CRITERIOS DE SELECCIÓN DE UN CORPUS LINGÜÍSTICO ESPECÍFICO PARA EL ESTUDIO DE LA METÁFORA EN LA CIENCIA

PILAR DURÁN ESCRIBANO (pilar.duran@upm.es)
Mª JOSÉ GÓMEZ ORTIZ (maria.gomez.ortiz@upm.es)
Universidad Politécnica de Madrid

RESUMEN. *Esta comunicación comienza con el concepto y la evolución histórica de la utilización de corpora en la investigación lingüística, para pasar a detallar los criterios seguidos en la selección de un corpus específico para el estudio de los términos metafóricos en la ciencia y la tecnología. Siguiendo a Biber (1993) y a Sinclair (1991), los criterios más importantes son: la autenticidad de la muestra, su representatividad y su especificidad, es decir, su adecuación al propósito del estudio. A ellos, Krishnamurthy (2001: 85) añade el equilibrio de la representatividad de los campos y/o géneros seleccionados. La comunicación termina relacionando dichos criterios con el tema de nuestra investigación: el uso de los términos metafóricos en la ciencia, para sentar las bases a las otras ponencias de la misma Mesa Redonda de la que forma parte, que presentarán pormenorizadamente el corpus utilizado en el proyecto META-CITEC[1] y aportarán datos concretos sobre el mismo.*

PALABRAS CLAVE: *lingüística de corpus, creación de corpus, discurso científico, metáforas en la ciencia.*

ABSTRACT. *This paper starts with the concept and historical evolution of the use of corpora in linguistic research, to continue with the criteria followed in the selection of a specific corpus to study metaphoric terms in science and technology. As Biber (1993) and Sinclair (1991) point out, the most important criteria are: authenticity, representativeness and specificity, in other words, its suitability to the purpose of study. To these criteria, Krishnamurthy (2001) adds the balance in representativeness of the different fields and genres selected. The paper ends linking those criteria with our research topic: the use of metaphoric terms in science. It sets the basis to the other papers that will be presented in the Round Table. These papers will deal with the META-CITEC corpus in detail, adding more specific information about the research project as a whole.*

KEY WORDS: *corpus linguistics, corpora development, scientific discourse, metaphors in science.*

1. Introducción

Para los trabajos en lingüística aplicada, el empleo de corpus, es decir de una base de datos integrada por un conjunto de textos digitalizados para su análisis, no es algo nuevo, aunque sí relativamente reciente. Sinclair (1991:171) define corpus como "*a collection of naturally-ocurring language texts, chosen to characterize a state or variety of a language*", es decir, un conjunto de textos reales, orales o escritos, que se han seleccionado para reflejar el lenguaje en uso dentro de un contexto determinado. Sinclair, cuatro años después, ofrece otra definición más completa en la que incluye la finalidad como criterio de selección: "*collections of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of language*" (1995:18).

Desde la aparición de los ordenadores y su uso en investigación lingüística, se produce el auge del corpus computerizado, que, a su vez, se revela como un importante método de análisis de la lengua (Leech 1991). Su impulso se acrecienta con el avance en tecnología y en herramientas de procesamiento de datos, que hacen posible almacenar enormes colecciones

de textos auténticos variados, fiables y complejos para realizar estudios empíricos, como en el caso del estudio de la metáfora en la ciencia.

2. Evolución histórica del empleo de corpora en la investigación lingüística

La descripción del lenguaje mediante estudios basados en pequeños corpus lingüísticos ya figura a finales del siglo diecinueve (McEnery y Wilson 1996), aunque los córpora empleados entonces eran simplemente colecciones de interacciones transcritas. Es mucho después cuando se tienen en cuenta otros criterios, como el equilibrio y la representatividad –"*balance and representativeness*" – (Krishnamurthy 2001: 85), al llevar a cabo estudios lingüísticos de cierta envergadura, como es el caso del proyecto COBUILD en el que participó.

Siguiendo la evolución histórica del uso de corpus, hay que decir que entre los años 1930 a 50, se realizan abundantes estudios basados en colecciones de textos. En la década de los 60 y los 70, este método decae considerablemente, debido a la impopularidad que tuvo por las teorías de Chomsky (1968/1977), que estaba más interesado en la facultad innata del hombre para el lenguaje y propuso que fuera la intuición el instrumento de análisis más apropiado, en contraposición a los estudios empíricos.

