Revisión y propuesta de clasificación de corpus

Goretti Faya Ornia
Universidad de Oviedo

Introducción

En este trabajo, (1) haremos un breve recorrido por algunas de las clasificaciones de corpus más amplias y que más repercusión han tenido en estudios posteriores, (2) comentaremos las ventajas que estas ofrecen así como los vacíos que presentan, y finalmente (3) daremos a conocer nuestra propia propuesta de clasificación, en donde hemos intentado abarcar los intereses tanto de la Lingüística Contrastiva como de los Estudios de Traducción.

Consideramos que contar con una clasificación de corpus clara, organizada e integradora es importante porque permite al investigador (1) localizar fácilmente el tipo de corpus que le pueda ofrecer más prestaciones para su estudio, y conocer qué otras alternativas puede tener a su disposición, así como (2) centrar su trabajo dentro de un marco general de corpus lingüísticos.

Clasificación

Existen diversas propuestas de clasificación de corpus, pero por los motivos expuestos anteriormente, aquí nos centraremos exclusivamente en estas cuatro: (1) la propuesta de Laviosa (1997:290–295), (2) la de Torruella y Llisterri (1999:53–57), (3) la de Corpas Pastor (2001:158) y (4) la de Granger (2003:21). Todas ellas son completas e integradoras, ya que tratan de representar los corpus utilizados en los dos enfoques comentados (la Lingüística Contrastiva y los Estudios de Traducción).

Propuestas existentes

Comenzaremos explicando la propuesta de Laviosa (1997), que atañe exclusivamente a textos traducidos a una misma lengua a partir de lenguas diferentes.

En una primera fase, la autora (1997:291–295) diferencia cinco grandes grupos en conformidad con los siguientes criterios. En primer lugar, según la extensión de los textos que integran el corpus ("texto completo", "de muestras", "mixto"

Babel 60:2 (2014), 234–252.
DOI 10.1075/babel.60.2.06fay ISSN 0521–9744 E-ISSN 1569–9668

y "monitor"). En segundo lugar, según la relevancia del aspecto temporal ("sincrónico" o "diacrónico"). En tercer lugar, según el grado de especialización ("general" o "terminológico"). En cuarto lugar, según el número de lenguas ("monolingüe", "bilingüe" y "multilingüe"). Y en quinto y último lugar, atendiendo al medio en el que se desarrollan ("escrito", "oral" o "mixto").

En una segunda etapa (1997: 292), se centra en el cuarto grupo (es decir, según el número de lenguas), y concreta aquellos corpus escritos en una sola lengua (monolingües). Sostiene que estos corpus pueden ser a su vez (1) "sencillos" (escritos en un solo idioma, y que pueden ser bien traducciones u originales) o (2) "comparables" (compuestos por dos grupos de corpus "sencillos" — uno de traducciones y otro de originales, que pueden ser dependientes, no dependientes, o independientes). A raíz de este segundo grupo (comparables), nos gustaría referirnos brevemente a la falta de consenso que existe entre Lingüística Contrastiva y los Estudios de Traducción en lo que atañe a los términos "corpus paralelos" y "corpus comparables". Laviosa resume este desbarajuste terminológico del siguiente modo:

> In corpus linguistics the term 'comparable corpus' is generally used to refer to a bi/multilingual corpus made up of two or more sets of texts from the same subject domain(s) [. . .], while the term 'parallel corpus' refers to a corpus of original texts in language A and their translations in language B. However, in translation studies and contrastive linguistics the terminology is not always consistent; some scholars use 'parallel corpus' to cover both types of bilingual corpora (Johansson and Hofland 1994; Hartmann 1994; Gellerstam 1996), while others follow the traditional terminology of contrastive analysis (Aijmer et al. 1996; Granger 1996) and differentiate between a 'translation corpus' (original texts in language A and their translations in language B) and 'parallel corpus' (original texts in language A and B). (Laviosa 1997: 292)

En su trabajo, la autora adopta la postura de la Lingüística Contrastiva para "corpus comparables"; pero no es en este subgrupo en el que profundiza sino que lo hace en los "corpus sencillos", y más exactamente en aquellos compuestos por textos traducidos ("*single*" > "*translational*"). En este nivel, diferencia varios subtipos en función de distintos parámetros: (1) desde el punto de vista de la lengua origen ("con una única lengua origen", "con dos lenguas origen" o "con varias lenguas origen"); (2) según el modo de traducción ("con un único modo de traducción", "con dos modos de traducción" o "con varios modos de traducción"; a su vez todos ellos pueden ser "escritos", "orales" o "interpretaciones"); (3) atendiendo al método de traducción ("con un solo método de traducción", "con dos métodos de traducción" o "con varios métodos de traducción"); (4) según la lengua meta ("a la lengua materna", "a una lengua extranjera", "a la lengua de uso habitual" o "estado mixto de lenguas meta"); (5) estableciendo el punto de mira en el traductor ("traducidos por un traductor profesional" o "traducidos por un estudiante", que se corresponde con lo que Baker denomina "*learner corpus*" en su obra de 2006); por último, (6) en lo que atañe a la publicación ("publicados" o "no publicados").

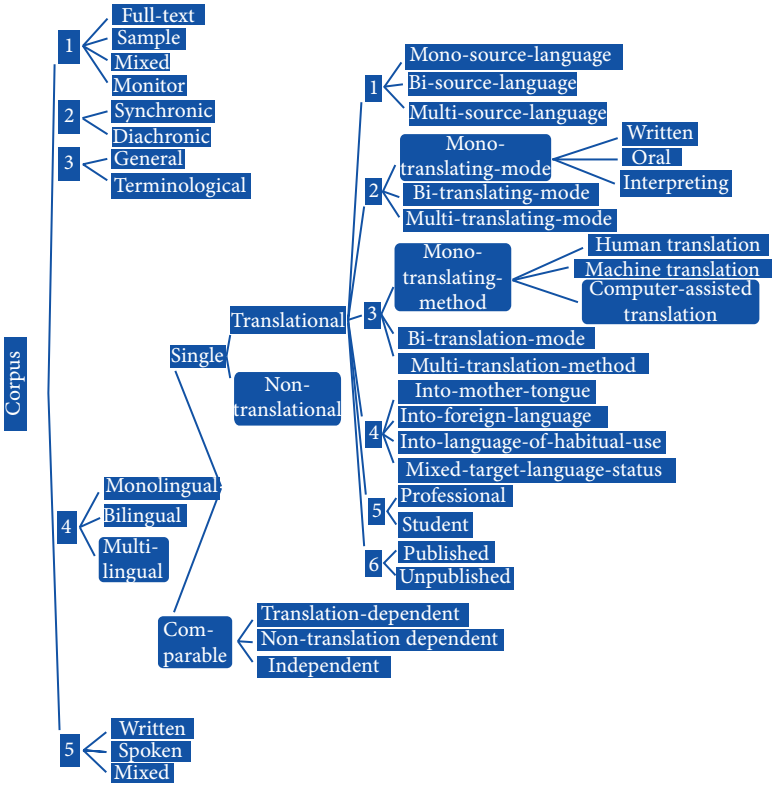

Figure 1.

