Revista de Lingüística y Lenguas Aplicadas
Vol. 13 año 2018, 59-73
EISSN 1886-6298
https://doi.org/10.4995/rlyla.2018.7912

DISEÑO Y COMPILACIÓN DE CORPUS PARALELOS ALINEADOS: DIFICULTADES Y (ALGUNAS) SOLUCIONES EN EL EJEMPLO DE UN CORPUS DE TEXTOS MUSEÍSTICOS TRADUCIDOS (INGLÉS-ESPAÑOL)

Jorge Leiva Rojo
Universidad de Málaga

Resumen: Los corpus de textos son herramientas de larga tradición y numerosas aplicaciones. De todos los tipos existentes, este trabajo se centra en uno en concreto: el corpus paralelo alineado. Tomando como punto de partida un corpus paralelo alineado de textos museísticos escritos originariamente en inglés y traducidos al español, se propone una metodología que pasa por cuatro fases esenciales. Con el apoyo de trabajos previos al respecto, y tras el recurso a programas de software —de pago y gratuitos, específicos y creados con otros fines—, se comprueba que, si bien es posible llevar a cabo la compilación del corpus, el camino está lleno de obstáculos, algunos salvables y otros no, como ha sucedido con la conservación de repeticiones en el corpus alineado.

Palabras clave: lingüística de corpus, textos museísticos, traducción, textos paralelos alineados, bitextos.

DESIGNING AND COMPILING PARALLEL ALIGNED CORPORA: PITFALLS AND (SOME) SOLUTIONS ON THE EXAMPLE OF A CORPUS OF TRANSLATED MUSEM TEXTS (ENGLISH-SPANISH)

Abstract: *Text corpora are tools having both a long tradition in research and a variety of applications. Of all existing types, this paper focuses specifically on parallel, aligned corpora. By taking one of this corpora as a starting point—a parallel, aligned corpus from museum texts originally written in English and subsequently translated into Spanish—, the aim of this article is to propose a methodology that consists of four basic stages. By the revision of previous literature on the topic, and by using multiple software programs—proprietary and free, specifically created for corpus compilation and created for other purposes—, it is concluded that, although the compilation of corpora such as the one that was intended is a feasible task, the procedure is full of obstacles. Some obstacles were overcome, while some were not; that is the case, for example, of the repetitions on the aligned corpus, which are not present in the corpus.*

Key *words: corpus linguistics, museum texts, translation, parallel-aligned texts, bitexts.*

> S[ource texts] and T[arget texts] are usually published in separate volumes, or on separate pages […] so that making comparisons between the two requires a good deal of patience. (Harris, 1988, p. 8)

1. INTRODUCCIÓN

La opinión de que la lingüística de corpus ha sido un campo de estudio prolífico desde hace tiempo es unánime. Son numerosos los trabajos que se consideran *clásicos* en este campo, como es el caso de las contribuciones de Sinclair (1991), Baker (1995), EAGLES (1996) y McEnery y Wilson (2001, primera edición publicada en 1996), por citar solo unos pocos.[1] Las investigaciones que hacen uso de los corpus como herramienta metodológica en los últimos años siguen siendo muy numerosas, como lo demuestran Zanettin, Saldanha y Harding (2015:180) en referencia a los estudios de traducción, ya que "in the last ten years or so about 1 out of 10 publications in the field has been concerned with or informed by corpus linguistics methods" (Fantinuoli y Zanettin, 2015:8). Un ejemplo más, igualmente ilustrativo, es el que se encuentra en esta revista: de los 153 artículos que han aparecido en los 12

[1] Esta lista, como es de esperar, es meramente enunciativa. Consúltense los trabajos de Parodi (2008) y Rojo (2008, 2015), para una visión general de la lingüística de corpus (con especial hincapié en el ámbito hispánico), y la exhaustiva lista de bibliografía sobre lingüística de corpus que proporciona Llisterri (2017).

To cite this article: Leiva Rojo, J. (2018). "Designing and compiling parallel aligned corpora: pitfalls and (some) solutions on the example of a corpus of translated musem texts (English-Spanish)". *Revista de Lingüística y Lenguas Aplicadas*, *13*, 59-73. https://doi.org/10.4995/rlyla.2018.7912

Received: 2017-07-26 Accepted: 2017-12-04

volúmenes publicados hasta julio de 2017, 38 de ellos versan sobre corpus o emplean corpus para la realización de las investigaciones.

Los corpus de textos, entendidos como "a large collection of authentic texts that have been gathered in electronic form according to a specific set of criteria" (Bowker y Pearson, 2002:9), pueden tener innumerables modalidades, atendiendo a criterios de muy diversa índole (véanse, por ejemplo, Hunston [2002:14-16] o la exhaustiva propuesta de clasificación de Faya Ornia [2015]). De todos los tipos de corpus, este trabajo hace referencia al corpus paralelo —o, según la terminología de Granger (2010:15), *translation corpus*—. Este puede definirse como el corpus que "consists of original, source language-texts in language A and their translated versions in language B" (Baker, 1995:230).

Con el objetivo de poder realizar estudios que permitan evaluar textos museísticos traducidos —desde distintos prismas: control de calidad, análisis de la traducción de unidades fraseológicas o estudios sobre variedades diatópicas, entre otros—, se planteó la conveniencia de compilar un corpus propio, habida cuenta de que hasta la presente no se han localizado corpus de este tipo que estén disponibles. Las distintas fases de que se compuso el proceso hicieron ver, de una parte, los numerosos escollos a los que es preciso enfrentarse en una tarea de este tipo y, de otra, el número limitado de soluciones que existen. Esto se debe no solo a que los corpus paralelos alineados no son con frecuencia objeto de estudio de investigaciones de la Lingüística de corpus, sino también al hecho de que las herramientas disponibles tienen funciones limitadas o se han desarrollado para proyectos o investigaciones específicas, con lo que su acceso es restringido.

Como se ha mencionado, el trabajo se centra en una modalidad de corpus paralelo, el corpus alineado, que se trata en el apartado que sigue, donde se aborda qué tipos de alineación se pueden dar, al tiempo que se mencionan las principales utilidades del corpus alineado. Seguidamente, en el apartado tercero, el central de este trabajo, se proporciona una propuesta de metodología para el diseño y la compilación de corpus paralelos alineados, tomando como modelo un corpus de textos museísticos. Para ello, se establece una división en cuatro grandes fases, a saber: selección de los textos, normalización gráfica y almacenamiento, alineación, y compilación.

2. MARCO TEÓRICO. DEFINICIÓN, DELIMITACIÓN Y UTILIDADES DEL CORPUS PARALELO ALINEADO

El corpus alineado es el corpus paralelo que resulta de un proceso denominado *alineación*, y que puede definirse como "finding correspondences, in bilingual parallel corpora, between textual segments that are translation equivalents" (Kraif, 2002:273). A pesar de no ser reciente, esta definición sigue manteniendo su vigencia, si bien debe hacerse la salvedad de que la alineación no tiene por qué ser solo bilingüe, pues el corpus paralelo puede conllevar más de dos lenguas (cf. EAGLES, 1996:9; Frankenberg-Garcia, 2009:57; Zanettin, 2012:150). En caso de que se sea un corpus bilingüe, como ocurre con el corpus referido en este trabajo, el texto resultante de la alineación, por lo tanto, sería un *bitexto*, de acuerdo con la terminología acuñada por Harris (1988) hace ya tres décadas. Es preciso mencionar, por otra parte, que el nivel de *paralelismo* puede variar también, ya que puede ir de la correspondencia en el nivel del texto a la correspondencia en el nivel de la palabra (Véronis, 2000b:6-ss.; Zanettin, 2012:149; Faya Ornia, 2015:352), aunque lo que se suele entender *de facto* por *alineación* guarda relación con el nivel oracional (Zanettin, 2012:155), o de segmento de traducción, siguiendo la modalidad de estructuración aplicada en las memorias de traducción.

Las investigaciones que abordan asuntos relativos a la alineación se muestran en la misma línea de lo mencionado anteriormente: desde la primera alineación en el nivel oracional, llevada a cabo por Kay y Röscheisen en 1987 (Kay, 2000:xv), hasta la actualidad, con procesos cada vez más automatizados, mucho se ha avanzado y publicado sobre esta fase, específica de este tipo de corpus. De estos estudios se desprende que la complejidad vendrá dada por el tipo de alineación que se busque o, en palabras de Frankenberg-Garcia (2009:61), "[t]he finer the alignment, the more complex it becomes". Trabajos como los de Álvarez Lugrís y Gómez Guinovart (2015), Serón Ordóñez (2015), Zubillaga, Sanz y Uribarri (2015) y Molés-Cases (2016), por mencionar únicamente algunos de los más recientes en el ámbito hispánico, abordan cuestiones sobre la problemática que plantea la alineación de corpus paralelos bilingües o multilingües de cierta extensión.

