10.18800/lexis.202001.003
Artículo

El discurso especializado del comercio en español: diseño de un corpus para su estudio diacrónico
The Specialized Language of Commerce in Spanish: Design of a Corpus for its Diachronic Study

Matteo De Beni[1]
Dunia Hourani-Martín[1]

[1]*Università degli Studi di Verona*

Resumen
El objeto del presente artículo es exponer las características de diseño del corpus *DIACOM*, una herramienta para el estudio del léxico especializado del comercio en español y francés que se está llevando a cabo en el Departamento de Lenguas y Literaturas Extranjeras de la Universidad de Verona (Italia) en el marco de un proyecto ministerial dedicado al desarrollo de las humanidades digitales aplicadas a los ámbitos lingüístico-filológicos. Se hará hincapié en las cuestiones de diseño del subcorpus español, sobre todo en aquellas que resultan especialmente problemáticas a la hora de compilar un corpus diacrónico del ámbito de especialidad en cuestión.

***Palabras clave*:** corpus especializado, terminología del comercio, lengua española, diacronía

Abstract
The purpose of this article is to present the design features of the *DIACOM* corpus, a tool for the study of the field-specific lexis related to trade and commerce in the Spanish and French language. This study, carried out by the Department of Foreign Languages and Literatures of the University of Verona (Italy) within the framework of a government-funded project, aims at developing the field of Digital Humanities and specifically its application in linguistic and philological research. The paper addresses the issues related to designing the Spanish subcorpus, particularly the ones which have proved especially problematic in the building of a diachronic corpus within the area taken into account.

***Keywords*:** specialized corpus, trade-and-commerce lexis, Spanish language, diachronic perspective

1. Introducción

El corpus *DIACOM*, nombre procedente del acrónimo de *diacronía* y *comercio*, pretende convertirse en una herramienta de referencia para el estudio del léxico especializado del comercio en español y francés con una perspectiva diacrónica, desde mediados del siglo XIX hasta nuestros días. Su realización se está llevando a cabo en el Departamento de Lenguas y Literaturas Extranjeras de la Universidad de Verona (Italia) en el marco de un proyecto ministerial dedicado al

desarrollo de las humanidades digitales aplicadas a los ámbitos lingüístico-filológicos: *Le Digital Humanities applicate alle lingue e letterature straniere*.

El objetivo final de nuestro trabajo es poner a disposición de la comunidad científica un corpus que proporcione una visión amplia y global del campo del comercio en francés y español y que, por lo tanto, permita múltiples posibilidades para el estudio y la descripción de fenómenos lingüísticos dentro de este ámbito de especialidad en el tiempo, en el espacio y en función de la situación comunicativa.[1] Permitirá, por ejemplo, la detección de neologismos, la observación de la evolución de formas lingüísticas, el análisis de la variación terminológica en todas sus vertientes y constituirá una ayuda para la elaboración de herramientas terminográficas y para la extracción de datos para la enseñanza de segundas lenguas y del lenguaje especializado del comercio.[2]

En lo que respecta al español, existen numerosos trabajos dedicados, desde la vertiente diacrónica o sincrónica, bien a la descripción de las características generales (p. ej. Mateo Martínez 2007; Álvarez García 2011) bien al análisis de un aspecto concreto del lenguaje económico y comercial (como Gómez de Enterría Sánchez 1992a y 1992b; Cassany 2004a; De Hoyos 2016), o que abordan su estudio desde una perspectiva aplicada —como la traducción en diferentes combinaciones lingüísticas (Mayoral Asensio 2007; Pizarro Sánchez 2010; Gallego Hernández 2012 y 2018; Álvarez García 2017, entre otros), la didáctica del español especializado (p. ej. Gómez de Enterría Sánchez 2009; Martínez Egido 2009) o la elaboración de herramientas terminográficas (Ramacciotti y Rodil 2006; De la Fuente Marina 2019)—.[3] Nobstante, las posibilidades de exploración de este lenguaje de especialidad no han sido aún agotadas dadas la multidisciplinariedad e interdisciplinariedad de este ámbito (*vid. infra*), que cuenta con una gran cantidad de subdominios, así como por las diferentes orientaciones a la hora de emprender su estudio y los diversos aspectos lingüísticos, terminológicos y fraseológicos de interés dentro de su discurso. Por lo tanto, *DIACOM* no solo se construye como una herramienta que permita múltiples posibilidades de análisis, sino también que contribuya a favorecer la sistematización de datos y la agrupación de resultados.

En la actualidad, nos encontramos en la fase de compilación del corpus *DIACOM*. A continuación, expondremos los criterios que hemos tenido en cuenta a la hora de diseñar el subcorpus español, a la vez que justificaremos las decisiones que hemos tomado y que han venido determinadas por las necesidades y los objetivos de nuestro proyecto. Para ello seguimos los parámetros de diseño

---

[1] Cabré Castellví y Estopà Bagot (2005: 69-70) describen los factores relevantes constantes que dan lugar a una situación comunicativa especializada: (i) un emisor que cuenta con un conocimiento importante —aprendido conscientemente— sobre el tema de comunicación; (ii) un receptor que tiene la expectativa de recibir información; (iii) una temática especializada, donde la representación y la transmisión del conocimiento tienen lugar respetando la estructura conceptual del dominio en cuestión; (iv) una función principalmente informativa; y (v) un entorno comunicativo profesional, en el que los expertos producen conocimiento dirigido a un público experto, semiexperto o lego. Este último rasgo determina la variación funcional dependiendo del nivel de especialización del texto —muy especializado, semiespecializado o de bajo nivel de especialización—, su grado de formalidad y la finalidad esencial del discurso, que es lo que se conoce en terminología como *variación vertical* (Cabré Castellví 2002: 13).

[2] Esta herramienta se construye, en primer lugar, como instrumento para la investigación del léxico especializado del comercio y de su historia, pero también servirá como fuente de documentación con una finalidad didáctica, en concreto, para su aplicación en la enseñanza en cursos universitarios como algunos de los que se ofertan en el Departamento de Lenguas y Literaturas Extranjeras de la Universidad de Verona en el ámbito del Grado en Lenguas y Culturas para el turismo y el comercio internacional y del Máster en Lenguas para la comunicación turística y comercial.

[3] En este ámbito, destaca el proyecto COMENEGO (Corpus Multilingüe de Economía y Negocios) (COMENEGO 2017), llevado a cabo en la Universidad de Alicante, cuyo objetivo es la creación de un corpus especializado de libre acceso en el ámbito de la economía y los negocios que permita diferentes tipos de explotación —con fines docentes, terminográficos o prácticos de traducción especializada— y, por lo tanto, sirva como una plataforma de transferencia de conocimiento entre el profesorado investigador, los traductores en formación y los profesionales de la traducción económica (Gallego Hernández 2013).

de corpus especializados expuestos en Bowker y Pearson (2002) y los completamos con los criterios de Torruella Casañas (2017) para la construcción de corpus diacrónicos.

2. El corpus *DIACOM* en español

2.1. En busca de la representatividad y el equilibrio

Todo corpus debe constituir un fiel reflejo de la realidad que pretende representar y los datos que de él se extraigan deben ser fiables y extrapolables a toda la población que este representa. Por lo tanto, la fase de diseño está siempre determinada por la necesidad de recoger una muestra representativa[4] y equilibrada de la población. Para ello, es necesario observar todos los parámetros que aseguren la representatividad y el equilibrio para, así, construir un corpus lo más neutro posible, es decir, "que recoja muestras proporcionales de todos sus aspectos (niveles, temáticas, registros, etc.)" (Torruella y Llisterri 1999: 46) de forma que se pueda analizar desde diferentes perspectivas y utilizar para diversas finalidades, así como que sea posible actualizarlo y reutilizarlo cuando se precise (*ibid.*).

En nuestro caso, *DIACOM* debe reflejar el lenguaje especializado del comercio internacional en español en tres periodos históricos (que justificamos *infra* en 2.3.8.): 1850-1914, 1945-1970 y 1990-2018. Teniendo este objetivo en mente, abordamos su diseño con el fin de construir un corpus representativo y equilibrado. Para ello nos planteamos una serie de cuestiones que nos ayudaran a garantizar estos aspectos. En primer lugar, y a diferencia de lo que ocurre en los corpus de lengua general, la adquisición de la representatividad y el equilibrio es una cuestión un tanto más simple (*cf.* Ahmad 1995: 73; Ahmad y Rogers 2001: 734), pues, en principio, el universo de estudio se restringe al dominio de especialidad y se simplifica la representación de las variedades, en concreto las referidas a los niveles diastráticos y diafásicos. No obstante, por un lado, el discurso especializado presenta una heterogeneidad de situaciones comunicativas (Cabré 1999: 118) y de tipos textuales (Ahmad 1995: 60), por lo que, para dar una cobertura adecuada del dominio objeto de estudio, hemos decidido incorporar al corpus diversos tipos textuales con diferentes grados de especialización de manera que se consiga representar de la forma más amplia posible la riqueza terminológica y conceptual del ámbito estudiado. Por otro lado, los campos de especialidad suelen ser interdisciplinares y multidisciplinares. Así, para adquirir el equilibrio en el corpus, ha sido necesario estructurar y establecer los límites del dominio del comercio para, de esta manera, seleccionar textos que pertenezcan