Para poder observar con facilidad la jerarquía de los distintos tipos de corpus, hemos realizado un esquema de su trabajo (Figura 1).

Debemos mencionar que a pesar de que esta propuesta no abarca todos los tipos de corpus, y es bastante exhaustiva y profundiza mucho en el tipo de corpus que analiza. Sin embargo, hay algunos aspectos que creemos que podrían mejorarse. En primer lugar, reconocemos que los criterios básicos que diferencia la autora son apropiados en una clasificación de corpus, pero debemos tener en cuenta que todos ellos están relacionados entre sí y no queda así reflejado en la propuesta de Laviosa. Consideramos por tanto que es mejor mostrar su jerarquía de otro modo diferente en donde se perciban las relaciones que los vinculan. Por otro lado, en nuestra opinión, sería adecuado clasificar el tipo de corpus "monitor" en otro bloque, ya que no está tan relacionado con la distribución de texto (como estarían los demás tipos del grupo, a saber: "texto completo", "de muestras" y "mixtos") como con la admisión de añadido de información. En tercer lugar, consideramos que la categoría "a la lengua de uso habitual" no debería aparecer en el mismo nivel que "a la lengua materna" o "a una lengua extranjera", sino que debiera considerarse un

subtipo dentro de estos dos grupos, es decir, debería detallarse en una etapa posterior. Por último, no compartimos la división que efectúa la autora entre "traducción humana" y "traducción asistida por ordenador", ya que en esta última también es necesario que intervenga un profesional y este simplemente utiliza las herramientas informáticas para agilizar su trabajo, pero el auténtico creador de la traducción es el propio traductor con su conocimiento e ingenio. Consideramos por tanto, que la traducción humana puede ser "asistida mediante ordenador" o "no asistida". Trataremos de dar respuesta a todos estos aspectos en nuestra propuesta.

A continuación, comentaremos brevemente la clasificación de Torruella y Llisterri (1999), que es más amplia que la anterior porque no se centra únicamente en textos traducidos, sino que tiene un enfoque más amplio. Estos autores establecen cinco grupos diferenciados: (1) según el porcentaje y la distribución de los diferentes textos que integran el corpus; (2) según la especificidad de los documentos contenidos en el corpus; (3) según la cantidad de texto que se recoge en cada uno de los documentos; (4) según la codificación y anotación de los corpus; y (5) según la documentación que acompaña a los textos. Puede verse de forma gráfica en la Figura 2.

Podemos observar que estos autores coinciden con Laviosa en el establecimiento de algún grupo, como el correspondiente a la extensión y distribución del texto, o a la especificidad de los textos. Sin embargo, ambas propuestas se diferencian en el número de subtipos que establecen. Corpas Pastor (2001) se basa en los mismos criterios expuestos por Torruella y Llisterri (1999) y completa su propuesta con las aportaciones anteriores de Baker (1995) y Johansson (1998). Puede verse gráficamente en la Figura 3.

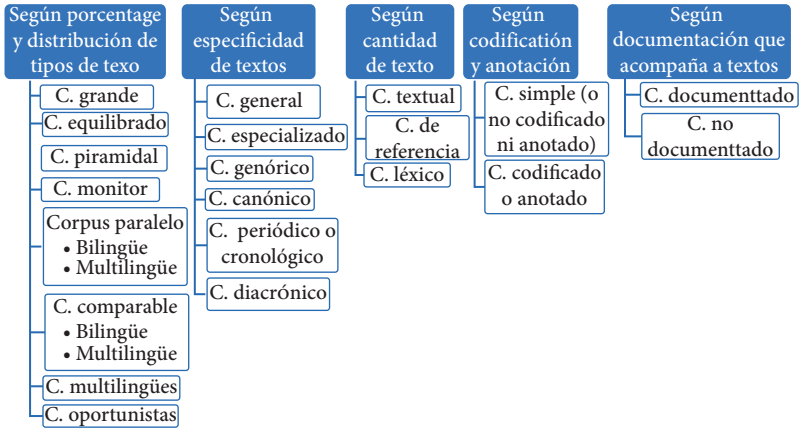

Figure 2.

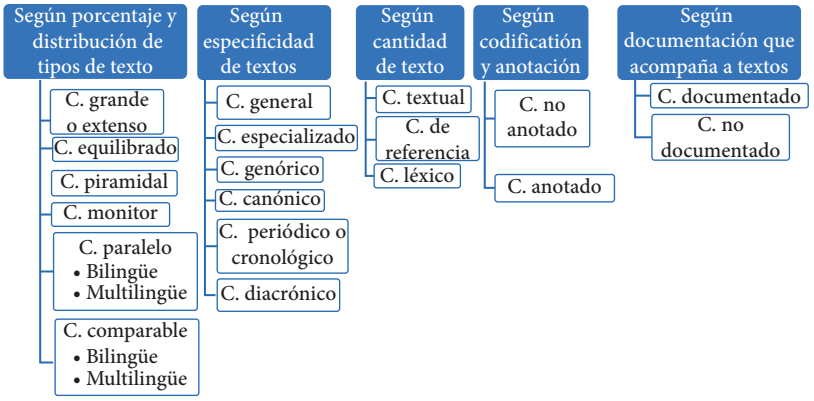

Figure 3.

Si comparamos su propuesta con la de Torruella y Llisterri, podemos observar que desaparecen las categorías de "corpus multilingües" y "corpus oportunistas", que Torruella y Llisterri habían propuesto para abarcar, respectivamente, a los corpus escritos en varias lenguas y a los corpus que se encuentran disponibles sin seguir ningún criterio de selección.

