LOS CORPUS LINGÜÍSTICOS Y LA ENSEÑANZA DEL INGLÉS

Noelia Ramón García
Universidad de León

Introducción

Los avances informáticos que se han venido produciendo en las últimas décadas, y especialmente en los últimos años, han desarrollado materiales y herramientas de diversos tipos que pueden ser extremadamente útiles para los investigadores en todos los campos lingüísticos, así como para la enseñanza de lenguas extranjeras. Una de estas herramientas, de la que hablaremos a continuación, son los corpus lingüísticos informatizados. "There is every reason to believe that language corpora will have a role of growing importance in teaching." (Leech, 1997: 1).

No existe una definición única de corpus. Básicamente, podemos decir que un corpus es un conjunto de textos, procedentes de diversas fuentes, de los que se puede extraer todo tipo de información lingüística. En el caso de los corpus informáticos, estos textos están almacenados en soporte informático, lo que facilita enormemente la tarea del investigador y le ahorra tiempo y esfuerzo.

Breve historia de los corpus

A continuación se mencionarán algunos de los corpus más importantes que se han recopilado hasta el momento.

Entre los años 30 y 60 de nuestro siglo, los estructuralistas norteamericanos, seguidores de Bloomfield ya empleaban corpus, aunque sus métodos de recopilación eran aún manuales y muy rudimentarios. Pretendían construir un corpus representativo de la lengua, para luego hacer un estudio descriptivo y sincrónico de ésta.

En 1960, en la UCL (University College London), R. Quirk y S. Greenbaum compilaron un corpus del inglés de un millón de palabras. Este corpus estaba aún recogido en papeletas, y era de uso manual y exclusivo para la universidad en la que se compiló.

En 1961 N. Francis y H. Kuchera crean el primer corpus informático en Estados Unidos, el Brown Corpus. Contenía 1 millón de palabras en inglés americano y estaba

minuciosamente anotado. Se puso a disposición de los lingüistas y alcanzó rápidamente una gran difusión.

De 1978 data el LOB (Lancaster-Oslo/Bergen) Corpus, que es bastante similar al Brown Corpus, aunque recopila textos exclusivamente escritos en inglés británico.

En los años 80, la industria informática comienza a desarrollarse a pasos agigantados. La editorial Harper/Collins y la Universidad de Birmingham comienzan un proyecto conjunto para la elaboración de un corpus general de la lengua inglesa, el Bank of English. En un primer momento, tenía 20 millones de palabras. En la actualidad alcanza casi 400 millones, y se añaden en torno a 10-12 millones de palabras nuevas al mes. En 1987 aparece el primer diccionario Collins Cobuild, pionero entre los diccionarios basados en información extraída de corpus informáticos.

Ya está también a disposición en Internet desde hace muy poco el CREA, Corpus de Referencia del Español Actual, con 100 millones de palabras.

Estos son sólo algunos de los ejemplos más importantes de los corpus existentes, pero existen muchos más. En los años 90, la mejora de los soportes informáticos y la siempre creciente memoria de los ordenadores, y su mayor velocidad facilitan la tarea de la compilación de corpus específicos por áreas de conocimiento, variedades regionales, tipos textuales, etc. En suma, hoy en día un corpus lingüístico es ya, por definición, informático.

Tipos de corpus

Guillermo Rojo (1998), impulsor desde la Universidad de Santiago de Compostela del proyecto que llevó a la realización del CREA, distingue, entre otros, los siguientes tipos de corpus:

1. sincrónicos / diacrónicos: Son corpus sincrónicos aquellos que recogen un estadio en particular de la lengua, y diacrónicos aquellos que recogen textos de diversas etapas en la evolución de una lengua. El corpus diacrónico inglés más conocido es el Helsinki Corpus, y el español el CORDE (Corpus Diacrónico del Español).
2. generales / especializados: Son corpus generales de la lengua los que como el Bank of English o el CREA aspiran a ser representativos de toda la lengua, y especializados aquellos que solamente recogen textos de un área concreta para

extraer resultados aplicables exclusivamente a ese campo. Existen, por ejemplo, corpus de la industria petroquímica, de economía, de política, etc.

