

Tipología de corpus: hacia una propuesta integradora

(Corpus typology: towards an integrative proposal)

VALERIA HERNÁNDEZ GARCÍA

<https://orcid.org/0000-0003-3453-6674>

vhergar@upo.es

Universidad Pablo de Olavide <https://ror.org/02z749649>

Fecha de recepción: 30 de agosto de 2023

Fecha de aceptación: 14 de noviembre de 2023

Resumen: La metodología lingüística basada en el trabajo con corpus se implementa en disciplinas tan dispares como la traducción (Baker 1993), el análisis del discurso, la dialectología y la variación, la enseñanza de idiomas y la lingüística, la cultura, la gramática, la estilística lingüística, la lexicografía, la lingüística histórica, la pragmática, la psicolingüística, la psicología social, la semántica, la lingüística textual, y la sociolingüística (McEnery y Wilson 2001: 104-130). Debido al amplio espectro de utilización de la lingüística de corpus, no se ha alcanzado un acuerdo absoluto en cuanto a los tipos de corpus existentes, su denominación y su definición. Por consiguiente, el objetivo del presente artículo es ofrecer una propuesta de clasificación de corpus aplicable a cualquier disciplina. Para tal propósito, de la multitud de categorizaciones encontradas, nos fundamentamos en aquellas formuladas por Sinclair (1996), Torruella Casañas y Llisterri Boix (1999), Laviosa (2002), Fernandes (2006), Borja Albi (2007), Cabré Castellví (2007), Villayandre Llamazares (2008) y Ornia (2015). De manera que delimitamos un total de dieciocho parámetros, y cuarenta y dos tipos de corpus en nuestra propuesta. En definitiva, este trabajo proporciona un instrumento que estandariza el proceso de selección de corpus para llevar a cabo estudios lingüísticos y la enseñanza/el aprendizaje de idiomas.

Palabras clave: Corpus. Tipología. Revisión. Propuesta. Interdisciplinariedad.

Abstract: Corpus-based linguistic approach is implemented in disciplines as dissimilar as translation (Baker 1993), discourse analysis, dialectology and variation, teaching of languages and linguistics, culture, grammar, stylistics, lexicography, historical linguistics, pragmatics, psycholinguistics, social psychology, semantics, text linguistics, and sociolinguistics (McEnery & Wilson

2001: 104-130). Due to the wide range of uses of corpus linguistics, an absolute agreement has been reached with regard to neither the existing types of corpus, nor their denomination and definition. Therefore, the aim of this article was to offer a proposal for corpus classification applicable to any discipline. To this end, amongst the multitude of categorisations found, those formulated by Sinclair (1996), Torruella Casañas and Llisterri Boix (1999), Laviosa (2002), Fernandes (2006), Borja Albi (2007), Cabré Castellví (2007), Villayandre Llamazares (2008) and Ornia (2015) were taken as a basis. In this way, eighteen parameters and forty-two types of corpus were delimited in our proposal. In conclusion, the present work provides a tool that standardises the corpus selection process so as to conduct linguistic studies and the teaching/learning of languages.

Keywords: Corpus. Typology. Review. Proposal. Interdisciplinarity.

1. Introducción

Un corpus consiste en una colección de materiales en formato electrónico, los cuales comparten rasgos definitorios al ser estos seleccionados y ordenados de acuerdo con unos criterios específicos comunes. De manera que dichos materiales representan el estadio o variedad de lengua objeto de estudio.

El auge de tales colecciones ha desembocado en la institución de una metodología con un estatus teórico propio, es decir, la lingüística de corpus. Esta incorpora métodos y principios atinentes a la implementación de corpus en estudios lingüísticos y en procesos de enseñanza/aprendizaje de idiomas (McEnery, Xiao y Tono 2006: 7).

La mencionada metodología se ejecuta en disciplinas tan dispares como la traducción (Baker 1993), el análisis del discurso, la dialectología y la variación, la enseñanza de idiomas y la lingüística, la cultura, la gramática, la estilística lingüística, la lexicografía, la lingüística histórica, la pragmática, la psicolingüística, la psicología social, la semántica, la lingüística textual, y la sociolingüística (McEnery y Wilson 2001: 104-130). Debido al amplio espectro de utilización de la lingüística de corpus, no se ha alcanzado un consenso con respecto a los tipos de corpus existentes, su denominación y su definición.

Así pues, el objetivo del presente artículo no es otro que el de ofrecer una propuesta de clasificación de corpus aplicable a cualquier disciplina.

2. Revisión de las clasificaciones de corpus existentes

Son numerosos los trabajos que han procurado poner orden entre la amplia gama de corpus teorizados¹. A continuación, nos limitaremos a revisar las tipologías de corpus sobre las que se cimienta nuestra propuesta de clasificación, la cual se expone en el apartado que sucede.

Los tipos de corpus se organizan esencialmente en torno a determinados criterios. Por su parte, Sinclair (1996) precisa cinco: 1) cantidad: corpus (a) grande, (b) monitor o (c) de referencia; 2) calidad: corpus (a) auténtico o (b) experimental; 3) simplicidad: corpus (a) de texto plano o (b) anotado; 4) documentación: corpus (a) documentado y (b) no documentado; y 5) muestras: corpus de (a) textos o (b) muestras. Asimismo, dicho autor (Sinclair 1996), aunque sin crear otro parámetro independiente, discierne entre corpus comparable y paralelo.

Torruella Casañas y Llisteri Boix (1999) establecen también otros cinco indicadores: 1) distribución y porcentaje de los distintos tipos de textos: corpus (a) grande, (b) equilibrado, (c) piramidal, (d) monitor, (e) paralelo, (f) comparable, (g) multilingüe o (h) oportunista; 2) especificidad de los documentos: corpus (a) general, (b) especializado, (c) genérico, (d) canónico, (e) periódico o cronológico, o (f) diacrónico; 3) cantidad de texto: corpus (a) textual, (b) de referencia o (c) léxico; 4) codificación y anotación: corpus (a) simple o no codificado ni anotado, o (b) codificado o anotado; y 5) documentación: corpus (a) documentado o (b) no documentado.

Fernandes (2006) señala siete parámetros: 1) tipo de relación entre los textos: corpus (a) comparable, (b) comparable multilingüe lingüístico o (c) paralelo; 2) área de estudio: corpus (a) lingüístico o (b) traslativo; 3) campo de investigación lingüística: corpus (a) general o (b) especializado; 4) modo: corpus (a) escrito, (b) oral, o (c) escrito y oral; 5) restricción temporal: corpus (a) diacrónico o (b) sincrónico; 6) número de lenguas: corpus (a) monolingüe, (b) bilingüe o (c) multilingüe; y 7) direccionalidad: corpus (a) unidireccional, (b) bidireccional o (c) multidireccional.

¹ Véase, *inter alia*, Stig Johansson, "On the role of corpora in cross-linguistic research". En: Johansson, Stig; Oksefjell, Signe (eds.), *Corpora and cross-linguistic research: theory, method, and case studies*. Ámsterdam: Rodopi, 1998, pp. 3-24; Gloria Corpas Pastor, "Compilación de un corpus 'ad hoc' para la enseñanza de la traducción inversa especializada". En: *TRANS: Revista de Traductología*, 5, 2001, pp. 155-184; Sylviane Granger, "The corpus approach: a common way forward for contrastive linguistics and translation studies?". En: Granger, Sylviane; Lerot, Jacques; Petch-Tyson, Stephanie (eds.), *Corpus-based approaches to contrastive linguistics and translation studies*. Ámsterdam y Nueva York: Rodopi, 2003, pp. 17-30.

