Sobre algunos conflictos en la 'pre'-lexicografía: la selección de corpus para la elaboración de un diccionario contrastivo alemán-español

Vanessa González Ribao

Universidade de Santiago de Compostela / Red de Lexicografía RELEX

On some conflicts in 'pre'-lexicography: Corpus selection for the development of a German-Spanish contrastive dictionary

Abstract

This article is concerned with the choice of a corpus to be used as the empirical basis of a bilingual, bidirectional and conceptual learner dictionary of German and Spanish. Several standard corpora as well as web corpora for German and Spanish will be compared with respect to their size, the variety of genres they contain, the time span and geographical areas covered and what kind of search facilities they allow (e.g. word queries based on lemmata rather than on word forms). It will be argued that, when standard corpora fail to meet a particular requirement, web data may provide a useful alternative for lexicographical purposes provided they are both linguistically (i.e. morpho-syntactically) and meta-linguistically tagged.

Keywords

Corpus linguistics, web as/for corpus, bilingual lexicography, German, Spanish.

1. Introducción

La incursión en lexicografía, o mejor dicho, en las fases previas a la elaboración del producto lexicográfico de temas y cuestiones propias de la lingüística de corpus encuentra su explicación en el intento de valerse de esta última para optimizar los recursos de la primera. En tal hecho encuentran su justificación estas líneas, que han sido concebidas con el propósito de hacer hincapié en las dificultades que las tareas 'pre'-lexicográficas conllevan y reivindicar la necesidad de profundizar en este área, especialmente en relación al trabajo con corpus lingüísticos. Para ello, se tomarán

como ejemplo las reflexiones surgidas a este respecto en el marco del proyecto de investigación DICONALE.¹

El objetivo de DICONALE es crear un diccionario conceptual para la producción a partir de un nivel B2 en español y alemán como lenguas extranjeras y para la traducción indirecta a las mismas. Al basarse en una ordenación semántico-conceptual del léxico (es decir, el léxico se estructura y organiza en distintos (sub)campos conceptuales y (mini)paradigmas léxico-semánticos), este diccionario pretende ser de especial utilidad en la desambiguación y correcta selección (y en consecuencia en el uso correcto) de un lexema determinado de entre la variedad de posibles lexicalizaciones existentes para el concepto que se quiere expresar. La investigación se lleva a cabo siguiendo la metodología propia de la lingüística de corpus (Katinas 2012: 78-86; Biber / Conrad / Reppen 1998: 21-54) con la intención de obtener, así, un reflejo “real” de la lengua en uso. Por un lado, los datos del corpus relativos a la frecuencia tanto de las verbalizaciones como de sus significados han de servir como criterio para delimitar y acotar el campo de estudio del proyecto; a pesar de que la selección provisoria (realizada *a priori*) de los lexemas verbales (tanto simples como complejos) que se han de incluir en el diccionario toma como referencia los trabajos lexicográficos ya existente en éste área. Por otro lado, el análisis de las realizaciones concretas extraídas del corpus ha de permitir identificar estructuras recurrentes, colocaciones y combinaciones frecuentes y tendencias de uso en general, discernir ente lo habitual y lo excepcional y, sobre todo, registrar lo idiomático para ofrecer un equivalente lingüístico adecuado a nivel de la producción. De este modo, la información proporcionada por el corpus hará posible cubrir ciertos vacíos observados en la lexicografía contrastiva español-alemán (cfr. Meliss 2013a, b, c) y que DICONALE se ha propuesto subsanar (cfr. Meliss / Sánchez Hernández 2014).

El *grosso* de este artículo lo constituyen, por tanto, las reflexiones en torno a los criterios de selección de un corpus lingüístico para el español y para el alemán, respectivamente, cuyas características reúnan las condiciones necesarias para poder sustentar el proyecto lexicográfico tomado como ejemplo; haciendo especial hincapié en el aspecto contrastivo del mismo. Concretamente, se comenzará por realizar un breve repaso al estado de la cuestión presentando una selección de los corpus actualmente disponibles para el español y el alemán, describiendo sus características y discutiendo sobre los pros y contras de su utilización para los fines del proyecto en cuestión. A continuación, se expondrán, por un lado, los principales problemas y dificultades surgidos en torno a la utilización y compatibilidad de uso de los distintos corpus lingüísticos y, por otro, se reflexionará sobre el empleo de la web con fines lexicográficos y la viabilidad de esta alternativa. Una recapitulación final cerrará el artículo.

---

¹ “Elaboración de un diccionario conceptual alemán-español con acceso online (Financiación: MINECO FFI2012-32658: FEDER); IP: Prof. Meike Meliss (USC) y vinculado a la red de lexicografía RELEX (Financiación: Xunta de Galicia CN2012/290); Coord. M^a Dolores Sánchez Palomino (UDC).

2. Estado de la cuestión

El primer paso ha sido buscar información en torno a los corpus actualmente disponibles para el español y para el alemán, que podrían ser relevantes para los fines aquí perseguidos.

2.1. Corpus para el español

La tabla 1 muestra la selección de corpus del español que ha sido sometida a evaluación en el marco del proyecto y sus principales características:

CORPORA	Tamaño	Tipo de texto	Época	Lematizado	Área geográfica
CREA	~ 86 Mi. Palabras	~ 40 Mi. Palabras (Prensa), ~ 44 (Libros ²), ~ 2.5 (Otros)	1975-2004	No	Todas (peninsular ³)
DAVIES (CdE)	~ 100 Mi. Palabras (s. XX: ~ 20,5 Mi.)	s. XX: ~ 5 Mi. Palabras (x 4: académico, prensa, ficción, oral)	S. XIII - XX	Si	Todas
ESCOW	V12 ⁴ : ~ 1,6 Bi. Tokens V11: ~ 1,3 Bi. Tokens	Textos Web: escrito 86%, oral 2,5%, espontáneo 3,5%, blogmix 8% Textos Web	2012 2011	Si: TreeTagger Si: FreeLing	Español peninsular

² El término “libros” se utiliza aquí en contraposición a “prensa” (periódicos, revistas y publicaciones similares), hace referencia tanto a textos escritos en prosa como en verso y abarca todo tipo de géneros de ficción y de no-ficción (académico, de carácter divulgativo, etc.).

³ El CREA registra textos escritos en las distintas variedades geográficas del español, no obstante, es posible utilizar tal criterio como filtro y restringir la(s) consulta(s) a la(s) variedad(es) deseada(s).

⁴ Entre la colección de corpus COW existen para el idioma español dos versiones. Una versión ESCOW2012 (V12) realizada en 2012 y otra de prueba ESCOW2011 (V11) en el 2011, de la que no hay información disponible en relación a los textos que la componen. Para mayor información véase Schaefer / Bildhauer (2012).

Uni- Leipzig_es	1 Mi. Oraciones	Textos Web (Prensa)	2009- 2011	No	Todas
Ngram: Google Books_es	~ 84 Bi. Tokens	Libros	1500 - 2008	Si	Todas
DAVIES (GBC_es)	~ 45 Bi. Palabras (s. XX: ~ 9 Bi.)	Libros	1500 - 2008	Si	Todas

Tabla 1: Selección de corpus lingüísticos actualmente disponibles para el español.