Sin embargo, aunque impopular, su práctica no cesó totalmente y, a finales de los años 60, encontramos el *Survey of English Usage* de material oral transcrito, seguido de otros tres córpora: el *Brown University Corpus,* en 1964, (un millón de palabras de textos publicados en los Estados Unidos), el *Lancaster Oslo/Bergen Corpus* o *LOB,* en 1978, (versión británica del *Brown Corpus*), y el *London-Lund Corpus of Spoken English,* de 1980, (versión electrónica del *Survey*). Estos tres córpora contienen textos en formato electrónico. A partir de los 80, se recupera con fuerza esta metodología de análisis lingüístico, pero la gran diferencia radica en el uso de herramientas de procesamiento de textos. La creación de los tres corpora citados inicia la Lingüística de Corpus, de la que Aijmer y Altemberg dicen: "*Corpus Linguistics can be described as the study of language on the basis of text corpora"* (1991:1).

Llegado este punto, resulta patente que la lingüística de corpus queda establecida como área de análisis y descripción del lenguaje (Flowerdew 1998: 541) y que con ella se abren nuevas perspectivas de investigación de carácter léxico-gramatical y lingüístico-textual, que permiten abordar tres aspectos fundamentalmente: la gramática sistémico-funcional (Ghadessy 1995), el análisis de género, estudiando las características según el movimiento (Connor 1999; Upton y Connor 2001), y el análisis del discurso (Ferguson 2001; Conrad 2002), con anotaciones y etiquetados (Meyer 2002). Ya entrado el siglo veintiuno, Mukherjee (2004: 112) afirma que la lingüística de corpus es más que una metodología de análisis y que ha quedado consolidada como una disciplina dentro de la lingüística.

3. Ventajas de la utilización de córpora en lingüística aplicada

3.1. Ventajas de los córpora digitalizados para la investigación

Muchas son las ventajas que ofrece este medio de estudio frente al análisis de textos tradicional. Según Biber (1995: 32) los córpora ofrecen las siguientes:

1- *The adequate representation of naturally occurring discourse, including representative text samples from each register,*
2- *the adequate representation of the range of register variation in a language,*

3- *the (semi-)automatic linguistic processing of texts, enabling analyses of much wider scope than otherwise feasible,*
4- *much greater reliability and accuracy for quantitative analyses of linguistic features; that is, computers do not become bored or tired – they will count a linguistic feature in the same way every time it is encountered, and*
5- *the possibility of cumulative results and accountability. Subsequent studies can be based on the same corpus of texts, or additional corpora can be analysed using the same computational techniques.*

Los fines de los estudios basados en corpus, según Biber, Conrad y Reppen (1998:5), incluyen el análisis empírico y sus aplicaciones en ámbitos que detallamos a continuación.

3.2 Ámbitos de aplicación

Los resultados de la investigación en lingüística de corpus se aplican en los siguientes campos, con los que mencionamos algunos de los autores más representativos:

- compilación de diccionarios y gramáticas (Sinclair 1987; Flowerdew 1998);
- estudios de géneros lingüísticos (Biber 1995; Connor 1999);
- diseño de syllabus (Willis y Willis 1989);
- diseño de material de enseñanza (Tribble y Jones 1990);
- corpus interlenguas para material pedagógico (Granger 1998);
- estudios sobre el aprendizaje y enseñanza de L2 (Johns 1997; Stubbs 1996; Sánchez 1995; Álvarez de Mon y Millán 2004), y
- análisis del discurso en LFE (Swales 2004a; Biber, Conrad y Repen 1998; Ferguson 2001).

4. El corpus para fines específicos

4.1. Tipos de corpora

Existen distintos tipos de córpora dependiendo de la finalidad de cada investigación. Bowker y Pearson (2002: 11-13) mencionan los corpus de referencia general versus corpus de referencia especial, para identificar las características de un lenguaje especializado; el corpus escrito versus el corpus oral; el monolingüe versus el multilingüe; el sincrónico versus el diacrónico; el abierto versus el cerrado; el corpus del alumnado (escritos de estudiantes de una L2) que se puede comparar con textos de nativos.

Lo ideal sería que un corpus fuera grande y representativo. Biber (1993: 243) define la representatividad como "*the extent to which a sample includes the full range of variability in a population*". En este caso, lo que diferencia un corpus para fines generales de uno para fines específicos son los rasgos específicos de los últimos: tamaño, número y tipo de textos, medio, materia, autoría, lengua y fecha de publicación (Bowker y Pearson 2002: 3).

4.2. Los corpora de lenguas para fines específicos (L.F.E.)

Sager, Dungworth y McDonald (1980) profundizan en la naturaleza de los lenguajes de especialidad que, perteneciendo al lenguaje natural, igual que el lenguaje general, tiene unos rasgos propios determinados por su uso: "*The difference between general and special language is a difference of degree rather than kind: the degree to which the fundamental characteristics of language are maximised or minimised in special language*" (1980: 17). Por

lo tanto, si la especificidad del lenguaje está vinculada a su uso y al contexto en el que la comunicación tiene lugar, lo mismo se puede aplicar a los corpora de LFE.