Aunque las dos últimas clasificaciones expuestas (Torruella y Llisterri y Corpas Pastor) son amplias, existen varios aspectos que en nuestra opinión, habría que matizar. En primer lugar, no se tiene en cuenta la importante distinción entre corpus monolingües y multilingües. En segundo lugar, no quedan reflejadas de forma clara las relaciones que existen entre los distintos corpus. En tercer lugar, no nos parece apropiada la clasificación de corpus paralelos y comparables bajo el epígrafe "Porcentaje y distribución de tipos de texto", ya que consideramos que tanto corpus paralelos como comparables más bien atienden a la relación existente entre los idiomas que componen el corpus. En cuarto lugar, el grupo de corpus grandes o extensos nos parece una categoría vacía, ya que se trata de un criterio subjetivo que no aporta información relevante (y que además a nuestro entender, los corpus deben ser amplios para poder se representativos). En quinto lugar, los conceptos de "corpus equilibrado" y "corpus piramidal" son poco claros y, en nuestra opinión, esta no es una distinción necesaria en una tipología. En sexto lugar, el desglose de las categorías de corpus genéricos, canónicos y periódicos (también llamados cronológicos) no nos parece adecuada, ya que perfectamente podrían pertenecer al grupo de textos especializados por centrarse en un determinado aspecto (ya sea, género, autor o periodo de tiempo) y no en la lengua general. En séptimo lugar, la inclusión de "corpus diacrónicos" dentro del bloque correspondiente al grado de especificidad de los textos, ya que sería conveniente que aparecieran claramente contrastados con los "corpus sincrónicos". En octavo lugar, a nuestro entender, y como adelantábamos al referirnos a la tipología de Laviosa, el grupo "monitor" de-

biera ir situado en otro grupo que lo distancie de la distribución de información y lo vincule más a la posibilidad de añadidos y actualización de información. En noveno lugar, los términos "corpus textual" y "corpus de referencia" pueden dar lugar a confusión, y por ello preferimos utilizar los términos propuestos por Laviosa, ya que los consideramos más esclarecedores: "corpus de textos completos" y "corpus de muestras" respectivamente). En décimo lugar, la inclusión de "corpus léxico". En una tipología general, nos parece imprescindible. En decimoprimer lugar, la distinción entre "corpus anotado" y "corpus no anotado", y "corpus documentado" y "no documentado" no la consideramos necesaria en una clasificación general, ya que únicamente se refiere al formato del corpus, y no aporta información relevante (no obstante, si se desea, es una categoría que se podría distinguir en cualquiera de los tipos de corpus). Y por último, en decimosegundo lugar, si se distingue entre "anotados" y "no anotados", creemos que sería necesario efectuar una división previa (por ejemplo, "electrónicos/papel" o "escritos/orales").

Para concluir esta primera parte, observaremos la clasificación de Granger (2003), quien establece una división bipartita de los corpus: (1) "corpus monolingües" y (2) "corpus multilingües". En base a esto, la autora (2003: 21) afirma que todos los corpus monolingües son comparables, mientras que los corpus multilingües pueden ser paralelos o comparables. Esta división de los corpus multilingües en paralelos y comparables es aceptada por varios autores (como Rabadán y Fernández Nistal 2002: 51 u Olohan 2004; aunque las primeras distinguen la categoría "bilingüe" en vez de "multilingüe"). Por un lado, los "corpus multilingües > paralelos", los desglosa en "unidireccionales" y "bidireccionales". Por otro lado, dentro de los "corpus multilingües > comparables", distingue entre "originales" y "traducidos". Por último, en lo que atañe a los monolingües, considera que pueden estar integrados por una "combinación de originales y traducciones" o ser "textos originales" (escritos por profesionales o por estudiantes). En la Figura 4, incluimos su propuesta (2003: 21) de forma gráfica.

En nuestra opinión, hay algunos vacíos que convendría completar. En primer lugar, con el fin de englobar todos los corpus existentes, opinamos que es necesario reflejar la distinción temporal entre "diacrónicos" y "sincrónicos". En segundo lugar, consideramos que, en una clasificación de corpus completa es recomendable dedicar también atención a la temática tratada y, por tanto, reflejar la diferencia que existe entre: "corpus de la lengua general (generalistas)" y "corpus centrados en un campo concreto de la lengua (especializados)". En tercer lugar, consideramos que el uso de sinónimos (como es el caso de "corpus de traducciones" y "corpus paralelos") puede dar lugar a confusión; y por ello, en nuestra opinión es más recomendable utilizar términos unívocos y prescindir del empleo de sinónimos. Nos decantamos por el término "corpus paralelos", ya que "corpus de traducciones" podría confundirse con la colección de corpus comparables formada por tex-

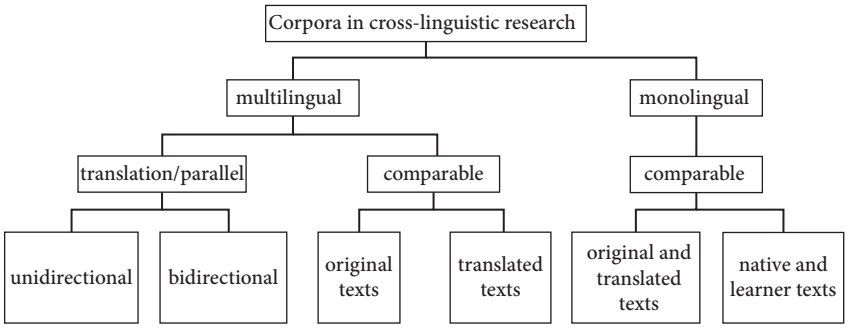

Figure 4.

tos traducidos, que hemos mencionado anteriormente al referirnos a la clasificación de Laviosa 1997. En cuarto lugar, consideramos que sería necesario incluir algunas categorías adicionales. Por un lado, (1) en el tercer nivel, en el grupo de "corpus paralelos", proponemos la categoría de corpus "multidireccionales" (que en nuestra opinión podría sustituir al término "bidireccionales") con el fin de incluir en la clasificación aquellos corpus formados por diversas lenguas en donde unos son traducciones de otros. Por otro lado, pensamos que también es necesario incluir (2) la categoría mixta de "textos originales y traducidos" en el grupo de corpus comparables multilingües, así como (3) la categoría de "únicamente textos traducidos" en el grupo correspondiente a corpus comparables monolingües. En quinto y último lugar, pensamos que en aquellos corpus en los que exista una relación directa entre los textos traducidos y los originales que componen el corpus (es decir, que sean traducciones unos de otros), sería necesario distinguir entre "alineados" y "no alineados".

Nuestra propuesta

Basándonos principalmente en las propuestas de Corpas Pastor (2001) y de Granger (2003), así como la clasificación que llevó a cabo Laviosa (1997) para los corpus comparables y teniendo en cuenta las sugerencias de mejora que hemos indicado, procedemos a elaborar nuestra propia propuesta de clasificación de corpus, que presentamos a continuación. Opinamos que es suficientemente amplia, ya que abarca tanto los intereses de la Lingüística Contrastiva como los de los Estudios de Traducción, e incluso los intereses de otras disciplinas afines que también trabajan con corpus lingüísticos.