Villayandre Llamazares (2008), al igual que Fernandes (2006), identifica siete criterios: 1) modalidad de lengua: corpus (a) escrito, (b) oral o (c) mixto; 2) número de lenguas: corpus (a) monolingüe, (b) bilingüe, (c) multilingüe, (d) comparable o (e) paralelo; 3) límites del corpus: corpus (a) cerrado, o (b) abierto o monitor; 4) especificidad de los textos: corpus (a) general o (b) especializado; 5) período temporal que comprenden los documentos: corpus (a) diacrónico o histórico, o (b) sincrónico; 6) tamaño de los textos: corpus (a) de referencia o (b) textual; y 7) tratamiento: corpus (a) simple, no anotado o no codificado, o (b) codificado o anotado.

Ornia (2015) determina once parámetros, a saber: 1) medio de difusión: corpus (a) oral, (b) escrito, o (c) oral y escrito; 2) grado de amplitud del análisis: corpus de (a) referencia o (b) no referencia; 3) momento temporal: corpus (a) sincrónico o (b) diacrónico; 4) posibilidad de actualización: corpus (a) monitorizado o (b) no monitorizado; 5) extensión de los textos: corpus (a) de textos completos, (b) de muestras o (c) mixtos; 6) publicación: corpus de textos (a) publicados, (b) no publicados, o (c) publicados y no publicados; 7) grado de especialización de los documentos: corpus (a) generalista o (b) especializado — (b.1) genérico, (b.2) canónico, (b.3) cronológico, (b.4) dialectal, (b.5) histórico, (b.6) nacional, (b.7) regional o (b.8) jurídico, entre otros—; 8) número de lenguas: corpus (a) monolingüe o (b) multilingüe; 9) composición de originales y traducción, y relación entre ellos: corpus (a) comparable —conformado a partir de (a.1) textos originales en una o varias lenguas, (a.2) archivos traducidos en diferentes lenguas, o (a.3) documentos originales y otros textos traducidos— o (b) paralelo —constituido por textos originales y sus respectivas traducciones—; 10) dirección de la traducción: corpus (a) unidireccional o (b) multidireccional; y 11) alineación: corpus (a) alineado o (b) no alineado.

Cabré Castellví (2007) da un paso más y, para organizar concretamente los textos de especialidad de un corpus, define seis indicadores: 1) tema: corpus (a) unidisciplinar o (b) pluridisciplinar; 2) nivel de especialización: corpus (a) de un nivel o (b) de distintos niveles de especialidad; 3) canal de transmisión: corpus (a) oral, (b) escrito, (c) audiovisual, o (d) oral, escrito y audiovisual; 4) género textual: corpus (a) de un género o (b) de distintos géneros textuales; 5) estrategia discursiva: corpus (a) heterogéneo o (b) homogéneo; y 6) lenguas: corpus (a) monolingüe, (b) bilingüe, (c) trilingüe o (d) paralelo.

Por su parte, Borja Albi (2007) cataloga los corpus atendiendo a dicotomías como: corpus (a) impreso o (b) electrónico, corpus (a) *online* o (b) *offline*, corpus (a) oral o (b) escrito, corpus de (a) textos completos o (b) fragmentos de textos, corpus (a) estático o (b) dinámico, corpus (a) sincrónico o (b) diacrónico, corpus

(a) grande o (b) pequeño, corpus (a) general o (b) especializado, corpus de fuentes (a) literarias o (b) no literarias, corpus (a) representativo o (b) no representativo, corpus (a) organizado en torno a un género o (b) no organizado en torno a un género, corpus (a) monolingüe o (b) multilingüe, corpus (a) paralelo multilingüe o (b) comparable multilingüe, corpus paralelo multilingüe (a) alineado o (b) no alineado, corpus (a) simple o (b) anotado, corpus para (a) uso profesional o (b) fines investigadores, corpus para fines investigadores (a) generales o (b) específicos, y corpus para uso profesional en traducción (a) general o (b) especializada.

Por añadidura, para terminar, otros teóricos se apoyan en niveles jerárquicos. Tal es el caso de Laviosa (2002), cuya clasificación de corpus alberga cuatro. El primer nivel consta de seis parámetros: 1) cantidad de texto: corpus (a) de textos completos, (b) de muestras, (c) mixto o (d) monitor; 2) período: corpus (a) sincrónico o (b) diacrónico; 3) tipo: corpus (a) general o (b) terminológico; 4) número de lenguas: corpus (a) monolingüe, (b) bilingüe o (c) multilingüe; 5) lenguas: corpus de, por ejemplo, (a) español o (b) inglés; y 6) modo: corpus (a) escrito, (b) hablado o (c) mixto. Los tres niveles restantes giran alrededor del criterio 4) *número de lenguas*. Con respecto al segundo nivel, Laviosa (2002) diferencia entre corpus (a) monolingüe (a.1) simple o (a.2) comparable, (b) bilingüe (b.1) paralelo o (b.2) comparable, o (c) multilingüe (c.1) paralelo o (c.2) comparable; con relación al tercer nivel, entre corpus (a) simple (a.1) traslacional o (a.2) no traslacional, (b) paralelo (b.1) monodireccional o (b.2) bidireccional, o (c) multilingüe (c.1) mono-lengua origen, (c.2) bi-lengua origen o (c.3) multi-lengua origen; y, en cuanto al cuarto nivel, entre corpus traslacional (a) mono-lengua origen, (b) bi-lengua origen o (c) multi-lengua origen.

A continuación, a título ilustrativo, se plasman las categorizaciones de corpus enunciadas en una tabla:

Clasificaciones de corpus existentes			
Autor		Parámetro	Tipos de corpus
Por parámetros	Sinclair (1996) (5 parámetros)	1) cantidad	3 (a) grande
			(b) monitor
			(c) de referencia
		2) calidad	2 (a) auténtico
			(b) experimental
		3) simplicidad	2 (a) de texto plano
(b) anotado			

Por parámetros	Sinclair (1996) (5 parámetros)	4) documentación	2	(a) documentado	
				(b) no documentado	
		5) muestras	2	(a) de textos	
			(b) de muestras		
	Aunque sin un parámetro delimitado, asimismo se distingue entre corpus (a) comparable y (b) paralelo.				
	Torruella Casañas y Llisterri Boix (1999) (5 parámetros)	1) distribución y porcentaje de los distintos tipos de textos		8	(a) grande
					(b) equilibrado
					(c) piramidal
					(d) monitor
					(e) paralelo
					(f) comparable
					(g) multilingüe
					(h) oportunista
		2) especificidad de los documentos		6	(a) general
					(b) especializado
					(c) genérico
					(d) canónico
					(e) periódico o cronológico
					(f) diacrónico
		3) cantidad de texto		3	(a) textual
(b) de referencia					
(c) léxico					
4) codificación y anotación		2	(a) simple o no codificado ni anotado		
			(b) codificado o anotado		
5) documentación		2	(a) documentado		
			(b) no documentado		
Fernandes (2006) (7 parámetros)	1) tipo de relación entre los textos		3	(a) comparable	
				(b) comparable multilingüe lingüístico	
				(c) paralelo	
	2) área de estudio		2	(a) lingüístico	
(b) traslativo					