Dentro de esta selección se pueden distinguir dos grupos: los corpus lingüísticos en línea “tradicionales” y los corpus originados siguiendo un sistema de *crawling* o araña web. El primer grupo lo integran el Corpus de Referencia del Español Actual (CREA) de la Real Academia Española (RAE) y el Corpus del Español de Mark Davies (Cde):

El CREA ha sido, hasta el momento, el corpus de referencia del español por excelencia para llevar a cabo estudios de la lengua escrita del s. XX y principios del s. XXI, debido a su tamaño y a la distribución relativamente equilibrada de sus textos. No obstante, presenta el inconveniente de haber sido cerrado en el año 2004, por lo que no resulta adecuado para realizar estudios de la lengua inmediatamente actual y la desventaja de no estar lematizado, lo que dificulta la búsqueda y consulta de ciertos fenómenos lingüísticos, sobre todo en cuanto a estructuras complejas o casos de apofonía. A pesar de ello, posee una interfaz agradable al uso que permite utilizar de forma sencilla filtros de selección que ayudan a especificar y restringir los criterios de búsqueda adaptándolos a las necesidades concretas de cada usuario.⁵

Por su parte el Cde también presenta una interfaz completa y flexible que sirve de especial utilidad para investigar aspectos gramaticales y sintácticos (p. ej. concordancias, coocurrencias o colocaciones), pues juega con la ventaja de estar lematizado. La selección y distribución de los textos que lo componen puede calificarse de variada y equilibrada; acorde con las premisas de la lingüística de corpus tradicional expuestas, entre otros, por Biber (1993) y Sinclair (1991) y revisadas por Leech (2007). Abarca el periodo comprendido entre el S. XV y el S. XX presentándose, así, como óptimo para llevar a cabo estudios de tipo diacrónico. Sin embargo, no permite investigar la lengua actual, ya que sus textos más recientes datan de 1999. Tampoco contempla la opción de seleccionar o distinguir *a priori*

⁵ Más información sobre CREA se encuentra disponible en <<http://corpus.rae.es/creanet.html>>.

entre las distintas variantes geográficas del español, por lo que no es la herramienta indicada para realizar estudios en los que la variedad lingüística sea relevante.⁶

El segundo grupo se presenta como la nueva generación de corpus lingüísticos. Se trata de recopilaciones de textos web, en el amplio sentido del término. La selección se realiza de forma automática mediante un *crawler* o araña web. Este método consiste en identificar y registrar textos web a partir de *seeds* previamente seleccionados. Los *seeds* son palabras clave que un motor de búsqueda genérico utiliza para identificar aquellas páginas web que establecen una conexión temática con los mismos e indexar posteriormente sus URLs. Para definir los *seeds* se suelen utilizar listas de palabras o vocabularios frecuentes extraídas de otros corpus, diccionarios o estudios piloto *ad hoc* (cfr. Fuertes-Olivera 2012:54). Este tipo de corpus se caracterizan por su extenso tamaño, lo que permite manejar gran cantidad de datos. Además, tienen la ventaja de trabajar con material muy actual, lo que también permite contemplar fenómenos *in statu nascendi*. Por el contrario, la calidad y cualidades de dicho material -textos web- son llevadas continuamente a discusión, especialmente en relación a su representatividad, debido a la imprecisión en la meta-información de las fuentes, así como a la dificultad de ordenar sus propiedades y características lingüísticas dentro de una tipología textual tradicional (vid. *infra* 4.1, 4.2, 5.).

Ejemplo representativo de este segundo grupo es el *Wortschatz* de la Universidad de Leipzig; proyecto a gran escala llevado a cabo por el departamento de ciencias computacionales de dicha universidad. Se trata de un portal lexicográfico basado en corpus monolingües generados automáticamente a partir de material web y que abarcan un amplio abanico de idiomas entre los que se encuentran el español y el alemán. El objetivo del portal es poner a disposición de cualquier usuario información lexicológica (paradigmática, sintagmática, sobre relaciones de sentido, coocurrencias y frecuencias de uso basadas en estadísticas, etc.) extraída y fundamentada en los mencionados corpus de material web (cfr. Quasthoff / Richter / Biemann 2006). Concretamente, para el español han recopilado gran cantidad de textos extraídos de periódicos en línea de habla española (ofrecen un corpus para el español de México y otro para el español de carácter internacional). No obstante, los principales inconvenientes a la hora de utilizar este recurso son: la imposibilidad de seleccionar *a priori* a la variante geográfica (dentro del corpus de habla española en general) y las dificultades que emanan del trabajo con textos no lematizados.

Dentro de este grupo también se inscriben los *corpora from the web* (COW) creados por el departamento de gramática alemana de la Universidad Libre de

---

⁶ En este punto puede llamar la atención que no se mencione dentro de este grupo al CORDE de la RAE. Esto se debe a que, a diferencia del *CdE* (el cual contiene textos que datan de finales del S. XX), el CORDE está constituido por textos anteriores a 1975 y por tanto, no se considera relevante en relación al proyecto utilizado aquí como ejemplo. Para más información sobre el *CdE* y el CORDE consúltese <<http://www.corpusdelespanol.org/>> y <<http://corpus.rae.es/cordenet.html>>, respectivamente.

Berlín. Los corpus COW siguen la metodología empleada en los proyectos BootCaT y WaCky (cfr. Schaefer / Bildhauer 2012:486) desarrollados en la Universidad de Trento por Marco Baroni y sus colaboradores (Baroni *et al.* 2004, 2009). Al igual que el proyecto de los colegas de Leipzig, se trata de un trabajo de gran envergadura y con amplias perspectivas de futuro, pues ya abarcan gran variedad de lenguas, entre las que se incluyen el alemán y el español, y pretenden extenderse aún más. Las ventajas que presenta el COW español (*ESCOW*) son las mismas que poseen todos los corpus de características similares: manejo de gran cantidad de datos totalmente actuales. En concreto, este corpus está compuesto por material web variado, como se puede observar en la Tabla 1. Está etiquetado y lematizado de forma automática, lo que permite realizar búsquedas en base a gran variedad de combinaciones morfosintácticas, aunque con cierto margen de error. Ha de ser consultado mediante algún programa de administración y análisis de corpus como *CWB* o similares. También se adapta fácilmente para trabajar de forma directa desde la interfaz de cualquier sistema Linux o MacOs. Recientemente han habilitado, además, un acceso online, el explorador COLiBrI, que permite consultar el corpus directamente desde la web sin necesidad de descargar ningún programa o aplicación específica, sin embargo, el uso de esta interfaz se encuentra restringido por el momento. Por las características arriba mencionadas, puede pronosticarse que los corpus COW serán en un futuro no muy lejano una herramienta idónea para llevar a cabo estudios contrastivos de la lengua actual. No obstante, el talón de Aquiles de dichos corpus reside, por un lado, en la cuestión de la representatividad de los textos que los componen (problema que se enmarca dentro de la discusión entorno a las propiedades y características lingüísticas del material web - *vid. infra* 4.1, 4.2, 5.) y por otro, en la imposibilidad de recuperación de las fuentes originales y otros metadatos (problema que, por norma general, surge al trabajar con gran cantidad de datos en línea extraídos por *crawling*).

Aparte de los corpus lingüísticos descritos, merecen mención los “productos google” que pueden ser utilizados con fines lingüísticos. Entre estos se distinguen *Google Ngram Viewer* (GNV) y *Google Ngram Corpus* (GNC), ambos basados en la recopilación de textos constituida por *Google Books* (GB). Estos recursos están constituidos por libros de diverso carácter que datan de entre el s. XV hasta el año 2008, por lo que podrían resultar de utilidad para un análisis diacrónico de este tipo de lengua escrita; sobre todo en relación a cuestiones de frecuencia de uso (cfr. Lin *et al.* 2012). No obstante, ha de tenerse en cuenta que se trata de una selección de textos que ha pasado por el filtro de Google, cuyos criterios de selección no son revelados al usuario (cfr. Leech 2007:144). Sin embargo, el lingüista Mark Davies ha creado para el español un corpus lingüístico basado en el *Google Books Spanish* (GBS) siguiendo la misma metodología y criterios científicos que caracterizan todos sus corpus.⁷ El material de GBS ha sido procesado desde la óptica lingüística de tal modo que al acceder a éste desde la interfaz del *Google Books Corpus* (GBC) para el español de Mark Davies puede consultarse utilizando sus opciones de búsqueda

---

⁷ Para más información al respecto véase <<http://corpus.byu.edu/>>.

avanzada. Además, como se trata de un corpus cerrado (45 billones de palabras) aporta fiabilidad a los estudios estadísticos que puedan realizarse en base al mismo.

2.2. Corpus para alemán

En cuanto a la selección de corpus realizada para el alemán (*vid.* Tabla 2), se puede efectuar la misma división que para el español. Por un lado, se encuentran los corpus lingüísticos con una relativamente larga tradición y por otro, los nuevos corpus de carácter más innovador constituidos con material en línea y basados en el método de *crawling*.

