ISSN: 1886-9300

Corpus y nuevas tecnologías aplicadas a la traducción: conceptos teóricos y prácticos para el trabajo de la competencia instrumental en el aula

Cristina Rodríguez-Faneca
Universidad de Córdoba
192rofac@uco.es

https://dx.doi.org.10.12795/futhark.2021.i16.12

Fecha de recepción: 09.09.2021
Fecha de aceptación: 13.12.2021

Resumen: El trabajo con corpus en el aula de traducción conlleva la asimilación previa de multitud de conceptos teóricos y prácticos. Por ello, en el presente trabajo se revisan los puntos clave del trabajo con corpus en el aula de traducción en el contexto del desarrollo de la competencia instrumental. Esta investigación tiene como objetivo principal reflexionar acerca de dichos contenidos en relación con la limitación temporal inherente al currículo formativo del traductor. De este modo, se presentan tanto los fundamentos teóricos como las nociones prácticas que suponen un sustrato básico para el trabajo en el aula: desde el concepto de corpus y los criterios básicos de selección de textos y su tipología hasta el proceso de compilación, análisis y gestión de corpus.

Palabras clave: corpus, competencia instrumental, didáctica de la traducción, documentación, nuevas tecnologías.

Corpora and emerging technologies in translation: theoretical and practical concepts for instrumental competence practice

Abstract: Working with corpora in the translation classroom implies the assimilation of many theoretical and practical concepts. This paper aims to review the key concepts for classroom practice related to the development of instrumental competence. The aim of this study is to reflect on these contents in relation to the time constraints of the translator training curriculum. Thus, the

theoretical and practical foundations representing a basic substratum for classroom practice are presented: the concept of corpus, the basic criteria for selecting texts, corpus typology and the process of compiling, analysing and managing corpora.

Keywords: corpora, instrumental competence, translation didactics, documentation, emerging technologies.

Sumario: 1. Introducción y objetivos. 2. Los corpus lingüísticos en la enseñanza de la traducción. 3. Conceptos clave para el trabajo en el aula. 3.1. Fundamentos teóricos. 3.1.1. Conceptos definitorios, clasificación y tipología. 3.1.2. Representatividad de los corpus. 3.1.3. Tipos de corpus de interés en el ejercicio professional de la traducción. 3.2. Nociones practices. 3.2.1. Compilación del corpus. 3.2.1.1. Selección de los textos. 3.2.1.2. Descarga y organización de los ficheros. 3.2.2. Análisis y gestión. 3.2.2.1. Software de trabajo con corpus. 3.2.2.2. Análisis básico del texto: composición del corpus e índices de frecuencias. 3.2.2.3. Palabras clave. 3.2.2.4. Concordancias. 3.2.2.5. Agrupamientos y n-gramas. 3.2.2.6. Lematización. 3.2.2.7. Otros aspectos funcionales de AntConc. 4. Conclusiones.

1. Introducción y objetivos

La importancia de las nuevas tecnologías en el ámbito de la traducción profesional y en sus tareas anexas —desde la propia labor de documentación hasta el contacto con el cliente, pasando por el resto de etapas inherentes a un encargo de traducción— ha sido ampliamente referida en la literatura en torno al tema (Rodríguez-Faneca y Maz-Machado, 2020). Es precisamente este contexto de sinergia tecnológica el que justifica el uso de los corpus lingüísticos electrónicos como apoyo a la labor de traducción y, por ende, la obligación de trasladar este tipo de contenidos a los planes de formación del traductor.

Por ello, el objetivo principal del presente trabajo es el de reflexionar acerca de la pertinencia de los contenidos teóricos y prácticos básicos del trabajo con corpus en el aula de traducción, debido precisamente a la limitación temporal inherente al currículo formativo del traductor. En este sentido, los contenidos teóricos y prácticos que se expondrán están orientados especialmente al desarrollo de la competencia instrumental del alumnado. Paralelamente, pretendemos sentar las bases de una futura propuesta didáctica que se apoye en el equilibrio entre la trasmisión de contenidos y la correcta temporalización de estos.

2. Los corpus lingüísticos en la enseñanza de la traducción

Existen multitud de investigaciones donde podemos asistir al uso de corpus como recurso pedagógico para la traducción, especialmente como herramienta para el desarrollo de las subcompetencias bilingüe y estratégica. En concreto, los corpus

lingüísticos se erigen como una herramienta didáctica para la enseñanza, sobre todo, de la traducción especializada. Así, por ejemplo, Corpas (2001) propone la compilación de un corpus *ad hoc* de textos originales comparables para la enseñanza de la traducción inversa especializada, mientras que Martínez y Arjonilla (2018), por su parte, presentan un corpus de prospectos de medicamentos como recurso didáctico para la traducción directa especializada.

Otras propuestas donde se insta a trabajar con corpus de un modo similar son los trabajos de Vicente (2010), López-Rodríguez y Buendía-Castro (2011; 2013), Seghiri (2012), Flores Acuña (2014) o Sánchez Ramos (2017b), por nombrar algunos. Algunos trabajos, como el de Jiménez-Crespo (2009), se centran en el uso de corpus como herramienta pedagógica para la localización. En el contexto pedagógico, el trabajo suele centrarse en los llamados corpus *ad hoc*.

Algunas características del trabajo con corpus *ad hoc* para la enseñanza de la traducción y de la interpretación han sido señaladas por Sánchez Ramos (2017a: 181):

> [...] responde a unas necesidades concretas, como tareas de documentación, o bien como recurso pedagógico para la elaboración de material didáctico, y cuya fuente no es otra que textos electrónicos extraídos de Internet y que han seguido una evaluación de dichos textos y un criterio o protocolo de compilación concreto.