3. cerrados / abiertos: Son corpus cerrados aquellos a los que ya no se continúan añadiendo nuevos textos, y abiertos aquellos que se renuevan continuamente.
4. monolingües / bilingües o multilingües: Un corpus monolingüe contiene textos en una sola lengua. Un corpus bilingüe textos en dos lenguas, y, por último, también existen corpus con textos en tres o más lenguas.
5. paralelos / comparables: Los corpus bilingües o multilingües pueden ser paralelos o comparables. Un corpus paralelo consta de textos originales en una lengua y su traducción a la otra o las otras. Son difíciles de recopilar y de conseguir, y casi siempre muy pequeños, pero muy útiles si la calidad de la traducción es buena. En cambio, en un corpus comparable se reúnen textos escritos originalmente en ambas lenguas y de una extensión, estilo y temática similar. Son muy prácticos para traductores, especialmente en campos técnicos.
6. etiquetados / no etiquetados: Algunos corpus tienen etiquetas en cada uno de sus elementos. Estas etiquetas pueden contener información de muchos tipos, normalmente relacionada con la categoría gramatical de la palabra, o su estatus semántico. Puesto que esto exige un gran esfuerzo, sólo los grandes corpus tienen sus palabras etiquetadas. Existe un programa, CLAWS (Constituent Likelihood Automatic Word-tagging System, que según Fligelstone, Rayson y Smith es "still one of the most accurate part-of-speech taggers available." (Fligelstone, Rayson & Smith, 1996: 181).

Características de los corpus

- Cantidad de palabras que contiene.

Por muy grande que sea un corpus y por muchos millones de palabras que contenga, nunca será lo suficientemente representativo de la lengua, porque no puede incluirlo todo. Un corpus es finito. Además, dependiendo de para qué se vaya a usar el corpus, quizás no sea necesario que sea muy amplio. Un profesor que quiera hacer un corpus de textos escritos por sus alumnos no necesita tener 400 millones de palabras. Quizás un corpus de relativamente pocas palabras puede ser suficientemente representativo en su caso concreto. Cuanto más grande sea un corpus, más

representativo será, pero también será más difícil de manejar por la cantidad de datos almacenados. Según B. Dodd,

> huge corpora are not necessary for language-teaching purposes. A modest corpus of a million or so words is certainly enough to make a valuable teaching aid, and is realistically within the reach of most teaching institutions. (Dodd, 1997: 131).

- Representatividad.

A la hora de compilar un corpus es muy difícil decidir qué introducimos y qué no. Los grandes corpus que pretenden reflejar la lengua general, como Cobuild/Bank of English y CREA, tienen un 10% de lengua oral, y el restante se reparte entre literatura, textos académicos, de prensa, publicidad, y otros textos escritos de todos los campos del saber. Debemos tener en cuenta que, dependiendo de nuestra elección, los resultados futuros serán unos u otros. Si utilizamos un corpus ya existente evitaremos estos problemas, pues otros han tomado las decisiones antes que nosotros.

- Disponibilidad real de textos.

Para compilar un corpus y poder publicarlo y citarlo posteriormente, es necesario tener la autorización escrita de todos los autores de todos los textos introducidos. Utilizar un texto sin autorización es ilegal. Un profesor que haga un corpus con textos de sus alumnos no tendrá estos problemas.

- Actualización de la información lingüística.

Para que un corpus continúe siendo siempre igual de representativo, deberá ser renovado continuamente. Esto exige un esfuerzo considerable y mucha constancia en el trabajo. Un profesor deberá introducir los textos producidos por sus alumnos a lo largo de varios años, y así conseguirá formar un corpus que será cada vez más representativo y que continuará siendo abierto y dinámico.

- ¿Cómo crear un corpus?