CORPORA	Tamaño	Tipo de texto	Época	Lematizado	Área geográfica
DeReKo	~ 6,1 Bi. Palabras	~ 3 Bi. Prensa (incl. Radio), ~ 1 Bi. Wikipedia, ~ 1,1 Bi. Libros, ~ 1 Bi. Otros	1964-2013	Si	Todas (Alemán de Alemania ⁸)
DWDS	Kernkorpus: ~ 123 Mi. Tokens Zeitungskorpus: ~ 1,7 Bi. Tokens	Kernkorpus: Ficción ~ 26%, Prensa ~ 27%, académico ~22%,propaganda ~ 20%, oral ~ 5%) Zeitungskorpus: Berliner Z. ~ 242 Mi. Tokens, Bild ~ 121 Mi., Süddeutsche Z. ~ 454 Mi., Tagesspiegel ~ 184 Mi., WELT ~ 238 Mi. ZEIT, ~ 417 Mi.	KK: S. XX (en construcción S. XXI) ZK: 1946-2009	Si	Todas

⁸ A través de COSMASII es posible predefinir un corpus de trabajo adaptado a las necesidades del usuario. De tal modo, es posible seleccionar aquellos textos y archivos que se corresponden con la variedad geográfica deseada, pudiendo trabajar así de forma independiente con cada una de ellas.

DECOW	V12 ⁹ : ~ 9,2 Bi. Tokens V11: ~ 1,2 Bi. Tokens	Textos Web: escrito 71%, oral 1%, espontáneo 22,5%, blogmix 4,5%	2012 2011	Si: TreeTagger Si: TreeTagger	Alemán de Alemania
Uni-Leipzig.de	~ 130 Mi. Oraciones	Textos Web (Prensa)	2005-2006	No	Alemán de Alemania
Ngram:Google Books.de	~ 65 Bi. Tokens	Libros	1500 - 2008	Si	Todas

Tabla 2: Selección de corpus lingüísticos actualmente disponibles para el alemán.

De entre los corpus de referencia del alemán destacan el *Deutsche Referenzkorpus* (*DeReKo*) del Instituto del Idioma Alemán (IDS) y el corpus del *Digitales Wörterbuch der deutschen Sprache* (*DWDS*) elaborado por la Academia de las Ciencias de Gotinga (AdW zu Göttingen) y la Academia de las Ciencias de Berlín y Brandemburgo (BBAW):

El corpus del IDS pretende servir de referencia y base para cualquier estudio en torno a la lengua alemana, por ello está compuesto por una extensa cantidad de textos del s. XX y XXI, además de una recopilación de obras de determinados escritores canónicos de la literatura de habla alemana, así como otros corpus específicos creados *ad hoc* para los distintos proyectos llevados a cabo en el IDS (*cfr.* Kupietz *et al.* 2010). En general, la composición de *DeReKo* es variada, aunque puede apreciarse una clara tendencia a la inclusión de material periodístico de tipo regional y cierta escasez en prosa de ficción; hechos que principalmente atienden a cuestiones prácticas relacionadas con los derechos de autor y la predisposición de los grandes grupos editoriales a ceder su material.¹⁰ Este corpus no solo está lematizado, sino que a él se asocia el sistema de búsqueda, administración y análisis de corpus del IDS: COSMAS II. Este recurso está dotado de una amplia oferta de filtros y criterios de selección que permite afinar detalladamente la búsqueda e incrementar la probabilidad de éxito y precisión de la consulta. Además, ofrece al usuario la posibilidad de configurar, a partir de los materiales que conforman los corpus de *DeReKo*, su propio corpus de trabajo adaptado a las exigencias concretas del mismo.

Por su parte el corpus del *DWDS* más que un corpus es, en realidad, un conjunto de corpus de entre los que destacan el *Kernkorpus* (corpus central o nuclear) y el *Zeitungskorpus* (corpus de periódicos). Por una parte, el *Kernkorpus* se compone por textos del S. XX y principios del XXI (en proyecto) cuya naturaleza está distribuida

⁹ Al igual que para el español conviven para el alemán una versión *DECOW2012* (V12) que data del 2012 y otra experimental del 2011 *DECOW2011* (V11) para la que no existe una descripción completa de sus propiedades (*cfr.* Schaefer / Bildhauer 2012).

¹⁰ Para más información al respecto consúltese <<http://www1.ids-mannheim.de/kl/projekte/korpora/>>.

de forma relativamente equitativa, como se muestra en la Tabla 2 (cfr. Geyken 2007: 25-30). Éste podría considerarse, dadas sus características, como el corpus de referencia alemán equivalente o más comparable al CREA, con la ventaja de que éste sí está lematizado, al igual que todos los corpus que componen el banco de datos del DWDS. Por otra parte, el *Zeitungskorpus* lo constituye una amplia variedad de textos extraídos de diversos periódicos, sobre todo de tirada suprarregional, lo que complementa la oferta de *DeReKo*. La consulta de los corpus del DWDS se realiza a través de una interfaz propia que permite filtrar la búsqueda en base a restricciones cronológicas, de autoría, tipo de texto, etc. Sin embargo, no es posible realizar un filtrado en base a las variedades geográficas del alemán y es que el punto débil de estos corpus radica precisamente en la infrarepresentación de las variedades suizas y austriacas (cfr. Geyken 2007: 30).

En lo relativo a los corpus del alemán elaborados principalmente por un *crawler* son válidas las mismas restricciones, ventajas e inconvenientes que se han comentado más arriba para sus homónimos españoles (cfr. 2.1), pues sus características son muy similares, a salvedad de que para el alemán no se ha creado aún ningún corpus específico basado en el *Google Books German* (GBG), como queda reflejado en la Tabla 2.

2.3. Breve recapitulación sobre el estado de la cuestión

De la comparación de los corpus seleccionados para el español y el alemán (*vid. supra* Tablas 1 y 2) se pueden extraer las siguientes conclusiones que sirven para cerrar este apartado sobre el estado de la cuestión. En resumen, los corpus tradicionales alemanes son corpus abiertos y ricos en material de prensa, mientras que en los corpus tradicionales españoles se observa una mayor presencia de textos literarios y a penas se registra material del S. XXI, pues están cerrados. Por su parte, los corpus “de nueva generación” se caracterizan, tanto en español como en alemán, por la actualidad de sus textos y su gran tamaño. Además juegan con la ventaja de poseer características similares para los dos idiomas aquí tratados, a pesar de que para el alemán no existe ningún corpus basado en GBG y el corpus de la Universidad de Leipzig para el español está menos trabajado que su homónimo alemán. No obstante, estos recursos no están salvos de inconvenientes, pues al estar generados de forma automática y mecánica presentan dificultades para recuperar información sobre las fuentes y otros metadatos, como por ejemplo es el caso de los corpus COW.

3. Dificultades en la búsqueda del corpus de trabajo adecuado

Requisito fundamental para llevar a acabo cualquier estudio contrastivo es que materiales, herramientas y condiciones de trabajo sean iguales para las distintas lenguas que se quieren comparar. Esto significa que los corpus han de poseer aproximadamente el mismo tamaño y enmarcarse cronológicamente en un mismo periodo. También la distribución y proporción de los distintos tipos de texto que los