[4] Torruella Casañas (2017: 136-142) estructura la cuestión de la representatividad atendiendo, por un lado, a la representatividad cualitativa, determinada por la calidad y la diversificación de las muestras que componen el corpus, y, por otro, la representatividad cuantitativa —conocida en la bibliografía especializada como equilibrio— que depende de la cantidad de las muestras incorporadas y sus porcentajes de distribución en la -estructura del corpus. El autor divide esta última en equilibrio externo o "relación entre el número de muestras y el total de la población" (*ibid*. 138) y el equilibrio interno, a saber, "el número de muestras seleccionadas para cada apartado del corpus" (*ibid*.). De igual modo, Seghiri (2011) considera que un corpus especializado puede ser representativo cualitativa y cuantitativamente. La representatividad cualitativa está determinada por la calidad del material lingüístico que compone el corpus y queda garantizada a partir de los parámetros de diseño y el protocolo de su compilación —búsqueda y acceso de la documentación, descarga de datos, normalización y almacenamiento—. La representatividad cuantitativa, por su parte, está condicionada por la cantidad de muestras que se ha incorporado al corpus y se alcanza cuando la terminología básica del dominio queda cubierta. Esta se puede verificar *a posteriori* mediante la aplicación informática ReCor (Seghiri 2017 y Seghiri 2015), que parte de la base de que el número de *types* no aumenta en proporción al número de *tokens* una vez que se ha incorporado al corpus una cantidad concreta de textos.

tanto al ámbito especializado en general como a los subdominios que abarca.

En segundo lugar, por cuanto respecta a los tres cortes temporales establecidos, es claro que tendremos limitaciones a la hora de garantizar su representatividad[5] y mantener el equilibrio entre ellos, en particular, porque para el primer periodo histórico solo podremos incorporar aquellos textos que hayan pervivido y sean fácilmente recuperables —cuya distribución, a su vez, probablemente no sea proporcional en todos los países hispanohablantes—, o también porque la evolución del dominio de especialidad, por ejemplo, tras el advenimiento de internet y las nuevas tecnologías, ha dado lugar a la creación de nuevos subdominios que lógicamente no eran posibles en el siglo XIX.

Para lidiar con estos obstáculos inevitables, hemos abordado la labor de diseño definiendo claramente los parámetros del corpus y los criterios de selección de los textos en los que se va a estudiar el lenguaje del comercio,[6] de forma que se asegure la -comparabilidad entre los tres cortes sincrónicos.[7] A continuación, se exponen los criterios de diseño de *DIACOM* en español y se justifican, al mismo tiempo, las decisiones metodológicas que se han tomado.

2.2. Delimitación del dominio de especialidad

La elección de los textos de un dominio de especialidad concreto puede resultar realmente complicada si el ámbito en cuestión es multidisciplinar o interdisciplinar, como es nuestro caso. El comercio es una actividad socioeconómica que consiste en el intercambio de bienes y servicios, que se desarrolla en diferentes sectores, pero en el que también se debe velar por la protección y la satisfacción de los consumidores, sin olvidar lo relacionado con la promoción del producto. El primer paso para delimitar el dominio de especialidad fue realizar un estudio del campo del comercio para, así, identificar sus características y establecer los límites que nos íbamos a fijar dentro de su estructura global y facilitar, de esta manera, la búsqueda y la selección de los textos. Para ello, elaboramos una clasificación temática a partir de bibliografía de referencia del ámbito (en particular, Buckley y Lessard 2005 y Zettinig y Vincze 2011) y, gracias a la consulta con un especialista,[8] establecimos una estructura conceptual del campo del comercio.

Debido al hecho de que se trata de un proyecto desarrollado en Italia y enfocado a contextos extranjeros, es decir, los países de habla española y francesa, hemos atendido a las clasificaciones propias del dominio del comercio internacional. Sin embargo, puesto que nuestro interés es histórico-terminológico, hemos decidido abarcar también textos sobre el comercio interior en los países

[5] En este punto, tenemos que aceptar que un corpus histórico totalmente representativo es "una construcción empíricamente imposible" (Kabatek 2013: 9), pues no se puede determinar el universo de estudio en función de los textos que se conservan, es decir, desde el punto de vista estadístico, un corpus histórico solo permitirá observar el comportamiento de la lengua de los textos que han pervivido y no de la lengua en general (Torruella Casañas 2017: 258).

[6] En la fase de diseño, hemos tenido en cuenta, asimismo, la posterior clasificación de los textos en función de sus rasgos definitorios, pues estos, como indica Torruella Casañas (2017: 137), "deben responder a las posibles variables que se quieran utilizar en las investigaciones", que en *DIACOM* serán la temporal, la geográfica y la textual.

[7] Como señala Enrique-Arias (2012: 96): "Una paradoja de la composición de los corpus diacrónicos es que, por un lado, deben ser heterogéneos (tienen que incluir textos de diferentes autores, épocas, géneros, registros, dialectos) y a la vez deben ser -homogéneos (es decir, los diferentes cortes sincrónicos representados en el corpus tienen que ser comparables entre sí)".

[8] Se trata del Dr. Fabio Cassia (Università degli Studi di Verona), a quien agradecemos su generosa ayuda a la hora de delimitar el dominio y establecer la tipología textual.

considerados. De hecho, creemos que los documentos vinculados con la realidad comercial de un país concreto del mundo hispánico proporcionan datos interesantes, por ejemplo, permiten comprobar casos de variación terminológica entre un país y otro, también a lo largo del tiempo. Además, es reseñable que nuestra propuesta abarca, entre otros, documentos de ámbito empresarial (*infra* 2.3.5.), lo cual permite ofrecer muestras textuales de las actividades comerciales entre empresas o entre una empresa y un particular, como pueden ser contratos o albaranes, por ejemplo.

Según Buckley y Lessard (2005), el ámbito del comercio internacional se vertebra en dos ejes, por un lado, los niveles de análisis y, por el otro, las disciplinas y temas:

(a) Los niveles de análisis principales son por lo menos cinco[9] y, atendiendo a una organización de lo general a lo particular (*cf.* -Zettinig y Vincze 2011), resultan ser:

1) macro (comercio global entre países, macroáreas, etc.);
2) sector (comercio e internacionalización de los sectores);
3) empresa (estrategias comerciales y de internacionalización de la empresa);
4) actividades y funciones para el comercio internacional y la internacionalización de la empresa (mercadotecnia, finanzas, etc.);
5) ejecutivos y empresarios (formación, competencias, etc.).

(b) Las disciplinas y temas principales, según la Academy of International Business (s/f), son 16:[10]

A. Economics
B. Finance
C. Accounting & Taxation
D. Organization
E. Management
F. Business Policy
G. Marketing
H. Human Resources & Industrial Relations
I. Law
J. International Relations and Political Science
K. Social Issues
L. Economic & Business History
M. Country or Area Study
N. Industry/Sectorial Study
O. Policy-Oriented Study
P. Education & IB [international business]

Compaginando los dos niveles propuestos en la bibliografía de referencia, hemos creado un primer árbol de campo del ámbito comercial:

[9] También existe un sexto nivel relacionado con los "procesos de internacionalización", que agrupa aportaciones teóricas nuevas y específicas sobre los mecanismos de internacionalización.

[10] La Academy of Internacional Business añade también en su clasificación la variable *Q. Research Areas Not Covered By Groupings*. Para cada uno de estos *discipline interests*, la Academy of International Business propone también subniveles relacionados con intereses de investigación concretos.

Niveles de análisis	Disciplinas y temas
Macro (mundo, países)	Relaciones internacionales / Política comercial
	Países / Áreas de estudio
	Macroeconomía
	Estudios orientados a las políticas
	Aspectos sociales
Sector	Estudios industriales y sectoriales
	Recursos humanos y relaciones industriales
Empresa	Políticas de la empresa
	Gestión de la empresa
	Historia de la economía y del comercio
	Organización
Actividades y funciones	Marketing
	Finanzas
	Contabilidad e impuestos
	Derecho
Ejecutivos y empresarios	Educación y comercio internacional

Dada la amplitud y la enorme diversidad que abarca el comercio (incluye aspectos económicos, legales, políticos, etc.), así como su capacidad de actualización a raíz de los avances sociales, técnicos y tecnológicos y la globalización, que no solo han dado lugar a la creación de nuevos subdominios (como el comercio electrónico), sino también a la modificación de las relaciones comerciales, la gestión de los procesos, las formas de contratación, los medios de transporte, etc., ha sido necesario, en una segunda fase, realizar un segundo árbol de campo simplificado a partir del primero. Así, hemos efectuado una estructuración del ámbito de especialidad, contrastada con el experto consultado, estableciendo tres grandes campos temáticos: "macroeconomía y economía internacional", "sectores" y "empresa", cada uno de los cuales se divide, a su vez, en diferentes subdominios.