Tipología de corpus: hacia una propuesta integradora

Por parámetros	Fernandes (2006) (7 parámetros)	3) campo de investigación lingüística	2	(a) general (b) especializado	
		4) modo	3	(a) escrito (b) oral (c) escrito y oral	
		5) restricción temporal	2	(a) diacrónico (b) sincrónico	
		6) número de lenguas	3	(a) monolingüe (b) bilingüe (c) multilingüe	
		7) direccionalidad	3	(a) unidireccional (b) bidireccional (c) multidireccional	
		Villayandre Llamazares (2008) (7 parámetros)	1) modalidad de lengua	3	(a) escrito (b) oral (c) mixto
			2) número de lenguas	5	(a) monolingüe (b) bilingüe (c) multilingüe (d) comparable (e) paralelo
	3) límites del corpus		2	(a) cerrado (b) abierto o monitor	
	4) especificidad de los textos		2	(a) general (b) especializado	
	5) período temporal que comprenden los documentos		2	(a) diacrónico o histórico (b) sincrónico	
	6) tamaño de los textos		2	(a) de referencia (b) textual	
	7) tratamiento		2	(a) simple, no anotado o no codificado (b) codificado o anotado	

Por parámetros	Ornia (2015) (11 parámetros)	1) medio de difusión	3	(a) oral
				(b) escrito
				(c) oral y escrito
		2) grado de amplitud del análisis	2	(a) de referencia
				(b) de no referencia
		3) momento temporal	2	(a) sincrónico
				(b) diacrónico
		4) posibilidad de actualización	2	(a) monitorizado
				(b) no monitorizado
		5) extensión de los textos	3	(a) de textos completos
				(b) de muestras
		(c) mixtos		
6) publicación	3	(a) de textos publicados		
		(b) de textos no publicados		
		(c) de textos publicados y no publicados		
7) grado de especialización de los documentos	2 ²	(a) generalista		
		(b) especializado		
		(b.1) genérico		
		(b.2) canónico		
		(b.3) cronológico		
		(b.4) dialectal		
		(b.5) histórico		
		(b.6) nacional		
		(b.7) regional		
		(b.8) jurídico		
8) número de lenguas	2	(a) monolingüe		
		(b) multilingüe		
9) composición de originales y traducción, y relación entre ellos	2	corpus (a) comparable		
		(a.1) de textos originales en una o varias lenguas		
		(a.2) de archivos traducidos en diferentes lenguas		

² Nótese que los subtipos de corpus no se han contabilizado a lo largo del presente artículo por el hecho de permanecer esta cuestión en un plano secundario.

Por parámetros	Ornia (2015) (11 parámetros)	9) composición de originales y traducción, y relación entre ellos	2	(a.3) de documentos originales y otros textos traducidos	
				(b) paralelo	
		10) dirección de la traducción	2	(a) unidireccional (b) multidireccional	
		11) alineación	2	(a) alineado (b) no alineado	
	Cabré Castellví (2007) (6 parámetros)	1) tema	2	(a) unidisciplinar (b) pluridisciplinar	
		2) nivel de especialización	2	(a) de un nivel (b) de distintos niveles de especialidad	
		3) canal de transmisión	4	(a) oral (b) escrito (c) audiovisual (d) oral, escrito y audiovisual	
		4) género textual	2	(a) de un género (b) de distintos géneros textuales	
		5) estrategia discursiva	2	(a) heterogéneo (b) homogéneo	
		6) lenguas	4	(a) monolingüe	
				(b) bilingüe (c) trilingüe (d) paralelo	
	Autor		Tipología de corpus		
	Por dicotomías	Borja Albi (2007) (18 dicotomías)	(a) impreso o (b) electrónico		
			(a) <i>online</i> o (b) <i>offline</i>		
(a) oral o (b) escrito					
(a) de textos completos o (b) de fragmentos de textos					
(a) estático o (b) dinámico					
(a) sincrónico o (b) diacrónico					
(a) grande o (b) pequeño					
(a) general o (b) especializado					
(a) de fuentes literarias o (b) de fuentes no literarias					
Por dicotomías		(a) representativo o (b) no representativo			
		(a) organizado en torno a un género o (b) no organizado en torno a un género			
		(a) monolingüe o (b) multilingüe			

	Borja Albi (2007) (18 dicotomías)	(a) paralelo multilingüe o (b) comparable multilingüe		
		(a) paralelo multilingüe alineado o (b) paralelo multilingüe no alineado		
		(a) simple o (b) anotado		
		(a) para uso profesional o (b) para fines investigadores		
		(a) para fines investigadores generales o (b) para fines investigadores específicos		
		(a) para uso profesional en traducción general o (b) para uso profesional en traducción especializada		
	Autor	Nivel	Tipos de corpus	
Por niveles jerárquicos	Laviosa (2002) (4 niveles jerárquicos)	Uno ³	14	(a) de textos completos, (b) de muestras, (c) mixto o (d) monitor
				(a) sincrónico o (b) diacrónico
				(a) general o (b) terminológico
				(a) monolingüe, (b) bilingüe o (c) multilingüe
				(a) escrito, (b) hablado o (c) mixto
Por niveles jerárquicos	Laviosa (2002) (4 niveles jerárquicos)	Dos	3	(a) monolingüe (a.1) simple (a.2) comparable
				(b) bilingüe (b.1) paralelo (b.2) comparable
				(c) multilingüe (c.1) paralelo (c.2) comparable
Por niveles jerárquicos	Laviosa (2002) (4 niveles jerárquicos)	Tres	3	(a) simple (a.1) traslacional (a.2) no traslacional
				(b) paralelo (b.1) monodireccional (b.2) bidireccional
				(c) multilingüe (c.1) mono-lengua origen (c.2) bi-lengua origen (c.3) multi-lengua origen

³ Cabe recordar que, tal y como se puede deducir de la teoría desarrollada más arriba, Laviosa (2002) solo se fundamenta en parámetros para delimitar los tipos de corpus relativos al primer nivel jerárquico. Por este motivo, no se han incluido tales parámetros en la tabla.

		Cuatro	3	(a) traslacional mono-lengua origen
				(b) traslacional bi-lengua origen
				(c) traslacional multi-lengua origen

Tabla 1: Clasificaciones de corpus existentes sobre las que se cimienta nuestra propuesta. Fuente: Elaboración propia

3. Nueva propuesta integradora de tipología de corpus

Las clasificaciones de corpus tratadas previamente han sido diseñadas tanto de forma general (Sinclair 1996; Torruella Casañas y Llisteri Boix 1999; Villayandre Llamazares 2008), como específicamente en el marco de disciplinas tan idóneas —desde el punto de vista del lenguaje— tales como la traducción (Laviosa 2002; Fernandes 2006; Borja Albi 2007; Ornia 2015) y la lingüística (Cabré Castellví 2007; Ornia 2015). Así pues, en aras de la completitud teórica y la interdisciplinariedad —a la vista de dichas clasificaciones—, presentamos seguidamente nuestra propuesta, en la cual se pueden reconocer dos tipos de parámetros. Por una parte, quince parámetros adoptados —y, en algunas ocasiones, adaptados— de las categorizaciones de corpus formuladas en el apartado que precede. Por otra parte, tres parámetros de nueva creación. Conviene apuntar que las denominaciones de corpus que aparecen en los subapartados siguientes se desprenden asimismo de la literatura desarrollada anteriormente, excepto: *corpus transcrito, multimodal, convencional, virtual y para fines docentes*. Nótese también que las definiciones que aportamos con posterioridad acerca de las variopintas clases de corpus no siempre concuerdan con aquellas proporcionadas por los teóricos arriba aludidos. Además, algunos tipos de corpus han sido ubicados en otros parámetros.