Los estudios mencionados se centran, principalmente, en algún campo de la traducción especializada, ya sea la económica, la científico-técnica o la jurídica; por ello, el foco de las aplicaciones didácticas propuestas no se encuentra en las operaciones de explotación de corpus y en su interpretación, sino en la mera extracción terminológica a partir de contextos. Por ello, y como se ha mencionado anteriormente, resulta pertinente indagar acerca de la idiosincrasia del trabajo con corpus en el aula cuando el foco de la formación está en el desarrollo de la competencia instrumental del alumnado o bien inserto en asignaturas de corte instrumental.

3. Conceptos clave para el trabajo en el aula

En este apartado nos disponemos a revisar los puntos clave del trabajo con corpus electrónicos en el aula de traducción en el contexto del desarrollo de la competencia instrumental. Para ello se presentarán, en primer lugar, los fundamentos teóricos que, a nuestro juicio, suponen un sustrato básico que permite entender la relación de la lingüística de corpus con otras disciplinas —y, por ende, con la propia traducción—: el concepto de corpus y los criterios básicos de selección de textos, los distintos tipos de corpus —en especial, los corpus

paralelos y comparables— y su finalidad. Por otro lado, resulta de especial interés el concepto de *representatividad,* ya que se encuentra estrechamente relacionado con la productividad y el factor temporal de la documentación en el ejercicio profesional de la traducción.

En segundo lugar, expondremos los contenidos prácticos de interés en relación con dos operaciones principales: la compilación del corpus —fase donde se seleccionan los textos que van a formar parte del corpus y se organizan los ficheros tras su descarga y conversión— y el análisis y gestión de corpus, fase donde se realizan una serie de operaciones sobre los textos con la asistencia de un programa informático de gestión de corpus.

3.1. Fundamentos teóricos

3.1.1. Conceptos definitorios, clasificación y tipología

Según el *Expert Advisory Group on Language Ingineering Standards* (EAGLES, 1996a: 4 *cit.* en Hernández, 2002) un corpus está compuesto por una recopilación de muestras lingüísticas, seleccionadas de acuerdo con una serie de criterios y con la finalidad de constituir una muestra representativa de la lengua. En otras palabras, un corpus es una colección de textos agrupados en torno a una serie de parámetros comunes con el fin de servir como referencia para una investigación lingüística.

Los criterios genéricos conforme a los cuales han de seleccionarse los textos que deben formar parte de un corpus son la cantidad, calidad, simplicidad, pertenencia al dominio de especialidad, actualidad, condición lingüística de los textos y factualidad (EAGLES, 1996b: 4). La calidad del corpus, por su parte, se mide por los criterios de inclusión de los propios textos que lo componen (Alvar Ezquerra y Corpas, 1994). Parodi (2008) apunta, por su parte, a una serie de principios o características relevantes en el momento de construir y comprender los alcances de un corpus: extensión, formato, representatividad, diversificación, marcado o etiquetado, procedencia, tamaño de las muestras y adscripciones de tipo disciplinar o temático.

Existen, además una serie de criterios específicos para la selección de textos que pueden subdividirse en dos grupos (Hernández, 2002): los criterios externos o no lingüísticos —como el tipo, género, modalidad, origen o finalidad de los textos que forman parte del corpus— y los criterios internos o lingüísticos, relacionados con las categorías lingüísticas presentes en los textos que forman parte del corpus.

Para abordar los parámetros clasificatorios de los corpus y, por ende, su diversa tipología, seguimos a Torruella Casañas (2017), quien clasifica los distintos

tipos de corpus atendiendo a la suma de varios parámetros (Torruella Casañas, 2017: 41-57):

- *Modalidad*, parámetro que determina si el corpus se encuentra vertebrado por discursos orales —basado en transcripciones y llamado corpus oral—, por discursos escritos —que consisten en colecciones de textos en formato electrónico, denominado corpus escrito—, o bien por una combinación de ambos, los conocidos como corpus mixtos.

- *Temática*, en virtud de la cual encontramos corpus generales o corpus especializados. Los primeros pretenden representar distintas variedades de la lengua de un modo amplio, mientras que los segundos pretenden exponer una variedad lingüística —dialecto, registro— o un sublenguaje especializado en virtud de una temática determinada.

- *Época*, parámetro que determina la limitación temporal establecida a la hora de compilar el corpus. Los corpus contemporáneos se centran en la lengua actual, mientras que los corpus históricos se centran en la lengua del pasado. Los corpus contemporáneos pueden ser diacrónicos —si los textos presentes en él se organizan en periodos— o sincrónicos, en caso contrario.

- *Temporalidad*, parámetro directamente relacionado con el anterior. Los corpus sincrónicos, como se ha mencionado arriba, recopilan textos enclavados en un momento lingüístico e histórico determinado. Los corpus diacrónicos recopilan textos de distintas etapas temporales sucesivas.

- *Magnitud*, parámetro que determina la envergadura del corpus y que hace referencia, por lo general, al número de palabras contenidas en él. Dentro de este parámetro se distinguen dos tipos de corpus: corpus grande y corpus restringido. Los corpus clasificados dentro del primer tipo pueden exceder los cien millones de palabras —siempre y cuando cumplan los criterios de compilación fijados— y no suelen concebirse con un límite de palabras a alcanzar. Sin embargo, los corpus restringidos sí se plantean con un límite de palabras, al estar insertos en un proyecto o al poseer finalidad muy concreta. Los corpus restringidos suelen llevar aparejados una posedición exhaustiva.

- *Evolución*. Si el corpus prevé la incorporación de nuevos textos estamos ante un corpus abierto, mientras que en un corpus cerrado encontramos un número finito de palabras. Como ocurre con los corpus restringidos, los corpus cerrados se encuentran insertos en un proyecto concreto y, por ello, el número de palabras que albergará se encuentra establecido de

antemano. Por otro lado, podemos hablar de los corpus monitor, cuyo volumen de palabras renueva periódicamente para incluir textos actuales.