componen han de ser similares. Para lenguas que presenten notables discrepancias entre sus variedades geográficas, como es el caso, es importante poder seleccionar la variedad con la que se quiere trabajar. Idealmente ambos corpus han de estar lematizados y poder ser analizados bajo criterios de búsqueda y filtrado comunes. En cuanto a la pareja de idiomas que atañe a este artículo, el principal problema para escoger un corpus adecuado reside precisamente en que los corpus lingüísticos ofertados para ambas lenguas (*vid. supra* 2.) son difícilmente comparables entre sí, tanto desde un punto de vista cuantitativo como cualitativo. Por un lado, el tamaño de los corpus es bastante dispar y por otro, la variedad y tipo de textos que los constituyen así como su proporción y distribución atienden a intereses distintos. Tampoco se puede establecer una comparación equilibrada en cuanto al periodo de tiempo o época en la que se inscriben los textos o en relación a la variedad lingüística que en estos se registra, sobre todo desde un punto de vista geográfico. Una posible solución sería valerse de los filtros y opciones de especificación de búsqueda que algunos corpus ofrecen, para así restringir el área de estudio a un grupo de textos de características similares y, por tanto, comparables. Por ejemplo, si se acota la búsqueda en el *Kernkorpus* del DWDS al período comprendido entre 1975 y 2004, entonces se obtiene un material de trabajo comparable al proporcionado por el CREA. De hecho, aún es posible afinar más la selección de textos. En el CREA podría efectuarse una búsqueda solo en aquellos textos que reúnan la siguiente combinación de cualidades: tipo = prensa, variedad lingüística = español peninsular, época = a partir del año 2000 y tema = todo (sin restricción). Si a su vez se restringe la búsqueda en *DeReKo* a textos de prensa alemanes datados del 2000 al 2004, se podrían utilizar sin problema los dos corpus para llevar a cabo un estudio contrastivo. No obstante, tal procedimiento suele generar nuevas dificultades a las que enfrentarse: por un lado, puede conducir a la obtención de un número de ejemplos insuficientes como para poder llevar a cabo cualquier estudio lingüístico y, por otro, es posible que la selección de textos obtenida tras el filtrado en el corpus español y la del corpus alemán no coincidan en su tamaño. En tal caso han de utilizarse otros criterios, por lo general de carácter arbitrario, que permitan restringir aún más la búsqueda en el corpus que corresponda, para poder trabajar con materiales cuantitativamente comparables. Por último, ha de tenerse en cuenta que las opciones de filtrado no siempre están disponibles para todos los corpus.

Otro problema añadido es la cuestión de la lematización, más concretamente, la ausencia de la misma. En muchas ocasiones los textos carecen de cualquier tipo de anotación morfosintáctica, lo que dificulta considerablemente su consulta. A pesar de ello, cada día los lingüistas idean nuevas estrategias de búsqueda valiéndose de los conocidos parámetros booleanos y de otros recursos estandarizados que, por norma general, casi todos los buscadores ofrecen. No obstante, en casos especialmente complicados, como son por ejemplo los verbos con prefijo separable en alemán, la apofonía o la metafonía, es necesario para el estudio de un mismo fenómeno realizar múltiples consultas en base a formas concretas y aisladas para combinarlas posteriormente. Este modo en que los datos son consultados implica un alto grado de reconstrucción y manipulación de los mismos condicionando, así, negativamente el valor de los resultados, sobre todo en relación a su objetividad y representatividad.

En lo que a estudios de tipo *corpus-driven* o con fines estadísticos se refiere, este método resulta definitivamente inviable.

Los siguientes ejemplos ilustran de forma clara esta problemática.¹¹ Para consultar en el CREA el verbo *difamar* con el fin de recuperar todas sus formas verbales para realizar un estudio sobre sus estructuras argumentales, se pueden seguir dos estrategias de búsqueda distintas. Una alternativa es utilizar la fórmula *difam** mediante la que se está pidiendo al buscador que muestre todas las formas registradas en los textos que comiencen por dicha combinación de caracteres. Esto significa que los datos devueltos por la máquina no solo contienen formas verbales, por lo que una revisión y limpieza manual de los datos es obligatoria. Concretamente, de dicha búsqueda en los textos de periódicos y revistas de España con la marca <CREA> se han recuperado 185 casos en 141 documentos. Tras la revisión manual de los mismos se obtienen 28 ejemplos que contienen el verbo *difamar*, de los cuales la mitad (14) se corresponden con la forma su infinitivo (incluida su combinación con todas las formas clíticas posibles). La segunda alternativa consiste en buscar por formas verbales concretas. Ciertamente es que se puede realizar una consulta para cada una de las formas conjugadas del verbo, sin embargo, lo más oportuno y rentable es establecer de antemano qué formas verbales concretas resultan de mayor relevancia para los fines del estudio en cuestión; teniendo siempre en cuenta la naturaleza del verbo y con vistas a obtener el mayor número posible de casos. De ahí que para el verbo *difamar* se hayan buscado en primer lugar las formas de presente y pretérito de indicativo de la tercera persona. Además de que por norma general esta persona verbal es la más recurrente,¹² en el caso concreto del verbo *difamar* carecería de sentido buscar formas en la primera persona, ya que resultaría verdaderamente extraño que alguien afirmase haber difamado a alguien, pues se estaría hablando entonces de un “suicidio elocutivo” (cfr. *illocutionary suicide* Searle / Vanderveken 1985). Pues bien, siguiendo los mismos criterios restrictivos de selección que para la consulta anterior, las formas encontradas para la tercera persona singular de presente (*difama*) han sido 2 y 3 más, para la forma de plural (*difaman*). Los casos obtenidos para el pretérito ascienden a 2, uno para *difamó* y otro para *difamaron* respectivamente. A mayores se ha consultado la forma *difamado* con vistas a obtener ejemplos en las formas de perfecto, sin embargo, al tratarse de una forma polisémica también se recuperan las formas masculinas de la pasiva y del adjetivo en singular. Finalmente, se puede contar con 3 ejemplos válidos para esta búsqueda, incluyendo casos en pasiva. Actuando de modo consecuente, se ha buscado también la forma pasiva de femenino *difamada*; lamentablemente sin obtener ningún resultado válido. La suma total de los casos hallados siguiendo este

¹¹ Se toma como ejemplo el verbo español *difamar* y su homólogo alemán *diffamieren* por considerarse verbos con una frecuencia de uso estándar en comparación con verbos de carácter más general como p. ej. *comunicar* / *mitteilen*. Para más información consúltense las listas de frecuencias del CREA y DWDS, respectivamente.

¹² Para verificar esta información también se han consultado *a posteriori* el resto de personas verbales en el modo y tiempos indicados. La hipótesis queda constatada, pues no se ha hallado ejemplo alguno.

método asciende a 10. Se puede concluir pues que ambos procedimientos conducen indistintamente a resultados similares: imprecisión de los datos recuperados y alto riesgo de falsos positivos, necesidad de revisión manual e insuficiente número de ejemplos, incluso como para llevar a cabo un *sample* o pequeño estudio piloto.

En cuanto a los corpus en alemán no lematizados, puede procederse del mismo modo. Por ejemplo, al realizar la consulta *diffamier** en el Corpus de la Universidad de Leipzig para el alemán de Alemania con la misma finalidad que la consulta anterior, se obtiene una lista con todas las formas registradas que empiezan por la secuencia *diffamier-* y el número de veces que éstas aparecen. Por supuesto, como se puede observar en la Figura 1, no todas las formas obtenidas son verbales e incluso se registran como formas léxicas casos en los que no se ha sabido identificar nombres propios, erratas u otros fallos tipográficos. Como ya se ha mencionado, esto se debe a que la búsqueda y selección de textos se realiza de forma automática y mecánica.

Figura 1: Formas recuperadas de la consulta *diffamier** en el Corpus de la Universidad de Leipzig para el alemán de Alemania.

Tras la eliminación de todos aquellos casos no relevantes para la investigación en cuestión, se procede a examinar detenidamente cada uno de los ejemplos que contienen las formas seleccionadas. El número final de casos válidos que resulta de

esta segunda criba puede ser muy variado. Ciertamente es que se podrían tomar todos los casos válidos sin restricción alguna, no obstante, para estudios contrastivos es recomendable establecer *numerus clausus* que permitan trabajar en ambas lenguas con un material comparable. Póngase a modo de ejemplo como cifra intermedia el 100. Es decir, que de todos los casos obtenidos se toman 100 por cada lexema consultado en cada idioma. Volviendo a los ejemplos anteriores, se puede observar que el hecho de establecer una restricción numérica conduce a dos posibles problemas. Por un lado, puede suceder que la suma total de todos los ejemplos recuperados para un lexema no alcance la cifra tope preestablecida, como ocurre con el caso de *difamar* en el CREA. Ante esta situación ha de prescindirse de estudiar tal lexema u optar por completar el material de trabajo utilizando otras fuentes, pues la cantidad inicial obtenida no se considera adecuada ni suficiente para llevar a cabo un estudio piloto o *sample* de calidad. Por otro lado, puede darse el caso contrario, que el número de casos recuperados sobrepase la barrera impuesta. Cuando se trabaja con corpus lematizados este hecho no suele suponer un grave problema, pues, por norma general, se ejecutaría un corte tras los 100 primeros ejemplos válidos recuperados sin afectar a la objetividad de la selección. No obstante, las dificultades se manifiestan cuando, como ocurre con el caso de *diffamieren* en el *Wortschatz* de la Universidad de Leipzig, los textos no están lematizados y se ha tenido que recurrir a la combinación de numerosas consultas para un mismo lexema. Por tanto, el resultado final no es una lista de ejemplos con formas mixtas distribuidas al azar, sino una amalgama de formas que se presentan por separado en diferentes listas. Esto obliga a efectuar una selección manual de los ejemplos, intentando respetar los criterios de representatividad y variedad en las formas. No obstante, siempre existen formas más frecuentes y recurrentes que otras dependiendo de la naturaleza del lema consultado, tal y como se puede observar en la Figura 1. La cuestión es entonces qué formas se han de escoger. En un intento de mantener las proporciones podría tratarse cada lista de forma independiente y realizar una selección porcentual en base al número de casos obtenidos por cada forma. Sin embargo, todavía queda abierta la duda de si este modo de proceder es realmente sinónimo de representatividad.

