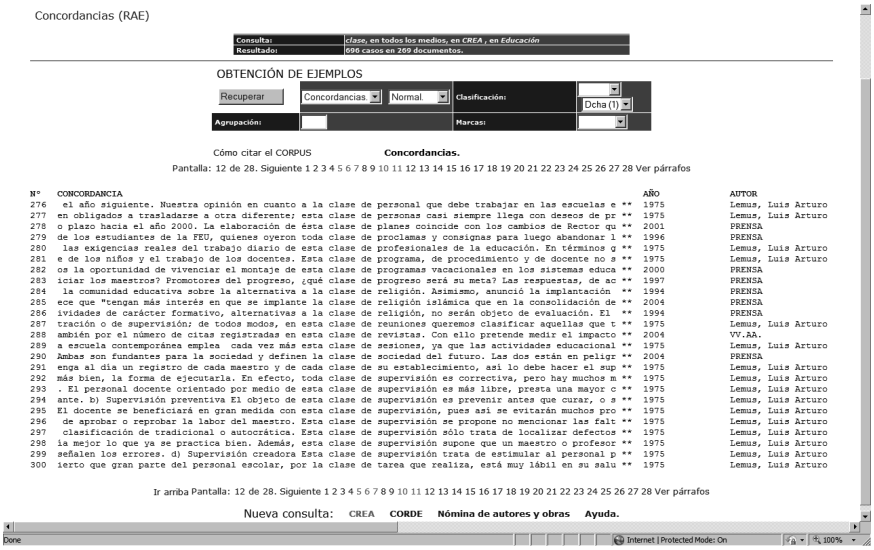

Fig. 6. Búsquedas en el CREA: reordenar las concordancias según la primera palabra a la derecha ("Dcha (1)").

A pesar de estas posibilidades, la mayor desventaja es que el CREA sigue (hasta el momento) sin lematizar. Globalmente, entre las ventajas y desventajas del CREA se destacan las siguientes:

1. *es menos fácil de usar*
2. *NO es multilingüe*
3. *se trata de textos sincrónicos y diacrónicos aparte (CREA y CORDE)*
4. *se trata de una selección de textos, la cantidad de textos es alta, la calidad mucho más alta que la de Google, la actualización mucho menos rápida*
5. se trata de un corpus NO lematizado
6. NO hay POS tagging
7. NO hay etiquetado semántico
8. *se toman en cuenta las mayúsculas, los acentos y la puntuación*
9. se puede buscar con comodines y operadores (Y, O, NO)
10. *hay varias herramientas para buscar combinaciones de palabras (comillas / reordenar las concordancias)*
11. *se puede limitar la búsqueda a ciertas áreas temáticas, tipos de textos y áreas geográficas*
12. no se pueden comparar distintos subcorpus
13. *se pueden sacar concordancias*

14. *se pueden reordenar las concordancias*
15. las informaciones sobre frecuencias y distribución de lo que se busca son más bien limitadas

3.5. Corpus del Español

El corpus que más posibilidades ofrece (y por lo tanto más habilidad y conocimientos requiere del usuario) es el Corpus del Español, de Mark Davies (www.corpusdelespanol.org). Es imposible describirlas todas aquí, pero para tener una idea basta tomar el "Tour de 5 minutos" bajo "Más información" a la derecha en la pantalla. Entre ellas destacamos las siguientes: *buscar palabras exactas, frases, colocaciones, etiquetas semánticas, lemas, categorías gramaticales (POS), períodos, géneros textuales o cualquier combinación de estos, además de comparar varias formas, colocaciones, períodos*: e.g. los adjetivos que aparecen cerca de "mujer" en textos académicos y orales; los verbos que comienzan con "des*" frente a los que empiezan por "in*"; las formas de "hacer" en el siglo XIII frente a las del siglo XX; adjetivos que van con "mujeres" frente a los que van con "hombres"; los sustantivos más frecuentes que ocurren con "limpio" frente a los que van con "puro"; todos los sinónimos o antónimos de "limpio".