- *Distribución*, parámetro relacionado con la organización interna de los textos integrantes del corpus. Los corpus proporcionales pretenden ajustar la cantidad de palabras o textos en relación con el fenómeno estudiado y la distribución de este en el total de la población —por ejemplo, en un muestreo del español actual en relación con los distintos dialectos. Los corpus equivalentes guardan equilibrio entre la cantidad de palabra contenida en cada apartado, sin tener en cuenta la incidencia dentro de la población del fenómeno estudiado.
- *Número de ediciones*. En los corpus monoedición se incluye una versión —o testimonio— de los textos que lo integran, mientras que en los corpus pluriedicion se incluyen varias versiones que ser de distinta naturaleza. Dentro de los corpus pluriedición encontramos los corpus comparables y los corpus paralelos. Dentro de estos, encontramos los corpus paralelos monolingües, plurilingües y alineados.
- *Número de lenguas*, parámetro que determina el número de lenguas presentes en los textos que integran el corpus, generando así una clasificación dicotómica: corpus monolingües y corpus plurilingües.
- *Tipo de edición*, parámetro que determina el tratamiento de los textos dentro del corpus en relación con las distintas modalidades de edición existentes —por ejemplo, facsímil o crítica— para generar corpus multiedición.
- *Muestras*, parámetro que determina la cantidad de texto asociado a cada unidad textual dentro del corpus. En los corpus textuales se recoge el texto íntegro de la unidad textual, mientras que en los corpus de referencia se recogen fragmentos representativos de dichas unidades textuales. En los corpus léxicos se recogen fragmentos de menor tamaño cuyo objetivo es el de proporcionar información lexicográfica.
- *Marcaje*, parámetro que determina las opciones de consulta del texto. En los corpus simples estas posibilidades son bastante reducidas puesto que la falta de etiquetas hace que, incluso, no puedan realizarse búsquedas dentro del corpus. Los corpus etiquetados sí ofrecen información adicional que permite personalizar la consulta del corpus. Dentro de este tipo de corpus encontramos los corpus codificados —cuyas etiquetas refieren a aspectos extralingüísticos— y los corpus anotados, cuyas etiquetas se refieren a aspectos lingüísticos. A su vez, los corpus anotados pueden ser corpus

anotados morfológicamente, corpus lematizados, corpus parentizados y corpus analizados.

Por otro lado, y según la finalidad del corpus (Torruella Casañas, 2017), podemos distinguir entre los corpus *ad hoc* o corpus universales. El corpus *ad hoc* tiene como finalidad servir como referencia para un proyecto o finalidad específica, mientras que el corpus universal puede ser usado para diversas finalidades. Dentro de la tipología desarrollada anteriormente merece la pena analizar más en profundidad los corpus paralelos y los corpus comparables, pues se trata de los dos tipos de corpus multilingües más usados para la documentación de encargos de traducción en la práctica profesional de la profesión y, por ende, los dos tipos de corpus más recurrentes y a los que se hará alusión reiteradamente en el aula. En relación con los corpus comparables podemos citar a Peters (1996: 60, *cit.* en Godínez, 2010):

> [...] son un conjunto de textos en más de una lengua que, sin ser traducciones, por coincidir en el tema, motivación situacional y función comunicativa, proporcionan una excelente base para la comparación de dos o más lenguas.

Los corpus comparables se erigen como una herramienta que permiten al traductor adquirir vocabulario de especialidad y familiarizarse con la estructura e idiosincrasia de determinadas tipologías textuales. Sirven, además, para observar el comportamiento de una lengua en diversas circunstancias comunicativas.

Los corpus paralelos, por otra parte, se componen de un conjunto de textos en más de una lengua, presentándose el original en una lengua y sus traducciones en otras. Su principal utilidad se basa en la posibilidad de hallar terminología o fraseología equivalente —así como el estudio de determinadas colocaciones—, si bien son también una materia prima necesaria para el desarrollo de la traducción automática, ya que los programas informáticos trabajan con datos probabilísticos que solo pueden obtenerse a partir de este tipo de corpus (Torruella Casañas, 2017: 52). Hoy en día, la presencia de este tipo de textos en línea es abundante, debido a la existencia de grandes instituciones a nivel europeo y global. No obstante, fuera de este ámbito la existencia de corpus paralelos previamente compilados es escasa.

Por último, hemos de consignar las características de los corpus paralelos alineados. Este tipo de corpus permite (Torruella Casañas, 2017: 52):

> […] la comparación entre las distintas versiones y la equiparación interlingüística de sus elementos, en este tipo de corpus los textos están dispuestos uno al lado de otro, en fragmentos por párrafos o por versículos o por fases o por versos, etc. El hecho de estar alineados por fragmentos textuales facilita enormemente la comparación entre distintas ediciones.

3.1.2. Representatividad de os corpus

La representatividad, la estandarización y la tipología de los corpus son tres de los temas más recurrentes en torno a este objeto de estudio (Hernández, 2002). La representatividad de un corpus hace referencia a su capacidad para exponer con solvencia la interacción o fenómeno que se desea mostrar. Para Biber (1993: 243), la representatividad mide «the extent to which a sample includes the full range of variability in a population».

En este sentido, se espera que la distribución de los textos del corpus siga un reparto proporcional o igualitario. Para que un corpus refleje a partir de las muestras que lo componen las características del total de la población, tiene que basarse en la representatividad de sus componentes (Torruella Casañas, 2016). La representatividad del corpus podría, así, medirse en virtud de dos ejes: cuantitativo —cantidad de textos del corpus— y cualitativo —características de las muestras textuales que lo componen: calidad, adecuación, etc.

Corpas y Seghiri (2006) han reflexionado acerca del número de muestras textuales necesarias para garantizar la representatividad y cientificidad del corpus. En este sentido, la mayoría de los autores apunta a que la cantidad de textos no es, en la mayoría de ocasiones, una garantía de representatividad, sino que la fijación de unos criterios textuales de compilación (Piñol, 2012) es la que puede dotar al corpus de la solvencia esperada. Sin embargo, aún coexisten las dos posturas en torno a esta problemática, con la premisa de «*there is no text like more text*» (Hernández, 2002).