Aún mayores dificultades presenta la consulta de formas compuestas o complejas en textos no lematizados. Por ejemplo, para buscar el verbo con prefijo separable *mit|teilen* ('comunicar') en el *Wortschatz* de la Universidad de Leipzig no es posible realizar una consulta en la que se especifique que se desea obtener todos aquellos casos en los que además de la raíz de verbo *teilen* al final de la frase aparezca el prefijo separable correspondiente, aquí la partícula *mit* (*teilen ... mit*). Por consiguiente, la única alternativa posible es llevar a cabo la tarea hercúlea de buscar todas las formas de la raíz verbal *teilen* (que no solo puede actuar como base combinable con otros prefijos separables como por ejemplo *einteilen* ('dividir'), *zuteilen* ('distribuir'), o *austeilen* ('repartir'), sino que por sí sola ya constituye un verbo pleno ('partir' o 'compartir') y revisar manualmente todos los casos extrayendo solo aquellos pertinentes. Queda pues de relieve que tal procedimiento no es en absoluto rentable dado el alto número de falsos positivos y la gran inversión de tiempo que ello exige.

Por último, merecen mención los problemas que emanan del trabajo con corpus generados automáticamente y que registran grandes cantidades de material web. Como ya se ha explicado, dentro de estas grandes bases de datos no suele ser posible escoger de antemano o filtrar *a posteriori* el material utilizando el criterio de la variedad lingüística. Para idiomas como el español y el alemán que no solo se extienden por diferentes zonas geográficas de carácter muy diverso, sino que además son lenguas oficiales de múltiples países distintos, puede recurrirse al dominio geográfico de la URL que localiza un texto determinado para, así, poder deducir dentro de qué variante se inscribe. No obstante, el hecho de que una página web posea un dominio “.es”, por ejemplo, no es sinónimo de garantía para poder asegurar que los textos contenidos en dicha página están escritos por un hispanohablante nativo que utiliza una variante estandarizada del español peninsular. La dudosa fiabilidad de la información proporcionada por las URLs también afecta a la recuperación de las fuentes, la datación y otras informaciones que atañen a las propiedades del texto.

4. Web como alternativa

En los últimos tiempos se ha visto en la Web una posible solución a las carencias que a menudo los corpus lingüísticos pueden presentar. La Web es una herramienta de gran tamaño y alcance que se encuentra en continuo movimiento y renovación. Se caracteriza por la innovación, inmediatez y gran variedad de sus recursos; además de por la fácil accesibilidad a los mismos. Por ello y como bien defiende Fletcher (2011), se ha instaurado recientemente como una concurrente alternativa dentro de la lingüística de corpus:

The Web corpus owes its popularity to its tremendous size; broad linguistic, geographic and social range; up-to-dateness; multimodality; and wide availability at minimal cost [...] The Web is the most effective resource to research and document linguistic reflections of contemporary culture. In contrast, any static corpus is cut off at the moment of its compilation [...] Besides Web-only content, such a corpus can contain material available in other forms, such as articles from newspapers and magazines, entire books and even movie subtitles (Fletcher 2011: 1339-1340).

A este respecto existen básicamente dos posibles vías para valerse de tal herramienta (*cfr.* Lüdeling / Evert / Baroni 2007: 8-10). Por un lado, es posible consultar la web tal y como se procedería a consultar cualquier otra base de datos lingüísticos y por otro, se pueden aprovechar los materiales lingüísticos que ésta contiene para moldear un corpus de uso propio a partir de los mismos.

La cuestión entonces es constatar el estatus de la web como alternativa, es decir, comprobar si la web es capaz de cubrir las exigencias y reunir las cualidades propias de corpus adecuado para un estudio lexicográfico contrastivo (*vid. infra* breve

digresión); y en caso afirmativo, determinar cuál de las dos metodologías a seguir ofrece mejores posibilidades de trabajo para los fines aquí expuestos.¹³

4.1. Breve digresión

Teniendo en cuenta las reflexiones de Lüdeling, Evert y Baroni (2007), puede decirse que son 5 los tipos de estudio básicos que un corpus destinado a sustentar un trabajo lexicográfico de las características aquí descritas debe permitir llevar a cabo:

(i) Estudio cualitativo: Descripción del paradigma de los diferentes campos léxicos, detección de diferencias en los significados o nuevos usos, combinaciones, etc. Para poder llevar a cabo este tipo de estudio es requisito fundamental que el corpus sea estable, permita la reproducibilidad de los datos y posterior verificación de los resultados obtenidos y, sobre todo, que sus textos estén anotados morfosintácticamente, de modo que sea posible su consulta a través de un motor de búsqueda adecuado que a su vez garantice cierto grado de precisión en la recuperación de los datos.

(ii) Estudio distribucional: Determinar en qué contextos de uso concretos tiene lugar una u otra expresión determinada, resolver a qué tipo de textos y registros pertenecen ciertos lexemas u acepciones o detectar restricciones de carácter pragmático. Requisito fundamental para estas tareas es la anotación y etiquetación de los textos tanto a nivel lingüístico como extralingüístico. El corpus debe ofrecer suficiente meta-información sobre los materiales que lo componen para poder evaluar y categorizar los datos también desde un punto de vista social o geográfico, por ejemplo.

(iii) Estudio cuantitativo: En relación a cuestiones como la reproductividad o frecuencia de uso queda nuevamente de relieve la importancia de la estabilidad en un corpus y la proporcionalidad balanceada y equitativa en la distribución de sus textos.

(iv) Estudio diacrónico: Para poder señalar cuándo, cómo y eventualmente por qué un fenómeno lingüístico determinado aparece en la lengua y poder rastrear la posterior evolución del tal fenómeno ha de remitirse, de nuevo, tanto a la estabilidad y reproducibilidad del corpus como a la necesidad de (meta-)anotación de sus textos.

(v) Estudio contrastivo: En cuanto a la observación y análisis de las divergencias así como semejanzas entre los fenómenos lingüísticos de distintas lenguas, con independencia del nivel de lengua tratado (semántico, pragmático, sintagmático, paradigmático, etc.), es esencial que los aspectos calificados en los puntos anteriores como “fundamentales” se den de igual manera en los distintos corpus para las determinadas lenguas y que los materiales que los componen sean comparables.

---

¹³ En este punto se considera necesario realizar una breve digresión del tema para explicar concretamente cuáles son las características que ha de reunir un corpus orientado al trabajo en lexicografía contrastiva y poder evaluar así si la web es capaz de reunir tales condiciones.

4.2. Web como corpus

La única vía posible para servirse de los datos web es a través de un motor de búsqueda comercial (*Search Engine*, SE). Entre los más conocidos destacan por ejemplo Google, Alta Vista, Bing o Yahoo. Estos bien pueden ser consultados directamente como es la forma usual o bien se puede efectuar una búsqueda en base a los mismos, pero por medio de un motor de búsqueda de pre- y/o post-procesamiento lingüístico específico del tipo WebCorp (Renouf *et al.* 2005), KWICFinder (Fletcher 2001) o Linguist's Search Engine (LSE) (Elkiss / Resnik 2004), de modo que tal consulta perfeccionada conduzca a obtener resultados más refinados y precisos.