Nuestra propuesta de clasificación se dividirá en dos fases principales: (1) la primera atañe a aspectos formales (medio de transmisión, momento temporal,

posibilidad de amplitud, extensión de textos, publicación de los textos y temática tratada), y (2) la segunda se centra en aspectos lingüísticos (originalidad, relación existente entre originales y traducciones, dirección de traducción y alineación). Veamos cada una de ellas con mayor grado de detalle:

En nuestra opinión, la primera división que debemos efectuar es si el corpus contiene textos escritos u orales, es decir cuál es el medio de difusión de los textos que lo integran. A estas dos categorías, añadimos una tercera, que comprenda aquellos corpus que contengan una combinación de textos escritos y orales. Los soportes en los que pueden encontrarse todos estos textos pueden ser extraordinariamente diversos y por ello no los detallaremos aquí con exhaustividad sino que tan solo citaremos algún ejemplo. Por un lado, los pertenecientes al ámbito escrito podrían estar en formato electrónico o papel. Podríamos distinguir a su vez diferentes tipos según el soporte: libros, periódicos, revistas, folletos, etc. Podríamos incluso englobar aquí algunos de los tipos comentados por autores como Laviosa (1977) o P. Baker (2006), como son los "pre-electrónicos" o los "no procesados". Por otro lado, dentro del ámbito oral podemos encontrar más variedades debido principalmente al desarrollo de los soportes multimedia. En este sentido, podemos hacer varias distinciones dependiendo de en qué aspecto situemos nuestro punto de mira. De este modo, por citar algunos ejemplos, podríamos clasificarlos según (1) el soporte en el que se encuentra el texto (CD, cinta de video o audio, reproductor MP3, medio electrónico, etc.), (2) el formato de grabación (.mp3, .mp4, .wma, etc.), (3) el momento de emisión (directo o diferido), (4) el canal de difusión (radio, televisión, rueda de prensa, Internet, videoconferencia, etc.), o incluso (5) si se han producido con preparación previa (conferencias, comunicados informativos, etc.) o si son improvisados (debate, rueda de prensa, etc.). Aunque no profundicemos en este aspecto por alejarnos de la intención de este artículo, debemos ser conscientes de su existencia, en caso de que en un trabajo de investigación determinado resultara necesario detallar esta información. De ser así, este sería el momento adecuado de la clasificación para indicarlo. Para ver un resumen del primer paso, obsérvese la Figura 5.

Asimismo, tal y como señalaba Laviosa (1997), en las primeras fases de la clasificación también debemos reflejar si el corpus está ligado a una época concreta

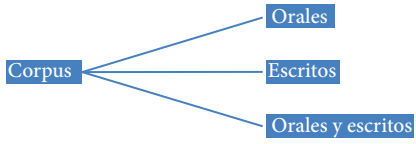

Figure 5.

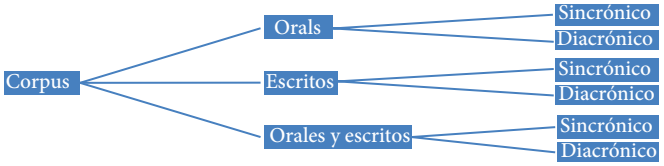

Figure 6.

(es decir, si es sincrónico) o si trata de ser representativo de una lengua (esto es, si es diacrónico[1]) (Figura 6).

Si un investigador estimara necesario indicar datos, como si es informatizado o electrónico, consideramos que debiera indicarlo a continuación. Sin embargo, no lo reflejamos en nuestro esquema por no considerarlo relevante en una clasificación general.

Posteriormente, debemos referirnos a las posibilidades de actualización y aumento de información de un corpus. Por ello, introduciremos el concepto de "monitorizados". Nótese que preferimos el término "monitorizado" a "monitor", y resaltamos igualmente que para nosotros no es imprescindible que mantenga un volumen textual constante, como sostenía Corpas (2001: 158), sino que lo realmente importante es que permita la actualización de información e incluso el añadido de datos. De este modo, en nuestra opinión, los corpus diacrónicos pueden dividirse en "monitorizados" o "no monitorizados", ya que al no ceñirse a un periodo restringido, podrán admitir la incorporación constante de nuevos corpus. Mientras que los "corpus sincrónicos", siempre serán "no monitorizados", ya que se crean para un proyecto concreto que se refiere a una determinada época con principio y fin. A continuación, reflejamos esto de forma gráfica. Téngase en cuenta que en este esquema y en los siguientes, ya no desglosaremos "corpus orales", "corpus escritos" y "corpus orales y escritos", ya que tienen las mismas divisiones, y ocuparía demasiado espacio desglosarlo por triplicado. En su lugar, nos referiremos a ellos de forma simplificada como "corpus orales/escritos" (Figura 7).

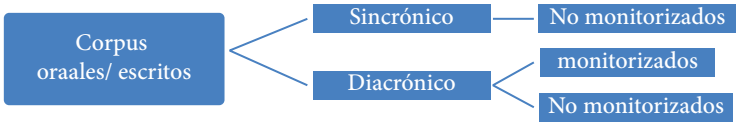

Figure 7.

1. En este trabajo, consideraremos sinónimos los términos "corpus de referencia" y "corpus diacrónico", ya que entendemos que si se refiere a un periodo completo es porque trata de ser representativo de la lengua en su conjunto. Para evitar confusiones, preferimos no utilizar sinónimos, y por ello nos decantamos por "corpus diacrónico" frente a "corpus sincrónico".

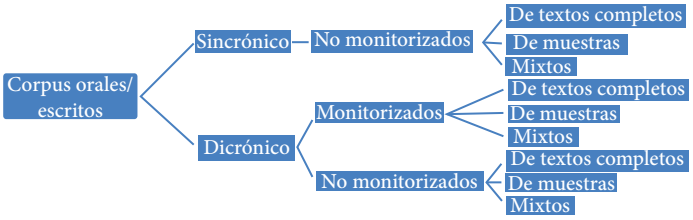

Figure 8.

Consideramos que en el siguiente paso es necesario especificar la extensión de los documentos que componen el corpus, y siguiendo a Laviosa (1997), podemos distinguir: (1) corpus de textos completos (que se corresponde con lo que Torruella y Llisterri y Corpas Pastor denominaban "corpus textual" en sus clasificaciones); (2) corpus de muestras ("corpus de referencia" según la terminología de Torruella y Llisterri y Corpas Pastor); y (3) corpus mixtos, en caso de que aparezcan combinados los dos anteriores. Como vemos, hemos eliminado de este grupo, los "corpus monitorizados", porque preferimos situarlos en otro nivel. Asimismo, podemos observar que no hemos incluido una categoría independiente para "corpus léxicos", que sí distinguían Torruella y Llisterri y Corpas Pastor (2001) y que entendían como aquellos formados por fragmentos muy pequeños de igual longitud, ya que en nuestra opinión este tipo de corpus podría incluirse dentro de los "corpus de muestras". Asimismo, aunque reconocemos su existencia, lo cierto es tan para aparecer en una clasificación general de corpus, y por ello, no lo reflejamos en el esquema, pero dejamos su uso a discreción de cada investigador en función del trabajo que lleve a cabo (Figura 8).