A la par que se espera que un corpus cumpla criterios de homogeneidad y heterogeneidad (Kabatek, 2016), aunque resulte paradójico, también se espera que puedan describir de manera genérica y precisa (Rodríguez-Faneca, 2019: 169) la interacción que se desea mostrar. A pesar de que esta paradoja —«la paradoja de Enrique»—, suele darse de un modo más claro en los corpus sincrónicos, merece reflexión, en cuanto que explica algunos de los criterios *sine qua non* para que un corpus sea representativo (Enrique-Arias, 2012: 96, *cit.* en Kabatek, 2016: 5):

> Una paradoja de la composición de los corpus diacrónicos es que, por un lado, deben ser heterogéneos (tienen que incluir textos de diferentes autores, épocas, géneros, registros, dialectos) y a la vez deben ser homogéneos (es decir, los diferentes cortes sincrónicos representados en el corpus tienen que ser comparables entre sí).

En relación con la representatividad, en definitiva, nuestra postura ha quedado ya establecida en trabajos anteriores (Rodríguez-Faneca, 2019: 169), donde manifestamos que, en lugar de establecer un volumen determinado de

palabras en un corpus, es indispensable prestar atención a la interacción quiere mostrarse y al propósito de la compilación del corpus.

3.1.3. Tipos de corpus de interés en el ejercicio profesional de la traducción

Acercándonos más a nuestro objeto de estudio en este epígrafe, y en relación con los corpus aplicados a un contexto puramente instrumental —donde, como hemos referido anteriormente, el foco principal se encuentra en la labor de documentación—, encontramos varias propuestas interesantes dentro de la literatura: los llamados *quick and dirty corpus* (Belchí, 2015), los *Do It Yourself Corpus* (DIYC) (Zanettin, 2002) y el propio concepto de *internet as corpus* (Zanettin, 2009), también referido como *web as corpus* por otros autores.

Estas tres propuestas se caracterizan, por lo general, por generar corpus de tamaño variable que resultan más funcionales que exactos. Si bien en la revisión de la literatura hemos encontrado algunas propuestas en las que no era posible discernir correctamente las diferencias entre estos tres tipos de corpus —en ocasiones, algunos autores consideran que se trata de una misma realidad—, consideramos que sí existen algunos elementos diferenciadores entre ellos tras una revisión profunda de dichas propuestas. Estos aspectos se consignan, a modo de confrontación, en la tabla 1.

Los *quick and dirty corpus* suponen una solución electrónica, rápida e informal para solventar problemas traductológicos a pequeña escala, normalmente a nivel microestructural. En palabras de Belchí (2015), los corpus de este tipo sirven para indagar o confirmar, sobre todo, información terminológica. Debido a que el problema más común de este tipo de corpus es el ruido generado en las búsquedas en línea (Belchí, 2015), existen diversas estrategias para realizar búsquedas más productivas y específicas en este contexto (Robb, 2003). Los *Do It Yourself Corpus* (DIYC) (Zanettin, 2002) pueden definirse como una «colección de documentos que provienen de internet o, más específicamente, páginas web en HTML» y que poseen las siguientes características (Zanettin, 2002: 4):

- Se ha creado ad hoc para la traducción de un texto específico.
- Es un corpus abierto. Se puede añadir más materiales si se necesitan.
- Es desechable (Varantola, 2000) o virtual (Ahmad *et al*., 1994). No está destinado a ser parte de un corpus de carácter permanente, y puede ser eliminado tan pronto como se complete la traducción. No se requieren derechos de Copyright.
- Al igual que los textos paralelos, pueden ser tanto comparables y bilingües como monolingües.

En relación con el uso de internet como corpus hemos de apuntar que, hoy en día, todavía no existe un consenso con respecto a su propio significado (López-Rodríguez y Buendía-Castro, 2011), ya que podemos acercarnos a este concepto desde tres perspectivas distintas (Bernardini, Baroni y Evert, 2006, *cit.* en López-Rodríguez y Buendía-Castro, 2011):

- La web como sustituta del corpus (*web as corpus surrogate*), enfoque que considera la web en sí misma como un gran corpus.
- La web como *mega-corpus* o *mini-web*, enfoque que se encuentra en proyecto, y que crearía un nuevo objeto con características similares a las webs y a los corpus, con material textual derivado de la red, implementando, al mismo tiempo, herramientas de análisis o anotación.
- La web como un supermercado de corpus (*corpus as supermarket*), enfoque que concibe la web como un modo de localizar textos mediante un motor de búsqueda para su posterior descarga.

Sin embargo, el uso de internet como corpus «no sigue unos criterios de diseño y en muchos casos falta información sobre el número y procedencia de los textos» (Llamazares, 2003: 211), por lo que en la mayoría de los casos no procede una descarga, procesado y compilación de los textos. El foco, por ello, se encuentra en la operación de consulta, para la que hay que emplear una serie de estrategias que garanticen una correcta recuperación de la información.

Zanettin (2009) considera que un corpus que haya sido compilado atendiendo a una serie de criterios, es decir, un corpus bien estructurado, puede verse como herramienta complementaria al uso de la web como corpus —y viceversa—, enriqueciendo así en horizonte documental del traductor. Este autor realiza una comparación entre el *British National Corpus* (BNC), el *Corpus of Contemporary American English* (COCA) y el uso de la web como corpus (2009: 218):

> While corpora such as the BNC and the COCA are certainly more reliable than the Web as concerns core patterns of language use, the sheer volume of the Web means that this source can cover the whole spectrum of possibilities: at one end no corpus can rival the lexical and terminological richness of the Web; at the other end only in the Web can very long stretches of discourse be found.

Como se ha mencionado con anterioridad, la diferencias entre estos tres tipos de corpus —los *quick and dirty corpus*, los DIYC corpus y el uso de internet como corpus— son sutiles, si bien es posible esbozar seis parámetros donde es posible

apreciar dichas diferencias: la evolución, compilación, foco de uso, carácter, tamaño y objetivo del corpus.