No obstante, como se puede apreciar en la fig. 7, este corpus no permite limitar a ciertas áreas geográficas (a pesar de lo que se anuncia en la leyenda) y tampoco permite reordenar las concordancias. En comparación con los corpus anteriores, las características del Corpus del Español son las siguientes:

		Google	*Webcorp*	*Wortschatz*	*CREA*	*Corpus del Español*
1.	facilidad de uso	+	+	+	+/-	+/-
2.	multilingüe	+	+	+	-	-
3.	histórico / sincrónico	-	-	-	+	+
4.	selección / calidad / cantidad	-/-/+	-/-/+	+/+/-	+	+
5.	lematizado	-	-	-	-	+
6.	POS-tagging	-	-	-	-	+
7.	etiquetado semántico	-	-	-	-	+
8.	mayúsculas / acentos / puntuación	-/+/-	+/+/-	+/+/-	+	+
9.	comodines y operadores	+	+	-	+	++
10.	combinaciones de palabras	+/-	+/-	+/-	+/-	+
11.	selección temática / geográfica / textual	-/+/-	+/+/-	-/-/-	+	+/-/+
12.	comparar subcorpus	-	-	-	+/-	+
13.	sacar concordancias	-	+	-	+	+
14.	reordenar concordancias	-	-	-	+	-
15.	frecuencia y distribución	+/-	-	-	+/-	+

Fig. 7. Comparación de Google, Webcorp, Wortschatz, CREA y Corpus del Español.

4. TRADUCCIÓN A LA CLASE DE ELE

En conclusión, cada corpus tiene ciertos puntos a favor y en contra, pero el que más se acerca al "corpus ideal" es el Corpus del Español, por lo que no debe sorprender que también es el más complejo para manejar. De ahí que aconsejemos introducir y enseñar gradualmente el uso de herramientas indispensables en los distintos niveles del Marco Común de Referencia Europea, además de usarlas tomando en cuenta los criterios expuestos anteriormente:

- A1: enseñar uso frecuente y *correcto* de los diccionarios uni y bilingües / uso de un verificador ortográfico
- A2: integrar instrumentos como *Wordreference, Google,* etc.
- B1: instrumentos como *Wortschatz* y *Webcorp*
- B2: instrumentos como el *CREA*
- C1-2: instrumentos como *Corpus del Español*

Nuestra experiencia demuestra que el uso de estas herramientas da lugar a una mejora sustancial de la calidad de la expresión de los aprendices; por otro lado, también demuestra que el uso de las herramientas sigue siendo insuficiente (y el progreso lingüístico también) si no enseñamos su uso al aprendiz y le "invitamos" a manejarlos. Un instrumento muy eficaz para hacerlo es la lista de control en un portafolio, donde el alumno debe contestar, en el marco de una tarea de redacción, a preguntas como las siguientes:

- ¿Has utilizado diccionarios monolingües? (para controlar traducciones literales, colocaciones, preposiciones fijas, género de palabras, falsos amigos...) ¿Para qué?
- ¿Has utilizado diccionarios multilingües? (mismos objetivos) ¿Para qué?
- ¿Has mirado toda la información en las entradas de los diccionarios (el género, la morfología, los distintos significados, las preposiciones fijas, las colocaciones, los modos, los ejemplos...)? ¿Para qué?
- ¿Has utilizado ciertos corpus (Wortschatz, CREA, Webcorp, Corpus del Español, Google...)? ¿Para qué?
- ¿Has utilizado un verificador ortográfico (en Word o en internet, como http://lomastv.com/free-online-spanish-spelling-grammar-checker.php)?

En otra ocasión aportaremos datos empíricos que demuestran que el uso de estas herramientas permite bajar sustancialmente el número de problemas (o "errores") formales en las redacciones de los alumnos.

BIBLIOGRAFÍA

Buyse, K. (2003): "Plataforma electrónica del español Económico dentro del proyecto ElektraVoc-II", en S. Posteguillo Gómez (ed.), *Internet and Language*, Castellón: Universidad de Castellón, 101-115.

Buyse, K., N. Delbecque y D. Speelman (2009): *Trampas y pistas para la expresión escrita de los neerlandófonos*, Averbode/Madrid: Averbode/Edelsa.

Buyse, K., A. Laffut, E. Saver y L. Vekemans (en prensa): "Urgentias, a lexical database for medical students in clinical placements: architecture, use and evaluation", en *Researching specialized languages*, Amsterdam: Benjamins [24 páginas].

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.