Cabe señalar que todos los corpus (pero especialmente los "de muestras" y los "mixtos") también podrían clasificarse atendiendo a los porcentajes de los textos que los compongan. Siguiendo a Torruella y Llisterri (1999: 10) y Corpas Pastor (2001: 158), destacamos los términos corpus equilibrados y corpus piramidales. Los primeros son aquellos que contienen diversas variedades de lengua en porcentajes similares; y los segundos son corpus que contienen textos distribuidos por niveles, los cuales se caracterizan por aumentar progresivamente la complejidad de las variedades temáticas, en detrimento del número de textos incluidos en cada variedad. Sin embargo, tal y como hemos mencionado anteriormente al referirnos a estos autores, no incluiremos este subgrupo en nuestra propuesta ya que no consideramos que se trate de una distinción relevante en una clasificación general, y creemos que la complicaría demasiado. No obstante, dejamos la puerta abierta a cualquier investigador que crea conveniente detallar este aspecto en su trabajo particular; en nuestra opinión, este sería el momento adecuado para señalarlo. En caso de que el investigador así lo decida, nos gustaría sugerir que en vez de detallar

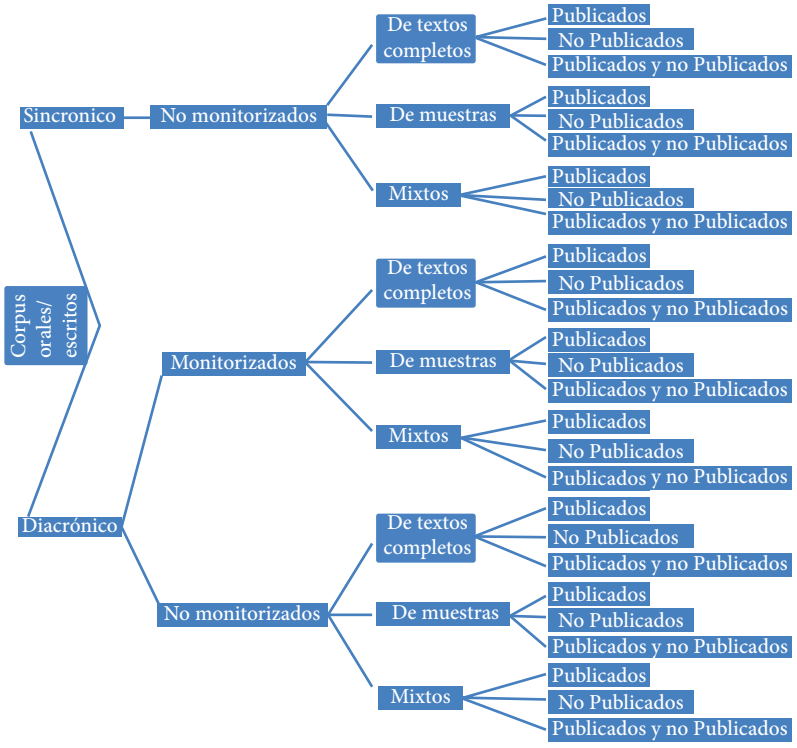

Figure 9.

porcentajes, quizás fuera conveniente efectuar una división más sencilla y general, como por ejemplo: "corpus equilibrados" (si tienen una extensión similar) y "corpus no equilibrados" (si no la tienen).

Ahora bien, retomando la explicación de nuestra propuesta de clasificación, siguiente paso atañe a la publicación de los textos, es decir, si el corpus consta de textos publicados, no publicados o una combinación de ambos (Figura 9).

En el siguiente estadio de la tipología, el último de la primera fase, prestaremos atención al grado de especialización de los documentos, es decir, determinaremos si pertenecen a la lengua general (corpus generalistas) o a la lengua especializada (corpus especializados). Entendemos por corpus generalistas aquellos que estudian la lengua en su conjunto, mientras que los especializados son aquellos que se centran en una parte concreta de la lengua (ya sea temática, campo del saber, autores, géneros, objetivos de cada proyecto, etc.). Dado el elevado número de alternativas que podrían existir para los corpus "especializados", no consideramos útil realizar una subdivisión estanca de los diversos tipos de corpus especializados, pero podemos citar algún ejemplo: corpus médicos, jurídicos, económicos, etc. (si

se atiende a la temática), genéricos (un determinado género), canónicos (un determinado autor), periódicos o cronológicos (según periodo de tiempo, y siguiendo la terminología de Corpas Pastor); dialectales (atendiendo al dialecto), históricos (en base al momento histórico al que hagan referencia), nacionales/regionales (atendiendo al ámbito), no estándares (si los datos que engloba no son estándares,

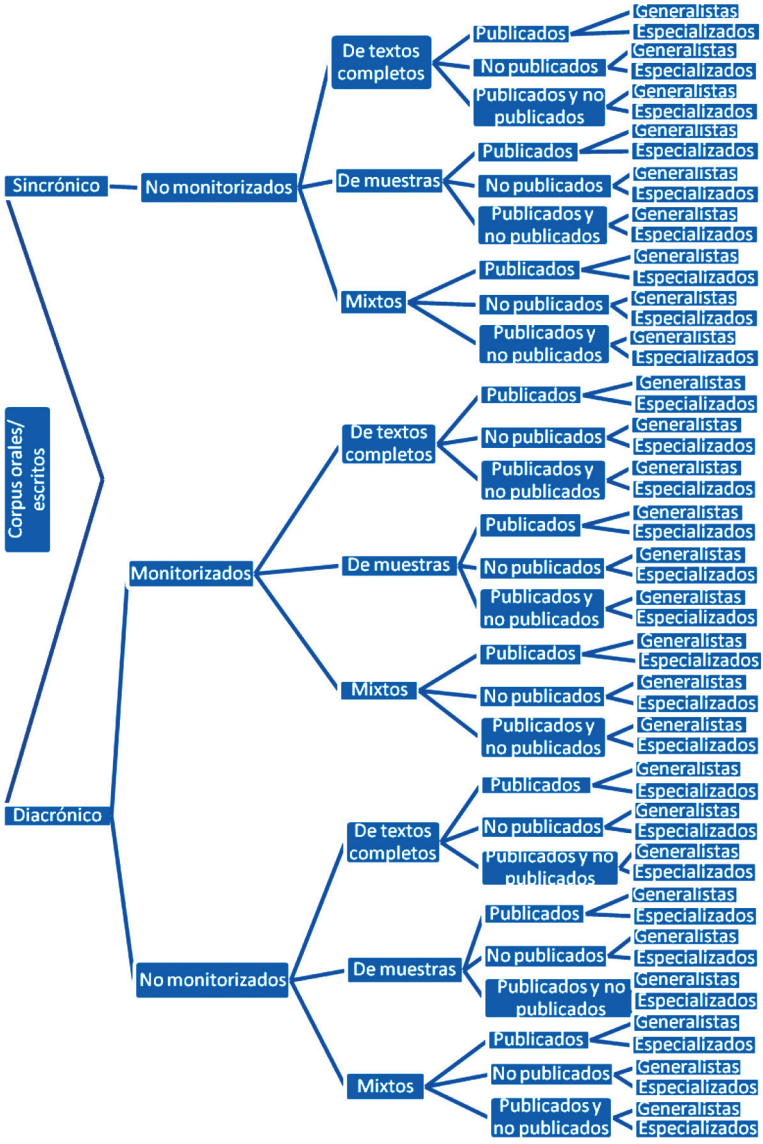

Figure 10.

y en relación con "corpus dialectal" y "corpus regional"), pedagógicos/formación (si tienen función didáctica), de segunda generación (creados con posterioridad a 1990, y también llamados "megacorpus" por su gran tamaño), entre otros, siguiendo la terminología de autores como Laviosa (1997) o P. Baker (2006). Teniendo esto presente, nuestra primera fase de clasificación queda como se indica en la Figura 10.