	Quick and dirty corpus	**DIYC**	**Internet as corpus**
Evolución	Abierto	Abierto	Abierto
Compilación	*Ad hoc*	*Ad hoc*	No hay compilación
Carácter	Desechable	Desechable	Virtual
Tamaño	Pequeño	Variable	Grande
Foco	**Compilación Consulta**	**Compilación Análisis**	**Consulta**
Objetivos	**Terminología**	**Conceptualización temática Terminología Tipología textual**	**Terminología**
Papel de la web	Obtener doc.	Obtener documentos	Consultar doc.

Tabla 1. Confrontación de distintos tipos de corpus. Fuente: elaboración propia.

3.2. Nociones prácticas

3.2.1. Compilación del corpus

3.2.1.1. Selección de los textos

Respecto a la preparación de los textos que han de formar parte del corpus, instamos a seleccionarlos de acuerdo con los parámetros consignados en epígrafes anteriores. En primer lugar, los criterios externos o no lingüísticos —como el tipo, género, modalidad, origen o finalidad de los textos—, así como los criterios internos o lingüísticos —relacionados con las categorías lingüísticas presentes en los textos, como la densidad terminológica o conceptual—, en línea con las propuestas de Hernández (2002), EAGLES (1996a; 1996b) y Parodi (2008). En segundo lugar, serán de aplicación, además, las premisas de representatividad desarrolladas en el epígrafe 3.1.2, a la par que se tiene en mente la creación de un

corpus con una tipología determinada, en la línea de lo señalado por Torruella Casañas (2017). En definitiva, y a la hora de preparar la selección de textos para un encargo determinado, las máximas a seguir deberán ser, como mínimo, las relativas a la temática, nivel de especialidad y relevancia para el proceso de traducción (Corpas, 2001).

3.2.1.2. Descarga y organización de los ficheros

Tras descargar los textos que van a formar parte del corpus debemos organizar los ficheros que los contienen. Resulta imprescindible prever la codificación de los caracteres de los documentos —es recomendable usar el formato UTF8 (*Unicode Transformation Format*)—, el formato de los ficheros y su propio nombre (Torruella, 2017: 166).

Es preferible que el formato de los ficheros que contienen los textos del corpus sea el mismo —ya sea *.doc, *.pdf, *.html, u otro formato— para que el procesado resulte más ágil. En caso de que no sea posible, existen varios conversores de archivo tanto online como de escritorio que es posible utilizar. En caso de que los textos se encuentren en *.html, es posible utilizar gestores de descarga de sitios web como *WinHTTrack Website Copier* (Sánchez Ramos, 2017a).

Posteriormente, independientemente del formato en el que se encuentren los textos, deberán convertirse a *.txt —texto plano— para su procesado por parte del programa informático de análisis de corpus. Es necesario, además, guardar una copia de seguridad de los archivos en formato original, de manera que pueda hacerse uso de ella si se detectase algún error tras la conversión. A la hora de nombrar los ficheros del corpus, el nombre debe estar compuesto solamente por caracteres alfabéticos latinos y números (Torruella, 2017: 166), dejando de lado las tildes.

Respecto al propio nombre, se recomienda incluir en él la fecha de inclusión en el corpus —p. ej.: «23-03-20»— junto con un identificador que dé cuenta del tipo de texto, lengua, etc. Algunos autores, como el propio Torruella (2017) recomiendan que los datos a consignar se expresen mediante una codificación determinada, ya que conviene que los nombres de los archivos generados sean cortos. Por ejemplo, durante la compilación de un corpus comparable bilingüe (inglés/español) de textos jurídicos, podríamos renombrar a una partida de nacimiento —otorgándole, por ejemplo, el distintivo «pn»— en español del siguiente modo: «23-03-20-pn-1-ES». Para asistirnos en este procedimiento existen, asimismo, programas para renombrar archivos en lote como *Personal Renamer, FastFile Rename* y *Lupass Rename* (Sánchez Ramos, 2017a).

3.2.2. Análisis y gestión

3.2.2.1. Software de trabajo con corpus

Genéricamente, podemos referirnos al *software* de trabajo con corpus como *herramientas de análisis, gestión o explotación de corpus*. En general, este tipo de programa informático permite desde un análisis básico del texto —dando información acerca de las estadísticas de composición del corpus o los índices de frecuencias— pasando por la determinación de sus palabras clave, hasta llegar a la codificación del corpus por medio del etiquetado de sus formas.

El uso principal de estos programas es generar concordancias, es decir, obtener todas las ocurrencias de una determinada palabra en un texto o colección de textos (Vivaldi Palatresi, 2009). Otras aplicaciones de los corpus incluyen el estudio de n-gramas —frecuencias de n palabras— u otros cálculos estadísticos de diversa índole.

AntConc es un programa de gestión y análisis de corpus creado por Laurence Anthony que permite obtener listados de palabras o de patrones colocacionales. Además, es gratuito y posee una interfaz intuitiva que permite un aprendizaje rápido. Por este motivo, se trata de nuestra elección para el trabajo en el aula, dejando de lado otros programas como *WordsSmith Tools*, *ParaConc* —el hecho de que sus licencias sean de pago no facilita su uso en el aula— *Simple Concordance Program*, WMkatrix o *Sketch Engine*.

3.2.2.2. Análisis básico del texto: composición del corpus e índices de frecuencias

Un análisis preliminar del texto en el que se obtenga información acerca de las estadísticas de composición del corpus es esencial para determinar rápidamente la red conceptual del texto sin tener que recurrir a su lectura, así como para determinar su temática y la adecuación del texto al corpus (Corpas, 2001). La información obtenida con un análisis básico del texto —índices de frecuencias del corpus— constituye en sí misma un resumen del mismo (Corpas, 2001). Por otro lado, la visualización de las unidades léxicas resultantes tras la aplicación del filtro ofrecen una panorámica de la temática del texto. De este modo, independientemente del uso que se le quiera dar al corpus, es recomendable realizar un análisis de este tipo, donde se dé cuenta de la composición del corpus y su tamaño (Hernández, 2002).