Esta segunda estructura conceptual —que es la que vamos a seguir en nuestro proyecto— simplifica los campos de los dominios y de los subdominios con el doble propósito de paliar los (inevitables) solapamientos en la clasificación y de crear una estructura clasificatoria de más fácil uso para el usuario externo. De hecho, esta clasificación basada en dos niveles se empleará para configurar la base de datos que va a almacenar los textos de *DIACOM* para que, una vez compilado el corpus, las categorías establecidas se puedan utilizar como variables que permitan múltiples posibilidades de análisis en función de si se interroga el corpus en general o si se restringe la consulta a campos más específicos del primer o segundo nivel. El resultado de este proceso taxonómico y conceptual es el siguiente árbol de campo, en su versión española:

Dominios (1.er nivel en la base de datos)	Subdominios (2.º nivel en la base de datos)
Macroeconomía y economía internacional (Comercio global entre países, macroáreas, comercio dentro de un país determinado, etc.)	Relaciones internacionales / Política comercial Países / Áreas de estudio Aspectos sociales
Sectores (Comercio e internacionalización de sectores específicos)	Productos Servicios
Empresa (Estrategias de internacionalización y comerciales de la empresa; actividades para el comercio internacional y la internacionalización de la empresa)	Administración (*management*) *Marketing* Logística Comercio electrónico Derecho

Esta simplificación temática, por un lado, facilita la búsqueda de los documentos y permite seleccionar, de forma controlada, solo aquellos que guarden relación directa con los subdominios implicados. Por otro lado, teniendo en cuenta que las disciplinas evolucionan y se desarrollan y, por lo tanto, la estructura que presentan en la actualidad no tiene por qué corresponder a la de épocas anteriores, se trata de una estructura conceptual lo suficientemente amplia para permitir mantener la homogeneidad temática necesaria —especialmente en el primer nivel— no solo entre los componentes del subcorpus español, sino también entre el subcorpus francés y el español.

Una vez restringido el dominio, hemos procedido al diseño del corpus definiendo sus características generales y los atributos textuales que deben contener las muestras que vamos a incorporar, pues, como se afirma en la bibliografía especializada, la calidad del proyecto terminográfico y, por ende, de sus resultados está directamente relacionada con la calidad de la documentación en la que se basa (Bowker 1996: 42; Meyer y Mackintosh 1996: 264, entre otros). Como explicamos más adelante (*infra* 2.3.5.), para representar la riqueza textual del conjunto de actividades que conforman el ámbito del comercio, recogeremos documentos de distinta índole.

2.3. Criterios de diseño del corpus *DIACOM*

Un corpus realmente útil es aquel que se compone al menos de dos elementos: el corpus textual en sí y su arquitectura[11] e interfaz de consulta (Davies 2009: 139). Por ello, la determinación de los parámetros en la fase de diseño no solo es fundamental para llevar a cabo las búsquedas y la selección de los textos que se van a incorporar al corpus, sino que también es esencial para la posterior gestión del material lingüístico que lo conforma: a partir de los parámetros se

[11] La *arquitectura* del corpus engloba tanto su *estructura*, a saber, los ejes del corpus, y sus divisiones, en torno a los cuales se disponen las muestras textuales, como la organización de los *parámetros de selección*, que permiten, en la herramienta de consulta, la selección y restricción de criterios para afinar las búsquedas (Torruella Casañas 2017: 69, n. 57).

configura la base de datos que almacena los textos y, una vez compilado el corpus, se emplean como variables para realizar búsquedas más o menos complejas de forma automática mediante un programa de gestión.

El diseño de *DIACOM* prevé, como aconseja Torruella Casañas (2017: 63), "una arquitectura multifuncional y con una base representativa suficiente de *muestras*, para que pueda ser consultado con diferentes fines". Así pues, con la intención de que se adapte lo mejor posible a las necesidades de múltiples usuarios, hemos tenido en cuenta la inexorable variación del lenguaje natural y, por ende, de los lenguajes de especialidad, manifestada en las dimensiones diatópica, diacrónica y comunicativa. De esta manera, prevemos un corpus que no solo refleje el uso de la lengua en el dominio de especialidad del comercio en toda su amplitud, sino que, a su vez, permita el manejo de datos a partir de búsquedas globales y específicas en dichos ejes, así como también su combinación.

2.3.1. Tamaño del corpus

No nos hemos fijado un número concreto de palabras al inicio del proyecto, sino que, una vez definidos los criterios específicos de diseño, se recopilarán las muestras que cumplan los requisitos estipulados. No obstante, consideramos que, teniendo en cuenta la delimitación del dominio de especialidad que nos proponemos estudiar y la disponibilidad de los textos, así como el tiempo y los recursos de los que dispongamos a lo largo del desarrollo del proyecto, *DIACOM* se clasificará como *corpus restringido*, es decir, aquel compuesto por "un número limitado de textos, bien estructurados y representativos, con la finalidad de que sean manejables y de poder desarrollar, con un coste razonable, procesos de post-edición (etiquetado, codificación, lematización, etc.)" (Torruella Casañas 2017: 47).

De hecho, no existen reglas estrictas ni fórmulas concretas, así como tampoco hay consenso en la bibliografía especializada (*cf*. Corpas Pastor y Seghiri Domínguez 2006) para determinar *a priori* el número de palabras que debe contener un corpus para constituir una muestra representativa del lenguaje que pretende reflejar. Si bien es cierto que, mientras que para algunos autores lo ideal sería construir un corpus lo más grande posible y que pudiese continuar creciendo (Sinclair 1991: 18), en los casos de los corpus especializados se acepta un tamaño más restringido e, incluso, si están bien diseñados, comienzan a ser útiles a partir de decenas de miles de palabras (Ahmad y Rogers 2001: 735-736; Bowker y Pearson 2002: 48).[12]

A este respecto, y a pesar de no marcar una cifra concreta al inicio del proyecto, sí tendremos en cuenta *a posteriori*, a partir de la observación del material que sea posible recuperar, la adquisición del equilibrio entre los componentes del corpus basado en el reparto equivalente de los textos.[13] Es decir, en la construcción de *DIACOM*, consideramos más importante mantener una distribución

[12] Para justificar este argumento, Ahmad y Rogers (2001: 735-736), por ejemplo, aluden, de una parte, a las características intrínsecas de los textos especializados: la autoridad de las fuentes, la menor variación léxica y gramatical, la restricción del ámbito de estudio y la mayor densidad terminológica de los lenguajes especializados; y, de otra, a las características del proyecto: los intereses propios de los terminólogos y los objetivos de la investigación.

[13] Este tipo de muestreo suele ser el idóneo para los corpus históricos dado que, a diferencia del *reparto proporcional*, donde sí se debe conocer o, por lo menos, intuir el total de la población para que la cantidad de muestras que componga cada apartado del corpus esté repartida en proporción numérica a su distribución real en la población (Torruella Casañas 2017: 140), el *reparto equivalente* "no requiere conocer el total de la población, ya que implica que las cantidades de palabras que componen cada apartado sean iguales o parecidas, prescindiendo de que haya correspondencia porcentual entre las diferentes partes en el corpus y entre las

textual apropiada en todos los apartados —ejes diatópico, diacrónico y tipológico— del corpus, siempre que la disponibilidad de los textos lo permita.

Asimismo, otra cuestión relacionada con el tamaño del corpus es su capacidad de actualización, es decir, en la fase de diseño es preciso establecer si el corpus será abierto o cerrado, a saber, si se trata de una colección de textos actualizable o finita, respectivamente. *DIACOM* se construye como corpus cerrado: su compilación finalizará en el marco del proyecto de investigación y se pondrá a disposición de la comunidad científica a finales de 2022, una vez comprobados los aspectos de representatividad y equilibrio. No obstante, el diseño del corpus no descarta su ampliación a partir de nuevos cortes temporales ni su actualización con nuevo material lingüístico en los tres periodos establecidos. Por ello, se está llevando a cabo una documentación detallada de cada una de las fases para garantizar no solo la fiabilidad del material lingüístico, sino también asegurar una futura actualización uniforme con material que se ciña a los criterios específicos de diseño propuestos.

2.3.2. Número de textos

Si bien el tamaño final de un corpus siempre se referirá a la cantidad de palabras que lo constituyen, en la fase de diseño también se debe tener en cuenta la selección de un amplio número de textos procedentes de diferentes autores como medio para asegurar la representatividad del corpus y mantener su equilibrio. Así, si un corpus está constituido por un número reducido de textos, un único texto puede condicionar los resultados de un análisis o, si solo se seleccionan textos de un único autor o un par de autores distintos, los resultados también quedan expuestos a su idiosincrasia lingüística. A este último respecto, también es preciso distinguir entre publicaciones escritas íntegramente por un solo autor —sea este individual o corporativo—, y aquellas que constan de diferentes apartados o capítulos, cada uno de ellos redactado por un autor diferente. Mientras que las primeras siempre serán consideradas un texto único, las segundas se pueden clasificar bien de forma conjunta, como un texto compuesto, bien de forma individual, donde cada una de sus secciones funciona como texto único (Pearson 1998: 60).

En definitiva, como sentencia Torruella Casañas (2017: 141), “tan importante como la frecuencia de un hecho lingüístico en los distintos apartados del corpus es su distribución entre las distintas obras del apartado”. Por lo tanto, a pesar de que, al igual que con la cantidad de palabras, tampoco nos hemos marcado en esta fase inicial una cifra de textos determinada, sí tenemos claro que se incorporará a *DIACOM* un amplio número de muestras escritas por diferentes autores como mejor opción para representar la heterogeneidad terminológica del lenguaje de especialidad estudiado. Al mismo tiempo, cada sección de las obras colectivas será clasificada como texto individual,[14] pues tanto el contenido como las particularidades lingüísticas del autor varían.

diferentes partes y el todo” (*ibid.* 141).