Ahora bien, tras haber especificado las características formales generales de los corpus en lo que atañe principalmente a medio de transmisión, referencia temporal, posibilidad de añadidos, extensión de los textos integrantes, estado de publicación y grado de especialización, procedemos a adentrarnos en la segunda fase de la clasificación, en donde analizaremos desde una vertiente más lingüística los textos que integran los corpus.

De este modo, la primera división que efectuaremos atañerá al número de lenguas que intervienen en el corpus, y consecuentemente distinguiremos entre (1) monolingües y (2) multilingües (Figura 11).

Tal y como hemos señalado al referirnos a la propuesta de Granger (2003), todos los corpus monolingües son comparables. Si por el contrario, nos centramos en los multilingües, adoptamos la misma postura que la autora y diferenciaremos entre (1) comparables y (2) paralelos. Como indicábamos anteriormente, ambos están formados por textos originales y traducciones, la única diferencia entre ellos radica en la relación que existe entre los textos que lo integran (originales y traduc-

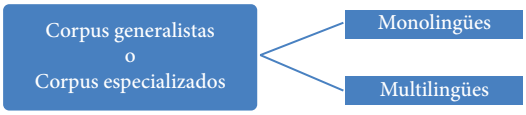

Figure 11.

ciones): los primeros (comparables) los textos son independientes, es decir, no son traducciones unos de otros, mientras que en los segundos (paralelos), contaremos con el original y su correspondiente traducción (Figura 12).

En otras palabras, podemos encontrar corpus comparables tanto en corpus monolingües como multilingües, y además pueden estar compuestos de diferentes conjuntos (1) textos originales en una misma lengua o en varias lenguas; (2) textos

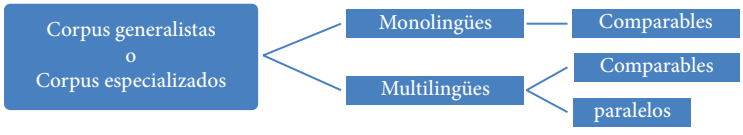

Figure 12.

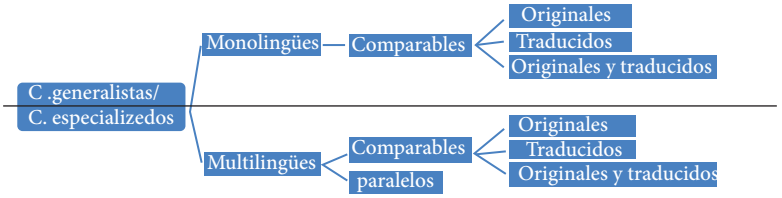

Figure 13.

traducidos (sin ser traducción de los primeros) en varias lenguas; y (3) una combinación de textos originales y traducidos (sin ser estos traducción de los primeros) (Figura 13).

Si dejamos a un lado los corpus comparables y atendemos a la relación existente entre los textos que integran los "corpus paralelos", no especificaremos si son originales o traducciones, ya que el término "paralelo" ya nos implica que está compuesto por "originales y sus correspondientes traducciones". De este modo, podremos distinguir entre "unidireccional" o "multidireccional" (en donde estarían incluidos los "bidireccionales", como indicábamos anteriormente, categoría diferenciada por algunos autores). Asimismo sostenemos que tanto originales como traducciones podrán estar alineados o no alineados (Figura 14).

En el siguiente paso de la clasificación, que los corpus de textos traducidos, tanto de corpus monolingües como multilingües, y ya sean "Comparables > Traducidos" o "Comparables > Originales y traducidos" (aunque en nuestro esquema solo lo indicaremos en los primeros para simplificar), podrían completase con la clasificación que proponía Laviosa (1997) y a la que nos hemos referido anteriormente, a saber:

En primer lugar, desde el punto de vista de la lengua origen, pueden ser "corpus con una única lengua origen" o "corpus con varias lenguas origen" (téngase en cuenta que no desglosamos la categoría de "corpus con dos lenguas origen", que distinguía Laviosa, ya que la englobamos en nuestro último grupo mencionado).

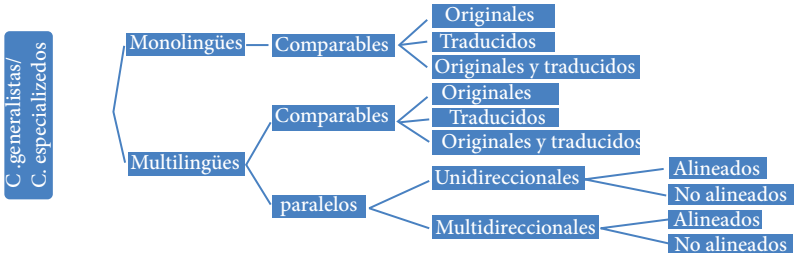

Figure 14.

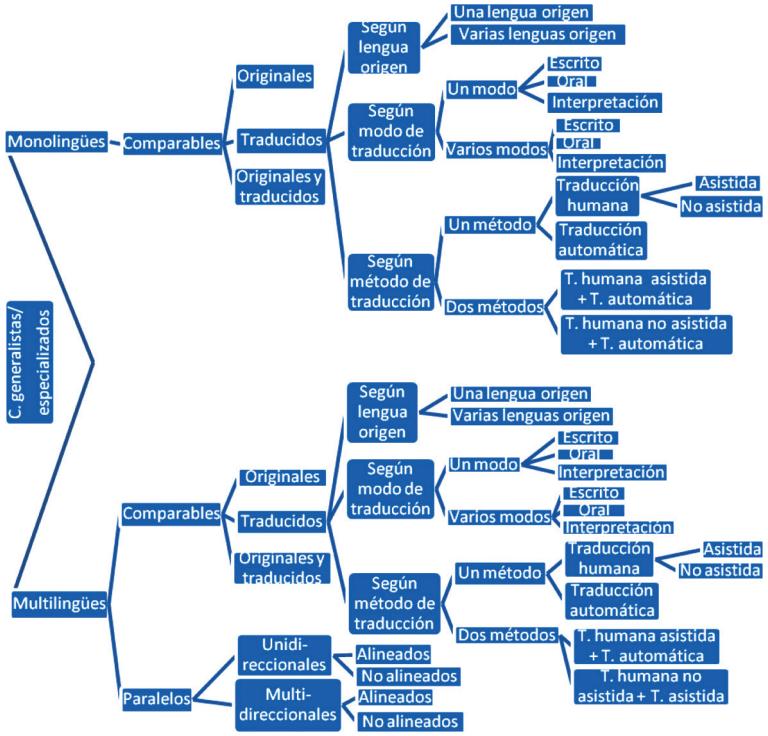

Figure 15.