Según Swales (2004b) el primer corpus importante en inglés científico data de 1971. Un buen medio de observar y describir las variaciones en el registro se basa en un número sustancial de textos representativos de una ciencia, de un campo profesional, de un registro o de un género particular. Una vez reunidos los textos y configurado el corpus específico, se pueden estudiar datos cuantitativos de frecuencia de co-ocurrencias y su distribución dentro y entre géneros, la co-variación de características lingüísticas con factores no lingüísticos, tales como los distintos autores, el canal empleado y el tipo de audiencia, (Biber, Conrad y Reppen 1998). En la misma línea, Ferguson (2001) habla de las indudables ventajas de los córpora electrónocos aplicados a LFE, pero sin excluir otros métodos de análisis complementarios.

4.3. Criterios para la selección de un corpus para fines específicos

La mayoría de los estudiosos están de acuerdo en que la definición más correcta de corpus es la que alude a los criterios seguidos para su formación; por ejemplo, Bowker y Pearson (2002: 9) lo definen así: "*a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria*". Aparece el concepto importante de que los textos se seleccionan de acuerdo con unos criterios.

Sinclair (1995:18) desarrolla los criterios mínimos de fiabilidad que deben seguirse para que un conjunto de textos en formato electrónico pueda cumplir la finalidad investigadora: cantidad, representatividad, variedad de muestras, equilibrio dentro de la variedad y documentación sobre su procedencia:

i. *the corpus should be as large as could possibly be envisaged,*
ii. *it should include samples from broad range of material in order to be representative,*
iii. *there should be an intermediate classification into genres between the corpus in total and the individual samples,*
iv. *the samples should be of an even size,*
v. *the corpus as a whole should have a declared provenance.*

Según (Bowker y Pearson 2002), los aspectos a tener en cuenta para la creación de un corpus con fines específicos son los siguientes:

- Tamaño: afirma que más grande no es siempre mejor, lo ideal es que esté bien diseñado según su finalidad; el corpus puede ir desde 10.000 palabras hasta varios cientos de miles.

- Textos: mejor completos y no fragmentos, ya que para la localización de los términos y conceptos dentro del texto, éstos pueden aparecer en cualquier parte.

- Número de textos: es importante considerar el número y cuántos han sido escritos por autores distintos para extraer los términos y conceptos más comunes.

- Medio: escrito, por ser más fácil de procesar, aunque no imprescindible.

- Materia, especialidad: para identificar los textos que pertenecen al dominio en cuestión.

- Tipo de texto: los textos escritos por expertos para expertos tienen un estilo y vocabulario distintos de los escritos por expertos para no expertos.

- Autoría: el autor debe ser un experto en la materia con base profesional adecuada.

- Lenguaje: textos escritos por nativos, y no nativos en caso de querer realizar estudios contrastivos.

- Fecha de publicación: indicativo del estado de la cuestión a nivel lingüístico y conceptual.

5. Comentarios finales

Hemos podido comprobar que los criterios anteriormente citados resultan válidos para recopilar un corpus específico para el estudio de la metáfora en la ciencia. En nuestro caso, al tratarse de un estudio de metáforas muertas, es decir, de metáforas lexicalizadas que se consideran términos específicos de uno o más campos de la ciencia y la tecnología, y que en su mayoría se pueden encontrar en los diccionarios, los miembros del GI DISCYT[2] que participamos en el proyecto META-CITEC[1] hemos acudido a las fuentes donde se pueden encontrar dichos términos con fiabilidad.

En la fase 1ª del estudio (búsqueda de términos metafóricos) hemos acudido a los diccionarios especializados. En la fase 2ª (contextualización de dichos términos), a los artículos de investigación e informes publicados en revistas periódicas especializadas y de prestigio, la mayoría en su edición electrónica, siguiendo los criterios que acabamos de exponer. Esto último con la finalidad de contextualizar dichos términos metafóricos, analizar su uso y poder concluir que los términos metafóricos identificados como tales por los miembros del GI DISCYT, a partir de los diccionarios, son utilizados como metáforas por los especialistas de las distintas ramas de la ciencia, en publicaciones especializadas.

Las otras ponencias de la Mesa Redonda, que vienen a continuación, presentarán pormenorizadamente el corpus utilizado en el proyecto META-CITEC y aportarán datos concretos sobre el mismo.

BIBLIOGRAFÍA

Aijmer, K. y Altenberg, B. eds. 1991. *English Corpus Linguistics.* Londres: Longman.