3.1. Número de lenguas: ¿monolingüe, bilingüe o multilingüe?

El criterio *número de lenguas* ha sido tomado de las clasificaciones de corpus realizadas por Laviosa (2002: 35), Fernandes (2006: 93), Cabré Castellví (2007: 91), Villayandre Llamazares (2008: 344) y Ornia (2015: 350). Este tiene que ver con la cantidad de idiomas implicados en el material reunido. De manera que resulta posible deslindar entre:

- Corpus monolingüe: incluye material en exclusivamente una lengua y, por consiguiente, comprende análisis intralingüales (Biel 2010: 3).
- Corpus bilingüe: incorpora material en dos idiomas distintos. Con lo cual, involucra estudios interlingüales (Biel 2010: 3).
- Corpus multilingüe: integra material en más de dos lenguas diferentes. Por lo tanto, tal y como el corpus bilingüe, abarca análisis interlingüales (Biel 2010: 3).

3.2. Medio de difusión del material: ¿textual, oral, transcrito o multimodal?

El parámetro *medio de difusión del material* se ha extraído de la categorización de corpus relativa a la autora Ornia (2015: 345). Dicho parámetro contempla el canal del material recopilado. Por ello, cabe delimitar cuatro tipos de corpus (Aldama *et al.* 2022: 3-4):

- Corpus textual: contiene textos provenientes de todas las modalidades de la lengua escrita (Marcos Marín 1994: 88).

- Corpus oral: consta de materiales en formato audio, esto es, de grabaciones de la lengua hablada (Aldama *et al.* 2022: 3).

- Corpus transcrito: está compuesto por transcripciones ortográficas —mediante las que se transliteran los enunciados de los hablantes utilizando la ortografía convencional (Santalla del Río 2005: 53)— y/o fonéticas y fonológicas de material fonético.

- Corpus multimodal: presta atención a “componentes de la comunicación que no son estrictamente el lenguaje verbal, pero que aportan significado al mismo” (Vacas Matos 2020: 21). Este tipo de corpus incluye normalmente vídeos que conjugan audio, imagen y texto (Aldama *et al.* 2022: 4).

3.3. Especificidad del material: ¿genérico o canónico?

El criterio *especificidad del material* ha sido recuperado de las tipologías de corpus trazadas por los estudiosos Torruella Casañas y Llisterra Boix (1999: 55), y Villayandre Llamazares (2008: 345). Tal criterio atiende a la uniformidad del material que conforma el corpus, pudiendo distinguirse entre:

- Corpus genérico: aloja materiales pertenecientes a la misma clase —por ejemplo, contratos— o, de forma estructurada, a diversas —verbigracia, contratos, testamentos y escrituras—. De modo que se disocia entre corpus genérico (a) homogéneo y (b) heterogéneo respectivamente (Cabré Castellví 2007: 91).

- Corpus canónico: recoge las obras completas de un autor (Corpas Pastor 2001: 158).

3.4. Relación entre el material: ¿comparable o paralelo?

El parámetro *relación entre el material* procede de la taxonomía de corpus de Fernandes (2006: 91). Este considera la conexión lingüística existente entre los materiales y, consecuentemente, alberga dos tipos de corpus:

- Corpus comparable: puede contener (a) material original en una o diversas lenguas —corpus comparable de material original (a.1.) monolingüe, (a.2) bilingüe o (a.3) multilingüe—, (b) material traducido en uno o varios idiomas —corpus comparable de material traducido (b.1) monolingüe,

(b.2) bilingüe o (b.3) multilingüe—, o (c) materiales originales y otros materiales traducidos —sin constituir estas traslaciones de los primeros— en una o múltiples lenguas —corpus comparable de materiales originales y otros materiales traducidos (c.1.) monolingüe, (c.2) bilingüe o (c.3) multilingüe—. Por ende, un corpus comparable nunca reúne materiales originales con sus respectivas traducciones (Ornia 2015: 341).

- Corpus paralelo: consta de materiales originales y sus traducciones (Granger 2003: 20). Esto conlleva la coexistencia de dos idiomas distintos como mínimo y, por tanto, la diferenciación entre corpus paralelo (a) bilingüe y (b) multilingüe. En función de la dirección de la traducción, dicha clase de corpus se cataloga como: 1) unidireccional: cuando aún materiales originales en la lengua origen —en adelante, LO— y sus correspondientes traducciones en la lengua meta —en lo sucesivo, LM—; 2) bidireccional: cuando incorpora, además de materiales originales en la LO y sus traducciones en la LM, materiales originales en la LM y sus traducciones en la LO; y 3) multidireccional: cuando participan más de dos idiomas y la dirección de la traducción no se centra en la LO, sino en la interacción de todas las lenguas que configuran el corpus (Zanettin 2000: 106). Con lo cual, un corpus paralelo bilingüe puede ser (a) unidireccional o (b) bidireccional; y un corpus paralelo multilingüe, (a) unidireccional, (b) bidireccional o (c) multidireccional.

3.5. *Naturaleza del material: ¿auténtico o experimental?*

El criterio *naturaleza del material* tiene su origen en el parámetro *quality* o “calidad” de Sinclair (1996: 7). Dicho criterio presta atención a la espontaneidad de las situaciones comunicativas en las que se generan los materiales recabados para el corpus. De ahí que discernamos entre:

- Corpus auténtico: incluye material obtenido de la comunicación genuina de las personas en sus ocupaciones habituales (Sinclair 1996: 7).

- Corpus experimental: comprende material producido en escenarios especiales a través de la intervención de lingüistas (Sinclair 1996: 7) —verbigracia, las entrevistas planificadas y los contratos modelo—.

3.6. *Fuente del material: ¿convencional o virtual?*

El parámetro *f fuente del material* se singulariza por su carácter neológico. Este se focaliza en la procedencia de los materiales recopilados y, por ello, se organiza en torno a la disyuntiva que sucede:

- Corpus convencional: contiene materiales que no han sido extraídos de Internet.

- Corpus virtual: se compone de materiales disponibles en Internet (Corpas Pastor y Seghiri 2009: 75). Tal tipo de corpus puede ser (a) preexistente en línea o (b) de nueva creación.

3.7. *Formato original del material: ¿impreso o electrónico?*

El criterio *formato original del material*, al igual que el parámetro anterior, ha sido propiamente acuñado, si bien sobre la base de la dicotomía corpus (a) impreso o (b) electrónico de Borja Albi (2007: 246). Dicho criterio valora las características de diseño originarias del material recuperado. Ciertamente, recoge las dos denominaciones de corpus indicadas por la autora mencionada previamente (Borja Albi 2007: 246), las cuales se definen a continuación:

- Corpus impreso: aloja material cuyo formato, a pesar de haber sido digitalizado, era originariamente impreso.
- Corpus electrónico: integra material que no ha necesitado ser digitalizado al constituir esta su forma originaria.

