LOS CÓRPORA BILINGÜES CORRESPONDIENTES A ÁREAS CIENTÍFICO-TÉCNICAS, QUE INCLUYEN LA LENGUA ESPAÑOLA[1]

AMELIA DE IRAZAZÁBAL, PURIFICACIÓN FERNÁNDEZ e ISABEL DE ANDRÉS

ITBYTE. Universidad de Valladolid

0. INTRODUCCIÓN

El despegue de la lingüística del corpus en los últimos años ha transformado en gran medida la investigación lingüística permitiendo el estudio empírico de las lenguas.

Muchos han sido los desarrollos y aplicaciones surgidos a la luz de las expectativas despertadas a raíz de la aparición en 1961 del Brown Corpus. No obstante, a pesar de la proliferación de córpora en distintas lenguas, la obtención de textos científico-técnicos en formato electrónico plantea serias dificultades en cualquiera de ellas.

Por otra parte, la existencia de megacórpora, fundamentalmente en lengua inglesa en unos primeros momentos, juntamente con la evolución de las aplicaciones que incorporan la inteligencia artificial al análisis de los mismos, permitió su utilización, ya en los años 90, en otras parcelas tales como la del análisis y confección de los sistemas conceptuales en el trabajo terminológico, así como para la extracción automática, el análisis, y el almacenamiento del comportamiento de los términos en una lengua dada.

El propio Sager se hace eco de este hecho al afirmar que:

> *Systematic terminology compilation is now firmly corpus-based, i.e. terminology is no longer from previous lists or by individual searches but from a* corpus *of material.*[2]

A modo de aclaración previa sería conveniente comenzar por precisar los conceptos de corpus y corpus bilingüe, incluidos en el título de nuestra ponencia.

Cuando hablamos de corpus, referido, claro está, a las disciplinas lingüísticas, este concepto hace referencia a material textual presentado necesariamente en soporte electrónico. De hecho, el término *electronic* corpus o corpus electrónico, que años atrás era

[1] Debemos expresar nuestro agradecimiento al Prof. J. M. Bravo de la Universidad de Valladolid por su asesoramiento en la elaboración de esta ponencia, así como por habernos facilitado la actualización de algunos de los datos en ella recogidos, procedentes del libro J. M. Bravo Gozalo, J. M. y P. Fernández Nistal: *Los Córpora lingüísticos y los Estudios de Traducción* (en prensa).

[2] J. C. Sager: *A Practical Course in Terminolgy Processing,* Amsterdam-Philadelphia, John Benjamins, 1990, p. 130.

Lengua y Cultura. IRAZAZÁBAL, Amelia de; FERNÁNDEZ, Purificación; ANDRÉS, Isabel...
Centro Virtual Cervantes

de uso frecuente, ahora resulta redundante. Si seguimos a Leech y Fligelstone[3] los córpora podrían definirse como *"bodies of natural language material* [...] *which are stored in machine-readable form"*.

Esta definición parece lo suficientemente amplia como para abarcar muchos y muy diversos tipos de corpus. Sin entrar aquí en una clasificación rigurosa de los mismos, baste con señalar que no todos los córpora existentes en la actualidad presentan planteamientos, diseños, objetivos, o estrategias textuales uniformes. Cuestiones como el tamaño, la representatividad, o la actualización de los mismos, por mencionar tan sólo algunos de los aspectos claves a la hora de confeccionar y/o de clasificar un corpus concreto, pueden dar lugar a un amplio desarrollo de la categorización de los mismos.

1. LA TENDENCIA BILINGÜE FRENTE A LA TRADICIÓN MONOLINGÜE Y/O MULTILINGÜE

Por lo que respecta a los córpora bilingües, también éstos pueden presentar formatos diferentes y, al acercarnos a la bibliografía especializada, es fácilmente observable la existencia de una cierta fluctuación terminológica respecto a la utilización de las denominaciones córpora bilingües y córpora paralelos, para designar realidades cercanas, pero diferentes. Este es el motivo por el que consideramos básico profundizar en las diferencias conceptuales entre ambos términos.