Una vez que los textos están localizados, llega la hora de introducirlos en el soporte informático. El método más fácil para pequeñas cantidades de texto es el del tecleo manual y directo. Esto facilita mucho la tarea, ya que se hace sobre un procesador de textos al que luego se le pueden aplicar sin problemas los programas de concordancias. En cambio, cuando se quiere ser muy representativo se necesitan grandes cantidades de textos, y no es fácil informatizarlos manualmente. Se pueden bajar

grandes cantidades de texto directamente de Internet, donde ya está el material informatizado. Si se necesitan algunos textos específicos, es necesario escanearlos y aplicarles el sistema ASCII de reconocimiento de caracteres.

- Programas informáticos específicos.

Una vez que se ha compilado el corpus, éste no tendrá ninguna utilidad si no se le aplica algún programa informático específico de alineamiento o de concordancias, para poder extraer la información que deseamos.

> It is widely acknowledged today that a corpus need the support of a sophisticated computational environment providing software tools both to retrieve data from the corpus and to process linguistically the corpus itself. (Leech, 1991: 22).

MicroConcord, por ejemplo, uno de los más básicos y baratos, sirve para buscar palabras o grupos de palabras, para sacarlas en su contexto, y para contar el porcentaje de frecuencia de aparición. Para conseguir una información más detallada, como categorías gramaticales, información sintáctica y semántica, hay que utilizar otros programas mucho más sofisticados.

Posibles aplicaciones de los corpus en la enseñanza

En principio, el profesor de inglés como lengua extranjera tiene dos grandes opciones para empezar a utilizar corpus en sus clases: puede utilizar corpus ya existentes, generales de la Lengua Inglesa o bilingües español-inglés, o puede crear su propio corpus.

1. En primer lugar, puede utilizar un corpus, ya sea general o específico, ya existente. Algunos se pueden comprar por precios módicos, como los corpus de ciertos periódicos, revistas científicas y otras publicaciones. A otros se accede on-line a través de Internet, y para acceder a los más grandes es necesario realizar una suscripción. En el caso de un profesor de inglés, emplearía corpus en lengua inglesa, que es la lengua meta de sus alumnos. Las posibles aplicaciones de estos corpus son múltiples y de ellas hablaremos más adelante.
2. También puede utilizar un corpus bilingüe inglés-español. En las clases de inglés, estos corpus son muy útiles para la enseñanza de técnicas de traducción, tanto directa como inversa. El profesor puede imprimir textos con palabras, expresiones u

oraciones enteras eliminadas en una y otra lengua, para que los alumnos los completen con ayuda del texto en la lengua opuesta.

3. Existen ya corpus de textos escritos por alumnos de inglés como lengua extranjera. El proyecto más claro en este sentido es el llamado ICLE (International Corpus of Learners' English) desarrollado por Sylviane Granger en Lovaina (Bélgica). La dirección de Internet para acceder a más información sobre este proyecto es: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html.

> The project aims at the collection of comparable corpora of learners' English from different native speaker populations: for example, the written English produced by French-speaking, Dutch-speaking and Chinese-speaking students. (Leech, 1997: 20).

La finalidad de estos corpus es observar las divergencias en los errores de los diferentes nativos de los diferentes idiomas, y cómo y hasta qué punto sus lenguas de origen influyen en su proceso de aprendizaje del inglés.

Existen otros dos corpus de textos escritos por no-nativos, como son el Longman Learners' Corpus y un corpus enorme de 5 millones de palabras compilado por John Milton en Hong Kong con textos de estudiantes chinos.

4. Por último, al profesor le queda la posibilidad de formar su propio corpus con los textos que le proporcionen sus alumnos cada año. La creación de un corpus puede llevar mucho tiempo si se quiere ser mínimamente representativo. Sin embargo, para el uso personal de un solo profesor o de un grupo relativamente reducido, el trabajo necesario no sería demasiado arduo y tendría muchas ventajas. El corpus estaría muy especializado para un tipo concreto de alumnos con un nivel determinado de inglés, y pronto se alcanzaría un número de textos representativo. Aplicando luego programas de concordancias como MicroConcord, el profesor podría establecer qué tipo de errores son los más frecuentes entre sus alumnos, qué construcciones emplean con más facilidad y frecuencia, qué expresiones no usan nunca, aunque teóricamente deberían saberlas, etc. Estos datos le pueden servir al profesor para decidir en qué puntos gramaticales o léxicos debe insistir más. También puede preparar ejercicios de detección y corrección de errores sacados de entre material auténtico, lo que lo hace mucho más fiable.