Para obtener la información deseada, recurriremos a la pestaña *Word List* de *AntConc*, donde obtendremos una lista de palabras ordenada alfabéticamente o por frecuencias. Sin embargo, antes de realizar este primer análisis es recomendable

crear una *stoplist*, es decir, una lista de palabras que contenga artículos, preposiciones y adverbios, de manera que se evite la aparición de palabras irrelevantes que generen ruido en el análisis de los textos (Sánchez Ramos, 2017a). Estas palabras, que se repiten continuamente pero que no poseen un significado léxico —aunque sí gramatical—, se denominan *stopwords*. Son también *stopwords* los verbos auxiliares, los pronombres o los determinantes. Alojando estas palabras manualmente en un fichero *.txt —creado *ad hoc* o descargado de internet— obtendremos una *stoplist* que podremos aplicar como filtro (figura 1).

```
a               here          some
a's             here's        somebody
able            hereafter     somehow
about           hereby        someone
above           herein        something
according       hereupon      sometime
accordingly     hers          sometimes
across          herself       somewhat
actually        hi            somewhere
after           him           soon
afterwards      himself       sorry
again           his           specified
against         hither        specify
ain't           hopefully     specifying
all             how           still
```

Figura 1. Extractos de una *stoplist* monolingüe (inglés).

Posteriormente, podemos obtener los índices de frecuencias del corpus, cálculos que se llevarán a cabo sin tener en cuenta las *stopwords* consignadas en la *stoplist*. En las figuras 2 y 3 se muestra el cálculo de frecuencias dentro de un corpus *ad hoc* sobre contaminación hídrica y salud pública en países en desarrollo. La información se muestra tanto con un filtro de *stopwords* (izquierda, figura 2) como sin dicho filtro (derecha, figura 3).

Rank	Freq	Word
1	829	water
2	245	drinking
3	186	health
4	172	pakistan
5	124	quality
6	101	contamination
7	99	samples
8	88	groundwater
9	81	pesticides
10	79	public

Rank	Freq	Word
1	1275	the
2	1208	of
3	1029	and
4	921	in
5	829	water
6	486	to
7	442	a
8	355	al
9	355	et
10	281	for

Figura 2. Cálculo de frecuencias. Filtro de *stoplist* activado.

Rank	Freq	Keyness	Effect	Keyword
1	829	+ 6452.59	0.0777	water
2	245	+ 2271.03	0.0271	drinking
3	172	+ 1614.7	0.0193	pakistan
4	88	+ 1030.29	0.01	groundwater
5	101	+ 1021.72	0.0115	contamination
6	186	+ 955.27	0.0182	health
7	99	+ 824.46	0.0111	samples
8	81	+ 801	0.0092	pesticides
9	124	+ 760.39	0.0133	quality
10	50	+ 629.32	0.0057	pcrwr

Figura 3. Cálculo de frecuencias. Filtro de *stoplist* desactivado.

Rank	Freq	Keyness	Effect	Keyword
1	829	+ 6452.59	0.0777	water
2	245	+ 2271.03	0.0271	drinking
3	186	+ 955.27	0.0182	health
4	172	+ 1614.7	0.0193	pakistan
5	124	+ 760.39	0.0133	quality
6	101	+ 1021.72	0.0115	contamination
7	99	+ 824.46	0.0111	samples
8	88	+ 1030.29	0.01	groundwater
9	81	+ 801	0.0092	pesticides
10	79	+ 253.67	0.0076	public

Además del listado, el análisis nos mostrará el número de palabras totales (los llamados *tokens*) y el número de formas distintas de cada una de ellas (los llamados *types*). Es posible calcular la ratio palabras/formas —y, por ende, la riqueza léxica del texto— de un modo básico con la siguiente fórmula (Hernández, 2002):

3.2.2.3. Palabras clave

La funcionalidad *Keyword List* permite extraer las palabras clave de un corpus. Para identificar las palabras clave, *AntConc* compara los patrones de frecuencia de aparición de una determinada palabra en el corpus compilado y en un corpus de mayor tamaño —corpus de referencia— que hemos de aportar. De esta manera, las palabras identificadas como clave no son necesariamente las que aparecen en un mayor número de ocasiones en el corpus, sino aquellas que muestran una frecuencia significativa en comparación con la del corpus de referencia (Hernández, 2002); por tanto, se tratará de aquellas palabras frecuentes en el corpus analizado e infrecuentes en el corpus de referencia.

En este caso, se realizó el análisis de las palabras claves para el corpus *ad hoc* sobre contaminación hídrica y salud pública en países en desarrollo aportando como corpus de referencia una muestra del *Corpus of Contemporary American English* (COCA) de 8,9 millones de palabras. Esta muestra incluía textos extraídos de noticias, documentos académicos, textos de ficción, textos orales —incluyendo programas de televisión— y textos extraídos de páginas webs de interés general o blogs, entre otros. Es posible listar la información tanto siguiendo un criterio de representatividad (*keyness*, figura 4) como por frecuencia (figura 5).

Figura 4. Ordenación. Criterio: representatividad.

Figura 5. Ordenación de las palabras clave. Criterio: alfabético.

3.2.2.4. Concordancias

La opción *Concordance* (figura 6) permite visualizar los términos en su contexto. Con esta opción, denominada también *keyword in context* (KWIC), es posible analizar el uso de una palaba concreta en el corpus. Las concordancias se erigen como una herramienta especialmente útil para el traductor, debido a que permiten indagar acerca de aspectos terminológicos. En palabras de Corpas (2001: 168), «las líneas de concordancia ofrecen al traductor documentación factual y terminológica fiable e inmediata».

Figura 6. Concordancias del término «water».