3.8. *Publicación del material: ¿publicado, no publicado o mixto?*

El parámetro *publicación del material* proviene de la clasificación de corpus elaborada por Ornia (2015: 348). Tal parámetro versa sobre la asequibilidad del material que conforma el corpus y, por consiguiente, diferencia entre:

- Corpus de material publicado: consta de materiales disponibles para cualquier público, ya sea en fuentes electrónicas o impresas y de pago o gratuitas.
- Corpus de material no publicado: alberga materiales que, frecuentemente, no se divulgan por su carácter confidencial y/o privado. De modo que el acceso a dichos materiales está subordinado a la voluntad de las partes intervinientes, quienes eligen si desean proporcionárselos a una persona ajena al procedimiento que las ocupa.
- Corpus de material mixto: combina material accesible para cualquier persona y material no difundido públicamente.

3.9. *Cantidad de material: ¿completo o de muestras?*

El criterio *cantidad de material* ha sido adoptado de las tipologías de corpus desarrolladas por Torruella Casañas y Llisterri Boix (1999: 55), y Laviosa (2002: 34). Este toma en consideración la completitud de los materiales reunidos. Así pues, cabe establecer los siguientes tipos de corpus:

- Corpus completo: se configura a partir de materiales completos con vistas a ofrecer una muestra representativa de la lengua.

- Corpus de muestras: incorpora fragmentos de materiales. El interés de tal corpus radica “en el estadio de lengua representado” (Corpas Pastor 2001: 158-159).

3.10. *Distribución y porcentaje del material: ¿grande, equilibrado o piramidal?*

El parámetro *distribución y porcentaje del material* se fundamenta en la categorización de corpus formulada por Torruella Casañas y Llisterri Boix (1999: 53). Dicho parámetro tiene en cuenta la correspondencia proporcional entre los materiales allegados y, en consecuencia, deslinda entre:

- Corpus grande: posee un número de palabras realmente elevado, por lo que obvia aspectos tales como la representatividad o el equilibrio. Su extensión se encuentra a menudo acotada (Torruella Casañas y Llisterri Boix 1999: 53).

- Corpus equilibrado: incluye diversas variedades de un tipo de material —textos, audios, transcripciones o vídeos, ya sea de forma fragmentada o completa— en porcentajes similares.

- Corpus piramidal: organiza sus materiales en niveles, “los cuales se caracterizan por aumentar progresivamente la complejidad de las variedades temáticas” (Corpas Pastor 2001: 158).

3.11. *Delimitación temporal del corpus: ¿diacrónico o sincrónico?*

El criterio *delimitación temporal del corpus* dimana de las clasificaciones de corpus realizadas por los autores Laviosa (2002: 35), Fernandes (2006: 93), Villayandre Llamazares (2008: 345) y Ornia (2015: 345). Este atiende al elemento cronológico de los materiales recabados y, por ende, separa las clases de corpus que suceden:

- Corpus diacrónico: aloja material restringido a un momento o período de tiempo específico (Ornia 2015: 345).

- Corpus sincrónico: está formado por material que no se halla delimitado temporalmente (Ornia 2015: 345).

3.12. *Área de estudio del corpus: ¿lingüístico o traslativo?*

El parámetro *área de estudio del corpus* se desprende de la tipología de corpus ideada por Fernandes (2006: 92). Tal parámetro evalúa la parcela del lenguaje sometida a análisis en el corpus. De ahí que distingamos entre:

- Corpus lingüístico: se ocupa del estudio de las lenguas (Fernandes 2006: 92).

- Corpus traslativo: examina los procesos y los productos de la traducción (Fernandes 2006: 92). Atendiendo al estatus del traductor, un corpus traslativo puede ser (a) profesional o (b) semiprofesional (Corpas Pastor 2008: 179).

3.13. *Propósito del corpus: ¿para fines profesionales, investigadores o docentes?*

El criterio *propósito del corpus*, tal y como aquellos designados *fuentes del material* y *formato original del material*, es de nueva acuñación, y se cimienta en la dicotomía corpus para (a) uso profesional o (b) fines investigadores de Borja Albi (2007: 246). Dicho criterio reflexiona sobre la finalidad con la que se ha diseñado el corpus y recoge las siguientes categorías:

- Corpus para fines profesionales: se recopila con miras a, verbigracia, desde la perspectiva lingüística, redactar un documento de acuerdo con las convenciones idiosincrásicas de su género textual o, desde el punto de vista traslativo, llevar a cabo un determinado encargo de traducción.

- Corpus para fines investigadores: se confecciona con vistas a, por ejemplo, desde la perspectiva lingüística, determinar los rasgos léxicos de un tipo concreto de lengua o, desde la vertiente traductológica, estudiar las técnicas de traducción implementadas con mayor asiduidad en un ámbito particular.

- Corpus para fines docentes: se compila a fin de, por ejemplo, desde el punto de vista lingüístico, impartir la gramática de un idioma o, desde la vertiente traslativa, promover la competencia traductora de los alumnos. A este respecto, conviene nombrar el corpus de estudiantes (Granger 2002: 7-11) por su gran utilidad en el ámbito de adquisición de segundas lenguas y lenguas extranjeras.

3.14. *Campo de investigación lingüística: ¿general o especializado?*

El parámetro *campo de investigación lingüística* se ha obtenido de la clasificación de corpus relativa a Fernandes (2006: 92). Este contempla el componente diafásico de los materiales recuperados y se bifurca en:

- Corpus general: permite la descripción general de una variedad lingüística o lengua (McEnery, Xiao y Tono 2006: 15). Consiste en un corpus de la lengua común que alberga multitud de materiales producidos en situaciones comunicativas habituales (Corpas Pastor 2001: 158).

- Corpus especializado: aborda un tipo particular de lengua (Zanettin 2011: 15). Se trata de un corpus de una variedad específica de lengua que incluye exclusivamente materiales pertenecientes a la mencionada variedad (Corpas Pastor 2001: 158).

3.15. *Temas del corpus: ¿unidisciplinar o pluridisciplinar?*

El criterio *temas del corpus* obedece a la tipología de corpus creada por Cabré Castellví (2007: 91). Tal criterio se centra en la diversidad temática del corpus. De modo que resulta posible hablar de:

- Corpus unidisciplinar: cuando comprende un campo de estudio, esto es, un ámbito completo (Cabré Castellví 2007: 91).

- Corpus pluridisciplinar: cuando abarca una parcela de un campo de estudio, es decir, solo una perspectiva de un ámbito —verbigracia, un corpus de derecho medioambiental— (Cabré Castellví 2007: 91).

3.16. Posibilidad de actualización del corpus: ¿cerrado o abierto?

El parámetro *posibilidad de actualización del corpus* ha sido extraído de la categorización de corpus según Ornia (2015: 346). Dicho parámetro versa sobre la opción de modificar el contenido del corpus y diferencia entre:

- Corpus cerrado: cuenta con un número limitado de palabras que se fija con carácter previo a su compilación, por lo que no puede actualizarse una vez alcanzado el tamaño finito preestablecido (Villayandre Llamazares 2008: 345).