En segundo lugar, según el modo de traducción (escrito, oral o interpretación), podemos distinguir "corpus con un único modo de traducción" o "corpus con varios modos de traducción" (nótese que no especificamos la categoría de "corpus con dos modos de traducción", que distinguía Laviosa, ya que la incluimos en este último grupo).

En tercer lugar, dependiendo del método de traducción, podrán ser "corpus con un solo método de traducción" o "corpus varios métodos de traducción" (no detallamos la categoría de "corpus con dos métodos de traducción", que distinguía Laviosa porque la englobamos en este último). Los métodos a los que nos estamos refiriendo son: traducción humana (que puede ser asistida o no asistida por ordenador) y traducción automática. Véase el esquema actualizado en la Figura 15.

Aquí finaliza nuestra clasificación. Podemos observar que hemos dejado a un lado las divisiones que atañen a la dirección de la lengua del traductor (a la lengua materna, lengua extranjera o estado mixto de lenguas meta) y a la categoría del traductor (estudiante o profesional, que distinguía Laviosa 1997), ya que no consideramos que sean aspectos relevantes en una clasificación general. No obstante, si un investigador se viera en la necesidad de especificar estos aspectos en su trabajo

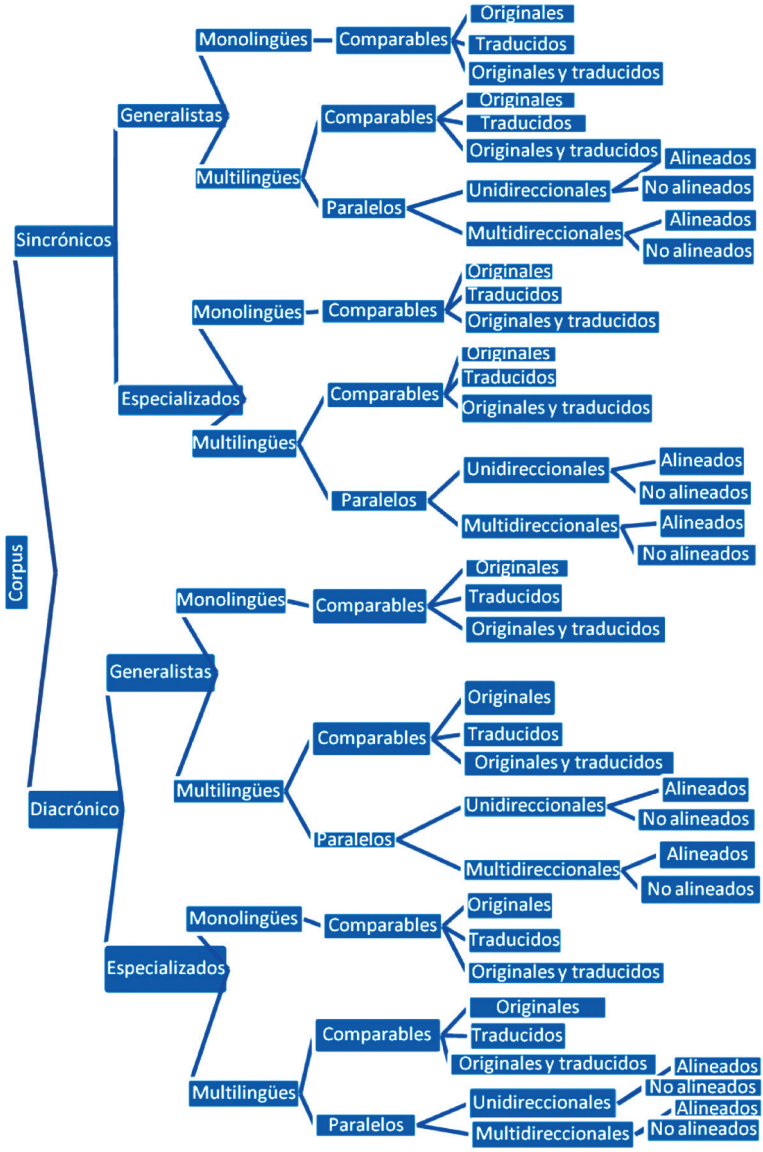

Figure 16.

debido a los objetivos particulares de su investigación, en nuestra opinión, este sería el lugar apropiado.

Asimismo, si en un proyecto concreto fuera necesario especificar otros datos que no se incluyan en la presente clasificación (por ejemplo, si el corpus es anotado o documentado), recomendamos hacerlo en último lugar.

A continuación, proporcionamos un esquema resumen de nuestra propuesta

de clasificación completa (incluyendo las dos fases), pero en donde únicamente resaltamos los estadios que consideramos de mayor relevancia, ya que de lo contrario, el esquema resultante sería demasiado extenso (Figura 16).

Conclusión

Cada tipo de corpus ofrece una información diferente, y por ello, la selección de uno u otro dependerá de los objetivos que se persigan en cada proyecto concreto, a saber: estudiar una clase de texto, género o tipo textual determinados, conocer las características de un campo del saber en concreto, observar textos de una época o lugar específicos, detectar diversas estrategias traslativas, etc.

Por este motivo, es imprescindible contar con una clasificación amplia e integradora de los corpus, ya que nos permitirá saber con precisión cuáles son los diferentes tipos de corpus que existen así como los rasgos principales de cada uno y las relaciones existentes entre ellos.

Consideramos que nuestra propuesta de clasificación (1) es más extensa y completa que las existentes, (2) tiene en cuenta todos los criterios que señalaba M. Baker (1995: 229); (3) es integradora porque combina varios enfoques, en concreto, atiende a las clasificaciones propuestas tanto desde la Lingüística Contrastiva como desde los Estudios de Traducción, y (4) es abierta ya que permite la inclusión de nuevos parámetros y etiquetas (por ejemplo, canónico, monitor, electrónico, etc.).

Referencias

Aijmer, Karin, Bengt Altenberg, y Mats Johansson. 1996. *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies.* Lund: Lund University Press.