Alvarez de Mon, I. y Millán, M. 2004. "El aprendizaje de IFE mediante la creación de un corpus y su tratamiento con la herramienta Wordsmith Tools" en Sanz, I. y Felices, A. eds. *Las nuevas tendencias de las lenguas de especialidad en un contexto internacional y multicultural.* Valencia: UPV, 539-547.

Biber, D. 1993. "Representativeness in corpus design". *Literary and Linguistic Computing*, Vol. 8 (4): 243-257.

Biber, D. 1995. *Dimensions of Register Variation. A Cross-Linguistic Comparison.* Cambridge: CUP.

Biber, D., Conrad, S. y Reppen, R. 1998. *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: CUP.

Bowker, L. y Pearson, J. 2002. *Working with Specialised Language. A Practical Guide to Using Corpora.* London: Routledge.

Chomsky, N. 1977. *El lenguaje y el entendimiento.* Barcelona: Seix Barrall. (traducido de *Language and Mind,* 1968).

Connor, U. 1999. "Grant Proposals: An Important Research and Practical Genre in EAP", *English for Specific Purposes* 18 (1): 47-62.

Conrad, S. 2002. "Corpus Linguistics approaches for discourse analysis", *Annual Review of Applied Linguistics* 22: 75-95.

Ferguson, G. 2001. "If you pop over there: a corpus-based study of conditionals in medical discourse", *English for Specific Purposes* 20 (1): 61-82.

Flowerdew, L. 1998. "Corpus linguistic techniques applied to text linguistics", *System*, 26: 541-552.

Ghadessy, M. 1995. "Thematic development and its relationship to registers and genres" en Ghadessy, M. ed. *Thematic Development in English Texts.* Londres: Pinter, 129-140.

Granger, S. ed. 1998. *Learner English in Computer.* London: Longman.

Johns, A.M. 1997. *Text, Role and Context. Developing Academic Literacies*. Cambridge: CUP.

Krishnamurthy, R. 2001. "The Science and Technology of Corpus. Corpus for Science and technology" en Aguado de Cea, G. y Durán, P. eds. *La investigación en lenguas aplicadas: enfoque multidisciplinar*. Madrid: Fundación Gómez-Pardo, Universidad Politécnica de Madrid, 79-114.

Leech, G. 1991. "The State of the Art in Corpus Linguistics" en Aijmer, K. y Altenberg, B. eds.. *English Corpus Linguistics.* Londres: Longman, 8-29.

McEnery, T y Wilson, A. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Meyer, C.F. 2002. *English Corpus Linguistics: An Introduction.* Cambridge: CUP.

Mukherjee, J. 2004. "The state of the art in corpus linguistics: three book-length perspectivas", *English Language and Linguistics* 8.1: 103-119.

Sager, J. Dungworth, D. y McDonald, P. 1980. *English Special Languages. Principles and Practice in Science and Technology.* Wiesbaden: Brandstetter.

Sanchez, A. 1995. *Corpus lingüístico del español contemporáneo: CUMBRE.* Madrid: SGEL.

Sinclair, J. 1987. *Looking up: An Account of the COBUILD Project in Lexical Computing.* London: Collins.

Sinclair, J. 1991. *Corpus, Concordance, Collocation.* Oxford : Oxford University Press.

Sinclair, J. 1995. "Corpus typology: a framework for classifications", en Melchers, P. and Buarren eds. 1995. *Studies in Anglistics.* Stockhomp: Almquist & Wiksell International, 17-34.

Stubbs, M. 1996. *Text and Corpus Analysis.* Oxford: Blackwell.

Swales, J. 2004a. *Research Genres. Exploration and Applications.* Cambridge: CUP.

Swales, J. 2004b. "Then and now: A reconsideration of the first corpus of scientific English", *IBERICA* 8: 5-21.

Tribble, C. y Jones, G. 1990. *The Lexical Syllabus.* London: Collins.

Upton, T.A. y Connor, U. 2001. "Using computerised corpus analisis to investigate the textlinguistic discourse moves of a genre", *English for Specific Purposes* 20: 313-329.

Willis, D. y Willis, J. 1989. *Collins COBUILD English Course. Books 1-3.* London: Collins.

NOTAS:

[1] La base de datos METACITEC es fruto de un Proyecto de Investigación financiado por la Comunidad de Madrid y la UPM. Está integrado en las líneas de I+D para la Creación y Consolidación de los Grupos de Investigación de la UPM ("IV Programa PRICIT").

[2] GI DISCYT, grupo de investigación formado por profesores del Dpto. de Lingüística Aplicada de la UPM.