Históricamente la mayoría de los datos demuestran que todos ellos están dedicados a una única lengua y, por lo general, podemos afirmar que la tendencia monolingüe es la más generalizada. Sólo recientemente la *Linguistic Data Corporation* (LDC) de los EE.UU. y la *European Language Resource Association*, por lo que a Europa se refiere, han empezado a distribuir productos textuales multilingües a mayor escala. Los datos consultados de la LDC ponen de manifiesto una presencia cada vez mayor de bases de datos textuales en español, fundamentalmente de países americanos, pero ocupando casi siempre una presencia minoritaria con respecto al inglés o al francés, por ejemplo, y con escasa o nula presencia de textos científicos que no sean de divulgación.

La proliferación de córpora multilingües, muy en boga en los últimos años, parece estar más orientada a la investigación en ingeniería lingüística que a su utilización para análisis lingüísticos complejos en los que intervendrían varias lenguas. También es de destacar la inclusión en los mismos de lenguas menos cercanas, tales como el cantonés, el japonés, el chino, etc.

Frente a esta orientación multilingüe ha existido una constante por parte de algunos sectores de la lingüística contrastiva, de la lexicografía y terminología bilingües, de los estudios de traducción, así como desde los estudios sobre bilingüismo en general, que han reclamado la necesidad de diseñar córpora bilingües que ayuden a desarrollar la investigación en las mencionadas áreas desde un punto de vista empírico.

La construcción de córpora bilingües,[4] es decir, aquellos que incluyen dos lenguas, puede adoptar diversas modalidades dentro de cada uno de los dos grandes grupos en

[3] G. Leech y S. Fligelstone: "Computers and Corpus Analysis", en Butler (ed.): *Computers and Written Texts,* Oxford, Blackwell, 1992, p. 115.

Lengua y Cultura. IRAZAZÁBAL, Amelia de; FERNÁNDEZ, Purificación; ANDRÉS, Isabel...
Centro Virtual Cervantes

los que se suelen clasificar: córpora paralelos o *parallel corpora* y córpora de traducciones o *translation corpora.*

Por lo que a los córpora paralelos se refiere, compuestos por textos escritos originalmente en las dos lenguas que integran el corpus, deberán proceder de entornos estilísticos semejantes y desarrollar una función comunicativa similar; ésta es la razón por la que, a veces, en algunos trabajos los expertos se refieren a ellos como *comparable corpora.* La disponibilidad de estos córpora es escasísima y aún más si necesitamos, sea cual sea el uso que pretendamos hacer del mismo, contar con que incluyan la lengua española. Sin embargo, éste es el tipo de córpora que, en principio, resultaría de mayor utilidad para el terminólogo (extracción automática, fraseología, etc.), para el traductor profesional (documentación paralela) y, por supuesto, para cualquier estudio serio de retórica comparada.

Entre las ventajas de dichos córpora, siempre que pertenezcan a un mismo campo de especialidad, destacaríamos el hecho de que nos permiten contar con textos comunicativos perfectamente comparables en los que las lenguas de ambos no se encuentran sometidas a la influencia propia de la lengua de partida, algo que ocurre con demasiada frecuencia en los córpora de traducciones.

La posibilidad de desarrollar una terminología del corpus en español, en áreas científico-técnicas a partir de este ideal de los córpora paralelos, se ve casi reducida a una mera declaración de intenciones, dada la dificultad, cada vez mayor, de encontrar textos de ciencia y tecnología del máximo nivel que estén escritos originariamente en nuestra lengua. La gravedad de la situación y las repercusiones de la misma al obstaculizar en gran medida el desarrollo del impacto social cualitativo a nivel mundial de las aportaciones que la ciencia y la tecnología españolas resultan bastante alarmantes.