[14] Los periódicos constituyen un caso aparte, ya que se clasificarán por lo general de manera colectiva. De hecho, hay que considerar que cuantiosos artículos periodísticos, sobre todo en el pasado, se publicaban sin firma y que, a menudo, los textos publicados eran muy breves. Por lo tanto, las publicaciones periódicas —o las secciones de diarios y revistas— dedicadas al comercio se clasificarán como conjunto, número por número.

Por último, y también en relación con la cantidad de muestras textuales, hemos tomado la decisión de incorporar todas las ediciones —no reimpresiones— de una misma obra, en caso de que se encuentren varias dentro de uno de los periodos temporales considerados. Esto permitirá estudiar los posibles cambios entre versiones derivadas de la evolución en el uso y/o significado de los términos, entre otros aspectos.[15]

2.3.3. Tamaño de los textos

El tamaño de las muestras que se incorporarán al corpus es también un aspecto que debe considerarse en la fase de diseño. Aquí la cuestión reside en si se van a recopilar textos completos o solo fragmentos de textos de una cantidad específica de palabras. A pesar de que el desarrollo informático y el subsiguiente aumento de la capacidad de los ordenadores en los últimos 30 años han favorecido la incorporación de textos completos y el almacenamiento de miles de millones de palabras, se sigue considerando la construcción de corpus a partir de fragmentos de obras[16] por las ventajas que presentan a la hora de mantener el equilibrio entre sus diferentes apartados.[17] Para la elaboración de corpus especializados, sin embargo, existe consenso en la necesidad de no limitar la muestra a un número determinado de palabras, sino de incluir el texto íntegro (Ahmad 1995: 61; Bowker 1996: 43; Meyer y Mackintosh 1996: 268; Pearson 1998: 59; Bowker y Pearson 2002: 49). La incorporación de textos completos es fundamental para el estudio de los lenguajes de especialidad, pues la información lingüística, conceptual y pragmática de las unidades terminológicas o fraseológicas puede aparecer en cualquier parte de un documento.

De acuerdo con lo anterior, las muestras que compondrán *DIACOM* serán textos íntegros. De esta manera, también se conseguirá un corpus "más abierto y apto para el estudio de un amplio abanico de aspectos lingüísticos" (Torruella y Llisterri 1999: 17). Esto implica, por lo tanto, que existan diferencias de tamaño entre las muestras textuales, pues su extensión variará de forma -significativa en función del contenido. No obstante, como se acaba de exponer, no existe motivo para justificar la necesidad de que los documentos sean uniformes, dado que esto podría desvirtuar su contenido terminológico.

Un último aspecto referido a la integridad de los textos de *DIACOM* es la decisión de prescindir, antes de incorporar de forma definitiva los textos al corpus, de algunas partes que no aporten datos de interés lingüístico ni conceptual para el análisis del corpus y que puedan descompensarlo.[18] Para ello, se ha creado

---

[15] Esta decisión, además, se toma con miras a futuras ampliaciones del corpus, en concreto, a la eventual incorporación de textos legislativos. Así, a pesar de que su validez jurídica solo responde a su periodo de vigencia, la inclusión de todas las versiones de, por ejemplo, un código de comercio puede ser fundamental para observar la evolución de términos jurídicos, tanto desde el punto de vista denominativo como conceptual.

[16] En terminología de Torruella y Llisterri (1999: 12), Corpas Pastor (2001: 158-159) y Torruella Casañas (2017: 53-55) este tipo de corpus se denomina *corpus de referencia*, frente a los *corpus textuales*, compuestos por obras completas, y los *corpus léxicos*, que recogen fragmentos muy pequeños de idéntica longitud y cuya finalidad es exclusivamente lexicográfica.

[17] Como señala Torruella Casañas (2017: 142, nota al pie 138), el problema del equilibro en un corpus pequeño tiene un valor diferente al que presenta en un corpus grande, ya que la cantidad de palabras establecida para cada uno de los apartados del corpus puede ocuparse con una única obra muy extensa, lo cual, al mismo tiempo, restaría representatividad al corpus.

[18] Así, por ejemplo, una tesis doctoral o un manual de comercio electrónico puede albergar un capítulo dedicado al nacimiento de internet o a su desarrollo. En estos casos, se elimina dicho capítulo pues, al mismo tiempo que aumenta el número de palabras del apartado en el que se clasifica, podría introducir una cantidad de nuevos *tokens* referidos al campo de la informática. Esta disciplina, si bien tiene relación directa con el comercio (interdisciplinariedad), no forma parte del dominio específico en sí.

un protocolo de limpieza de textos con los siguientes elementos:

- portada;
- índices;
- encabezados y pies de página;
- numeración de los apartados;
- figuras, tablas, esquemas, cuadros de texto y gráficos, así como sus títulos y pies;
- notas al pie de página, con sus números dentro del cuerpo textual, cuando estas incluyan referencias bibliográficas u otros datos que no sean de interés terminológico;
- fórmulas matemáticas;
- referencias bibliográficas;
- datos del autor e información editorial;
- firmas (p. ej. países que suscriben los tratados internacionales);
- anexos;
- resúmenes que estén en idiomas diferentes del español.

No obstante, estos criterios de eliminación son indicativos y no estrictos, pues es posible que encontremos excepciones con valor terminológico.[19]

2.3.4. Medio de producción de los textos

Atendiendo al medio de producción de los textos, *DIACOM* va a ser un *corpus escrito*, es decir, solo va a estar compuesto de material escrito. A pesar del interés lingüístico que pueda suscitar el análisis de conversaciones entre expertos o entre expertos y legos, lo más habitual en la construcción de corpus especializados es recopilar muestras escritas. Esto se justifica por dos motivos fundamentales: (i) el tiempo y el esfuerzo que requiere compilar un corpus oral debido al proceso de grabación y la posterior transcripción (sea esta ortográfica, fonética o fonológica) de las muestras orales y (ii) las dificultades asociadas a la recolección del material oral que señalan Bowker y Pearson (2002: 50): “For example, if the speakers are aware that you are recording their conversation, they may be intimidated and be careful about what they say. This means that your language sample may not be completely natural. In contrast, if you were to record a conversation without the speakers’ knowledge, you would get a more natural sample, but this type of practice raises many ethical questions”.

La decisión de descartar la recopilación de material oral se justifica, asimismo, por la necesidad de mantener el equilibrio entre los tres cortes temporales considerados, pues conseguir material no escrito sería imposible en los periodos referidos al siglo XIX y a buena parte del XX.

2.3.5. Tipos de textos

Debido al amplio espectro de manifestaciones textuales en el ámbito del comercio (Álvarez García 2017: 119), así como a la necesidad de establecer una tipología que fuese útil para los propósitos de nuestra investigación y válida para

[19] Por ejemplo, las notas al pie de página que contengan información de interés se mantendrán, pero se emplazarán al final del documento a modo de lista con el propósito de que no interrumpan la coherencia textual.

todos los cortes temporales, centramos nuestro diseño en cuatro tipos textuales,[20] a saber: textos institucionales, textos empresariales, textos académicos[21] y textos periodísticos. Esta clasificación se ha establecido teniendo en cuenta el emisor del texto (institución pública, empresa privada, ámbito académico y prensa) y se ha constatado su idoneidad y utilidad con el experto en la materia consultado. No se trata, por lo tanto, de una tipología textual exhaustiva —pues nuestra intención no reside en analizar los tipos de textos que se producen en el ámbito comercial ni establecer un inventario textual en función de, entre otros, la situación comunicativa y el contenido (para ello, *vid.* Álvarez García 2017)—, sino de una clasificación funcional de los textos producidos por la comunidad de expertos para facilitar la selección de las muestras dentro de nuestro ámbito de especialidad y que, al mismo tiempo, permita su clasificación exclusiva en uno de los tipos considerados, así como su ampliación *a posteriori* a través de la inclusión de otras tipologías textuales sin mayores modificaciones en los parámetros de diseño —por ejemplo, con textos legislativos—.

Somos conscientes de que los cuatro tipos textuales considerados no son estancos y que, a la hora de clasificar los documentos, se podría producir, en principio, algún solapamiento. Para paliar este aspecto, atenderemos a determinados criterios. En primer lugar, las publicaciones periódicas de corte académico (por ejemplo, una revista de *marketing* internacional) se clasificarán como textos académicos, ya que mantienen dos características fundamentales de dicha categoría, a saber, el público meta (la comunidad científica) y el formato de la publicación (el artículo científico). En segundo lugar, por lo que se refiere al caso —más problemático— de los textos institucionales, clasificaremos bajo dicha categoría todos los documentos internos de las entidades en cuestión (como los -memorándums), además de los documentos que se dirigen al público externo si forman parte del cometido y de la responsabilidad social de la institución que los produce: es este el caso, por ejemplo, de un informe de una cámara de comercio publicado en su página web. En cambio, las publicaciones periódicas patrocinadas por instituciones que recogen artículos divulgativos, reportajes etc. se clasificarán como textos periodísticos por compartir con estos el formato.