3.2.2.5. Agrupamientos y n-gramas

La función agrupamientos y n-gramas (*Clusters and n-grams*) permite analizar patrones de comportamiento de una determinada palabra cuando aparece junto a otras del corpus. Los denominados n-gramas hacen referencia a grupos de palabras, donde n puede ser cualquier número, para así formar bigramas, trigramas o tetragramas, dependiendo del número de palabras que estén posicionadas de modo consecutivo en el desarrollo del texto.

En las figuras 7 y 8 se muestran bigramas y trigramas asociados al término «water» en un corpus *ad hoc* sobre contaminación hídrica y salud pública.

Figura 7. Bigramas asociados al

Figura 8. Trigramas asociados al

término «water». término «water».

A pesar de esto, es muy común que los n-gramas —más frecuentemente, bigramas— de interés se encuentren en posiciones intermedias de la lista generada por *AntConc*, debido a que la mayoría de bigramas están formados por la combinación de una *stopword* y un determinado término.

3.2.2.6. Lematización

A través de la lematización —es decir, la segmentación de unidades léxicas en lexemas y afijos— es posible identificar todas las variantes de un determinado término que aparecen en el texto para, así, poder cuantificarlas correctamente mediante el análisis de frecuencias, ya que todas las variantes morfológicas de una misma palabra —por ejemplo, «vivir», «viviendo» o «vivido», distintas formas del verbo «vivir» con distintas flexiones— se cuantificarán como una sola a partir del lema asignado. La lematización de un corpus con *AntConc* se lleva a cabo a partir de la anexión de un documento *.txt donde, previamente, se hayan asignado distintos lemas a la palabra principal.

3.2.2.7. Otros aspectos funcionales de AntConc

Por último, nos parece pertinente consignar algunos aspectos funcionales y aspectos de la configuración de *AntConc* que resultan de utilidad a la hora de trabajar con cualquiera de las operaciones mencionadas con anterioridad. En concreto, describiremos el uso de comodines en la herramienta, el tratamiento de mayúsculas y minúsculas y referiremos algunos aspectos relacionados con la codificación del corpus.

El uso de comodines —cuyo uso en *AntConc* queda reservado al símbolo «*»— permite ahondar en lo relativo a la información conceptual y sinonímica que puede aportar el corpus (Sánchez Ramos, 2017a). El uso de comodines a la hora de recuperar información dentro del corpus permite, por ejemplo, obtener resultados de la palabra en singular y plural, eliminar afijos de la búsqueda en pos de resultados menos restrictivos e incluso obtener información definitoria del término a través de la visualización de distintos contextos (figura 9).

Figura 9. Uso del comodín «*» asociado a la raíz «bio».

Otro punto en el que hemos de incidir es en el tratamiento de mayúsculas y minúsculas a la hora de realizar búsquedas en el texto, ya que el programa dispone de una casilla —denominada «case»— cuya activación regirá una búsqueda que respete el uso de mayúsculas previsto en la caja de búsqueda y, en caso de desactivarse, regirá la omisión de estas en la búsqueda. La utilidad de esta función radica en la posibilidad de encontrar unidades que —más allá, por ejemplo, de hallarse al principio de una frase— posean connotaciones distintas al resto, como pueden ser nombres propios o de organizaciones (figura 10).

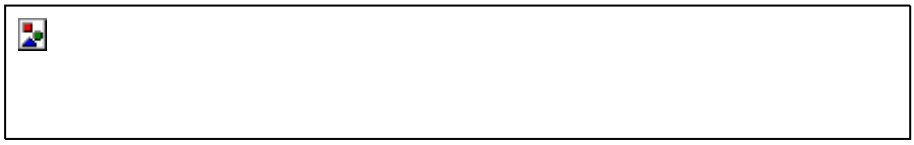

Figura 10. Aplicación del filtro «case» en la palabra «water».

Por último, la codificación del corpus permite añadir información metatextual de diversa índole para estudiar más profundamente los textos que lo componen (Hernández, 2002) y para poder precisar las búsquedas del corpus. Esta información se añade a través de códigos simplificados. Los tipos de anotación más comunes son la etiquetación morfológica —*part of speech tagging*—, el análisis sintáctico —*parsing*— y la lematización —*lemmatisation*—. En relación con el procedimiento de etiquetado o anotación, Leech (1993) advierte que la anotación debe poder sustraerse del texto sin alterarlo y que se ha de desarrollar una metodología de etiquetado, es decir, que la anotación debe responder siempre a las mismas reglas.

4. Conclusiones

El trabajo con corpus en el aula de traducción conlleva la asimilación previa de multitud de conceptos teóricos y prácticos. Por ello, en este trabajo hemos tratado de ofrecer una visión panorámica sobre los contenidos que, a nuestro juicio, son básicos a la hora de abordar el trabajo con corpus electrónicos en el aula de traducción desde un punto de vista instrumental —en asignaturas centradas en informática, terminología, etc.—, en contraposición al trabajo con corpus en asignaturas de traducción —donde la labor del docente se centra, más frecuentemente, en impulsar las competencias bilingüe y estratégica del alumnado. Es precisamente en el contexto de las asignaturas de corte instrumental donde la limitación temporal obliga a seleccionar cuidadosamente los contenidos en relación

con la planificación del curso. Así, resulta imprescindible delimitar qué contenidos son esenciales y qué contenidos conforman un objetivo agradecido dentro de la tríada temporalización-dificultad-productividad.

Como futura línea de investigación derivada de este trabajo se encuentra la aplicación de los contenidos aquí consignados a una propuesta didáctica real en el marco de la temporalización de una asignatura concreta de corte instrumental.

Referencias bibliográficas

ALVAR EZQUERRA, M. y CORPAS, G. (1994). Criterios de diseño para la creación de córpora. En: Manuel Alvar Ezquerra y Juan Andrés Villena Ponsoda (eds.). *Estudios para un corpus del español* (pp. 31-40). Málaga: Universidad de Málaga.

BELCHÍ, E. M. (2015). Recursos en línea sobre corpus y su utilidad para la traducción de unidades fraseológicas. En: *Enfoques actuales para la traducción fraseológica y paremiológica: ámbitos, recursos y modalidades* (pp. 85-96). Centro Virtual Cervantes.