- Corpus abierto: posee un volumen textual susceptible de renovación —a través de la introducción y la supresión paralela de materiales nuevos y antiguos correspondientemente (Corpas Pastor 2001: 158)— o ampliación. Por lo tanto, un corpus abierto puede ser (a) actualizable o (b) ampliable.

3.17. Documentación de los materiales: ¿documentado o no documentado?

El criterio *documentación de los materiales* se deriva de las clasificaciones de corpus sugeridas por Sinclair (1996: 8), y Torruella Casañas y Llisterri Boix (1999: 57). Este gira en torno a la inclusión de archivos que describan los materiales allegados y, consecuentemente, encierra dos tipos de corpus:

- Corpus documentado: incorpora materiales con archivos DTD (*Document Type Definition*) o cabeceras descriptivas de su filiación y sus constituyentes (Torruella Casañas y Llisterri Boix 1999: 57).

- Corpus no documentado: está formado por materiales carentes de archivos DTD o cabeceras descriptivas (Corpas Pastor 2001: 159).

3.18. Codificación y anotación: ¿anotado o no anotado?

El parámetro *codificación y anotación* se debe a la tipología de corpus enunciada por Torruella Casañas y Llisterri Boix (1999: 56). Tal parámetro parte de la marcación de los materiales recabados y deslinda entre:

- Corpus anotado: aloja materiales que han sido etiquetados manual, semiautomática o automáticamente. Dependiendo del tipo de etiqueta, cabe disociar entre:

1) Corpus anotado textualmente: consta de códigos sobre aspectos estructurales tales como cabeceras, párrafos, títulos, límites de oraciones o citas (Lavid López 2005: 311).

2) Corpus anotado extratextualmente: integra los metadatos (Carriazo Ruiz 2021: 290), a saber, información —situada normalmente en las cabeceras

descriptivas (Lavid López 2005: 311)— acerca de los materiales —como los títulos, el contenido, la variedad o las variedades lingüísticas, la temática general en la que se insertan y las fechas de publicación o producción— y el autor o los autores —como, por ejemplo, su edad y su sexo— (Carriazo Ruiz 2021: 290).

3) Corpus anotado lingüísticamente: contiene información explícita referente a todas las palabras de los materiales reunidos (Aguado de Cea, Álvarez de Mon y Rego, y Pareja Lora 2009: 222). Dentro de la marcación lingüística, hallamos asimismo diversos tipos de anotación en función del medio de difusión del corpus, es decir, del canal del material allegado.

Por una parte, los corpus textuales y transcritos, aunque también los multimodales por su heterogénea inclusión de componentes comunicativos, se marcan fundamentalmente mediante:

3.1) Lematización: a través de la cual se circunscriben las variantes de una palabra a la forma básica recogida en los diccionarios, esto es, a su lema.

3.2) Anotación fonética y fonológica: a través de la cual se definen las palabras desde el punto de vista fonético y fonológico.

3.3) Anotación prosódica: a través de la cual se precisan los acentos, las pausas en la producción, etc. (Pérez Guerra 1999: 39).

3.4) Anotación ortográfica: a través de la cual se señalan las funciones de los símbolos y los signos gráficos de los materiales recuperados —como de las comillas dobles, el punto, las comillas simples y la letra itálica— (Sekhar Dash 2005: 106).

3.5) Anotación gramatical: a través de la cual se le atribuye a cada palabra una marca con su categoría gramatical, pudiendo incluir esta a su vez “información morfosintáctica y, en ocasiones, rasgos sintácticos-semánticos” (Lavid López 2005: 314).

3.6) Anotación sintáctica: a través de la cual, sobre la base de la marcación gramatical, se someten a análisis las oraciones (McEnery, Xiao y Tono 2006: 36).

3.7) Anotación discursiva: a través de la cual se localizan las conexiones referenciales —es decir, anafóricas y catafóricas—.

3.8) Anotación semántica: a través de la cual, a partir de un inventario con significados potenciales, se le asigna a cada palabra la acepción que mejor se ajusta al contexto en el que se enmarca (Gries y Berez 2017: 385).

3.9) Anotación estilística: a través de la cual se asocian valores —como estilo directo libre, pensamiento indirecto, estilo indirecto, estilo indirecto libre, narración de pensamiento, etc. (Pérez Guerra 1999: 41)— que posibilitan, frecuentemente, identificar al autor de los materiales.

Por otra parte, los corpus orales, al igual que los multimodales, son etiquetados principalmente —además de por medio de la lematización, y la

anotación fonética y fonológica, prosódica, gramatical, sintáctica, discursiva, semántica y estilística— mediante:

3.10) Anotación pragmática: a través de la cual —de acuerdo con Pérez Guerra (1999: 41), quien se sustenta en el modelo de Stiles que abordan Leech, McEnery y Wynne (1997: 85-101)— se atribuyen códigos relativos a los significados literales y pragmáticos de los enunciados. De tal manera que se codifica información objetiva, así como percepciones, pensamientos, sentimientos e intenciones.

3.11) Anotación del lenguaje de signos y gestos: a través de la cual se detecta el lenguaje no verbal que figura en los materiales audiovisuales (Gries y Berez 2017: 389).

3.12) Anotación de interacciones: a través de la cual se introducen símbolos con respecto a la temporalidad o la secuencialidad de los enunciados —como la sucesión de turnos y la superposición del habla—, el ritmo —lento o rápido—, la entonación —por ejemplo, el énfasis y las subidas o las bajadas de tono—, las disfluencias, la presencia y la duración de las pausas, etc. (Schegloff 2007: 265-269).

En suma, existen corpus anotados lingüísticamente sobre (a) fonética y fonología, (b) prosodia, (c) ortografía, (d) gramática, (e) sintaxis, (f) discurso, (g) semántica, (h) estilística, (i) pragmática, (j) signos y gestos, (k) interacciones, y (l) mediante lematización.

- Corpus no anotado: consta de materiales en formato ASCII (*American Standard Code for Information Interchange*) —igualmente denominado **.txt* o texto plano—. Esto significa que están desprovistos de etiquetas y que poseen un alto grado de simplicidad (Corpas Pastor 2001: 159).