Cada uno de los cuatro tipos textuales incluidos en nuestro corpus, a su vez, albergará diferentes géneros que, siguiendo a Cassany (2004b: 42-43), se definen, dentro de los lenguajes de especialidad, como aquellas unidades de comunicación desarrolladas sociohistóricamente en el ámbito de una actividad laboral específica y que presentan una serie de rasgos léxicos, gramaticales, discursivos y pragmáticos que las caracterizan. No obstante, en esta primera fase de diseño, no se han determinado los géneros que se van a incorporar al corpus, ya que nuestra intención es mantener este campo lo más abierto posible para dar cabida a los diferentes géneros representativos dentro de las tipologías textuales en los tres cortes temporales.[22] Esta decisión se justifica, a su vez, por el interés en dar

[20] Sin profundizar en la controversia que presenta esta noción en la lingüística textual (*vid*. Heinemann y Heinemann 2002: 140-156 y Ciapuscio 2005), distinguimos entre *tipo textual* y *género textual* (Heinemann y Heinemann 2002, Heinemann y Viehweger 1991). Recordemos que por *tipo textual* se entiende aquella categoría ligada a una teoría que sirve para clasificar científicamente los textos; se refiere a una forma específica de textos, y se describe y define dentro del marco de una tipología textual o discursiva (Heinemann y Viehweger 1991: 144).

[21] Los textos académicos, debido a la interdisciplinariedad del ámbito de especialidad, en ocasiones son redactados por economistas, juristas, historiadores, etc. No obstante, por tratarse de emisores especializados en un dominio íntimamente conectado con el comercio, el grado de especialización del texto, con la consiguiente adecuación terminológica y conceptual, está asegurado.

[22] Como advierte Enrique-Arias (2012: 96), "a la hora de diseñar un corpus debemos asegurarnos de que, para cada periodo representado, estamos caracterizando estados de la lengua y no meras tipologías textuales".

cobertura a la mayor variedad de géneros textuales que hagan posible una descripción amplia del lenguaje especializado del comercio en todas sus facetas y en el que estén representadas las diversas situaciones comunicativas que se producen en el dominio: el nivel más alto de especialización (informes, artículos científicos, etc.), el nivel semiespecializado (p. ej. manuales) y el nivel de baja especialización (prensa).[23] De esta manera, se podrá garantizar el equilibrio del corpus, pues a pesar de las diferencias internas que se puedan dar entre periodos o países, sí será posible mantener el equilibrio entre tipologías textuales. Por último, hemos optado por esta clasificación abierta y flexible debido a la necesidad de que la representación textual en todas las épocas sea lo más homogénea posible. Es claro que no siempre se podrá disponer de la misma cantidad de géneros[24] en todos los periodos, pero, por lo menos, los tipos textuales sí son habituales en todos ellos.

2.3.6. Autoría y autoridad de las fuentes

Para garantizar la calidad de un corpus especializado es imprescindible asegurar la autenticidad del material lingüístico que se va a incorporar al mismo. Así, los autores de los textos que se recojan deben ser expertos en el ámbito de estudio, es decir, deben tener la formación académica adecuada y/o la experiencia profesional en la materia y, además, disfrutar del reconocimiento de otros compañeros de profesión (Pearson 1998: 60; Bowker y Pearson 2002: 51). Al mismo tiempo, como señalábamos *supra*, es igualmente necesario seleccionar textos que hayan sido redactados por un amplio número de autores para, así, evitar preferencias particulares o idiosincrasias lingüísticas. Nuestro diseño trata de cumplir con estos criterios de calidad. Para ello hemos seleccionado tipos textuales que, con toda certeza, fueron elaborados por expertos en la materia[25] y que permiten, a su vez, la incorporación de textos redactados por múltiples autores. Dentro de los textos empresariales e institucionales, en ocasiones, es difícil establecer la identidad del autor o de los autores, pues están suscritos por un organismo o institución. No obstante, la calidad de los mismos no se puede cuestionar, dado que han atravesado un proceso de elaboración, en el que probablemente habrán intervenido distintos profesionales, hasta desembocar en el texto final publicado que se incorporará al corpus.

El parámetro anterior (autoría) se encuentra intrínsecamente relacionado con el criterio de autoridad de las fuentes de las que se seleccionan los textos: no solo el autor debe ser un experto reconocido en el ámbito en el que escribe, sino que las fuentes de las que se extraigan los documentos deben gozar también de cierta reputación. Los tipos de textos que vamos a incorporar a *DIACOM* cumplen con este criterio, pues todos han sido sometidos —en el caso de los documentos

---

[23] Dentro de la prensa, no obstante, podemos encontrar textos más o menos especializados, pues el discurso de un periodista que escribe artículos de interés comercial en un diario generalista no es el mismo que el de un experto que escribe en una revista especializada de comercio.

[24] Como es consabido, los géneros textuales no son estáticos, sino que surgen y varían a lo largo de los años en función de las necesidades comunicativas.

[25] Debido a la multidisciplinariedad e interdisciplinariedad del campo estudiado, en algunos casos, los autores de los textos académicos son economistas, juristas, ingenieros, políticos, etc. que se dedican al comercio en alguno de sus subdominios o, incluso, las revistas de las que se extraen los textos se insertan en otros ámbitos de especialidad. En estos casos, también se han seleccionado los documentos, pues no se puede cuestionar el nivel de experticia de sus autores.

institucionales, académicos y periodísticos— a un filtro editorial o corporativo de algún tipo. En el caso de los textos empresariales, son todos documentos auténticos que se utilizan o se utilizaron en el ámbito estudiado. La selección de las muestras se realizará, asimismo, a través de páginas electrónicas de bibliotecas digitales, instituciones y organismos nacionales e internacionales, revistas y publicaciones periódicas, etc., de manera que también se podrá validar la fiabilidad de las fuentes.

2.3.7. Lenguas y procedencia geográfica de las fuentes

Las lenguas que contemplamos en la construcción de *DIACOM* son el español y el francés. En concreto, se trata de un corpus bilingüe comparable que recoge textos con características similares en estos dos idiomas. Así, cada uno de los subcorpus monolingües que conforma *DIACOM* se ha diseñado de acuerdo con criterios semejantes de selección de muestras referidos al dominio de especialidad, tipo de texto y fecha de publicación.

Los textos que incorporamos suelen ser textos originales en una de estas dos lenguas. No obstante, no se descartan los textos traducidos por tres motivos: (i) porque muchos de los textos especializados de los siglos pasados, a veces de carácter fundacional en el ámbito de una disciplina, han entrado en el español como traducciones de otras lenguas, aunque cabe considerar que no siempre el idioma del texto que se traducía era el en que se había redactado el texto original (de hecho a menudo el francés sirvió, sobre todo en los siglos XVIII y XIX, como lengua de mediación); (ii) las traducciones también son una fuente de entrada de neología (*vid.*, entre otros, Gómez de Enterría 1999); (iii) los textos procedentes de instituciones supranacionales o internacionales suelen ser redactados en una de las lenguas procedimentales de la institución y, posteriormente, traducidos a las demás. Por este motivo, consideramos la distinción entre texto original, texto traducido y, para casos de difícil constatación, desconocido; así los usuarios podrán seleccionar las opciones que más les convengan en la herramienta de consulta del corpus.

En cuanto al subcorpus español, se ha previsto incorporar textos de España y de todos los países de lengua española en América. A su vez, este subcorpus también podrá considerarse un corpus comparable en sí mismo,[26] pues recopila textos en diferentes variedades con características similares (EAGLES 1996: 12; Torruella y Llisterri 1999: 11; Hunston 2002: 15) y podrá utilizarse para llevar a cabo análisis intralingüísticos de diversa índole.

Para su almacenamiento en la base de datos, así como la posterior consulta del corpus a través de la herramienta de búsqueda, hemos establecido una doble clasificación de los documentos según el país de procedencia del autor, de una parte, y, de otra, el lugar de publicación/redacción del texto,[27] siempre que su

[26] En la bibliografía especializada podemos encontrar numerosas terminologías para denominar estos componentes monolingües*: corpus* (Hunston 2002), *subcorpus* (McEnery *et al*. 2006) o, de forma genérica, *componentes* (*ibid*.). En cualquier caso, la elección de un término u otro depende del diseño que lleve a cabo el investigador que compile el corpus.