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

CORPAS, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. *TRANS. Revista de traductología*, (5), 155-184.

CORPAS, G. y SEGHIRI DOMÍNGUEZ, M. (2006). *El concepto de representatividad en la Lingüística del Corpus: aproximaciones teóricas y metodológicas. Documento Técnico.* Disponible en: <https://bit.ly/3IYCYLq> [Consultado el 16/05/21].

EAGLES (1996a). *Text Corpora Working Group Reading Guide*. Documento EAGLES (Expert Advisory Group on Language Engineering).

EAGLES (1996b). *Preliminary Recommendations on Corpus Typology*. Documento EAGLES (Expert Advisory Group on Language Engineering).

ENRIQUE-ARIAS, A. (2012). Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad. *Scriptum digital*, 1, 85-106.

FLORES ACUÑA, E. (2014). El corpus como herramienta para la traducción especializada italiano/español: una experiencia con textos de la industria cosmética. *Philologia Hispalensis, 28*(3-4), 155-182.

GODÍNEZ, J. C. (2010) *El corpus ad hoc como herramienta de traducción*. Memorias del VI foro de estudios en lenguas internacional, 73-95.

HERNÁNDEZ, M. C. P. (2002). Terminografía basada en corpus: aspectos fundamentales de la gestión terminológica. *Estudios de lingüística del español,* 18. Disponible en: <https://bit.ly/3txLkf2> [Consultado el 16/05/21].

HONORÉ, T. (1979). Some simple measures of richness of vocabulary. *ALLC Bulletin, 7*(2), 172-7.

JIMÉNEZ-CRESPO, M. A. (2009). El uso de corpus textuales en localización. *Tradumàtica: traducció i tecnologies de la informació i la comunicació,* (7), 1-15.

KABATEK, J. (2016). Un nuevo capítulo en la lingüística histórica iberorrománica: el trabajo crítico con los corpus. Introducción a este volumen. En: *Lingüística de corpus y lingüística histórica iberorrománica* (pp. 1-18). Berlín: De Gruyter.

LEECH, G. (1993). Corpus Annotation Schemes. *Literary and Linguistic Computing 8*(4), 275-281.

LLAMAZARES, M. V. (2003). Internet como corpus: el caso de" Bibidí". *Contextos*, 41, 205-231.

LÓPEZ-RODRÍGUEZ, C. I. y BUENDÍA-CASTRO, M. (2011). En busca de corpus online a la carta en el aula de traducción científica y técnica. *Trans-Kom (Journal for Translation and Technical Communication Research),* 4(1), 1-22.

LÓPEZ-RODRÍGUEZ, C. I. y BUENDÍA-CASTRO, M. (2013). Aplicación de la lingüística de corpus en la didáctica de la traducción científica y técnica. En: *Tome VIII* (pp. 205-216). Berlín: De Gruyter.

MARTÍNEZ, M. C. R. y ARJONILLA, E. O. (2018). El corpus de prospectos farmacéuticos como recurso didáctico en el aula de traducción especializada francés-español. *MonTI. Monografías de Traducción* e *Interpretación*, (10), 117-140.

PARODI, G. (2008). Lingüística de corpus: una introducción al ámbito. *RLA. Revista de lingüística teórica y aplicada, 46*(1), 93-119.

PIÑOL, M. C (2012). *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: Arco Libros.

RAMOS SÁNCHEZ, Mª. D. M (2017a). Compilación y análisis de un corpus ad hoc como herramienta de documentación electrónica en Traduccción e Interpretación en los Servicios Públicos (TISP). *Estudios de Traducción*, 7, 177-190.

RAMOS SÁNCHEZ, Mª. D. M (2017b). 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, 137-150.

ROBB, T. (2003). Google as a quick 'n dirty corpus tool. *The Electronic Journal for English as a Second Language,* 7(2). Disponible en: <https://bit.ly/3yYdDUO> [Consultado el 16/05/21].

RODRÍGUEZ-FANECA, C. (2019) Detección y planificación de contenidos problemáticos: el caso de las preposiciones locativas en el aprendizaje de lenguas afines (italiano-español). *Futhark: Humanities and Social Sciences Review,* 14, 167-180.

RODRÍGUEZ-FANECA, C. y MAZ-MACHADO, A. (2020). Las TIC en el currículo formativo del traductor de italiano. En: *Claves para la innovación pedagógica ante los nuevos retos: respuestas en la vanguardia de la práctica educativa* (pp. 3048-3052). Barcelona: Octaedro.

SEGHIRI, M. (2012). El corpus comparable para la didáctica de la traducción jurídica inversa. En: Cruces, S., del Pozo, Luna, A. y Álvarez, A. (Eds.). *Traducir en la Frontera* (pp. 815-830). Granada: Atrio.

TORRUELLA CASAÑAS, J. (2016). Tres propuestas en el ámbito de la lingüística de corpus. En: *Lingüística de corpus y lingüística histórica iberorrománica* (pp. 9-112). Berlín: De Gruyter.

TORRUELLA CASAÑAS, J. (2017). *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación científica*. Frankfurt am Main: Peter Lang.

VICENTE, C. (2010). Lingüística de corpus y traducción especializada: aplicaciones a la traducción francés-español de la economía. En: *Tome I-VII* (pp. 1-691). Berlín: De Gruyter.

VIVALDI PALATRESI, J. (2009). Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de corpus textuales. *Tradumàtica*, (7), 1-9.

ZANETTIN, F. (2002). DIY Corpora: The WWW and the Translator. En: Maia, J., Haller, J., Urlrych, M. (eds.). *Training the Language Services Provider for the New Millennium* (pp. 1-10). Porto: Facultade de Letras, Universidade do Porto.

ZANETTIN, F. (2009). Corpus-based translation activities for language learners. *The Interpreter and Translator Trainer, 3*(2), 209-224.