Una vez llegados a este punto, a modo aclaratorio, vertemos nuestra propuesta de clasificación de corpus en la tabla que sigue:

Propuesta integradora de clasificación de corpus	
Parámetro	Tipos de corpus
1) número de lenguas	3 (a) monolingüe (b) bilingüe (c) multilingüe
2) medio de difusión del material	4 (a) textual (b) oral (c) transcrito (d) multimodal
3) especificidad del material	2 (a) genérico (a.1) homogéneo (a.2) heterogéneo

		(b) canónico
4) relación entre el material	2	(a) comparable (a.1) de material original (a.1.1) monolingüe (a.1.2) bilingüe (a.1.3) multilingüe (a.2) de material traducido (a.2.1) monolingüe (a.2.2) bilingüe (a.2.3) multilingüe (a.3) de materiales originales y otros materiales traducidos (a.3.1) monolingüe (a.3.2) bilingüe (a.3.3) multilingüe
		(b) paralelo (b.1) bilingüe (b.1.1) unidireccional (b.1.2) bidireccional (b.2.) multilingüe (b.2.1) unidireccional (b.2.2) bidireccional (b.2.3) multidireccional
5) naturaleza del material	2	(a) auténtico (b) experimental
6) fuente del material	2	(a) convencional (b) virtual (b.1) preexistente en línea (b.2) de nueva creación
7) formato original del material	2	(a) impreso (b) electrónico
8) publicación del material	3	(a) de material publicado (b) de material no publicado (c) de material mixto
9) cantidad de material	2	(a) completo (b) de muestras
10) distribución y porcentaje del material	3	(a) grande (b) equilibrado (c) piramidal
11) delimitación temporal del corpus	2	(a) diacrónico (b) sincrónico
12) área de estudio del corpus	2	(a) lingüístico

Tipología de corpus: hacia una propuesta integradora

		(b) traslativo (b.1) profesional (b.2) semiprofesional
13) propósito del corpus	3	(a) para fines profesionales (b) para fines investigadores (c) para fines docentes
14) campo de investigación lingüística	2	(a) general (b) especializado
15) temas del corpus	2	(a) unidisciplinar (b) pluridisciplinar
16) posibilidad de actualización del corpus	2	(a) cerrado (b) abierto (b.1) actualizable (b.2) ampliable
17) documentación de los materiales	2	(a) documentado (b) no documentado
18) Codificación y anotación	2	(a) anotado (a.1) textualmente (a.2) extratextualmente (a.3) lingüísticamente (a.3.1) sobre fonética y fonología (a.3.2) sobre prosodia (a.3.3) sobre ortografía (a.3.4) sobre gramática (a.3.5) sobre sintaxis (a.3.6) sobre discurso (a.3.7) sobre semántica (a.3.8) sobre estilística (a.3.9) sobre pragmática (a.3.10) sobre signos y gestos (a.3.11) sobre interacciones (a.3.12) mediante lematización (b) no anotado

Tabla 2: Propuesta integradora de clasificación de corpus

Fuente: Elaboración propia

Conclusiones

La lingüística de corpus constituye una metodología lingüística con aplicación en multitud de disciplinas, fenómeno que ha conducido a la falta de consenso en lo que concierne a los tipos de corpus existentes, su denominación y su definición.

El presente artículo pretende buscarle una solución a este problema, ya que ofrece una posible propuesta de clasificación de corpus interdisciplinaria a partir de, principalmente, las categorizaciones de Sinclair (1996), Torruella Casañas y Llisterri Boix (1999), Laviosa (2002), Fernandes (2006), Borja Albi (2007), Cabré Castellví (2007), Villayandre Llamazares (2008) y Ornia (2015).

Dichas categorizaciones poseen aspectos negativos y positivos a escala individual. Por ejemplo, aquella formulada por Sinclair (1996) no contempla tipos de corpus en torno a criterios tan esenciales como el *número de lenguas* o el *medio de difusión del material*. Si bien es cierto que recoge un parámetro importante ausente en el resto de las clasificaciones mencionadas, es decir, el indicador *calidad* —que adoptamos y adaptamos en nuestra propuesta como *naturaleza del material*— junto con la distinción entre *corpus auténtico* o *experimental*. En lo atinente a la taxonomía de Torruella Casañas y Llisterri Boix (1999), curiosamente también prescinde de los mismos criterios esenciales que la clasificación de Sinclair (1996). Si cotejamos ambas, la primera (Torruella Casañas y Llisterri Boix, 1999) aporta como novedad, y con gran acierto, los parámetros *distribución y porcentaje del material*, y *especificidad del material*. Con relación a la categorización en niveles jerárquicos de Laviosa (2002), se halla inoportunamente acotada al ámbito de la traducción. Empero, esta contribuye a la introducción de la variable *direccionalidad* en los tipos de corpus y, además, conforma la primera taxonomía de las consultadas en considerar el parámetro *número de lenguas*. Por lo que se refiere a la clasificación de Fernandes (2006), se encuentra limitadamente en la misma línea que la realizada por Laviosa (2002), aunque dicha clasificación añade, de manera conveniente, los criterios *área de estudio* y *campo de investigación lingüística*. En cuanto a las dicotomías de Borja Albi (2007), se estiman verdaderamente ilustrativas, a pesar de que algunas, como *corpus de fuentes (a) literarias* o *(b) no literarias*, carezcan de practicabilidad debido a su extrema especificidad. Por lo que respecta a la categorización por parámetros de Cabré Castellví (2007), innova con la inclusión de tipos de corpus como *(a) unidisciplinar* o *(b) pluridisciplinar*, o *(a) heterogéneo* o *(b) homogéneo*; sin embargo, deja atrás otros como el *corpus comparable*. En lo tocante a la taxonomía de Villayandre Llamazares (2008), resulta general y breve en comparación con aquella ideada por Ornia (2015), la cual configura la clasificación por parámetros más amplia de las abordadas en el artículo que nos ocupa. No obstante, aun así, esta olvida criterios específicos relevantes como, entre otros, la *documentación de los materiales*, y la *codificación y anotación*.

Habida cuenta de todo lo expuesto, mediante la fusión selectiva, reorganización y ligera ampliación de todas estas categorizaciones, se delimitaron un total de dieciocho parámetros, y cuarenta y dos tipos de corpus en nuestra

propuesta. De modo que este trabajo proporciona un instrumento que estandariza el proceso de selección de corpus para llevar a cabo estudios de corte lingüístico y la enseñanza/el aprendizaje de idiomas. Es decir, brinda una herramienta que facilita la fase de diseño de corpus y que, por ende, procura la máxima explotación de estos.

La limitación de la tipología de corpus presentada —tal y como cualquier otra— radica en la vinculación de la lingüística de corpus con los cambios tecnológicos, los cuales repercuten en la estabilidad de dicha metodología y provocan que esta se halle en continua transformación.

Referencias bibliográficas

- AGUADO DE CEA, Guadalupe; ÁLVAREZ DE MON Y REGO, Inmaculada; PAREJA LORA, Antonio, “Una visión interdisciplinar de la anotación semántica”. En: Alcina Caudet, María Amparo; Valero Doménech, Esperanza; Rambla Mulet, Elena (eds.), *Terminología y sociedad del conocimiento*. Berna: Peter Lang, 2009, pp. 219-254.
- ALDAMA, Nuria *et al.*, “Anotación de corpus lingüísticos: metodología utilizada en el Instituto de Ingeniería del Conocimiento (IIC)” (White paper). Instituto de Ingeniería del Conocimiento de la Universidad Autónoma de Madrid, 2022, pp. 1-15. Disponible en línea en: <https://www.iic.uam.es/pdf/ anotacion-corpus-linguisticos.pdf>
- BAKER, Mona, “Corpus linguistics and translation studies: implications and applications”. En: Baker, Mona; Francis, Gill; Tognini-Bonelli, Elena (eds.), *Text and technology: in honour of John Sinclair*. Ámsterdam y Filadelfia: John Benjamins, 1993, pp. 233-250.
- BIEL, Lucja, “Corpus-based studies of legal language for translation purposes: methodological and practical potential”. En: Heine, Carmen; Engberg, Jan (eds.), *Reconceptualizing LSP: online proceedings of the XVII European LSP Symposium 2009*. Aarhus: Aarhus University, 2010, pp. 1-14.
- BORJA ALBI, Anabel, “Corpora for translators in Spain. The CDJ-GITRAD Corpus and the GENTT Project”. En: Anderman, Gunilla; Rogers, Margaret (eds.), *Incorporating corpora: the linguist and the translator*. Clevedon: Multilingual Matters (Series Translating Europe), 2007, pp. 243-265.
- CABRÉ CASTELLVÍ, María Teresa, “Constituir un corpus de textos de especialidad: condiciones y posibilidades”. En: Ballard, Michel; Pineira-Tresmontant, Carmen (eds.), *Les corpus en linguistique et en traductologie*. Arras: Artois Presses Université, 2007, pp. 89-106.