Según se desprende de los estudios sobre este tipo de corpus, tanto teóricos como descriptivos, el hecho de que el corpus sea alineado parece tener consecuencia directa sobre dos fenómenos: lo limitado de su disponibilidad y un tamaño más reducido en comparación con los monolingües y comparables. Sobre el primero de los elementos, aunque existen algunos corpus paralelos accesibles, la disponibilidad de corpus alineados es, en líneas generales, limitada (Granger, 2010:5; Zanettin, 2012:153-154), por el mayor coste que suele llevar aparejada su compilación (Véronis, 2000b:15, cf. también Sharoff, 2002:449). En cuanto a su tamaño, este viene motivado, entre otros motivos, porque el tiempo que se precisa invertir es mayor, no solo por tener que localizar los textos, sino también por precisar de un procesamiento laborioso: la alineación. El tamaño, además, suele ser más reducido cuanto más especializado sea el dominio del corpus (Zanettin, 2012:152). También viene motivado el tamaño, por último, por

la propia naturaleza de los textos, ya que, recordemos, se trata de traducciones (Nádvorníková, 2017:3). Como muy gráficamente afirma Frankenberg-Garcia:

> Only a very small part of what people in general say or write ever gets to be translated, which seriously limits the number and types of texts available for the compilation of parallel corpora. Indeed, this is one of the main reasons why parallel corpora are usually much smaller in scale than monolingual corpora. (Frankenberg-Garcia, 2009:60)

En lo que respecta a la utilidad de los corpus alineados, una de las más evidentes es la aplicación directa a la traducción, donde, según palabras de Baker (1995:231), "they support a shift of emphasis, from *prescription* to *description*". Ya en el año 1992, por su parte, veía Isabelle en el uso sistemático de los textos traducidos unas posibilidades para la solución de problemas de traducción que no encuentran parangón en otros recursos:

> **La masse des traductions produites chaque année contient infiniment plus de solutions a plus de problemes que tous les outils de référence existants et imaginables!** Si seulement les services de traduction pouvaient considérer que ces masses de textes qu'ils génerent constituent une richesse qu'ils pourront réexploiter… (Isabelle, 1992:726; negrita en el original)

Adicionalmente, en los trabajos de EAGLES (1996:9), Véronis (2000a:IX, 12), Zanettin (2012:153), Fantinuoli y Zanettin (2015:3) y Nádvorníková (2017:2), entre otros, se señala que algunos estudios abordan la alineación de los corpus como elemento para la alimentación de motores de traducción automática. Igualmente, su empleo también puede ir orientado a una mejora de la calidad de las traducciones automáticas.

Aparte de las mencionadas, son infinidad las aplicaciones adicionales que se han identificado, como se ha puesto de manifiesto en numerosos estudios (Kay, 2000:XVII; Véronis, 2000a:XI, 2000b:2-3, 8-9, 12, 14; Granger, 2010:18; Zanettin, 2012:12). La lista es prácticamente interminable, pues va desde el aprendizaje de idiomas y la lingüística contrastiva hasta los estudios de traducción, tanto teóricos como descriptivos y de investigación en didáctica de la traducción, pasando por la traducción asistida por ordenador, con la compilación de memorias de traducción como máximo exponente. Su aportación a la lexicografía tampoco deja de ser desdeñable, ya que encuentra aplicaciones para la lexicografía y terminología multilingües y para la elaboración de listas terminológicas.

3. MARCO EMPÍRICO. EL CORPUS MUSA16 COMO PROPUESTA DE METODOLOGÍA DE DISEÑO Y COMPILACIÓN DE UN CORPUS PARALELO ALINEADO

El corpus al que se hace referencia en el presente trabajo se denomina MUSA16 por la suma de *museums*, *USA* y el último año del que se incluyen textos, 2016. Tiene como objetivo recoger todos los textos (completos, con independencia de su extensión) escritos originariamente en lengua inglesa y traducidos a la lengua española que se hayan publicado en las páginas web de museos y centros de arte del área metropolitana de Nueva York entre el año 1999 y el mes de julio de 2016. Se trata de un corpus especializado, paralelo (inglés-español), unidireccional (Frankenberg-Garcia, 2009:58; Faya Ornia, 2015:352) y alineado. Es, además, un *sample corpus* (Biber, 1993:243) o, más concretamente, un corpus multilingüe de tipo A (McEnery y Hardie, 2012:19).

La temática en que se basa, los textos museísticos, ha sido objeto de estudio de contadas aunque interesantes contribuciones que tienen como fin evaluar la calidad en la traducción de este tipo de textos —destacan sobre todo los trabajos de Neather (2008) y Jiang (2010)—. Otros estudios que abordan la traducción de textos museísticos incluso mencionan la conveniencia de recurrir a corpus (Neather, 2012:207), mientras que otros sí los emplean en efecto (Guillot, 2014:82; Valdeón, 2015:364), aunque no especifican ni sus características ni la forma en que se han compilado. En los últimos años, además, está surgiendo cierto interés por el estudio de la traducción de textos museísticos, como se pone de manifiesto en el número creciente de investigaciones que han ido apareciendo. Muestra de ello, además, es el hecho de que en el IATIS 6th International Conference, que se celebrará en Hong Kong en julio de 2018, una de las dieciocho sesiones temáticas previstas lleva por título "Museum translation: Encounters across space and time".

La compilación de MUSA16, por lo tanto, puede resultar de utilidad para la investigación en un ámbito con un interés científico creciente. Además, el aumento de los textos disponibles en campos específicos, unido a los avances técnicos (McEnery y Hardie, 2012:4; Fantinuolli y Zanettin, 2015:1), hace que crear un corpus como el que se menciona aquí sea una tarea más asumible que hace unos años. Su tamaño, dada su condición de ser especializado y alineado, será inferior a otros corpus generales o monolingües,[2] como se ha mencionado en el apartado anterior. En cualquier caso, la tendencia actual parece hablar de una menor preponderancia de la cuestión del tamaño del corpus

[2] Véase a este respecto la interesante revisión que hacen Vaughan y Clancy (2013:54-57) de lo que se entiende por *pequeño* en la lingüística de corpus.

—y, por ende, de la representatividad[3]—, como se pone de manifiesto en numerosos trabajos recientes (Anthony, 2013:146; Scheer, 2013:1; Rojo, 2015:376; Nádvorníková, 2017:3-ss.). De hecho, con total claridad, hablan Arbach y Ali de la "nécessité de corpus moins grands, plus spécifiques" (2013:11).

Habida cuenta de lo complicado del proceso de compilación del corpus paralelo, con algunos procedimientos que además son exclusivos de este tipo de corpus, es preciso establecer unas líneas claras que permitan poder realizar con éxito esta tarea. Si bien existen trabajos que establecen un protocolo para la compilación de corpus —caso del trabajo de Seghiri (2011), referido a corpus comparables—, no resulta fácil encontrar contribuciones en las que se detallen los pasos y se especifiquen las principales dificultades encontradas en el caso específico de los corpus alineados. Por ese motivo, en los apartados que siguen, divididos en función de las distintas etapas esenciales del procedimiento de compilación del corpus, se indicarán cuáles han sido los principales obstáculos encontrados y la forma en que se han resuelto estas complicaciones. Para ello, se han empleado como modelo, aparte del trabajo de Seghiri mencionado, los procedimientos recogidos con menor o mayor exhaustividad en las contribuciones de Frankenberg-Garcia (2009), Fantinuoli y Zanettin (2015), Serón Ordóñez (2015), Zubillaga, Sanz y Uribarri (2015) y Molés-Cases (2016). Cuando la solución a un problema ha radicado en una herramienta informática, no se ha recurrido a herramientas elaboradas expresamente para este trabajo, sino que se ha intentado emplear en la medida de lo posible software que sea gratuito y, sobre todo, de fácil manejo, para prevenir la situación mencionada por Zanettin, quien afirma que las metodologías y herramientas empleadas habitualmente para la compilación de corpus "require considerable computational expertise, which most researchers in descriptive translation studies and translators probably do not possess" (2012:161).

3.1. Identificación de museos y localización de los textos

Una deficiente selección de las fuentes, por imprecisa, errónea o incompleta, puede arruinar el trabajo de horas (en el mejor de los casos) del investigador. A ello se une, además, la complicación que señala Granger con respecto a los corpus paralelos, alineados o no:

> As regards translation corpora, however, electronic resources are scarce. It is not always possible to find translations of all texts, either because of the text type – letters and e-mail messages, for instance, are not usually translated – or because there are more translations in one direction (English to Chinese, for instance) than in another (Chinese to English). (Granger, 2010:5)

Una vez planteado el diseño general del corpus, se procedió a identificar los museos y centros de arte en los que más adelante se indagaría sobre la existencia de traducciones al español. Para ello, se recurrió al índice temático de la 5.ª edición de la prestigiosa guía *Blue Guide New York* (Wright, 2016), lo que permitió identificar un total de 77 instituciones. Seguidamente, se realizó un análisis pormenorizado de las páginas de los 77 museos, para localizar texto traducido al español, ya que la compilación hubo de plantearse partiendo del texto de destino (TD). En primer lugar, se recurrió a la navegación por la página de cada museo y a las opciones de búsqueda que ofrecen dichas páginas.