No ocurre lo mismo con textos científicos de divulgación dada la enorme demanda social de publicaciones de este tipo, lo que ha obligado a los editores a incluir secciones específicas en los periódicos de mayor tirada nacional, cuando no a publicar "versiones" españolas de prestigiosas revistas escritas en inglés, tales como *National Geographic, Scientific American,* etc.[5]

La gran dificultad técnica que plantea este tipo de córpora es la de su explotación automática dado que, a pesar de utilizar una terminología y fraseología comparables, no permiten su alineamiento, por lo que tendremos que aplicar técnicas diferentes para su estudio, técnicas en las que se está investigando en la actualidad.

Si los córpora paralelos que incluyen la lengua española son prácticamente inexistentes, la situación es aún más complicada en el caso de los córpora de traducciones; es decir, córpora formados por un conjunto de textos de partida y sus respectivas traducciones. Ni que decir tiene que su utilidad dependerá directamente de la calidad de las mismas, pero ofrecerían una situación ideal para analizar traducciones, recoger logros alcanzados por traductores especialmente capaces, contrastar equivalentes y conceptos en ambas lenguas, etc.; por no hablar de su utilidad en el desarrollo de herramientas para la enseñanza de la traducción y de los estudios contrastivos, entre otros.

---

[4] Cf. J. M. Bravo Gozalo y P. Fernández Nistal, o. cit.

[5] Otra fuente de obtención de textos paralelos en varias lenguas sería a través de las agencias de prensa: Associated Press, Reuter, etc.

Lengua y Cultura. IRAZAZÁBAL, Amelia de; FERNÁNDEZ, Purificación; ANDRÉS, Isabel...
Centro Virtual Cervantes

El gran inconveniente de este tipo de córpora radica en la gran dificultad que existe para hacerse con ellos o para construirlos, debido a los problemas de carácter técnico que supondría el alineamiento y actualización a gran escala, por el hecho de que, a diferencia de lo que ocurre con los textos originales de los autores, quienes no suelen tener inconveniente en ceder el uso de sus obras para la confección de un corpus electrónico y poder ponerlos así al servicio de la comunidad científica, las agencias de traducción, quizás por temor a facilitar la labor de la competencia, tienden a guardar celosamente las traducciones que producen.

Si bien algunos organismos internacionales tales como las Naciones Unidas, o la Unión Europea son grandes productores de traducciones, la realidad es que tampoco éstos son fáciles de adquirir para tareas exclusivamente de investigación; es decir, sin fines lucrativos. Por otra parte, para que fueran rentables exigirían un trabajo enorme de clasificación temática, depuración y alineamiento.[6]

Así pues, la obtención de textos traducidos a gran escala que incluyan la lengua española resulta muy dificultoso.

2. LA ORIENTACIÓN ESPECIALIZADA FRENTE A LOS CÓRPORA GENERALISTAS

Tradicionalmente, el debate respecto a las características que deben definir un corpus, sea éste del tipo que sea, se ha centrado, ante todo, en su representatividad. A los especialistas en el diseño y construcción del corpus les ha interesado prioritariamente ofrecer una perspectiva globalizadora de la lengua recogida en el mismo —inglesa, española, francesa o alemana, por citar algunos ejemplos— y no tanto destacar las peculiaridades de un corpus en función de las características estilísticas diferenciadoras de textos representativos de distintos registros dentro de un mismo campo de especialidad.[7]

Esfuerzos pioneros como los que dieron lugar a proyectos para la lengua inglesa tales como el Brown (1961-64), el LOB (1970), o las grandes bases de datos textuales que componen el *British National* Corpus, o el *International* Corpus *of English*, entre los más utilizados y conocidos, presentan una estructura claramente orientada a la confección de un producto "representativo" de lo que en un momento dado podríamos describir como la lengua inglesa, ya sea en su variedad regional británica o norteamericana.