Posibilidades pedagógicas de los grandes corpus generales de la lengua inglesa

> Perhaps the most obvious pedagogic use of corpora is to treat them as sources of classroom materials which the teacher can select from and adapt according to requirements. (Aston, 1997: 52).

Las ventajas de trabajar con el material que nos proporcionan los corpus generales son muchas. En primer lugar, se trata de textos reales producidos en origen por hablantes nativos de inglés. Así, los alumnos tendrán acceso al uso auténtico de la lengua tal y como es en realidad en los países donde se habla, y no tratarán exclusivamente con materiales preparados, artificiales, y específicamente diseñados para extranjeros. Además estos corpus tienen anotaciones gramaticales muy útiles para los profesores, y también existen numerosos programas informáticos en la actualidad con los que se pueden extraer multitud de datos interesantes en la enseñanza del inglés como lengua extranjera. Todas las actividades presentadas a continuación se basan en el uso del corpus Cobuild.

Es importante señalar que con un programa de concordancias sencillo como MicroConcord se pueden hacer búsquedas diversas. La más simple es la búsqueda de una palabra concreta que aparecerá rodeada de su contexto más inmediato. Pero también se pueden buscar grupos de palabras seguidas con simplemente añadir el símbolo + entre ellas, o palabras que comiencen por un determinado grupo de letras para buscar todas las ocurrencias de un verbo (*driv**), por ejemplo. También se pueden buscar palabras por su combinación con una categoría gramatical concreta, por ejemplo *rather+JJ* (rather + adjetivo). Y hay muchas más posibilidades de búsqueda. Por supuesto, el usuario puede decidir cuántas entradas de su palabra quiere ver, puede elegir una para ver su contexto más ampliamente, y también puede imprimir los datos que aparezcan en su pantalla.

Una de las actividades más sencillas de preparar tomando como base un corpus del inglés general, como por ejemplo, Cobuild, es el ejercicio del tipo *fill-in*. No hay más que extraer e imprimir listas de concordancias de determinadas palabras problemáticas. El profesor escogerá entre los diversos ejemplos de texto real, aquellos que le parezcan más representativos, y los imprimirá. Los ejercicios de *fill-in* se pueden referir a prácticamente todos los aspectos gramaticales, desde distinguir entre diversas

formas de verbos irregulares, hasta vocabulario que genera problemas frecuentemente, diferencias entre gerundios e infinitivos, etc.

Otra de las aplicaciones más útiles de los corpus es su explotación como fuente inagotable de ejemplos para todo tipo de explicaciones gramaticales, ejemplos que además son lenguaje auténtico que un hablante nativo ha producido alguna vez de forma espontánea. Es suficiente con teclear la palabra *would* para que aparezcan cientos de oraciones condicionales de todos los tipos y en contextos tan diversos que aportan mucha mayor variedad que los tradicionales ejemplos en los que aparece constantemente el mismo vocabulario: libros, amigos, perros y vecinos. Lo mismo sucede en cualquier área de la gramática inglesa. Tecleando *said+that* se obtendrán numerosos ejemplos de oraciones en estilo indirecto que podrán servir de ejemplo o como ejercicios de inversión a estilo directo.

Conclusión

A pesar de que es cierto, como aseguran muchos autores, que los datos obtenidos a partir de corpus lingüísticos son especialmente útiles en una etapa avanzada del aprendizaje de una lengua extranjera, también es verdad que todos los datos pueden ser posteriormente manipulados por el profesor adaptándolos éste a las necesidades concretas de sus alumnos. Si existieran suficientes medios electrónicos disponibles, la situación ideal sería la de permitir que los propios estudiantes aprendieran a manejar ciertos corpus para buscar información lingüística determinada y sacar conclusiones gracias a su propio esfuerzo.