Puesto que los resultados fueron muy escasos (y, como se comprobó posteriormente, incompletos), se amplió la búsqueda con la opción de búsqueda avanzada de Google. Dentro de la opción "Búsqueda avanzada", se limitó la búsqueda —en el campo "sitio o dominio"— al dominio del museo en cuestión (por ejemplo, moma.org para el Museum of Modern Art) y, dentro de la opción "Buscar páginas con…", se introdujo en el campo "cualquiera de estas palabras" la cadena "spanish español espanol", con objeto de localizar resultados que incluyeran alguna de estas palabras. Esta búsqueda dio resultados mucho más positivos que la exploración manual de las páginas web. Se descartaron aquellos textos de los que no se pudo localizar el otro componente del bitexto o en los que no quedaba claro que la dirección de la traducción era inglés-español.

La fase de selección y localización dio el siguiente resultado: de los 77 museos estudiados, solo 21 ofrecían algún tipo de texto traducido en su página web (véase Figura 1). En estos 21 museos, por su parte, los registros que se consideraron válidos para el corpus fueron un total de 151, de los que se dejó constancia en un archivo de Excel creado como repositorio de metadatos para su codificación en XML en un futuro.

[3] No se abordará la cuestión de la representatividad del corpus en el presente trabajo por motivos de espacio y por tratarse MUSA16 de un corpus que no aspira a ser representativo, sino a recopilar la totalidad de textos de un período y un lugar concretos. No obstante, se remite al trabajo de Arbach y Ali (2013), para una visión general de la representatividad en las últimas décadas.

Museo	Archivos	Palabras
Museum of Modern Art	28	121 793
New-York Historical Society	22	95 380
American Museum of Natural History	6	62 202
National Museum of the American Indian	6	44 875
Metropolitan Museum of Art	15	32 445
Salomon R. Guggenheim Museum	8	16 868
Queens Museum	20	12 002
The Frick Collection	5	11 708
El Museo del Barrio	22	10 770
American Folk Art Museum	2	8307
Brooklyn Musem	2	7926
Skyscraper Museum	1	3562
Museum of Jewish Heritage	4	3099
Museum of Arts and Design	1	2063
The Jewish Museum	2	1345
New York Hall of Science	1	602
Nicholas Roerich Museum	1	483
Lower East Side Tenement Museum	1	432
Dyckman Farmhouse Museum	2	354
New York City Police Museum	1	194
National September 11 Memorial	1	97

Figura 1. Museos de Nueva York que cuentan con textos traducidos al español, ordenados según número de palabras del subcorpus MUSA16EN.

3.2. Normalización gráfica y almacenamiento

En la compilación de MUSA16 no ha habido que realizar digitalización como tal al estar todos los textos en formato electrónico, pero sí ha sido necesario tratar prácticamente todos los textos para poder alinearlos posteriormente. En el caso de este corpus, ha sido una tarea laboriosa que ha requerido de muchas horas y del recurso a una serie de herramientas de software. Habida cuenta de la naturaleza de los textos (de los 151 textos, 3 están en formato DOC, 71 en HTML y 77 en PDF) y del propio tipo de corpus, por ser paralelo, el tratamiento de los archivos que componen el corpus exige de un esfuerzo adicional.

Los archivos en formato DOC, solo tres, son los que precisan de menor tratamiento, puesto que se pueden almacenar como se encuentran en la página web. El grueso de los archivos se encuentra en formato PDF y HTML; dos de los del último formato, por añadidura, contienen animaciones Flash con contenido no copiable,[4] lo que complica la extracción del texto. Además de los programas habituales para este tipo de tareas —navegador de internet, lector de archivos PDF, procesador y editor de textos—, se han empleado los programas y las extensiones, todos gratuitos, que aparecen en la Figura 2.

Tarea	Aplicación	Programa o extensión
Realización de capturas de pantalla	Imágenes con texto en PDF o texto de animaciones Flash	Greenshot
Conversión de texto en imágenes a texto editable	Capturas de pantalla con texto (opción «Copiar texto de la imagen»)	Microsoft OneNote
Localización de líneas de texto repetidas	DOC, PDF y HTML extensos	Remove Duplicate Lines de TextMechanic
Descarga de múltiples enlaces HTML	Página web con múltiples enlaces de descarga de un mismo directorio	DownThemAll
Fusión en un único archivo HTML	Conjunto de archivos HMTL descargados con DownThemAll	HTMLMerge

Figura 2. Listado de programas y extensiones empleados e indicación de la tarea y finalidad en las que se han empleado.

[4] En concreto, se trata del material relativo a dos exposiciones del National Museum of the American Indian: bitextos NMAIEN01-NMAIEN01 y NMAIEN04-NMAIES04. Para una descripción de los registros aludidos como ejemplo en este trabajo, véase el Apéndice A.

Una vez que el texto de cada registro se encontraba en formato editable, en Word se han llevado a cabo las siguientes operaciones, similares a las que registra Frankenberg-Garcia (2009:63): eliminación de encabezados repetidos, números de páginas, llamadas de nota al pie y referencias bibliográficas si se encontraban sin traducir, y replicación del formato en los dos bitextos del mismo registro.

Los textos provenientes de archivos en PDF han planteado problemas adicionales al trasladar su contenido a Word, puesto que este programa interpreta en ocasiones, según la forma en que se haya editado el PDF, que todo fin de reglón en PDF es un retorno de carro. Esto, que no es grave en corpus no alineados, puesto que no afecta a las búsquedas lingüísticas, se debe solucionar en el caso de MUSA16 para que la alineación se pueda llevar a cabo de forma correcta. Son dos las soluciones posibles:

1. Comprobar manualmente que todas las marcas de párrafo en el documento DOC procedente del PDF se corresponden con saltos de carro reales, tarea que es la más recomendable para registros cortos.

2. Hacer uso de la opción "Buscar y reemplazar" de Word para automatizar en la medida de lo posible la limpieza de los registros. Aunque la automatización completa es difícil de lograr, la tarea de revisión posterior que habrá que realizar es en cualquier caso más rápida que la opción planteada en el apartado anterior si el texto es extenso.

Además, se han registrado en ocasiones problemas de selección del texto en los archivos PDF, por la distinta configuración de selección de los originales y sus traducciones en el momento de maquetar los textos. Por este motivo, al copiar los dos componentes de un bitexto y pegarlos en Word, la disposición de sus componentes era distinta en ocasiones (por ejemplo, los títulos aparecían en su orden lógico en el TO pero en otra disposición en el TD), por lo que ha habido que revisar estos casos manualmente. Aunque una digitalización realizada con estos o similares procesos supone una inversión de tiempo considerable, no cabe duda de que la revisión pormenorizada de los documentos resultantes es el paso previo necesario para una alineación con éxito.

3.3. Alineación

Se trata de la operación, sin lugar a dudas, más complicada del trabajo, para la que se precisa que los textos se hayan normalizado correctamente, con objeto de que la pérdida de contenido y el tiempo invertido sean lo más reducidos posible. Con un corpus del tamaño de MUSA16, esta fase requiere necesariamente del empleo de herramientas informáticas, de las que hay un amplio catálogo, aunque no existe hasta la fecha una herramienta que sea plenamente satisfactoria, como se verá más adelante. Como consecuencia, el componente manual, según afirma Zanettin, "cannot be bypassed completely if the desired outcome is accurate alignment" (2012:161).[5] El hecho de que sea necesario, al menos, revisar manualmente los resultados de la alineación viene motivado, además, por la propia naturaleza de los textos que componen el corpus:

> The creation of robust and reliable parallel corpora for descriptive translation studies is demanding and laborious work. The high quality needed for descriptive translation research can only be obtained through manual alignment editing, as opposed to corpus-based machine (assisted) translation […]. This is especially true for corpus based studies of genres such as fiction and news writing, whose language is often rather "noisy", that is, resistant to automatic alignment. (Zanettin, 2013:30)

En los textos que componen MUSA16, se dan situaciones como las recogidas en la cita anterior o en Kraif (2002:275). Ello es debido a la configuración de los TO y sus respectivas traducciones, con una disposición del texto que a veces no es similar y donde, además, en el TD se omite en ocasiones texto, aparecen fragmentos sin traducir o se muestra contenido resumido —caso de METMEN08-METMES08, donde el TO casi dobla en extensión al TD—. En otros casos, incluso, los textos no son totalmente idénticos, sino que se trata de versiones distintas porque se ha actualizado el documento de uno de los idiomas, pero el otro no (como sucede con NMAIEN02-NMAIES02). Una pérdida o alineación errónea de, por ejemplo, el 5 % del contenido de un corpus de varios millones de palabras destinado a sistemas de recuperación para traducción automática sería algo aceptable, pues en estos casos la cantidad puede primar sobre la precisión (Zanettin, 2012:164). En el caso de corpus como el que se describe en este trabajo, sin embargo, supondría una pérdida que consideramos inasumible, no solo por la merma en la extensión del corpus, sino también por la finalidad para la que se compila, ya que una alineación incorrecta desvirtuaría los resultados del corpus en los estudios descriptivos realizados con él.

[5] No es este el único autor que habla del componente manual en la alineación de corpus paralelos. Tal es el caso, por ejemplo, de Lahausois y Séverine (2012:35), McEnery y Hardie (2012:20-21), Zubillaga, Sanz y Uribarri (2015:83) o Molés-Cases (2016:156).

Siguiendo las recomendaciones de Zanettin y habida cuenta del origen de los textos, en primer lugar, se realizó una revisión manual de cada bitexto para realizar una alineación cotejada por párrafos, ya que ello "allows for smoother alignment at sentence level" (Zanettin, 2012:156).