Baker, Mona. 1995. "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research." *Target* 7 (2): 223–243. DOI: 10.1075/target.7.2.03bak

Baker, Paul, Andrew Hardie, y Tony McEnery. 2006. *A Glossary of Corpus Linguistics*. Edimburgo: Edinburgh University Press.

Blecua, José Manuel, Gloria Clavería, Carlos Sánchez, y Joan Torruella. 1999. *Filología e Informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Editorial Milenio.

Fries, Udo, Gunnel Tottie, y Peter Schneider. 1994. *Creating and Using English Language Corpora*. Ámsterdam: Editions Rodopi B. V.

Gellerstarm, Martin. 1996. "Translations as a Source for Cross-Linguistic Studies." En *Languages in Contrast. Papers from a Symposium on Text-based Cross-linguistic Studies*, de Karin Aijmer, Bengt Altenberg, y Mats Johansson, 53–62. Lund: Lund University Press.

Granger, Sylviane. 1996. "From CA to CIA and Back: An Integrated Approach to Computerized Bilingual and Learner Corpora." En *Languages in Contrast. Papers from a Symposium*

on Text-based Cross-linguistic Studies, de Karin Aijmer, Bengt Altenberg, y Mats Johansson, 37–51. Lund: Lund University Press.

Granger, Sylviane. 2003. "The Corpus Approach: A Common Way Forward for Contrastive Linguistics and Translation Studies?" En *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, de Sylviane Granger, Jacques Lerot, y Stephanie Petch-Tyson, 17–30. Ámsterdam y Nueva York: Editions Rodopi B.V.

Granger, Sylviane, Jacques Lerot, y Stephanie Petch-Tyson. 2003. *Corpus-based approaches to Contrastive Linguistics and Translation Studies*. Ámsterdam y Nueva York: Editions Rodopi B.V.

Hartmann, Reinhard R. K. 1994. "The Use of Parallel Text Corpora in the Generation of Translation Equivalents for Bilingual Lexicography." En *Proceedings of the Euralex Congress (Vrije Universiteit)*, 291–298. Ámsterdam: VLL.

Johansson, Stig. 2003. "Contrastive Linguistics and Corpora." En *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*, de Sylviane Granger, Jacques Lerot, y Stephanie Petch-Tyson, 31–44. Ámsterdam y Nueva York: Editions Rodopi B. V.

Johansson, Stig. 1998. "On the Role of Corpora in Cross-Linguistic Research." En *Corpora & Cross-Linguistic Research: Theory, Method, & Case Studies*, de Stig Johansson y Signe Oksefjell, 3–24. Ámsterdam y Atlanta: Editions Rodopi B. V.

Johansson, Stig, y Knut Hofland. 1994. "Towards an English-Norwegian Parallel Corpus." En *Creating and Using English Language Corpora*, de Udo Fries, Gunnel Tottie, y Peter Schneider, 25–37. Ámsterdam: Editions Rodopi B. V.

Johansson, Stig, y Signe Oksefjell. 1998. *Corpora & Cross-Linguistic Research: Theory, Method, & Case Studies*. Ámsterdam y Atlanta: Editions Rodopi B. V.

Laviosa-Braithwaite, Sara. 1997. "How Comparable can 'Comparable Corpora' Be?" *Target (John Benjamins Publishing Company)* 9 (2): 289–319.

Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. Oxford: Routledge.

Pastor, Gloria Corpas. 2001. "Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada." *Trans* 5: 155–184.

Rabadán, Rosa, y Purificación Fernández Nistal. 2002. *La traducción inglés-español: fundamentos, herramientas, aplicaciones*. León: Universidad de León.

Torruella, Joan, y Joaquim Llisterri. 1999. "Diseño de corpus textuales y orales." En *Filología e Informática. Nuevas tecnologías en los estudios filológicos*, de José Manuel Blecua, Gloria Clavería, Carlos Sánchez, y Joan Torruella, 45–81. Barcelona: Editorial Milenio.

Abstract

In this article we first focus on some of the most important classification systems of linguistic corpora and discuss both the strengths and weaknesses. We pay close attention to the work of Sara Laviosa (1997), Joan Torruella y Joaquim Llisterri (1999), Gloria Corpas Pastor (2001) and Sylviane Granger (2003), not forgetting the contributions by Mona Baker (1995), Stig Johansson (1998 and 2003), Rosa Rabadán y Purificación Fernández Nistal (2002), Maeve Olohan (2004) and Paul Baker (2006). However, despite great advantage of all previous work, our study identified gaps for which we would respectfully suggest some solutions. Finally, we present our classification system in which we attempt to reflect clearly the hierarchical relationships that exist between different types of corpora. The ultimate goal is to offer a wide, comprehensive

and flexible classification, which can be easily adapted to the needs of each research work and meet the requirements of linguistic corpora analysis, particularly in the field of Applied Linguistics and Translation Studies.

Keywords: linguistic corpora, classification of corpora, types of corpora, hierarchy of corpora, relationships between corpora

Résumé

Dans cet article nous nous concentrons d'abord sur quelques systèmes de classification de corpus linguistiques importants et nous en examinons tant les atouts que les faiblesses. Ainsi, nous analysons principalement les travaux de Sara Lavitoutsosa (1997), Joan Torruella et Joaquim Llisterri (1999), Gloria Corpas Pastor (2001) et Sylviane Granger (2003), sans vouloir oublier les réflexions de Mona Baker (1995), Stig Johansson (1998 et 2003), Rosa Rabadán et Purificación Fernández Nistal (2002), Maeve Olohan (2004) et Paul Baker (2006).

Tout en tirant profit de ces travaux, nous nous permettons cependant de soumettre des propositions visant à combler des lacunes que nous avons constatées lors de notre étude des différents systèmes. Finalement, nous présentons notre propre classification, dans laquelle nous tentons d'accentuer les relations hiérarchiques qui existent entre deux types de corpus. L'objectif final est d'offrir un système de classification approfondi, complet et flexible qui puisse s'adapter aux nécessités de tout type de recherche et répondre aux exigences d'analyses de corpus linguistiques, en particulier dans les domaines de la Linguistique Appliquée et des Études de Traduction.

Mots-clés: corpus linguistiques, classement de corpus, types de corpus, hiérarchie de corpus, relations entre corpus

About the author

Goretti Faya Ornia graduated in Translation and Interpretation (English–Spanish & German–Spanish) from the University of Valladolid-Soria. She completed also a master degree in Medical Translation (English–Spanish, University of Jaume I) and in Medical–Technical Translation (German–Spanish, University of Córdoba). She is currently a lecturer at the University of Oviedo.
Address: Faculty of Philosophy and Letters, Department of English, German and French Philology, University of Oviedo, Teniente Alfonso Martínez, Oviedo 33011, Principality of Asturias, Spain.
E-mail: fayagoretti@uniovi.es; gorettifayaornia@gmail.com