A pesar de ello, los datos han de ser interpretados con cautela (*cfr.* Meyer *et al.* 2003: 253), pues como ya se ha mencionado con anterioridad, el usuario ignora los algoritmos empleados por los buscadores comerciales no pudiendo evadir la influencia que los intereses ajenos a la lingüística (principalmente económicos y comerciales) ejercen sobre los datos recuperados en cualquier consulta. Y aún así, ¿es posible cumplir con los requisitos arriba mencionados (*vid.* 4.1) al utilizar la web como corpus?

De esos requisitos se deduce que un corpus lingüístico debe ser ante todo estable; no tanto en cuanto la representatividad y proporcionalidad o equilibrio de sus materiales, sino en relación a la reproducibilidad de sus datos (*cfr.* Lüdeling / Evert / Baroni 2007: 10-11). Por norma general, un corpus lingüístico es validado por un contra-corpus que contrasta y legitima los resultados obtenidos o por el contrario funciona como indicador de las carencias de sus datos ayudando a completarlos. Es evidente que para la Web como corpus no existe una "Web como contra-corpus" que actúe como agente controlador. De facto, entre las principales objeciones al uso de la Web como corpus destacan las dificultades existentes en torno a la recuperación e identificación de los datos y meta-datos de las páginas web (*cfr.* Kilgarriiff / Grefenstette 2003: 344-345). Un buen corpus, pues, no solo debe permitir la consulta de datos, sino también posibilitar la categorización y valoración de la información obtenida a partir de los mismos.

A este respecto, la utilización de la Web como corpus no aporta realmente una solución alternativa a los corpus presentados en los apartados 2.1 y 2.2, pues los textos web no suelen estar lematizados, de modo que, al igual que en los corpus lingüísticos que presentan este mismo inconveniente, la precisión en la recuperación de datos es relativamente baja e inversamente proporcional al alto número de falsos positivos que se obtienen, exigiendo, de igual modo, una gran inversión de tiempo y trabajo manual. Además, la Web es relativamente inestable, de ahí que los datos no sean siempre recuperables, la calidad y cualidad de sus textos sean en ocasiones difícilmente identificables y la meta-información en las páginas web, en caso de ser ofrecida, sea con frecuencia dudosa y poco fiable. En resumen, haciéndose eco de las palabras de Meyer *et al.*, puede afirmarse que "when conducting a search on the Web, we have no idea what kinds of texts our search results have been taken from" (2003: 241). Ciertamente es que por su naturaleza resultaría inapropiado establecer valoraciones sobre la web como corpus en base a criterios propios de los corpus

tradicionales (*cfr.* Schaefer / Bildhauer 2012: 248), no obstante, aún teniendo presente esta consideración, la respuesta a la pregunta arriba formulada sigue siendo negativa.

Una solución intermedia entre los corpus lingüísticos específicos y el uso directo de la Web, serían los bancos de datos en línea de tipo BDS, ADESSE, WiSo, LexisNexis, Lemmurproject, Wikipedia, entre otros. A pesar de que muchos de estos bancos de datos no están concebidos con fines exclusivamente lingüísticos, sus textos suelen estar anotados, sus buscadores permiten, por norma general, restringir y precisar las consultas por medio de distintos filtros y criterios de selección y el aspecto más decisivo: se puede identificar qué textos concretos los conforman y, por tanto, determinar exactamente su tamaño. No obstante y de forma sorprendente, el idioma español se encuentra en desventaja cuantitativa en lo que a este tipo de recursos se refiere, lo que para el estudio contrastivo español-alemán aquí tratado supone volver a enfrentarse al problema de la escasez de datos e imposibilidad de comparar los materiales disponibles para ambas lenguas; único obstáculo que con el uso directo de la Web como corpus se había salvado con éxito.

4.3. Web para corpus

Confeccionar un corpus de trabajo propio diseñado a medida valiéndose de los materiales disponibles en la Web parece la opción más propicia. Sin embargo, no es ni la más rápida ni la más sencilla. Por un lado, se ha de disponer de los recursos técnicos y económicos necesarios para extraer y procesar los textos (*cfr.* Lüdeling / Evert / Baroni 2007: 18), así como enfrentarse a cuestiones de autoría, derecho a la propiedad intelectual, plagio, etc., cuya frontera hacia la transgresión es, especialmente en relación al material web, de carácter difuso (Wilks 2004). Por otro lado, se ha de reflexionar sobre qué criterios generales y específicos se han de seguir para llevar a cabo la confección de un corpus determinado (Sinclair 1991, Biber 1993). Generalmente, estos criterios están regidos por la finalidad que el corpus está destinado a seguir; con independencia de si se trata de un corpus “tradicional” o de uno elaborado con material web (Leech 2007). Por tanto, lo que se ha de comprobar en este caso es si un corpus compuesto por material web es capaz de cubrir los criterios y requisitos descritos anteriormente (*vid. supra* breve digresión).

Para extraer material web se puede proceder bien realizando una selección previa de determinadas páginas, sitios o dominios web escogidos en base a la finalidad del futuro corpus o bien utilizando un *crawler* (*vid. supra*). En ambos casos lo que se está haciendo en realidad es “an enormous sample of the web” (Leech 2007: 144), una instantánea de la web que inmortaliza un momento o estado concreto de ésta, en otras palabras, un corpus offline (*cfr.* Claridge 2007: 89). Esto implica trabajar con un conjunto de materiales acotado, medible y reproducible, ganando así en estabilidad. Por un lado, un corpus offline determinado puede ser validado por un corpus offline sucesivo, es decir, por otras capturas de la web realizadas bajo las mismas condiciones pero en momentos posteriores; lo que al mismo tiempo significa que los corpus offline permiten realizar estudios desde una perspectiva

diacrónica (cfr. Claridge 2007: 89). Por otro lado, el éxito y precisión en la recuperación de los datos ya no depende de terceros, sino que la responsabilidad recae sobre el propio lingüista, pues es éste el que debe decidir bajo qué criterios y con qué recursos (programas de conversión, anotación, búsqueda, etc.) va a procesar y analizar sus textos. A pesar de que mediante este método es posible alcanzar, en gran medida, los requisitos arriba mencionados (*vid. supra* breve digresión), aún siguen surgiendo dificultades en torno a la categorización y valoración de los tipos de textos que componen un corpus offline determinado. Para poder realizar una distribución proporcional y balanceada de los textos que han de componer un corpus offline dado, ha de poderse llevar a cabo, en primer lugar, una etiquetación de los mismos. El sistema de etiquetado de las páginas web o “meta-tags” es de naturaleza extralingüística y atiende generalmente a razones económicas o comerciales, de ahí que no siempre resulten de utilidad para fines lingüísticos. Por su parte, los criterios de etiquetación lingüísticos que tradicionalmente definen las categorías textuales no siempre son aplicables a los textos web o en palabras de Karlsgrén (2000: cap. 15.1) “there is no well-established genre palette for Internet materials”. He precisamente aquí dónde, en base, reside el problema de la cuestión.

Como bien apunta Storrer (cfr. 2012: 277), la web ha evolucionado recientemente de tal manera que ha pasado de ser un medio de recepción a convertirse en un potente medio de comunicación. Conforme al desarrollo de nuevas formas de comunicación en la web, han surgido, de igual modo, dentro de la misma nuevos tipos de texto. En este sentido puede considerarse a la web como objeto de estudio en si misma. Ya no se trata solamente de una herramienta o recurso que permite generar un conjunto de textos, sino que también es posible entender la web como un tipo de discurso propio con sus respectivos géneros y categorías textuales. De ahí que haya que distinguir entre los conceptos *web corpus* (corpus web, ergo, de material web) y *web-based genre corpus* (corpus de la web, ergo, de géneros web como por ejemplo chat, foros, blogs, wikipedia, etc. - cfr. Fuertes-Olivera 2012: 62; Beißwenger *et al.* 2012: 1 y Gamallo / González 2010). En este marco, Beißwenger *et al.* (cfr. 2012: 2) demandan una colaboración interdisciplinar encomendada a la elaboración de modelos de descripción específicos y herramientas de análisis adecuadas para el estudio de estas nuevas formas de comunicación con el fin de entenderlas mejor; requisito previo fundamental para poder incluir de forma constitutiva en la investigación lingüística dichos géneros que cada vez están más presentes en las actividades diarias de nuestra sociedad (cfr. Storrer 2010).