En lo que respecta a las modalidades de herramientas para la alineación, descartadas la alineación manual y la completamente automática, son dos las opciones posibles:

1. Sistemas de alineación integrados en programas específicos de gestión de corpus. Su uso, no obstante, suele presentar complicaciones. Así ocurre, por ejemplo, con ParaConc o WordSmith Tools, herramientas que, de acuerdo con Zanettin (2012:158-159, 162) plantean el problema de que consideran que ambos integrantes del bitexto tienen el mismo número de oraciones y párrafos, asunción que ya mencionaba Véronis (2000b:9) como fuente de problemas en la alineación.

2. Herramientas creadas, en su mayoría, para la traducción asistida por ordenador (fundamentalmente, memorias de traducción o aplicaciones de alineación para conversión en memorias de traducción), cuya finalidad es realizar la alineación automática de dos documentos que luego puede revisarse de forma manual. El recurso a estas herramientas viene motivado por el hecho de que el corpus paralelo alineado va a guardar similitudes con las memorias de traducción (Zanettin, 2012: 169-ss.; Álvarez Lugrís y Gómez Guinovart, 2015:35). Algunas de las herramientas gratuitas más conocidas son BiText2tmx (versión 1.0), LF Aligner (4.1) y la herramienta de alineación de OmegaT (4.0). De entre las que son de pago, son numerosas las existentes (véase Molés-Cases, 2016:155), pero se mencionan aquí solamente dos: SDL Trados Studio y ABBYY Aligner 2.0.

La opción elegida ha sido la segunda, tanto por cuestiones de disponibilidad y facilidad de uso de la herramienta como por motivos de conocimiento previo del manejo de memorias de traducción y aplicaciones para la traducción asistida por ordenador. Del análisis de las herramientas de alineación gratuitas se desprende que su empleo plantea inconvenientes para los fines de este trabajo: o su uso es un tanto rudimentario (BiText2tmx, LF Aligner) o presenta escasas opciones de desplazamiento de las unidades de traducción a otros lugares del documento —caso de OmegaT, que, por el contrario, presenta en su versión 4.0, de 2016, una forma de identificación de unidades alineadas de uso muy sencillo por parte del usuario—. Por este motivo, ha sido necesario recurrir a herramientas de pago (SDL Trados Studio y ABBYY Aligner 2.0), puesto que el volumen de segmentos que ha habido que alinear desaconsejaba recurrir a la mejor alternativa gratuita según nuestro juicio, LF Aligner.

La primera herramienta de pago que se seleccionó para la alineación fue SDL Trados Studio en sus versiones 2014 y 2017 (SDL Group, 2013, 2016), herramienta de traducción asistida por ordenador de amplísima difusión. Se trata, además, de un tipo de herramienta, la memoria de traducción, que se ha empleado anteriormente en la alineación de corpus paralelos, como sucede con los compilados por Álvarez Lugrís y Gómez Guinovart (2015), Buján Otero (2015), Serón Ordóñez (2015) y Molés-Cases (2016), por citar algunos de los más recientes del ámbito hispánico.

Como sucede con las herramientas anteriores, el uso de SDL Trados Studio no implica que no conlleve dificultades u obstáculos: a pesar de su robustez y sencillo manejo, este programa presenta algunas limitaciones que ya se han superado en otros programas, algunos incluso gratuitos: la opción de alineación de documentos no permite realizar búsquedas por palabras ni ofrece la opción de ir a un segmento en concreto, sino que el desplazamiento debe hacerse mediante la navegación con la barra de desplazamiento. A mayor abundamiento, la herramienta no permite vincular más de cuatro unidades en total a la vez entre los dos textos ni permite tampoco filtrar la visualización de unidades según su estado, lo cual facilitaría mucho la alineación de aquellos textos en los que la información no se encuentre organizada de la misma forma en ambos. Igualmente, se echa en falta la posibilidad de poder anular la conexión de varios segmentos a la vez mediante una opción de selección múltiple, tarea esta que ahorraría una cantidad considerable de tiempo. SDL Trados Studio fue la primera de las herramientas empleadas, con un funcionamiento correcto en textos con un nivel de similitud alto entre ambos integrantes del bitexto. Sin embargo, al afrontar la alineación de un texto con una distribución muy dispar entre el TO y el TD (AMNHEN03-AMNHES03, Figura 3), se descartó esta herramienta y se intentó localizar otra.

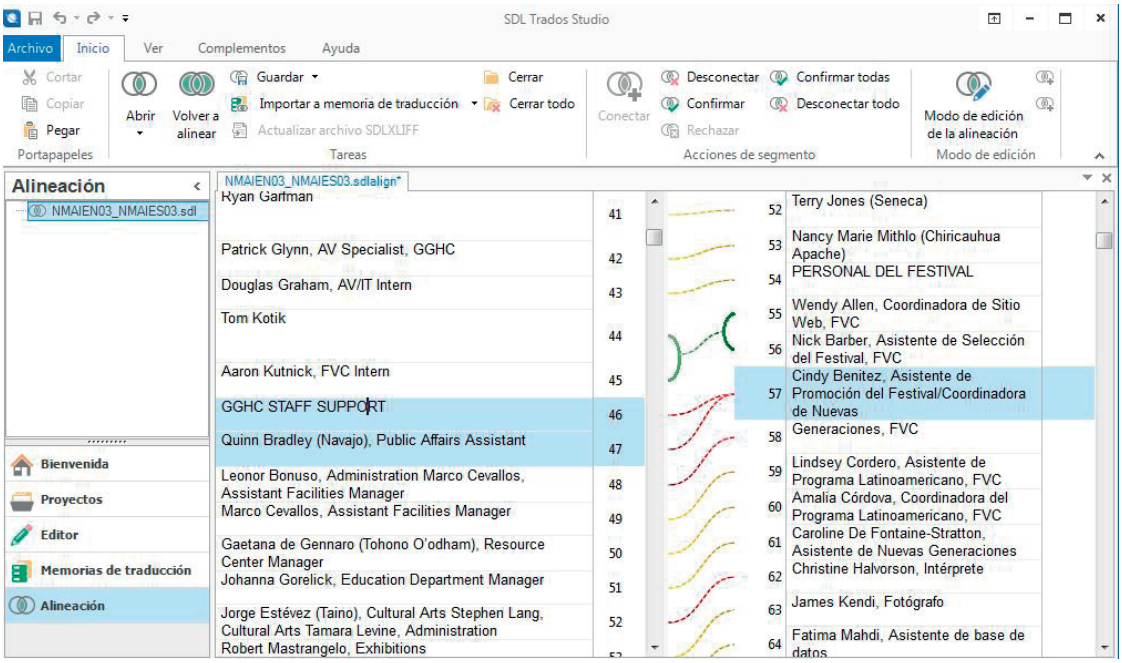


Figura 3. Alineación del bitexto AMNHEN03-AMNHES03 con SDL Trados Studio 2017.

La segunda herramienta empleada fue ABBYY Aligner 2.0 (ABBYY Software, 2011, Figura 4), cuya versión Freelance —la empleada en este trabajo— está disponible de forma gratuita para usuarios que dispongan de una cuenta de smartCAT, también sin coste alguno. Se trata de una herramienta no muy empleada para la compilación de corpus, aunque existen algunas excepciones, como el caso de Ginezi (2014:183), quien utiliza una versión anterior en línea. El empleo de ABBYY Aligner 2.0 es extremadamente sencillo, y su configuración y disposición es muy similar a la de las hojas de Excel. Incluye métodos abreviados de teclado para las distintas opciones de menú, lo que permite un empleo más eficiente del programa. Al igual que sucede con Excel, permite copiar tantos segmentos (o celdas) como se precise, sin la limitación de cuatro segmentos de SDL Trados Studio. Dispone además de una opción de búsqueda de texto, algo de gran utilidad para textos con configuración muy dispar entre TO y TD.[6] Otra ventaja muy grande con la que cuenta es la posibilidad de realinear el fragmento que se desee, lo cual es una buena opción para textos con distinta disposición de sus elementos textuales en el TO y el TD: una vez localizados y reubicados, solo hay que volver a alinear el fragmento en cuestión. En el lado de los debe, destacan los siguientes: no tiene opción de bloqueo de segmentos, por lo que la opción de realineación puede modificar lo que se había editado anteriormente. Plantea el inconveniente de que la alineación elimina todos los elementos de diacrisis tipográfica (cursiva, negrita, distintos tamaños de letra), lo que afecta, por ejemplo, a cómo se muestran nombres de exposiciones o cuadros. Por último, los formatos de salida son muy limitados: solamente ATA (específico del programa) y TMX, el estándar de las memorias de traducción. Esta limitación, por tanto, debe tenerse en cuenta antes de compilar el corpus, puesto que solamente podrán emplearse gestores de concordancias que admitan la extensión TMX.