[27] Se ha desechado la clasificación por variedades o dialectos no solo por las dificultades aducidas por Torruella Casañas (2017: 86-91), hecho que implica "un conocimiento histórico y sociológico de los textos y de cada centro de producción de documentos y/o de cada área geográfica establecidos en la concepción de los diferentes apartados diatópicos definidos en el corpus" (*ibid.* 87), sino porque la compilación del corpus *DIACOM* no se emprende con la pretensión de contribuir a la descripción de todos los rasgos lingüísticos del español en general, sino solo de aquellos que terminológicamente tienen relevancia para el dominio de especialidad estudiado y, por lo tanto, esta clasificación metodológica nos parece la más eficaz para nuestros propósitos.

determinación sea posible.[28] Así, el usuario podrá utilizar la variable que mejor se acomode a la finalidad de su estudio y, al mismo tiempo, podrá también definir sus consultas en función de si le interesa obtener resultados globales u observar la distribución de un fenómeno en un espacio geográfico determinado. Será posible, por ejemplo, realizar sondeos sobre las zonas lingüísticas habituales para el español en América: México y Centroamérica, el área caribeña (las Antillas), la del Caribe continental (Colombia y Venezuela), la andina, la chilena y el Río de la Plata.[29]

No obstante, existen ciertas dificultades asociadas a estos criterios, que exponemos a continuación. Por un lado, la procedencia del autor no siempre es conocida y, al mismo tiempo, no se puede tomar como argumento absolutamente fiable para la clasificación del documento dentro de un país concreto. Para el caso de los textos institucionales y empresariales, así como de los textos periodísticos, se espera que mantengan las convenciones terminológicas establecidas en el país en el que se publican: los primeros porque se emiten en el seno de una institución, organismo oficial o entidad privada y, por lo tanto, reflejan la terminología normalizada en el ámbito de especialidad en ese territorio;[30] los segundos, ya que se redactan en función de un público meta que se encuentra en un país o región concretos. Más problemático es el caso de los textos académicos, principalmente en el periodo de 1990 a 2018. Así, la adscripción universitaria del autor o el lugar de publicación de la revista no siempre se corresponden con la procedencia del autor y, por lo general, no suelen coincidir.[31] En estos casos, los textos se clasifican en la base de datos de acuerdo con la adscripción universitaria del autor, pues sobrentendemos que este se adecuará terminológicamente al ámbito donde ejerce su profesión.[32]

Por otro lado, la delimitación político-administrativa de los diferentes estados ha sufrido cambios a lo largo de la historia, por lo que el territorio que comprende un estado no siempre es el mismo en las tres franjas temporales estudiadas. Así, para organizar los textos en nuestra base de datos de forma que fuese posible recuperar la información de la manera más sencilla, se nos planteaba el problema de su clasificación teniendo en cuenta (i) el territorio que comprenden los estados en la actualidad (solución a), (ii) la denominación actual del territorio en el que se ha publicado/redactado el texto (solución b), o (iii) la denominación del territorio —lugar de publicación/redacción— tal y como consta en el texto (solución c). Por ejemplo, si trasladamos estos planteamientos a un caso concreto como el de la República de Colombia (*vid*. Palacios y Safford 2002), encontramos que, en el primer corte temporal estudiado (1850-1914), se produjeron numerosos cambios en su territorio y denominación oficial: República de la Nueva Granada (1831-1858), Confederación Granadina (1858-1863), Estados Unidos de Colombia

[28] Los textos en los que la procedencia del autor o el lugar de publicación/redacción no pueda ser establecidos con seguridad, se clasificarán bajo la variable "desconocido".

[29] Se trata de las "zonas lingüísticas habituales" consideradas para Hispanoamérica en el *Corpus del Español del Siglo XXI* (*CORPES XXI*).

[30] Para la clasificación de textos procedentes de instituciones internacionales o supranacionales se tendrán en cuenta los países miembros del organismo en cuestión, pues su validez se extiende solo a estos.

[31] Sin olvidar, por supuesto, el idiolecto de los autores que implica numerosas variables de las que no podemos tener control: estudios superiores cursados en una universidad distinta a la de su país de origen, preferencia por extranjerismos, etc.

[32] En los casos de trabajos en coautoría se toma en consideración la adscripción académica del primer autor.

(1863-1886) y República de Colombia (desde 1886). Si tuviésemos en cuenta el territorio que comprenden los estados en la actualidad (solución a), se deberían clasificar los textos extraídos entre 1850 y 1858 procedentes de la República de la Nueva Granada, por ejemplo, tanto en Colombia como en Panamá. Por lo tanto, para los campos referidos a Colombia y a Panamá en la base de datos deberían constar varios subapartados en función de la denominación —junto con las fechas de duración— que recibiese el territorio históricamente. Si se clasificase el texto en función de la denominación que recibe el territorio actual correspondiente al lugar de publicación original (solución b), se estaría ignorando el ámbito de validez del documento en el momento de su redacción. De esta manera, si el lugar de publicación original se sitúa en la actual Panamá, en la base de datos se clasificaría en este campo, pero se omitiría el hecho de que también era válido en el territorio de la actual Colombia. Por último, la clasificación en función de la denominación político-administrativa del territorio vigente en la época de publicación del texto (solución c), opción por la que nos hemos decantado, permite recoger como campos independientes en la base de datos los estados según la denominación que consta como lugar de publicación/redacción del documento. Creemos que esta categorización es la que menos problemas metodológicos plantea[33] y la que más opciones de consulta permite al usuario, quien podrá interrogar el corpus bien en su totalidad bien combinando los criterios que más se adapten a la finalidad que persigue con su investigación.

2.3.8. Periodos de tiempo

DIACOM recogerá tres cortes temporales: 1850-1914, 1945-1970 y 1990-2018. Dichos cortes cronológicos tienen una motivación extralingüística a partir de hitos históricos y eventos sociales de gran repercusión internacional. El primer periodo (1850-1914) coincide con la llamada segunda revolución industrial, caracterizada por cambios técnico-logísticos (nuevos medios de transporte, como el avión, y de comunicación, como el teléfono o la radio) y también por los avances en el aprovechamiento de recursos energéticos (gas, petróleo, electricidad). Además, en la segunda mitad del siglo XIX inicia sus primeros pasos lo que más tarde se denominaría *globalización*:

> Benché, nel dibattito corrente, si tenda a considerarla una novità assoluta della nostra era, fra gli storici economici vi è ampia convergenza nell'affermare che la globalizzazione, nella sua essenza, non è un fenomeno del tutto nuovo. Più precisamente – circoscrivendo il campo d'indagine al capitalismo moderno e considerando la dinamica dei *flussi migratori*, delle *esportazioni* e degli *investimenti diretti all'estero* – si evidenzia come l'economia mondiale abbia vissuto tre fasi di globalizzazione (Collier y Dollar 2003):
>
> - la prima coincidente con il periodo 1870-1914;
> - la seconda con gli anni 1945-1980;
> - la terza, quella attualmente in corso, con la fine del ventesimo secolo.
>
> (Valdani y Bertoli 2014: 5, cursivas en el texto).

[33] Teniendo también en cuenta que, para futuras eventuales incorporaciones de textos legislativos, todos los países, aunque compartan la misma lengua, tienen un ordenamiento jurídico propio, por lo que las regulaciones en materia de comercio vienen impuestas fundamentalmente desde instancias nacionales —pero también por organismos internacionales o supranacionales—, al mismo tiempo que los sectores y/o productos con los que se comercia dependen principalmente de cada estado.

El segundo y el tercer corte cronológico considerados en nuestro corpus, así pues, coinciden en gran medida con la segunda y la tercera fase de la globalización según las periodizaciones aceptadas entre los historiadores de la economía, además de apuntalarse en grandes acontecimientos como la posguerra (1945-1970) y el afianzamiento de la revolución digital y de internet (1990-2018).[34]

Queda claro que los procesos histórico-sociales que tomamos como puntos de referencia no se desarrollaron con la misma intensidad ni al mismo tiempo en todos los países hispánicos, pero el mismo inconveniente se daría con cualquier otra periodización basada en motivaciones extralingüísticas. En nuestro caso, este posible problema lo palia la propia amplitud cronológica de las tres franjas temporales. Además, gracias al amplio consenso existente entre los especialistas en historia del comercio al respecto de los hitos históricos en cuestión, los tres cortes cronológicos se pueden compartir con el subcorpus francés del *DIACOM*.

Por consiguiente, nuestro corpus no sigue criterios estrictos de división por cuartos de siglo, por ejemplo, porque —como bien señala Torruella Casañas (2017: 78)— la evolución lingüística no se produce de acuerdo con el calendario. Por lo tanto, en cuanto al periodo de tiempo que contempla, podrá clasificarse como un corpus diacrónico,[35] dado que recogerá textos de franjas temporales en tres siglos sucesivos, si bien cada uno de sus componentes -temporales podrá considerarse sincrónico, pues reflejará la lengua de especialidad de un periodo concreto. Este diseño permite observar tanto la evolución de los fenómenos lingüísticos como su estado en cada uno de los cortes temporales.

2.4. Marcaje de los textos

DIACOM va a ser un corpus etiquetado con información tanto metalingüística como lingüística. En el nivel de etiquetado metalingüístico se tendrán en cuenta todos aquellos aspectos de los documentos incorporados al corpus que permitan, posteriormente, una consulta específica o combinada por periodo de tiempo, tipo textual, dominio y subdominio, así como país de procedencia del autor o lugar de publicación/redacción del texto. En cuanto al nivel de anotación, el proyecto contempla —por lo menos en la primera fase en la que nos encontramos— la anotación morfológica y la lematización.

En este último caso, los textos se incorporarán al corpus manteniendo sus peculiaridades gráficas —es decir, no se llevará a cabo la normalización según los usos ortográficos actuales— y el lema que albergará todas las formas de una palabra, así como de su paradigma flexivo, se seleccionará siguiendo las normas ortográficas vigentes (p. ej. *esportar* / *exportar* → *exportar*). De esta manera, *DIACOM* no solo permitirá efectuar análisis terminológicos y fraseológicos, sino que, al mantener las convenciones gráficas de las diferentes épocas, también

[34] La revolución informática empieza a mediados del siglo XX, pero es en los años ochenta cuando los procesos de informatización se afianzan y se produce la eclosión de los *personal computers*.