De cualquier modo, independientemente del aprovechamiento que se quiera realizar de la web (bien como recurso o bien como fuente), queda de relieve la necesidad de identificar las características propias de los textos web para poder finalmente determinar, en términos lingüísticos, qué es lo que una página web contiene en realidad (cfr. Biber / Kurjian 2007: 110-112), pues una vez solventado este obstáculo se podrá retomar la discusión en torno al uso de dichos materiales con fines lexicográficos.

5. Recapitulación y desiderata

Se ha comprobado que el trabajo cooperativo entre la lingüística de corpus y la lexicografía ha favorecido el desarrollo recíproco de ambas disciplinas (*cfr.* Katinas 2012: 83-86). Desde hace unas décadas, al instaurarse la web como potente herramienta de trabajo e inmensa fuente de recursos también han querido estas dos disciplinas participar de las posibilidades que las nuevas tecnologías ofrecen. De ahí que actualmente ambas se ocupen de forma conjunta con cuestiones como las aquí planteadas en Meyer *et al.*:

is [it] possible to gauge the size of the Web and the rate at which its size is increasing, to determine with any certainty the kinds of texts that exist on the Web, to calculate which portion of the Web commonly available search engines actually obtain results from, and to describe the very real limitations of search engines in terms of what they can or cannot search for [?] (Meyer *et al.* 2003:241-242).

El deseo común es, por tanto, lograr responder positivamente a estas cuestiones en el futuro, pues si las barreras que presentan actualmente los materiales web fuesen vencidas, si

We could parse the Web. Web searches could be specified in terms of lemmas, constituents (e.g., noun phrase), and grammatical relations rather than strings. The way would be open for further anatomizing of Web text types and domains. Thesauruses and lexicons could be developed directly from the Web (Kilgariff / Grefenstette 2003: 345).

Para cubrir las expectativas arriba formuladas por Kilgariff y Grefenstette, debe llevarse a cabo un estudio multidimensional que atienda a criterios tanto lingüísticos como situacionales obteniendo, de este modo, una clasificación del material web (*cfr.* Biber / Kurjian 2007: 112). Por un lado, se han de analizar e identificar los parámetros lingüísticos de variación subyacentes en los textos. Por otro lado, se han de distinguir las características situacionales y de uso propias de cada contexto comunicativo para establecer un espectro de géneros o registros. La suma de ambos procedimientos definirá las distintas categorías textuales en base a las que se podrá elaborar un sistema de etiquetado adecuado a las necesidades del lingüista que permita a su vez optimizar los “meta-tags” de las páginas web y anotar los textos lingüísticamente de forma detallada (*cfr.* Meyer *et al.* 2003: 253). Dicha tarea ha de realizarse con la ayuda de los recursos y procedimientos metodológicos propios de la lingüística de corpus, de la lingüística computacional, de la lexicología, de la semántica, así como de la sociolingüística entre otras (*cfr.* Beißwenger *et al.* 2012: 2).

En este marco interdisciplinar, la clave para evitar o solventar gran parte de las dificultades está en que las relaciones entre las disciplinas sean establecidas correctamente. A este respecto, Fuertes-Olivera (2012: 67) advierte que “the corpus is fundamental in lexicography if it is not taken as an act of faith, but as a tool for performing certain tasks”, pues las tareas lexicográficas le corresponden al lexicógrafo y es éste el encargado de valorar, en última estancia, cómo y en qué medida ha de ser utilizada dicha herramienta.

Es, precisamente, en este contexto en el que se debe entender la noción de representatividad. Ciertamente es que tal noción es severamente cuestionada dentro de la lingüística de corpus y que en relación al material web goza aún de una mayor controversia a la habitual debido al desconocimiento de la naturaleza de dicho material. Sin embargo, tal y como explican Kilgariff y Grefenstette, la web puede ser tan representativa como cualquier otro corpus:

The Web is not representative of anything else. But neither are other corpora, in any well-understood sense. 'Text type' is an area in which our understanding is, as yet, very limited. Although further work is required irrespective of the Web, the use of the Web forces the issue. Once we move to the Web as a source of data, and our corpora have names like 'April03-sample77' the issue of how the text type(s) can be characterized demands attention (Kilgariff / Grefenstette 2003: 243).

Como se ya ha explicado, un corpus debe ser confeccionado teniendo en cuenta la finalidad que está destinado a seguir y es, por tanto, en base a dicha finalidad que su representatividad debe ser medida. En este sentido, la web ofrece la posibilidad de configurar un corpus *X* que sea representativo de *Y* para la finalidad *Z*. Por consiguiente, se puede afirmar que la web, a pesar de las dificultades, sí es una alternativa válida, es decir, la web como o para corpus puede ser útil para optimizar las tareas (pre-)lexicográficas. Esta argumentación es, por lógica extensión, igualmente válida para los corpus de nueva generación originados a partir de material web.

A lo largo del artículo se ha observado que el trabajo con los corpus tradicionales disponibles así como con los corpus de la Universidad de Leipzig es costoso y poco rentable para los fines aquí perseguidos (*vid. supra* 3.) y que los recursos Google y otras herramientas lingüísticas (*vid. supra* 2.3, 4.1), a pesar de ser de gran utilidad para cuestiones concretas, no sirven para llevar a cabo un estudio como el aquí planteado. También se ha comprobado que el uso de material web para fines lingüísticos, a pesar de ser polémicamente discutido, resulta útil y ventajoso (*vid. supra* 4., 5.). De ahí que este texto concluya con un alegato en pro de los corpus COW, pues son totalmente comparables entre sí tanto en tamaño y composición como en relación a la variedad geográfica y al intervalo cronológico que abarcan. Su gran extensión y la actualidad de sus textos no solo permiten establecer frecuencias de uso e identificar fenómenos lingüísticos y estructuras recurrentes, sino también registrar aquellos casos excepcionales y poco estudiados; y además, aunque automáticamente, están etiquetados (*vid. supra* Tabla 1 y 2). No obstante, el desiderátum final aboga por la mejora en el aprovechamiento de dichos corpus, en la medida en que se haga posible recuperar información sobre las fuentes y otros metadatos, se revise la anotación morfosintáctica y se optimicen las condiciones de uso, búsqueda y, sobre todo, acceso a los mismos.