Aunque ABBYY Aligner 2.0 ha facilitado mucho el trabajo de alineación, no ha sido tarea ni sencilla ni rápida. Sirva como ejemplo el bitexto NYHSEN01-NYHSES01, que, de los más de 1900 segmentos al principio, ha pasado a tener solamente algo más de 900 como consecuencia de la reestructuración de segmentos y la eliminación de texto sin traducir en el TO. No obstante, el empleo de la herramienta de ABBYY ha permitido recuperar registros que se descartó alinear con SDL Trados Studio por resultar muy penoso de alinear (el caso de AMNHEN03-AMNHES03 mencionado anteriormente). En cualquier caso, también ha habido alineaciones que se han podido realizar con gran rapidez, sin lugar a dudas como resultado de cómo estaba realizada la traducción —SKYMEN01 y SKYMES01, por ejemplo, apenas ha precisado de edición de los resultados—.

[6] Sirva como ejemplo de ello el bitexto MOMAEN26-MOMAES26, uno de los más extensos del corpus: cinco segmentos concretos aparecían sobre la celda 1200 en el TO y alrededor de la celda 200 en el TD. La opción "Find" del programa permitió alinear en segundos algo que habría supuesto mucho más esfuerzo realizar con otros programas.

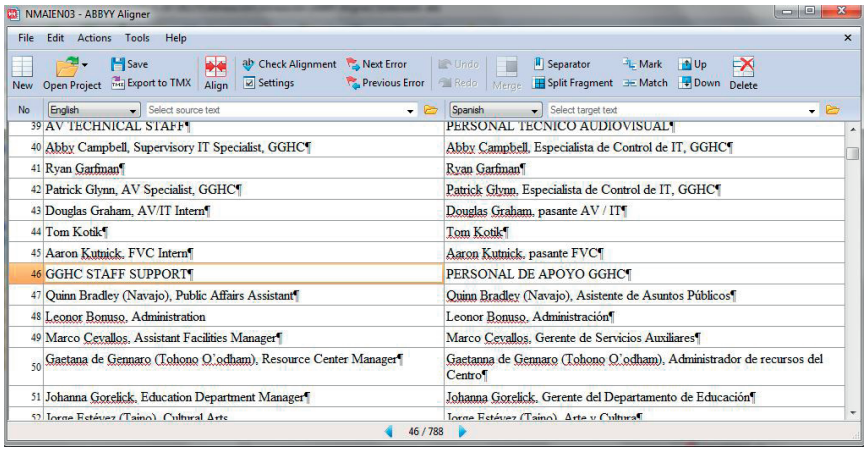


Figura 4. Alineación del bitexto AMNHEN03-AMNHES03 con ABBYY Aligner 2.0.

La alineación de los registros del corpus ha dado como resultado 151 bitextos, 49 de ellos alineados con SDL Trados Studio y 102 con ABBY Aligner 2.0. En total, se ha revisado manualmente la alineación de 35 223 segmentos de traducción. Se ha procurado que en cada alineación haya solo un segmento, aunque se han fusionado algunos segmentos concretos —fundamentalmente citas, para mantener su integridad, y en obras de arte (autor, obra, técnica y dimensiones), para ayudar a su contextualización—.

3.4. Compilación y gestión del corpus

Una vez completada la alineación de los registros del corpus, es preciso procesar los bitextos para que puedan realizarse consultas lingüísticas en ellos. Para la alineación de los bitextos del corpus MUSA16, como se ha mencionado en el apartado anterior, el recurso a ABBYY Aligner 2.0 en gran parte de los registros conlleva la limitación de que solamente se podrán emplear herramientas de gestión que admitan archivos en formato TMX, lo cual restringe el número de opciones disponibles. Esto quiere decir que herramientas específicas para corpus paralelos, tales como AntPConc, ParaConc o WordSmith Tools, entre otras herramientas habituales para la alineación y gestión de corpus paralelos (Torrellas Castillo, 2009:11; Zanettin, 2012:180; Vigier y Sánchez, 2017:265), no podrán usarse para el presente corpus. En cuanto a otras herramientas que sí admiten archivos en formato TMX, caso de Corpus Query Processor (CQP), InterText o NoSketchEngine (Čermák y Rosen, 2012; Vondřička, 2014; Molés-Cases, 2016:161), se ha descartado su empleo por precisar de trabajo informático especializado para la creación del corpus.

Como consecuencia, y a la luz de los gestores de concordancias que se ha decidido emplear, se han procesado los bitextos como sigue:

1. Los archivos alineados con ABBYY Aligner 2.0 (formato ATA) se han convertido a archivos con formato TMX (opción del menú "File" del propio programa).

2. Los archivos alineados con SDL Trados Studio (formato SDLALIGN), se han convertido cada uno de ellos a una memoria de traducción, lo que ha dado como resultado sendos archivos SDLTM. Estos se han importado a memorias de traducción, una por archivo, con extensión TMX (ventana "Memorias de traducción" > "Inicio" > "Importar").

En lo que respecta a los programas para la gestión de concordancias, se propone la utilización de las siguientes herramientas:

1. Xbench (ApSIC, 2001, Figura 5), para el que solo se precisa cargar los archivos TMX individuales (en un proyecto creado, "Project properties" > "Add" > selección de "TMX Memory"). Se trata de un programa que dispone de una versión gratuita (2.9.474) y que, si bien su propósito fundamental es el control de calidad en traducciones, se ha empleado en investigaciones como gestor de concordancias (Buján Otero, 2015:119-ss.; Martínez Belchí, 2015:93). Aunque presenta el inconveniente de no ofrecer datos y aplicaciones de utilidad para trabajos basados en corpus —tales como listas de frecuencia o de palabras clave—, tiene la ventaja de que realiza las búsquedas con gran rapidez, de que permite saber el archivo de procedencia de cada aparición y de que "dispone de una sintaxis de búsqueda extremadamente avanzada que nos da la posibilidad de realizar búsquedas complejas" (Martínez Belchí, 2015:93), por lo que es una buena opción gratuita. Una posible solución para salvar la ausencia de elementos propios de la búsqueda de concordancias puede ser recurrir simultáneamente a AntConc (Anthony, 2016), aunque mediante consulta solo como corpus monolingües.

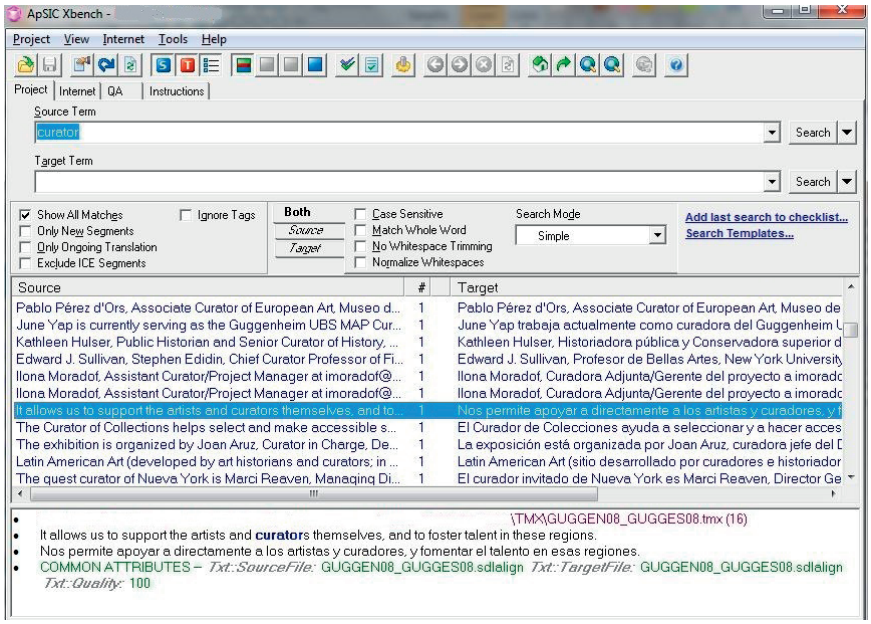


Figura 5. Ventana de búsqueda de Xbench. Apariciones de *curator* y sus traducciones en el corpus paralelo

2. Sketch Engine (Kilgarrif *et al.,* 2014, Figura 6), herramienta de análisis textual de pago que permite consultar corpus (monolingües y paralelos) y también crear corpus propios. Aunque se trata de una herramienta diseñada específicamente para la gestión de corpus, en el caso de los corpus paralelos no se permite cargar archivos individuales.[7] Por ese motivo ha sido necesario cargar un único archivo TMX, compuesto mediante la importación en SDL Trados Studio 2017 de las 151 memorias de traducción (una por bitexto) a un único archivo TMX.[8] La limitación de poder cargar solamente un archivo ha supuesto una dificultad insalvable: al importar las 151 memorias de traducción en una sola, SDL Trados Studio —al igual que sucede con otras herramientas de traducción asistida por ordenador—no mantiene las repeticiones de segmentos que son totalmente idénticos en el TO y en el TD, sino que solamente conserva una de las apariciones.