Sin embargo, puesto que esta situación es todavía bastante utópica, especialmente en lo que se refiere a niveles elementales en el aprendizaje, animo a todos los profesores de lenguas extranjeras a que se familiaricen con estas nuevas herramientas de nuestro tiempo, que son extraordinariamente útiles, y que en el caso del inglés, están ya en un estadio avanzado de desarrollo tecnológico.

Bibliografía

AIJMER, K. & ALTENBERG, B. (eds.) (1991): *English Corpus Linguistics*. London: Longman.

AIJMER, K.; ALTENBERG, B. & JOHANSSON, M. (eds.) (1996): *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press.

ASTON, G. (1997) "Enriching the Learning Environment: Corpora in ELT". En: WICHMANN, A. et al. (eds.) *Teaching and Language Corpora*. Londres: Longman, 51 - 64.

BIBER, D. et al. (1998) *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

DODD, B. (1997) "Exploting a Corpus of Written German for Advanced Language Learning". En WICHMANN, A. et al. *Teaching and Language Corpora*. Londres: Longman, 131 - 145.

FLIGELSTONE, S.; RAYSON, P. & SMITH, N. (1996) "Template Analysis: Bridging the Gap between Grammar and the Lexicon". En THOMAS, J. & SHORT, M. (eds.) *Using Corpora for Language Research*. Londres: Longman, 181 - 207.

LEECH, G. (1997) "Teaching and Language Corpora: A Convergence". En WICHMANN, A. et al. *Teaching and Language Corpora*. Londres: Longman, 1 - 23.

ROJO, G. (1998) "La Lingüística Basada en el Análisis de Corpus y el Español". En el *Curso Emilio Alarcos*. Universidad de León.

THOMAS, J. & SHORT, M. (eds.) (1996) *Using Corpora for Language Research*. Londres: Longman.

WICHMANN, A. et al. (1997) *Teaching and Language Corpora*. Londres: Longman.

Universidad de Valladolid (25-27 de marzo, 1999)

LOS CORPUS LINGÜÍSTICOS Y LA ENSEÑANZA DEL INGLÉS

Noelia Ramón García
Universidad de León

Tipos de corpus

- Sincrónicos / diacrónicos
- Generales / especializados
- Cerrados / abiertos
- Monolingües / bilingües / multilingües
- Paralelos / comparables
- Etiquetados / no etiquetados

Cobuild Direct / Bank of English
e-mail: direct@cobuild.collins.co.uk
URL: http://titania.cobuild.collins.co.uk

CREA (Corpus de Referencia del Español Actual)
e-mail: corpus@rae.es
URL: http://www.rae.es/CREA.HTM

ICLE (International Corpus of Learners' English)
Sylviane Granger
Université Catholique de Louvain (Lovain-la-Neuve, Belgique)
URL: http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/cecl.html

Bibliografía

AIJMER, K. & ALTENBERG, B. (eds.) (1991): *English Corpus Linguistics*. London: Longman.

AIJMER, K.; ALTENBERG, B. & JOHANSSON, M. (eds.) (1996): *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*. Lund: Lund University Press.

ASTON, G. (1997) "Enriching the Learning Environment: Corpora in ELT". En: WICHMANN, A. et al. (eds.), 51 - 64.

BIBER, D. et al. (1998) *Corpus Linguistics. Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

DODD, B. (1997) "Exploting a Corpus of Written German for Advanced Language Learning". En WICHMANN, A. et al. (eds.), 131 - 145.

FLIGELSTONE, S.; RAYSON, P. & SMITH, N. (1996) "Template Analysis: Bridging the Gap between Grammar and the Lexicon". En THOMAS, J. & SHORT, M. (eds.), 181 - 207.

LEECH, G. (1997) "Teaching and Language Corpora: A Convergence". En WICHMANN, A. et al. (eds.), 1 - 23.

ROJO, G. (1998) "La Lingüística Basada en el Análisis de Corpus y el Español". En el *Curso Emilio Alarcos*. Universidad de León.

THOMAS, J. & SHORT, M. (eds.) (1996) *Using Corpora for Language Research*. Londres: Longman.

WICHMANN, A. et al. (eds.) (1997) *Teaching and Language Corpora*. Londres: Longman.