[35] Coincidimos con Torruella Casañas (2017: 46) en que la consideración de un corpus como diacrónico o sincrónico en su diseño no se refiere tanto al periodo de tiempo que este abarque como a la determinación de fraccionar o no ese periodo de tiempo en función de la perspectiva que adopte la investigación. Asimismo, este autor (*ibid.* 45) establece como criterio la época, además de la temporalidad —corpus sincrónico y diacrónico—, con una distinción entre *corpus contemporáneo* y *corpus histórico*, para aquellos que recogen textos de la lengua actual o de una o diversas del pasado, respectivamente. Nosotros prescindimos de etiquetar nuestro corpus como *contemporáneo* o *histórico*, pues a pesar de que uno de sus componentes refleja el estado actual de la lengua de especialidad, se van a incorporar textos de otros dos periodos históricos.

hará posible observar otros fenómenos lingüísticos desde mediados del siglo XIX a la actualidad.[36]

2.5. Búsqueda y selección de textos

Tras la primera fase de diseño del corpus, le sigue una segunda de compilación teniendo en consideración todos los parámetros que garanticen, tanto en este primer estadio como en los posteriores, su representatividad y equilibrio. Debido a la interdisciplinariedad y la multidisciplinariedad del dominio de especialidad estudiado, así como al propósito por el que se compila *DIACOM*, a saber, el análisis del lenguaje especializado del comercio internacional desde diferentes aproximaciones, nuestro interés se centrará en mantener la homogeneidad temática en los subdominios acotados.[37] A diferencia de los grandes corpus de lengua general, donde los desajustes de representatividad y equilibrio acaban siendo compensados con la inmensa cantidad de datos que incorporan, en un corpus más restringido, se contrarrestan con la calidad del material lingüístico que lo compone y que, en nuestro caso, vendrá determinada por la pertenencia de los textos al dominio de especialidad y por la organización de los datos, esto es, su clasificación a partir de diferentes parámetros para su posterior gestión y extracción en la herramienta de consulta. Así, además de recuperar datos globales, el usuario podrá ajustar sus variables de búsqueda por país, periodo o tipo textual y combinarlas para observar la distribución de los fenómenos lingüísticos de la manera más acertada para su investigación.

Atendiendo a los criterios de diseño expuestos, iniciaremos la búsqueda de los textos. Se recopilarán las muestras en función de criterios externos, pues estos pueden establecerse *a priori* (Biber 1993: 245), sin necesidad de leer el texto (Atkins *et al*. 1992: 5). En concreto, seleccionaremos los documentos que cumplan los requisitos estipulados referidos al ámbito de especialidad, la lengua y el tipo textual.

Para la obtención de los textos en los sitios web recurriremos a la búsqueda temática por la palabra clave “comercio”. No obstante, no se trata de hacer acopio de todos los textos fruto de la búsqueda por dicha palabra clave, puesto que su amplitud puede provocar desajustes entre los componentes del corpus. De hecho, de los resultados que muestre la búsqueda se seleccionarán solo aquellos que se inserten en uno de los subdominios en los que hemos estructurado el ámbito para su estudio. Para ello, hemos determinado una serie de palabras clave —simplificadas siguiendo la bibliografía especializada— que nos servirán tanto para la selección de los textos como también para su posterior clasificación y organización en la base de datos. Solo de esta manera podremos garantizar la homogeneidad temática necesaria para alcanzar el equilibro del corpus.

Así, nuestra búsqueda se restringe, para los tres cortes cronológicos, a documentos en formato digital —bien hayan pasado estos por un proceso de digitalización bien sea su soporte original el electrónico—. Esta decisión viene motivada

[36] En cuanto a la ortografía, sobra destacar que las oscilaciones más evidentes se apreciarán en las muestras recogidas para el primer corte temporal (1850-1914), no solo por ser la época más lejana en el tiempo, sino también por las diferencias que se dieron entre los países hispanófonos, en particular porque algunos adoptaron durante un tiempo —como es el caso de Chile, por ejemplo— normas propias.

[37] La homogeneidad, así como la comparabilidad de los subcorpus, se comprobará con el programa ReCor (*vid.* nota al pie 5).

por la disponibilidad de los documentos en formato digital, cuyo acceso es sencillo y rápido, de forma que el proceso de tratamiento informático de los datos se agiliza. A este respecto, también hemos tenido en cuenta las cuestiones relacionadas con los derechos de autor o *copyright*. Si bien, por un lado, se trata de textos de acceso libre —algunos de ellos exentos de derechos de autor— y gratuito, y, por otro, nuestro proyecto no tiene fines comerciales ni prevé facilitar al usuario final el acceso al texto completo, nuestra intención es establecer colaboraciones y pedir permisos, pues como señala Torruella Casañas (2017: 155), los derechos morales pueden ser igual de importantes o, incluso, más que los derechos legales.

3. Desarrollo futuro del proyecto

Como se ha visto a lo largo del trabajo, la construcción de cualquier corpus conlleva la toma de múltiples decisiones que deberán justificarse y ponerse a disposición de la comunidad científica para que los investigadores puedan hacer uso de la herramienta con la confianza de que esta proporcione resultados fidedignos. Consideramos, pues, que la documentación de los criterios de diseño y la justificación de las decisiones que hemos tomado a la hora de diseñar *DIACOM* conforman una fase fundamental del proyecto que estamos llevando a cabo, no solo para informar a los futuros usuarios de su contenido, sino también para evaluar su calidad, así como para que pueda servir de modelo para otras investigaciones. En esta fase inicial hemos prestado especial atención a los criterios generales de diseño del corpus para garantizar, *a priori*, la representatividad y el equilibrio del corpus.

No obstante, no dejamos de lado que la composición de un corpus debe ser realista y que tendrá limitaciones como parte finita de un universo infinito (Parodi 2010: 24; McEnery y Hardie 2012: 15); por ello, la compilación de un corpus representativo es un proceso cíclico (Biber 1993: 256), dado que no siempre se pueden determinar todos los criterios que influirán en su construcción al inicio del trabajo. En el caso de *DIACOM*, por tratarse de un corpus especializado y diacrónico, enfrentamos tres cuestiones clave: la menor disponibilidad de textos especializados, la supervivencia de muestras representantes del dominio y el desigual volumen de producción y conservación de documentos por los países estudiados. Conscientes de estas limitaciones, nuestra intención ha sido disponer unos parámetros de diseño bien planteados y flexibles que se puedan reajustar *a posteriori*.

Para ello, se ha previsto la creación de un corpus piloto, ya marcado, que permita revisar la distribución de los documentos en la estructura del corpus y detectar carencias en su composición, para, posteriormente, realizar los ajustes y las actualizaciones convenientes en los parámetros de diseño y la distribución textual (*cf*. Biber 1993: 256; Torruella Casañas 2017: 147-149). Además, la creación de este corpus piloto permitirá probar la base de datos y la herramienta de consulta diseñadas para *DIACOM* antes de ponerlas a disposición de los usuarios externos al proyecto.

Referencias Bibliográficas

1. Academy of International Business. Research Codes. Consultado: 15 de mayo de 2019. https://aib.msu.edu/membership/Research_Codes.pdf

2. Ahmad, K. (1995). Pragmatics of Specialist Terms: The Acquisition and Representation of Terminology. En: Machine Translation and the Lexicon: Proceedings of the Third International EAMT Workshop, Heidelberg, Germany, April 1993.. Ed., Petra Steffens. Berlín/Heidelberg: Springer. (pp. 51-76) https://doi.org/10.1007/3-540-59040-4_20
3. Ahmad, K. y Rogers, M. (2001). Corpus Linguistics and Terminology Extraction. En: Handbook of Terminology Management. Vol. 2. Application-Oriented Terminology Management. Eds., Sue Ellen Wright y Gerhard Budin. (pp. 725-760). Ámsterdam/Filadelfia: John Benjamins, https://doi.org/10.1075/z.htm2.28ahm
4. Álvarez, C. (2011). Estudio del lenguaje de especialidad económico: el lenguaje del comercio internacional. Entreculturas, (3), 279-290.
5. Álvarez, C. (2017). Los textos en el ámbito del comercio exterior: una taxonomía para la formación de traductores. Sendebar, (28), 113-133.
6. Atkins, S. , Clear, J. y Ostler, N. (1992). Corpus Design Criteria. Literary and Linguistic Computing,7(1), 1-16. https://doi.org/10.1093/llc/7.1.1
7. Biber, D. (1993). Representativeness in Corpus Design. Literary and Linguistic Computing, 8(4), 243-257. https://doi.org/10.1093/llc/8.4.243
8. Bowker, L. (1996). Towards a Corpus-Based Approach to Terminography. Terminology, International Journal of Theoretical and Applied Issued in Specialized Communication, 3(1), 27-52. https://doi.org/10.1075/term.3.1.03bow
9. Bowker, L. y Pearson, J. (2002). Working with Specialized Language: A Practical Guide to Using Corpora. Londres: Routledge. https://doi.org/10.4324/9780203469255
10. Buckley, P. J. y Lessard, D. R. (2005). Regaining the Edge for International Business Research. Journal of International Business Studies, 36, (6), 595-599. https://doi.org/10.1057/palgrave.jibs.8400170
11. Cabré, M.a T. (1999). La terminología: representación y comunicación. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
12. Cabré, M.a T. (2002). Textos especializados y unidades de conocimiento: metodología y tipologización. En: Texto, terminología y traducción. Eds., Joaquín García Palacios y M. T. Fuentes Morán. (pp. 15-36). Salamanca: Almar.
13. Cabré, M.a T. y Estopá, R. (2005). Unidades de conocimiento especializado: caracterización y tipología. En: Coneixement, llenguatge i discurs especialitzat. Eds., M.a Teresa Cabré Castellví y M.a del Carme Bach Martorell. (pp. 69-93). Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
14. Cassany, D. (2004a). Explorando los discursos de las organizaciones. En: Textos y discursos de especialidad. El español de los negocios. Dir., Andreu van Hooft Camajuncosas. (pp. 49-60). Ámsterdam: Rodopi.
15. Cassany, D. (2004b). La lectura y escritura de géneros profesionales en EpFE. En: Español para fines específicos. Actas del II CIEFE (Congreso Internacional de Español para Fines Específicos). (pp. 40-64). Madrid: Ministerio de Educación y Ciencia,
16. Ciapuscio, G. (2005). La noción de género en la Lingüística Sistémico Funcional y en la Lingüística Textual. Revista Signos, 38(57), 31-48. https://doi.org/10.4067/s0718-09342005000100003