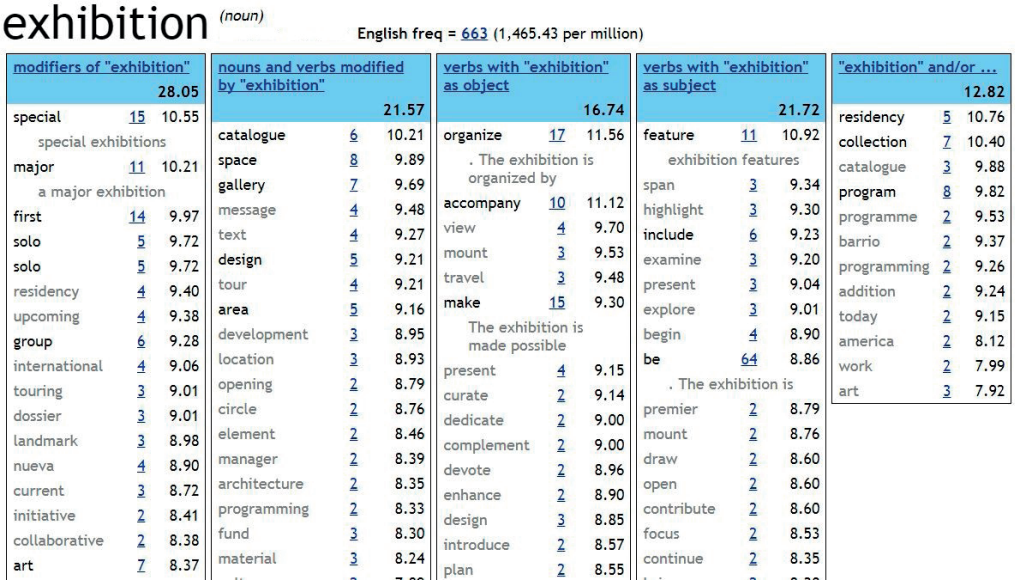
Figura 6. Ventana de concordancias de Sketch Engine. Apariciones de *exposición individual* y sus traducciones en el corpus paralelo.

Esto ha supuesto no solo una pérdida en cuanto a la extensión del corpus, sino también con respecto a cuestiones de frecuencia y distribución de los elementos del corpus, aunque solo afecta a segmentos totalmente idénticos en el TO y el TD. Es de desear que este problema pueda solventarse en futuras versiones de Sketch Engine. En otro orden de cosas, al tratarse de un único archivo, Sketch Engine no permite distinguir a qué registro del corpus corresponde cada aparición, como sí sucede con Xbench, pero, por el contrario, permite ampliar el

[7] Téngase en cuenta, además, que el procedimiento explicado aquí es para corpus en los que se da correspondencia 1:1, es decir, a un segmento en el TO corresponde solamente uno en el TD. Para aquellos casos en los que se dé una equivalencia m:n, es posible realizar la compilación del corpus desde junio de 2017 mediante un método para usuarios avanzados, según se indica en la guía del usuario de la herramienta.

[8] Existe también la opción de emplear Xbench, que permite exportar todos los archivos cargados en un proyecto —en este caso, los 151 archivos TMX— a un solo archivo TMX mediante el uso de la opción "Tools" > "Export Items". Para evitar desajustes en la alineación en Sketch Engine, se recomienda, en "Filtering", dejar activada "Remove repetitions in exported file", puesto que ha planteado problemas de compilación en MUSA16.

contexto con gran facilidad, a la vez que dispone de datos sobre el corpus —recogidos en las opciones "Word list" y "Word sketch"— que resultan de gran utilidad para realizar estudios descriptivos con el corpus (Figura 7).



exhibition *(noun)*

English freq = 663 (1,465.43 per million)

modifiers of "exhibition"		28.05
special	15	10.55
special exhibitions		
major	11	10.21
a major exhibition		
first	14	9.97
solo	5	9.72
solo	5	9.72
residency	4	9.40
upcoming	4	9.38
group	6	9.28
international	4	9.06
touring	3	9.01
dossier	3	9.01
landmark	3	8.98
nueva	4	8.90
current	3	8.72
initiative	2	8.41
collaborative	2	8.38
art	7	8.37

nouns and verbs modified by "exhibition"		21.57
catalogue	6	10.21
space	8	9.89
gallery	7	9.69
message	4	9.48
text	4	9.27
design	5	9.21
tour	4	9.21
area	5	9.16
development	3	8.95
location	3	8.93
opening	2	8.79
circle	2	8.76
element	2	8.46
manager	2	8.39
architecture	2	8.35
programming	2	8.33
fund	3	8.30
material	3	8.24

verbs with "exhibition" as object		16.74
organize	17	11.56
. The exhibition is organized by		
accompany	10	11.12
view	4	9.70
mount	3	9.53
travel	3	9.48
make	15	9.30
The exhibition is made possible		
present	4	9.15
curate	2	9.14
dedicate	2	9.00
complement	2	9.00
devote	2	8.96
enhance	2	8.90
design	3	8.85
introduce	2	8.57
plan	2	8.55

verbs with "exhibition" as subject		21.72
feature	11	10.92
exhibition features		
span	3	9.34
highlight	3	9.30
include	6	9.23
examine	3	9.20
present	3	9.04
explore	3	9.01
begin	4	8.90
be	64	8.86
. The exhibition is		
premier	2	8.79
mount	2	8.76
draw	2	8.60
open	2	8.60
contribute	2	8.60
focus	2	8.53
continue	2	8.35

"exhibition" and/or ...		12.82
residency	5	10.76
collection	7	10.40
catalogue	3	9.88
program	8	9.82
programme	2	9.53
barrio	2	9.37
programming	2	9.26
addition	2	9.24
today	2	9.15
america	2	8.12
work	2	7.99
art	3	7.92

Figura 7. Patrones sintácticos de *exhibition* en el subcorpus MUSA16EN, opción "Word sketch" de Sketch Engine.

Como resultado de la compilación del corpus en Sketch Engine se han obtenido los siguientes datos de los dos subcorpus de que se compone MUSA16:

- MUSA16EN, subcorpus de textos escritos originariamente en lengua inglesa, compuesto por 151 archivos, 381 074 palabras y 452 424 tokens.

- MUSA16ES, subcorpus de textos traducidos a la lengua española, con 151 archivos, 429 019 palabras y 499 895 tokens.

El corpus resultante, como ya es sabido, es un corpus paralelo, completamente alineado, que se encuentra preparado para su consulta tanto en Sketch Engine como en Xbench (y también para su consulta, aunque no como corpus alineado, en AntConc). En lo que respecta a su división por años, es la que se puede ver en la Figura 8, donde se puede apreciar que los años de origen de gran parte de los textos son 2010 y 2011. En otro orden de cosas, bajo el elemento "Otros" se han incluido textos de los que no sabe con certeza su año, si bien están vigentes en la actualidad (por ejemplo, por concordar el precio de la entrada al museo con el precio en vigor; 2757 palabras), se tiene constancia de que son posteriores a 2004 (5813 palabras) o aparecen en una página cuyo dato del año se ha extraído del copyright (2016; 5183 palabras). En cuanto al último año del que hay datos, 2016, téngase en cuenta que no incluye el año completo, puesto que se han recogido solamente textos hasta el mes de julio de 2016.

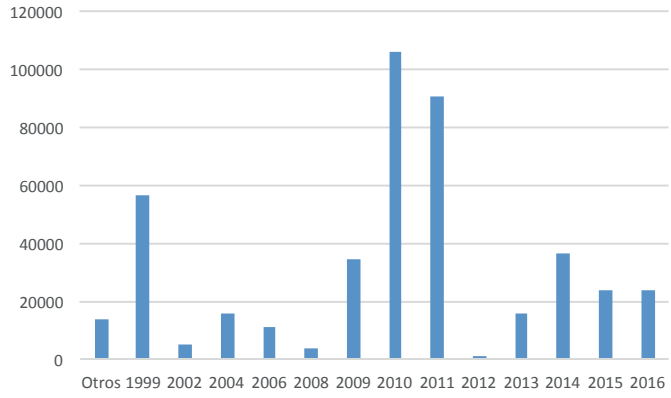


Figura 8. Número de palabras del subcorpus MUSA16EN, distribuido según año.

4. CONCLUSIONES

Los corpus de textos paralelos y alineados son herramientas de gran valor y utilidad potencial para la investigación y la enseñanza y práctica profesional de la traducción, entre muchas aplicaciones. El recurso al corpus paralelo, aunque incentivado en estudios e investigaciones, sigue teniendo el inconveniente de que su disponibilidad es escasa, por lo que en ocasiones el investigador tendrá que aventurarse a diseñar, alinear y compilar su propio corpus paralelo alineado. Esas son las premisas que han motivado la compilación del presente corpus, ya que, para el estudio de textos museísticos y sus traducciones, no hay disponibles corpus de textos en el momento actual. La futura explotación del corpus que se presenta aquí se pretende que pueda servir de base para trabajos descriptivos de diversa índole, tales como control de calidad de traducciones, estudio de la traducción de unidades fraseológicas o análisis de las variedades diatópicas del español que se emplean en los textos de museos de Nueva York traducidos a la lengua española.

En lo concerniente a la creación de un corpus de este tipo, son tres las dificultades metodológicas a las que será necesario enfrentarse: en primer lugar, la localización de los textos precisa de un esfuerzo adicional por parte de quien compile el corpus; en segundo, la alineación los textos y sus traducciones es un proceso laborioso y arduo. En tercer lugar, y no por ello menos importante, el software de que se dispone en la actualidad no viene desprovisto de inconvenientes, ya que no satisface todas las necesidades que se precisa para la compilación con éxito de un corpus paralelo. Esto último viene motivado, además, por el hecho de que a veces el software al que se ha de recurrir no es específico para esta tarea.