Dando por hecho la existencia de diferencias discursivas entre los que comúnmente se denominan "texto literario" y "texto científico", los diseñadores de estos córpora suelen escoger ciertos índices de lo que consideran representativo de los distintos géneros y actividades humanas, adscritas, por lo general, a alguna de las dos orillas de esta

---

[6] Tal vez el gran ejemplo de córpora bilingües de traducciones que existe en la actualidad sea el de los CANADIAN HANSARDS, con algo más de 100 millones de palabras, la mitad en inglés y la otra mitad en francés, incluyendo los temas más variopintos que puedan aparecer en los debates parlamentarios de ese país.

A menor escala cabría mencionar los informes anuales de la Unión de Bancos Suizos que se publican en inglés, francés y alemán.

[7] Existen, por supuesto, excepciones y en esta línea podría interpretarse la subclasificación textual del ICE.

Lengua y Cultura. IRAZAZÁBAL, Amelia de; FERNÁNDEZ, Purificación; ANDRÉS, Isabel...
Centro Virtual Cervantes

clasificación bipartita. Sin embargo, una vez dicho esto, sorprende comprobar hasta qué punto la actividad científico-tecnológica, tan trascendente y representativa desde un punto de vista socio-económico y de vital importancia en el desarrollo del léxico de las lenguas en la actualidad, se recoja en todos ellos de manera marginal.[8]

Sin olvidar nunca los derechos de autor, la única solución inmediata a nuestro alcance es la de iniciar la construcción de lo que podríamos denominar un "minicorpus" a partir de la utilización del CD-ROM, la extracción de textos de Internet, el escaneado, el tecleado masivo en aquellos casos en los que no exista otra posibilidad, por medio de sistemas de reconocimiento de voz, etc.

Por lo demás, desde el punto de vista de la investigación terminológica, el tipo de córpora que necesitaríamos presentaría una serie de características y restricciones específicas tales como: estar circunscrito a un(as) área(s) temática(s), la necesidad de constar de textos en la LO, su carácter dinámico —la arquitectura informática deberá permitir su revisión y actualización constantes—, etc.. En el caso de un corpus de español que responda a estas características deberá incluir, además, las muestras correspondientes a otras variedades regionales de nuestra lengua, debidamente etiquetadas con las anotaciones diatópicas pertinentes.[9]

La inclusión de traducciones científico-técnicas de calidad en córpora orientados a la investigación terminológica exigiría un tratamiento específico de las mismas, de tal manera que los usuarios potenciales del corpus tuvieran constancia del carácter de LT que están consultando y/o utilizando. Por otra parte, podrían resultar de gran utilidad para estudios terminológicos puntuales tales como el análisis de la sistematización en el uso de la terminología en servicios de traducción de grandes instituciones y de empresas en las que intervienen varios profesionales; desarrollar criterios de evaluación de las mismas; búsqueda de posibles soluciones dadas en el caso de equivalentes terminológicos nuevos o de difícil fraseología en la LT, etc.[10]

Si tenemos en cuenta los datos recogidos en el *Informe sobre recursos lingüísticos para el español (II).*Corpus *escritos y orales disponibles y en desarrollo en España*, llevado a cabo por el Observatorio Español de Industrias de la Lengua del Instituto Cervantes en 1996, sólo cuatro estarían dedicados de forma global a textos técnicos:

---

[8] The British National Corpus: *Natural and pure science 5%; applied science 5%.*

The International Corpus *of English*: de las 31 subdivisiones de textos escritos y orales sólamente seis de ellas (las relativas a textos impresos etiquetados como "informational: learned" e "informational: popular") podrían añadirse a las correspondientes subdivisiones dentro de los textos orales ("legal presentations" y las "business transactions"), lo que sumaría un total de 8, quedando una relación de 8/31. Cf. G. Leech: "100 Million Words of English", *English Today* (1993), 33, January, p. 13; y S. Greenbaum: "ICE: the International Corpus of English", *English Today* (1991), 28, October, p. 5.