17. Collier, P. y Dollar, D. (2003). Globalizzazione, crescita economica, povertà. Rapporto della Banca Mondiale. Bologna: 2 Mulino.
18. COMENEGO = Gallego Hernández, D. et al. (2017). COMENEGO: Corpus Multilingüe de Economía y Negocios. Valencia: Universidad de Alicante. Consultado: 15 de mayo de 2019. https://dti.ua.es/es/comenego/comenego-corpus-multilingue-de-economia-y-negocios.html
19. Corpas, G. (2001). Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada. TRANS, Revista de Traductología, (5), 155-185. https://doi.org/10.24310/trans.2001.v0i5.2916
20. Corpas, G. y Seghiri, M. (2006). El concepto de representatividad en la lingüística del corpus: aproximaciones teóricas y metodológicas. Documento técnico BFF2003-04616 MCYT/TI-DT-2006-1. Consultado: 15 de mayo de 2019. http://www.uma.es/hum892/Publicaciones/Corpas_Segh iri_2006i.pdf
21. CORPES XXI = Real Academia Española. Parámetros de selección de textos. Corpus del Español del Siglo 21 (CORPES 21). Consultado: 27 de diciembre de 2019. https://www.rae.es/publicaciones/parametros-de-seleccion-de-textos
22. Davies, M. (2009). Creating useful historical corpora: a comparison of CORDE, the Corpus del español, and the Corpus do português. En: Diacronía de las lenguas iberorrománicas: nuevas aportaciones desde la lingüística de corpus. Ed., Andrés Enrique-Arias. (pp. 137-166). Madrid/Frankfurt am Main: Iberoamericana/Vervuert.
23. De Hoyos, J. C. (2016). El léxico de la Economía: etimología, historia y lexicografía. En: Etimología e historia en el léxico del español: estudios ofrecidos a José Antonio Pascual (Magister bonus et sapiens). Eds., Mariano Quirós García et al. (pp. 499-516). Madrid/Frankfurt am Main: Iberoamericana/Vervuert.
24. De La Fuente, B. (2019). El nuevo ecosistema de financiación del emprendimiento (Business angels, crowdfunding, mercados alternativos...). Glosario ES-EN-FR. Granada: Comares.
25. EAGLES = Expert Advisory Group on Language Engineering Standards. (1996). Preliminary recommendations on corpus typology. EAG- TCWG-CTYP/P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale. http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html
26. Enrique-Arias, A. (2012). Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad, (1), 85-106. Scriptum Digital.
27. Gallego, D. (2012). Traducción económica y corpus: del concepto a la concordancia. Aplicación al francés y al español. Alicante: Universidad de Alicante.
28. Gallego, D. (2013). Proyecto COMENEGO: algo más que un corpus multilingüe de economía y negocios. En: 10 Jornadas de redes de investigación en docencia universitaria. La participación y el compromiso de la comunidad universitaria. Coords., María Teresa Tortosa Ybáñez, José Daniel Álvarez Teruel y Neus Pellín Buades. (pp. 2242-2251).Alicante: Universidad de Alicante.
29. Gallego, D. (ed.) (2018). Nuevos estudios sobre traducción para el ámbito

institucional y comercial / New Approaches to Translation in Institutional and Business Settings. Berna: Peter Lang.
30. Gómez de Enterría, J. (1992a). Las siglas en el lenguaje de la economía. Revista de Filología Románica, (9), 267-274.
31. Gómez de Enterría, J. (1992b). Neología y préstamo en el vocabulario de la economía. Anuario de estudios filológicos, (15), 97-106.
32. Gómez de Enterría, J. (1999). Las traducciones del francés. Cauce para la llegada a España de la ciencia ilustrada. Los neologismos en los textos de botánica. En: La traducción en España (1750-1830). Lengua, literatura, cultura. Ed., Francisco Lafarga. (pp. 143-155). Lleida: Edicions de la Universitat de Lleida.
33. Gómez de Enterría, J. (2009). El español lengua de especialidad: enseñanza y aprendizaje. Madrid: Arco/Libros.
34. Heinemann, M. y Heinemann, W. (2002). Grundlagen der Textlinguistik. Tübingen: Max Niemeyer.
35. Heinemann, W. y Viehweger, D. (1991). Textlinguistik: eine Einführung. Tübingen: Max Niemeyer.
36. Hunston, S. (2002). Corpora in Applied Linguistics. Cambridge: Cambridge University Press.
37. Kabatek, J. (2013). ¿Es posible una lingüística histórica basada en un corpus representativo?. Ibero, (77), 8-28.
38. Martínez, J. J. (2009). El léxico del español de los negocios: propuesta de análisis para su enseñanza y aprendizaje. En: Estudios de lingüística: investigaciones lingüísticas en el siglo 21. Eds., Juan Luis Jiménez Ruiz y Larissa Timofeeva. (pp. 169-187). Alicante: Universidad de Alicante.
39. Mateo, J. (2007). El lenguaje de las ciencias económicas. En: Las lenguas profesionales y académicas. Eds., Enrique Alcaraz Varó, José Mateo Martínez y Francisco Yus Ramos. (pp. 191-203). Barcelona: Ariel.
40. Mayoral, R. (2007). La traducción comercial. En: Problemas lingüísticos en la traducción especializada. Coord., Pedro Antonio Fuentes Olivera. (pp. 33-48). Valladolid: Universidad de Valladolid.
41. McEnery, T. y Hardie, A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511981395
42. McEnery, T., Xiao, R. y Tono, Y. (2006). Corpus-Based Language Studies. Londres/Nueva York: Routledge.
43. Meyer, I. y Mackintosh, K. (1996). The Corpus from a Terminographer's Viewpoint. International Journal of Corpus Linguistics, 1(2), 257-285.
44. Palacios, M. y Safford, F. (2002). Colombia: país fragmentado, sociedad dividida, su historia. Traducción de Angela García. Bogotá: Norma.
45. Parodi, G. (2010). Lingüística de corpus: de la teoría a la empiria. Madrid/Frankfurt am Main: Iberoamericana/Vervuert.
46. Pearson, J. (1998). Terms in Context. Ámsterdam/Filadelfia: John Benjamins.
47. Pizarro, I. (2010). Análisis y traducción del texto económico (inglés-español). La Coruña: Netbiblo.
48. Ramacciotti, S. B. y Rodil, M. V. (2006). Economics: Glossary of Metaphorical Usage / Glosario económico-financiero: uso metafórico de voces. Buenos Aires: Quorum/Universidad del Museo Social Argentino.
49. Seghiri, M. (2011). Metodología protocolizada de compilación de un

corpus de seguros de viajes: aspectos de diseño y representatividad. RLA, Revista de Lingüística Teórica y Aplicada, 49,(2), 13-30. https://doi.org/10.4067/s0718-48832011000200002
50. Seghiri, M. (2015). Determinación de la representatividad cuantitativa de un corpus ad hoc bilingüe (inglés-español) de manuales de instrucciones generales de lectores electrónicos. En: Corpus-Based Translation and Interpreting Studies: From description to application. Ed., María Teresa Sánchez Nieto. (pp. 125-146). Berlín: Frank & Timme.
51. Seghiri, M. (2017). Metodología de elaboración de un glosario bilingüe y bidireccional (inglés-español/español-inglés) basado en corpus para la traducción de manuales de instrucciones de televisores. Babel, 63(1), 43-64. https://doi.org/10.1075/babel.63.1.04seg
52. Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press. https://doi.org/10.2307/330144
53. Torruella, J. (2017). Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación en lingüística. Frankfurt am Main: Peter Lang.
54. Torruella, J. y Llisterri, J. (1999). Diseño de corpus textuales y orales. En: Filología e informática: nuevas tendencias en los estudios filológicos. Eds., José Manuel Blecua, Gloria Clavería, Carlos Sánchez y Joan Torruella. (pp. 45-77). Lleida: Milenio/Universidad Autónoma de Barcelona. http://latel.upf.edu/traductica/lc/material/torruella_llisterri_99.pdf
55. Valdani, E. y Bertoli, G. (2014). Marketing internazionale. Milano: EGEA.
56. Zettinig, P. y Vincze, Z. (2011). The Domain of International Business: Futures and Future Relevance of International Business. Thunderbird International Business Review, 53(3), 337-349.

Recibido: 15/07/2019
Aceptado: 29/10/2019