El corpus sobre el que versa este trabajo, MUSA16, se compone de textos de museos y centros de arte de la ciudad de Nueva York, escritos originariamente en inglés y traducidos al español y publicados entre 1999 y 2016. Para su compilación, y siguiendo metodologías de otros trabajos previos, se han establecido las siguientes fases: 1) selección de fuentes y localización de los textos, que ha permitido acotar una base textual no excesivamente extensa, pero sí exhaustiva; 2) normalización gráfica y almacenamiento de los registros, fase de gran laboriosidad, imprescindible para el éxito de la tercera fase; 3) alineación, la fase más complicada por la dificultad que ha supuesto localizar una herramienta para la alineación que diera resultados satisfactorios; y 4) compilación y gestión del corpus, fase en la que han surgido obstáculos conllevados por el uso combinado de SDL Trados Studio y Sketch Engine.

A lo largo del apartado tercero de este trabajo, que compone su parte central, se registran herramientas de gran utilidad para la extracción del texto (DownThemAll, Greenshot, HTMLMerge, Microsoft OneNote, Remove Duplicate Lines), así como otras específicas para la compilación y gestión de concordancias (AntConc, Sketch Engine) y otras ideadas en realidad para la traducción asistida por ordenador (ABBYY Aligner 2.0, SDL Trados Studio, Xbench). Como se planteaba al principio de este trabajo, se pretendía dar con herramientas y métodos que facilitaran la compilación de un corpus paralelo y alineado y que solventaran sus principales escollos. Según se ha podido ver, siguen existiendo algunas complicaciones no resueltas aún, siendo la más significativa la pérdida de texto repetido a la hora de compilar el corpus paralelo. La introducción de mejoras en el software (sobre todo, gestión de repeticiones en SLD Trados Studio y en Xbench, así como posibilidad de carga de más de un archivo en Sketch Engine) supondría una compilación y una gestión mucho más sencillas y exhaustivas de un elemento, el corpus alineado, crucial para la práctica profesional e investigadora de la traducción y de las lenguas aplicadas.

Al principio del artículo se hizo hincapié en un dato relativo a esta revista: un alto número de las contribuciones habidas hasta la fecha, 38, se basan en corpus y lingüística de corpus. Un análisis detallado muestra lo siguiente: de las 38 contribuciones, solo 1 aborda de forma específica cuestiones metodológicas de compilación y consulta de corpus (Sánchez Ramos, 2017), y 3 utilizan el corpus paralelo como elemento de estudio (Rioja, 2010; Argüelles Álvarez y Muñoz Muñoz, 2012; Hui-Chuan y Hsueh, 2012), mientras que ninguna hace mención alguna a los corpus alineados o a la alineación. La metodología que se describe aquí tomando como base la aplicada para el corpus MUSA16 pretende ser una aportación a un campo tan fértil como la lingüística de corpus —aunque con terrenos aún por terminar de explorar— y se ofrece como ejemplo que pueda servir de orientación a quienes se aventuren en la laboriosa tarea, pero gratificante al final, de diseñar, alinear y compilar un corpus de este tipo.

AGRADECIMIENTOS

El autor desea agradecer los valiosos comentarios y observaciones realizados por los dos revisores anónimos del presente trabajo.

REFERENCIAS

ABBYY Software (2011). *ABBYY Aligner 2.0* (versión 1.0.6.59) [Software]. Milpitas: ABBYY Software.

Álvarez Lugrís, A. y Gómez Guinovart, X. (2015). "Lexicografía bilingüe práctica basada en corpus: planificación y elaboración del *Diccionario Moderno Inglés-Galego*" en M.J. Domínguez Vázquez, X. Gómez Guinovart y C. Valcárcel Riveiro (eds.). *Lexicografía de las lenguas románicas. II: Aproximaciones a la lexicografía moderna y contrastiva*. Berlín: De Gruyter, 31-47.

Anthony, L (2016). *AntConc* (versión 3.4.4.) [Software]. Tokyo: Waseda University. http://www.laurenceanthony.net/.

Anthony, L. (2013). "A critical look at software tools in corpus linguistics". *Linguistic Research, 30*/2: 141-161.

ApSIC (2011). *Xbench* (versión 2.9.474) [Software]. Barcelona: ApSIC.

Arbach, N. y Ali, S. (2013). "Aspects théoriques et méthodologiques de la représentativité des corpus". *Corela, 13*: 1-16. https://doi.org/10.4000/corela.3029

Argüelles Álvarez, I. y Muñoz, A. (2012). "An insight into Twitter: a corpus based contrastive study in English and Spanish". *Revista de Lingüística y Lenguas Aplicadas, 7*/1: 37-50. https://doi.org/10.4995/rlyla.2012.1123

Baker, M. (1995). "Corpora in Translation Studies: An overview and some suggestions for future research". *Target, 7*/2: 223-243. https://doi.org/10.1075/target.7.2.03bak

Biber, D. (1993). "Representativeness in corpus design". *Literary and Linguistic Computing, 8*/4: 243-257. https://doi.org/10.1093/llc/8.4.243

Bowker, L. y Pearson, J. (2002). *Working with specialized language. A practical guide to using corpora*. Londres: Routledge. https://doi.org/10.4324/9780203469255

Buján Otero, P. (2015). *A tradución da fraseoloxía no xénero textual MANUAL DE INSTRUCIÓNS na combinación lingüística alemán-español* (tesis doctoral). Vigo: Universidad.

Čermák, F. y Rosen, A. (2012). "The case of InterCorp, a multilingual parallel corpus". *International Journal of Corpus Linguistics, 17*/3: 411-427. https://doi.org/10.1075/ijcl.17.3.05cer

EAGLES (1996). *Preliminary recommendations on Corpus Typology* (EAGLES Document EAG-TCWG-CTYP). Pisa: Consiglio Nazionale delle Ricerche.

Fantinuoli, C. y Zanettin, F. (2015). "Creating and using multilingual corpora in translation studies" en C. Fantinuoli y F. Zanettin (eds.). *New directions in corpus-based translation studies* Berlín: Language Science Press, 1-10.

Faya Ornia, G. (2015). "Propuesta de clasificación de corpus textuales" en M.T. Sánchez Nieto *et al.* (eds.). *Metodologías y aplicaciones en la investigación en traducción e interpretación con corpus*. Valladolid: Universidad, 339-356.

Frankenberg-Garcia, A. (2009). "Compiling and using a parallel corpus for research in translation". *International Journal of Translation, 21*/1-2: 57-71.

Ginezi, L. (2014). "Desafios para a construção de um corpus de aprendizes de Interpretação Simultânea". *TradTerm, 23*: 165-191. https://doi.org/10.11606/issn.2317-9511.tradterm.2014.85576

Granger, S. (2010). "Comparable and translation corpora in cross-linguistic research: Design, analysis and applications". *Journal of Shanghai Jiaotong University, 2*: 14-21.

Guillot, M.-N. (2014). "Cross-cultural pragmatics and translation: The case of museum texts as interlingual representation" En J. House (ed.). *Translation: A multidisciplinary approach*. Hampshire: Palgrave, 73-95. https://doi.org/10.1057/9781137025487_5

Harris, B. (1988). "Bi-text, a new concept in translation theory". *Language Monthly, 54*/marzo: 8-10.

Hui-Chuan, L. y Hsueh, L. (2012). "Estudio del uso del artículo a partir de un corpus paralelo de aprendices, CPATEI". *Revista de Lingüística y Lenguas Aplicadas, 7*/1: 193-202. https://doi.org/10.4995/rlyla.2012.1135

Huston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Isabelle, P. (1992). "La bi-textualité: vers une nouvelle generation d'aides à la traduction et la terminologie". *Meta, 37*/4: 721–737. https://doi.org/10.7202/003228ar

Jiang, C. (2010). "Quality assessment for the translation of museum texts: Application of a systemic functional model". *Perspectives, 18*/2: 109126. https://doi.org/10.1080/09076761003678734

Kay, M. (2000). "Preface" en J. Véronis (ed.). *Parallel text processing: Alignment and use of translation corpora*. Dordrecht: Kluwer, xv-xx. https://doi.org/10.1016/S0891-5520(05)70246-4

Kilgarriff, A., *et al.* (2014). "The Sketch Engine: ten years on". *Lexicography, 1*/1: 7-36. https://doi.org/10.1007/s40607-014-0009-9

Kraif, O. (2002). "Translation alignment and lexical correspondences: A methodological reflection" en B. Altenberg y S. Granger (eds.). *Lexis in contrast: Corpus-based approaches*. Ámsterdam: John Benjamins: 271-289. https://doi.org/10.1075/scl.7.19kra

Lahaussois, A. y Séverine, G. (2012). "A viewing and processing tool for the analysis of a comparable corpus of Kiranti mythology" en R. Rapp *et al.* (eds.). *Proceedings of LREC2012.* Estambul: European Language Resources Association: 33-41.

Llisterri, J. (2017). "Corpus linguistics and written language resources — Bibliography". http://liceu.uab.es/~joaquim/language_resources/lang_res/biblio_corpus.html#General_references_on_corpus_linguistics.