Referencias bibliográficas

- Baroni, Marco / Bernardini, Silvia / Ferraresi, Adriano / Zanchetta, Eros (2009): «The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora», *Language Resources and Evaluation*, 43.3, 209-226.
- Baroni, Marco / Bernardini, Silvia (2004): «Bootcat: Bootstrapping corpora and terms from the web», *Proceedings of LREC 2004*, 1313-1316.
- Beißwenger *et al.* (2012): «A TEI Schema for the Representation of Computer-mediated Communication», *Journal of the Text Encoding Initiative (TEI and Linguistics)*, Noviembre 2012/3, 1-85. Disponible en <<http://jtei.revues.org/476>> [Consulta: 24/09/2014].
- Biber, Douglas (1993): «Representativeness in Corpus Design», *Literary and linguistic computing*, 8.4, 246-257.
- Biber, Douglas / Conrad, Susan / Reppen, Randi (1998): *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.
- Biber, Douglas / Kurjian, Jerry (2007): «Towards a taxonomy of web registers and text types: a multidimensional analysis», en Marianne Hundt *et al.*, eds., *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 109-131.
- Claridge, Claudia (2007): «Constructing a corpus from the web: message boards», en Marianne Hundt *et al.*, eds., *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 87-108.
- Elkiss, Aaron / Resnik, Philip (2004): *The Linguist's Search Engine User's Guide*. Disponible en <<http://lse.umiaccs.umd.edu/lseuser.pdf>> [Consulta: 17/09/13].
- Fletcher, William H. (2011): «Corpus Analysis of the World Wide Web», en Carol A. Capelle (2011), ed., *Encyclopedia of Applied Linguistics*. Wiley-Blackwell, vol. III, 1339-1347.
- (2001): «Concordancing the Web with KWicFinder», *Proceedings of the 3rd North American Symposium on Corpus Linguistics and Language Teaching*, Boston. Borrador. Disponible en <<http://kwicfinder.com/FletcherCLLT2001.pdf>> [Consulta: 17/09/13].
- Fuertes-Olivera, Pedro Antonio (2012): «Lexicography and the Internet as a (Re-)source», en Rufus Hjalmar Gouws *et al.*, eds., *Lexicographica. International Annual for Lexicography / Revue Internationale de Lexicographie / Internationales Jahrbuch für Lexikographie*, 28, 49-70.
- Gamallo, Pablo / González, Isaac (2010): «Wikipedia as Multilingual Source of Comparable Corpora», *Proceedings of the 3rd Workshop on Building and Using Comparable Corpora*, LREC 2010, Malta, 21-25.
- Geycken, Alexander (2007): «The DWDS corpus: a reference corpus for the German language of the twentieth century», en Christiane Fellbaum, ed., *Idioms and collocations: corpus-based linguistic and lexicographic studies*. London: Continuum, 23-40.
- Hundt, Marianne / Nesselhauf, Nadja / Biewer, Carolin, eds. (2007): *Corpus Linguistics and the Web*. Amsterdam: Rodopi.
- Karlsgrén, Jussi (2000): *Stylistic Experiments for Information Retrieval*. Stockholm: SU.
- Katinas, Daumantas (2012): *Zur Position und Methodik der kontrastiven Lexikologie: eine Untersuchung am Beispiel des lexikalischen Feldes Relief im Litauischen und im Deutschen*. Marburg: Tectum.
- Kilgariff, Adam / Grefenstette, Gregory (2003): «Introduction to the Special Issue on the Web as Corpus», *Computational Linguistics*, 29.3. Reeditado en Thierry Fontenelle (2008), ed., *Practical Lexicography: a Reader*. Oxford: OUP, 89-101.
- Kupietz, Marc / Belica, Cyril / Keibel, Holger / Witt, Andreas (2010): «The German Reference Corpus DeReKo: A primordial sample for linguistic research» en Nicoletta Calzolari *et al.*, eds., *Proceedings of the 7th conference on International Language Resources and Evaluation*, LREC 2010, Malta: European Language Resources Association (ELRA), 1848-1854.

- Leech, Geoffrey (2007): «New resources, or just better old ones? The Holly Grail of representativeness», en Marianne Hundt *et al.*, eds., *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 133-149.
- Lin, Yuri *et al.* (2012): «Syntactic Annotations for the Google Books Ngram Corpus», *50th Annual Meeting of the Association for Computational Linguistics (ACL2012)*, Korea. Disponible en <<http://www.petrovi.de/data/acl12b.pdf>> [Consulta: 24/09/2014].
- Lüdeling, Anke / Evert, Stefan / Baroni, Marco (2007): «Introduction Using web data for linguistic purposes», en Marianne Hundt *et al.*, eds., *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 7-24.
- Meliss, Meike (2013a). «Das zweisprachige Wörterbuch im bilateralen deutsch-spanischen Kontext. Alte und neue Wege», en M^a José Domínguez Vázquez, ed., *Trends in der deutsch-spanischen Lexikographie* (Sapnische Akzent 1), 61-87.
- (2013b): «Online-Lexikographie im DaF-Bereich: Eine erste kritische Annäherung: Bestandsaufnahme – Nutzen – Perspektiven», *Real Revista de Estudos Alemães*, 4, 176-199.
- (2013c): «Das verbale Kombinationspotenzial in einsprachigen DaF-Lernerwörterbüchern: Kritische Bestandsaufnahme – Neue Anforderungen», *ZDaF* (en prensa).
- Meliss, Meike / Sánchez Hernández, Paloma (2014): «Theoretical and methodological foundations of the DICONALE project: a conceptual dictionary of German and Spanish», en João Paulo Silvestre *et al.*, eds., *Dicionários que não existem. Tipologias lexicográficas*. Lissabon. (en prensa)
- Meyer *et al.* (2003): «The World Wide Web as Linguistic Corpus», en Pepi Leistyna, Charles F. Meyer, eds., *Corpus Analysis. Language Structure and Language Use*. Amsterdam/New York, 241-254.
- Quasthoff, Uwe / Richter, Matthias / Biemann, Christian (2006): «Corpus Portal for Search in Monolingual Corpora», *Proceedings of the LREC 2006*, Genova, 1799-1802.
- Renouf, Antoinette / Kehoe, Andrew / Banerjee, Jay (2005): «The WebCorp Search Engine: a holistic approach to Web text Search», *Electronic Proceedings of CL2005*, University of Birmingham. Disponible en <http://rdues.bcu.ac.uk/publ/WebCorp_holistic_approach.pdf> [Consulta: 24/09/2014].
- Schaefer, Roland / Bildhauer, Felix (2012): «Building Large Corpora from the Web Using a New Efficient Tool Chain», en Nicoletta Calzolari *et al.*, eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, ELRA, 486-493.
- Searle, John R. / Vanderveken, Daniel (1985): *Foundations of Illocutionary Logic*. Cambridge: CUP.
- Sinclair, John McHardy (1991): *Corpus, concordance, collocation*. Oxford: OUP.
- Storrer, Angelika (2012): «Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia», en Juliane Köster, Helmuth Feilke, eds., *Textkompetenzen für die Sekundarstufe II*. Freiburg: Fillibach, 277-304.
- (2010): «Über die Auswirkungen des Internets auf unsere Sprache», en Hubert Burda, Mathias Döpfner, Bodo Hombach, Jürgen Rüttgers, eds., *2020 – Gedanken zur Zukunft des Internets*. Essen: Klartext Verlag, 219-224.
- Wilks, Yorick (2004): «On the Ownership of Text», *Computers and the Humanities*, 38,2, 115-127.

Corpus y recursos lingüísticos

ADESSE. *Base de datos de Verbos, Alternancias de Diátesis y Esquemas Sintáctico-Semánticos del Español*: <<http://adesse.uvigo.es/>>.

BDS. *Base de Datos Sintácticos del español actual*: <<http://www.bds.usc.es/>>.

BootCat. *Simple Utilities to Bootstrap Corpora And Terms from the Web*: <<http://bootcat.sslmit.unibo.it/>>.

CORDE. *Corpus diacrónico del español*: <<http://corpus.rae.es/cordenet.html>>.

COW. *Corpora from the Web*: <<https://hpsg.fu-berlin.de/cow/>>. Consultado con COLiBrI, COW's Light Browsing Interface: <<http://hpsg.fu-berlin.de/cow/colibri/>>.

CREA. *Corpus de referencia del español actual*: <<http://corpus.rae.es/creanet.html>>.

CWB. *The IMS Open Corpus Workbench*: <<http://cwb.sourceforge.net/>>.

Davies Corpus: CdE, *Corpus del Español y GBS*, Google Books Spanish: <<http://corpus.byu.edu/>>.

DeReKo. *Deutsches Referenzkorpus*: <www.ids-mannheim.de/DeReKo>. Consultado con COSMAS II, *Corpus Search, Management and Analysis System*: <<http://www.ids-mannheim.de/cosmas2/>>.

DWDS. *Das Digitale Wörterbuch der deutschen Sprache*: <<http://www.dwds.de/ressourcen/korpora/>>.

GNV. *Google Ngram Viewer*: <<https://books.google.com/ngrams>>.

KWiCFinder. *Key Word in Context search tool for the web*: <<http://www.kwicfinder.com/KWiCFinder.html>>

Lemurproject. *The Lemur Project*: <<http://www.lemurproject.org/>>.

LexisNexis. *Deutschesprachige Quellen*: <<http://www.lexisnexis.de/quellen/deutschsprachige-quellen>>.

LSE. *Linguist's Search Engine*: <<http://lse.umiaccs.umd.edu/>>.

WaCky. *The Web-As-Corpus Kool Yinitiative*: <<http://wacky.sslmit.unibo.it/doku.php>>.

WiSo. *Datenbank für Hochschulen*: <<http://www.wiso-net.de/webcgi?START=03A&SEITE=amedien.tin>>.

WebCorp. *Concordance the web in real-time*: <<http://www.webcorp.org.uk/live/>>.

Wortschatz Uni Leipzig: <<http://corpora.informatik.uni-leipzig.de/>>.