[9] Cf. A. Sánchez y otros: CUMBRE. Corpus *lingüístico del español contemporáneo. Fundamentos, metodología y aplicaciones*, Madrid, SGEL, 1995, pp. 25-37. Dado el carácter general de la descripción del material que compone el corpus, nos ha resultado imposible precisar el porcentaje exacto de textos (orales o escritos) de carácter científico, a pesar de que se mencionen los epígrafes "técnica", "arquitectura" "ciencia" "medio ambiente", etc., tanto en el apartado dedicado al lenguaje oral como al escrito y no sólo para España, sino también para Hispanoamérica.

[10] Muy cercano a este uso de los córpora se encuentra el campo de las memorias de traducción, las cuales, en un determinado momento, siempre podrían incluirse en un corpus *bilingüe de traducciones alineadas*, sometiéndolas para conseguir dicha presentación al procesamiento adecuado a base de utilizar alguna de las herramientas informáticas existentes a tal efecto; por ejemplo BITEXXT.

Lengua y Cultura. IRAZAZÁBAL, Amelia de; FERNÁNDEZ, Purificación; ANDRÉS, Isabel...
Centro Virtual Cervantes

Corpus textual plurilingüe especializado (IULA-UPF), CRATER (Lancaster University/Universidad Autónoma de Madrid), LEGEBIDUN (Universidad de Deusto) y LAN (SPAL y DISUF) (Micro Focus SA.), todos ellos de tamaño, disponibilidad y características diferentes, pero que poseen la peculiaridad de incluir el español juntamente con otras lenguas en el caso de los tres primeros.

Una consulta breve hecha recientemente en Internet, utilizando el buscador español Olé[11] ha arrojado los siguientes datos:

1. Abundancia de textos científicos en español en el área de las telecomunicaciones, la física, la psiquiatría, la economía, la astronomía.

2. Textos bilingües (a veces trilingües) en inglés y en español (a menudo traducciones que no especifican la lengua de partida) en los campos de la medicina (cirugía pediátrica, psiquiatría), la farmacia, la astronomía, la informática (usuarios de ordenadores), la música (músicas alternativas), los deportes (golf, mundo del caballo), la economía, la enología, el turismo (guías), la tecnología del láser, los lenguajes de programación (Visual Basic, LISP).

3. Salvo contadas excepciones los textos pertenecían a revistas dirigidas no a expertos, sino al público en general.

3. CONCLUSIÓN

Después de describir un panorama bastante desolador, especialmente si lo comparamos con la disponibilidad de córpora en otras lenguas, parece necesario el dedicar mayores esfuerzos a la confección de córpora de áreas científico-técnicas que, desde una perspectiva bilingüe permitan, por una parte, el desarrollo de herramientas de etiquetado, extracción, análisis y actualización que faciliten la utilización de los mencionados córpora desde las más variadas parcelas del saber relacionadas con la/s lengua/s de la ciencia y de la tecnología en el mundo actual.

Juntamente con las dificultades añadidas para consultar algunos de estos córpora, debemos no olvidar el grave problema que plantean los derechos de autor a la hora de intentar confeccionar nuestro propio corpus a partir de los recursos disponibles que acabamos de mencionar.

Para terminar, permítasenos la licencia de expresar una opinión personal: nuestro total convencimiento de la necesidad de que trabajos y proyectos en esta línea deberían entusiasmar a jóvenes investigadores y animar a las autoridades a dedicar mayores recursos para su desarrollo, de forma que el mundo de las comunicaciones globales que se acerca, nos encuentre mejor preparados para la era en la que el procesamiento de los lenguajes naturales constituirá la clave del papel más o menos protagonista que una sociedad determinada jugará en las próximas décadas.

[11] Olé: 13/12/97 (palabras claves: publicaciones y periodismo/revistas). La utilización de otros buscadores no nos facilitó resultados significativos y ésta es la razón de no incluirlos aquí.

Lengua y Cultura. IRAZAZÁBAL, Amelia de; FERNÁNDEZ, Purificación; ANDRÉS, Isabel...
Centro Virtual Cervantes