Martínez Belchí, E. (2015). "Recursos en línea sobre corpus y su utilidad para la traducción de unidades fraseológicas" en G. Conde Tarrío *et al.* (eds.). *Enfoques actuales para la traducción fraseológica y paremiológica*. Madrid: Centro Virtual Cervantes, 85-96.

McEnery, T. y Hardie, A. (2012) *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T. y Wilson, A. (2001). *Corpus linguistics: An introduction* (2.ª ed.). Edimburgo: Edinburgh University Press.

Molés-Cases, T. (2016). "Compilación y análisis de un corpus paralelo para la investigación en traducción. Proyecto con Déjà Vu, TreeTagger e IMS open corpus Workbench". *Revista de Lingüística Teórica y Aplicada, 54*/1: 149-174. https://doi.org/10.4067/S0718-48832016000100008

Nádvorníková, O. (2017). "Pièges méthodologiques des corpus parallèles et comment les éviter". *Corela, 21*: 1-28. https://doi.org/10.4000/corela.4810

Neather, R. (2008). "Translating tea: On the semiotics of interlingual practice in the Hong Kong Museum of Tea Ware". *Meta, 53*/1: 218-240. https://doi.org/10.7202/017984ar

Neather, R. (2012). "Intertextuality, translation, and the semiotics of museum presentation: The case of bilingual texts in Chinese museums". *Semiotica, 192*: 197-218. https://doi.org/10.1515/sem-2012-0082

Parodi, G. (2008). "Lingüística de corpus: Una introducción al ámbito". *Revista de Lingüística Teórica y Aplicada, 46*/1: 93-119. https://doi.org/10.4067/S0718-48832008000100006

Rioja, M. (2010). "English-Spanish translations of narrative texts under Franco. Findings from corpus TRACEni (1962-1969)". *Revista de Lingüística y Lenguas Aplicadas, 5*/1: 177-194. https://doi.org/10.4995/rlyla.2010.768

Rojo, G. (2008). "Lingüística de corpus y lingüística del español" en *Actas del XV Congreso de la ALFAL*. Montevideo: ALFAL, 31 págs.

Rojo, G. (2015). "Hispanic corpus linguistics" en M. Lacorte (ed.). *The Routledge handbook of hispanic applied linguistics*. Nueva York: Routledge, 371-387.

Sánchez Ramos, M. (2017). "Metodología de corpus y formación en la traducción especializada (inglés-español): una propuesta para la mejora de la adquisición de vocabulario especializado". *Revista de Lingüística y Lenguas Aplicadas, 12*/1: 137150. https://doi.org/10.4995/rlyla.2017.6969

Scheer, T. (2013). "The corpus: A tool among others". *Corela, 13*: 1-14. https://doi.org/10.4000/corela.3006

SDL Group (2013). *SDL Trados Studio 2014* [software]. Maidenhead: SDL Group.

SDL Group (2016). *SDL Trados Studio 2017* [software]. Maidenhead: SDL Group.

Seghiri, M. (2011). "Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad". *Revista de Lingüística Teórica y Aplicada, 49*/2: 13-30. https://doi.org/10.4067/S0718-48832011000200002

Serón Ordóñez, I. (2015). "Cómo crear y analizar corpus paralelos. Un procedimiento con *software* accesible y económico y algunas sugerencias para *software* futuro" en M.T. Sánchez Nieto (ed.). *Corpus-based translation and interpreting studies*. Berlín: Frank & Timme, 167-190.

Sharoff, S. (2002). "Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics" en *Proceedings of LREC02*. Las Palmas de Gran Canaria: European Language Resources Association, 447-452.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Torrellas Castillo, M. (2009). "Corpus bilingues massifs et mémoires de traduction : la version espagnole des textes juridiques de l'UE". *Revue Française de Linguistique Appliquée, 24*/1: 83-92.

Valdeón, R.A. (2015). "Colonial museums in the US (un)translated". *Language and Intercultural Communication* 15/3: 362-375. https://doi.org/10.1080/14708477.2015.1015351

Vaughan, E. y Clancy, B. (2013). "Small corpora and pragmatics" en J. Romero-Trillo (ed.). *Yearbook of Corpus Linguistics and Pragmatics 2013: New Domains and Methodologies.* Dordrecht: Springer, 53-73. https://doi.org/10.1007/978-94-007-6250-3_4

Véronis, J. (2000a). "Foreword" en J. Véronis (ed.). *Parallel text processing: Alignment and use of translation corpora*. Dordrecht: Kluwer, XI-XII. https://doi.org/10.1007/978-94-017-2535-4

Véronis, J. (2000b). "From the Rosetta stone to the information society: A survey of parallel text processing" en J. Véronis (ed.). *Parallel text processing: Alignment and use of translation corpora*. Dordrecht: Kluwer, 1-24. https://doi.org/10.1007/978-94-017-2535-4_1

Vigier, F.J y Sánchez, M. del M. (2017). "Using Parallel Corpora to Study the Translation of Legal System-Bound Terms: The Case of Names of English and Spanish Courts" en R. Mitkov (ed.). *Europhras 2017*, LNAI 10596, 260–273. https://doi.org/10.1007/978-3-319-69805-2_19

Vondřička, P. (2014). "Aligning parallel texts with InterText" en N. Calzolari *et al.* (eds.). *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), 18751879.

Wright, C.V. (2016). *Blue Guide New York* (5.ª ed.). Londres: Somerset.

Zanettin, F. (2012). *Translation-driven corpora*. Mánchester; St. Jerome.

Zanettin, F. (2013). "Corpus methods for descriptive translation studies". *Procedia, 95*: 20–32. https://doi.org/10.1016/j.sbspro.2013.10.618

Zanettin, F., Saldanha, G. y Harding, S.A. (2015). "Sketching landscapes in translation studies: A bibliographic study". *Perspectives, 23*/2: 161-182. https://doi.org/10.1080/0907676X.2015.1010551

Zubillaga, N., Sanz, Z. y Uribarri, I. (2015). "Building a trilingual parallel corpus to analyse literary translations from German into Basque" en C. Fantinuoli y F. Zanettin (eds.). *New directions in corpus-based translation studies.* Berlín: Language Science Press, 61-81.

APÉNDICE A. DESCRIPCIÓN DE LOS REGISTROS DE MUSA16 MENCIONADOS

Código	Museo	Título (documento inglés)	Año	Descripción
AMNH03	American Museum of Natural History	Floor plan	2011	Mapa e información práctica
	URL (inglés)	http://www.amnh.org/content/download/81303/1518508/file/Evergreen_floorplan_2011_6-16-11%202.pdf		
	URL (español)	http://www.amnh.org/content/download/81292/1518453/file/Evergreen_floorplan_2011_ES.pdf.		
METM08	Metropolitan Museum of Art	Metropolitan Museum Announces Gift of Major Cubist Collection	2013	Comunicado de prensa
	URL (inglés)	http://www.metmuseum.org/press/news/2013/lauder-announcement		
	URL (español)	http://www.metmuseum.org/press/news/2013/lauder-announcement-spanish		
MOMA26	Museum of Modern Art	The impact of the development of museum programs	2011	Informe ejecutivo
	URL (inglés)	http://www.moma.org/momaorg/shared/pdfs/docs/meetme/Resources_AudienceFocus_Evaluation.pdf		
	URL (español)	http://www.moma.org/momaorg/shared/pdfs/docs/meetme/Resources_AudienceFocus_Evaluation_es.pdf		
NMAI01	N'I Museum of the American Indian	Ancient Mexican Art	2002	Página web de exposición
	URL (inglés)	http://nmai.si.edu/exhibitions/ancient/english/index.html		
	URL (español)	http://nmai.si.edu/exhibitions/ancient/spanish/flash_espanol.swf		
NMAI02	N'I Museum of the American Indian	Film and Video Preview Submission Form	2011	Formulario de registro
	URL (inglés)	http://nmai.si.edu/sites/1/files/nativenetworks/preview_sub_form_11.pdf		
	URL (español)	http://nmai.si.edu/sites/1/files/nativenetworks/preview_sub_form_sp.pdf		
NMAI04	N'I Museum of the American Indian	Beauty surrounds us	2011	Página web de exposición
	URL (inglés)	http://nmai.si.edu/static/exhibitions/beauty_surrounds_us/english/		
	URL (español)	http://nmai.si.edu/static/exhibitions/beauty_surrounds_us/spanish/		
NYHS01	New-York Historical Society	Nueva York (1613-1945)	2010	Texto de exposición
	URL (inglés)	https://www.nyhistory.org/web/PDF/nuevayork/NuevaYork_LARGEPRINT_EN-FINAL.pdf		
	URL (español)	https://www.nyhistory.org/web/PDF/nuevayork/NuevaYork_LARGEPRINT_SP_FINAL.pdf		
SKYM01	Skyscraper Museum	Downtown New York	1999	Texto de exposición
	URL (inglés)	http://skyscraper.org/EXHIBITIONS/DOWNTOWN_NEW_YORK/CONTENT/dny_text.htm		
	URL (español)	http://skyscraper.org/EXHIBITIONS/DOWNTOWN_NEW_YORK/CONTENT/dny_texts.htm		