- CARRIAZO RUIZ, José Ramón, “La investigación del significado: la lingüística de corpus”. En: Carriazo Ruiz, José Ramón; Julià Luna, Carolina (eds.), *Manual de semántica de la lengua española*. Madrid: Ramón Areces, 2021, pp. 279-299.
- CORPAS PASTOR, Gloria, “Compilación de un corpus ‘ad hoc’ para la enseñanza de la traducción inversa especializada”. En: *TRANS: Revista de Traductología*, 5, 2001, pp. 155-184. Disponible en línea en: <https://doi.org/10.24310/TRANS.2001.v0i5.2916>
- , *Investigar con corpus en traducción: los retos de un nuevo paradigma*. Fráncfort del Meno: Peter Lang, 2008.
- CORPAS PASTOR, Gloria; SEGHIRI, Miriam, “Virtual corpora as documentation resources: translating travel insurance documents (English-Spanish)”. En: Beeby, Allison; Rodríguez Inés, Patricia; Sánchez-Gijón, Pilar (eds.), *Corpus use and translating: corpus use for learning to translate and learning corpus use to translate*. Ámsterdam y Filadelfia: John Benjamins, 2009, pp. 75-108.
- FERNANDES, Lincoln, “Corpora in translation studies: revisiting Baker’s typology”. En: *Fragmentos*, 30, 2006, pp. 87-95.
- GRANGER, Sylviane, “A bird’s-eye view of learner corpus research”. En: Granger, Sylviane; Hung, Joseph; Petch-Tyson, Stephanie (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Ámsterdam y Filadelfia: John Benjamins, 2002, pp. 3-33.
- , “The corpus approach: a common way forward for contrastive linguistics and translation studies?”. En: Granger, Sylviane; Lerot, Jacques; Petch-Tyson, Stephanie (eds.), *Corpus-based approaches to contrastive linguistics and translation studies*. Ámsterdam/Nueva York: Rodopi, 2003, pp. 17-30.
- GRIES, Stefan; BEREZ, Andrea, “Linguistic annotation in/for corpus linguistics”. En: Ide, Nancy; Pustejovsky, James (eds.), *Handbook of linguistic annotation*. Dordrecht: Springer, 2017, pp. 379-409.
- JOHANSSON, Stig, “On the role of corpora in cross-linguistic research”. En: Johansson, Stig; Oksefjell, Signe (eds.), *Corpora and cross-linguistic research: theory, method, and case studies*. Ámsterdam: Rodopi, 1998, pp. 3-24.
- LAVID LÓPEZ, Julia, *Lenguaje y nuevas tecnologías: nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI*. Madrid: Cátedra, 2005.
- LAVIOSA, Sara, *Corpus-based translation studies: theory, findings, applications*. Ámsterdam y Nueva York: Rodopi, 2002.
- LEECH, Geoffrey; MCENERY, Tony; WYNNE, Martin, “Further levels of annotation”. En: Garside, Roger; Leech, Geoffrey; McEnery, Tony (eds.), *Corpus annotation: linguistic information from computer text corpora*. Londres: Longman, 1997, pp. 85-101.

- MARCOS MARÍN, Francisco Adolfo, *Informática y humanidades*. Madrid: Gredos, 1994.
- MCENERY, Tony; WILSON, Andrew, *Corpus linguistics* (2ª ed.). Edimburgo: Edinburgh University Press, 2001.
- MCENERY, Tony; XIAO, Richard; TONO, Yukio, *Corpus-based language studies: an advanced resource book*. Londres/Nueva York: Routledge, 2006.
- ORNIA, Goretta Faya, “Propuesta de clasificación de corpus textuales”. En: Sánchez Nieto, María Teresa; Álvarez Álvarez, Susana; Arnáiz Uzquiza, Verónica; Ortego Antón, María Teresa; Santamaría Ciordia, Leticia; Fernández Muñiz, Rosa (eds.), *Metodologías y aplicaciones en la investigación en traducción e interpretación con corpus*. Valladolid: Universidad de Valladolid, 2015, pp. 339-356.
- PÉREZ GUERRA, Javier, “Estándares de anotación en lingüística de corpus”. En: *Revista Española de Lingüística Aplicada*, 1, 1999, pp. 25-52.
- SANTALLA DEL RÍO, María Paula, “La elaboración de corpus lingüísticos”. En: Cal Varela, Mario; Núñez Pertejo, Paloma; Palacios Martínez, Ignacio Miguel (eds.), *Nuevas tecnologías en lingüística, traducción y enseñanza de lenguas*. Santiago de Compostela: Universidad de Santiago de Compostela, 2005, pp. 45-63.
- SCHEGLOFF, Emanuel Abraham, *Sequence organization in interaction: a primer in conversation analysis* (vol. I). Nueva York: Cambridge University Press, 2007.
- SEKHAR DASH, Niladri, *Corpus linguistics and language technology: with reference to Indian languages*. Nueva Delhi: Mittal Publications, 2005.
- SINCLAIR, John, “Preliminary recommendations on corpus typology”. En: *Expert Advisory Group on Language Engineering Standards* (EAGLES), 1996, pp. 1-13.
- TORRUELLA CASAÑAS, Joan; LLISTERRI BOIX, Joaquim, “Diseño de corpus textuales y orales”. En: Blecua Perdices, José Manuel; Clavería Nadal, Gloria; Sánchez Lancis, Carlos; Torruella Casañas, Joan (eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Editorial Milenio, 1999, pp. 45-77.
- VACAS MATOS, Marta, “El aprendizaje de la pragmática a través de un corpus multimodal de nativos y aprendientes de español”. En: *Revista Nebrija de Lingüística Aplicada a la Enseñanza de las Lenguas*, 14, 29, 2020, pp. 17-32. Disponible en línea en: <https://doi.org/10.26378/rnlael1429421>
- VILLAYANDRE LLAMAZARES, Milka, “Lingüística con corpus”. En: *Estudios Humanísticos. Filología*, 30, 2008, pp. 329-349. Disponible en línea en: <http://dx.doi.org/10.18002/ehf.v0i30.2847>
- ZANETTIN, Federico, “Parallel corpora in translation studies: issues in corpus design and analysis”. En: Olohan, Maeve (ed.), *Intercultural faultlines. Research*

VALERIA HERNÁNDEZ GARCÍA

models in translation studies I: textual and cognitive aspects. Manchester: St. Jerome, 2000, pp. 105-118.

———, “Translation and corpus design”. En: *Synaps: a Journal of Professional Communication*, 26, 2011, pp. 14-23. Disponible en línea en:
<http://hdl.handle.net/11250/2394011>