3. Estilística de corpus

La estilística de corpus es una disciplina de estudio que se encarga de analizar cuestiones de estilo empleando metodologías lingüísticas de tipo cuantitativo. El mayor filón de este tipo de enfoques es su capacidad de detectar patrones de los que no siempre somos conscientes, pero que pueden contribuir a configurar algunos de los efectos que percibimos como lectores. Se trata, además, de una disciplina que permite manejar cantidades de texto inasumibles para otras disciplinas más tradicionales, como la estilística (en su vertiente tradicional) o la crítica literaria. Esta es la razón por la que los estudios de estilística de corpus suelen tener como objeto de estudio el género novelesco, pues la masa de texto de una novela suele ser mayor que la de una obra de teatro o un poema. Este es precisamente el punto de partida de este libro. En este capítulo nos ocupamos de definir algunos presupuestos teóricos fundamentales de la estilística de corpus, prestando atención a aquellos aspectos que la distinguen de otras disciplinas (apartados 3.1.1 y 3.1.2). Además, ofrecemos un breve repaso de los antecedentes teóricos que se relacionan con la idea analítica de este libro (la valoración exegética de Pérez Galdós) (apartado 3.2), desgranaremos algunos de los principales puntos de partida de análisis (apartado 3.3) y nos detendremos en cuestiones relacionadas con la compilación de corpus de estudio y de referencia, explicando qué textos hemos seleccionado (y por qué lo hemos hecho) para llevar a cabo nuestro análisis (apartado 3.4). Todo ello nos permitirá no solo acotar nuestro análisis dentro de un marco de estudio bien definido, sino que servirá para explicar el potencial de los estudios de estilística de corpus en el ámbito hispánico, un área de estudio todavía escasa de este tipo de análisis.

https://dx.doi.org/10.5209/ling.004.03
Humanidades digitales y análisis estilístico: la lengua de Benito Pérez Galdós. Pablo Ruano San Segundo y Guadalupe Nieto Caballero. © Ediciones Complutense, 2025.

3.1. Definición

Para definir la estilística de corpus es necesario ofrecer una contextualización dentro del marco de los estudios literarios y lingüísticos en los que se imbrica. Desde un punto de vista conceptual, la estilística de corpus se sitúa en una posición intermedia entre la lingüística de corpus y la estilística tradicional. Sin embargo, es importante incidir que la estilística de corpus es una disciplina de estudio autónoma, y no una rama subsidiaria de la estilística o de la lingüística de corpus: aunque se apoya en presupuestos teóricos y de dimensión práctica de cada una de estas disciplinas, es importante decir que también se distingue de ambas en gran medida. Para explicarlo, a continuación abordamos la relación de la estilística de corpus con cada una de estas disciplinas, poniendo de manifiesto tanto las similitudes como las diferencias de la estilística de corpus con ambas.

3.1.1. Un enfoque basado en presupuestos de la lingüística de corpus y la estilística tradicional

Para comprender y explicar la relación de la estilística de corpus con la lingüística de corpus debemos acudir al plano metodológico y a los presupuestos que la definen. La lingüística de corpus, como apunta Mahlberg (2013, 1) «has provided the methodology to investigate repeated patterns in a new light, revealing facts about language that have largely remained hidden from human observation». La estilística de corpus, gracias al uso de enfoques procedimentales basados en metodologías de la lingüística de corpus, es capaz de llevar a cabo análisis que, debido a la falta de herramientas adecuadas, no han podido ser identificados con anterioridad –o al menos no han podido ser analizados de forma metódica– e incluso descubrir «meanings of literary texts that cannot be detected either by intuitive techniques as in literary studies» (Fischer-Starcke 2010, 2). Para comprender el sincretismo metodológico entre la estilística de corpus y la lingüística de corpus es necesario detenerse en la importancia de la repetición en los estudios de lingüística de corpus, trasladada a los estudios de estilística de corpus para explicar hábitos estilísticos y ofrecer una valoración artística del texto literario. Así, en los estudios de lingüística de corpus, lo que se considera normativo se basa en generalizaciones construidas sobre la repetición de patrones lingüísticos en distintos textos.

Esta premisa ha sido adoptada en la estilística de corpus, en donde el procesamiento de textos con herramientas adecuadas «makes it possible to observe repeated patterns, and the patterns in turn serve as the basis for the repetition of repeatedly expressed meanings» (Mahlberg 2013, 1). La importancia de la repetición que comparten la lingüística de corpus y la estilística de corpus se puede explicar recurriendo al cambio de paradigma que trajo consigo el empleo de herramientas de corpus como las mencionadas en el apartado 2.2.2 con respecto a otras formas de análisis más tradicional. Tognini-Bonelli (2001, 3) lo sintetizó con precisión cuando aseguró que las metodologías de corpus (en un sentido general) hicieron posible un cambio de paradigma en el análisis de cualquier texto (sea cual sea el propósito): además de la lectura horizontal tradicional de los textos, la muestra de resultados en los *softwares* de concordancias permite una lectura vertical de los datos. Sin duda, esta esta lectura vertical puede ser la llave para un estudio cualitativo imposible de llevar a cabo de forma manual –en el apartado 4.4.1 mostramos un ejemplo relacionado con la caracterización de Clara en *La Fontana de Oro*–.

Por su parte, la relación de la estilística de corpus con la estilística tradicional se encuentra en la motivación teórica de ambas disciplinas, que consideran el texto literario como una creación artística –al contrario de lo que ocurre en los estudios de lingüística de corpus–. Dicho de otro modo, mientras que en su análisis de textos literarios la lingüística de corpus se centra en cuestiones de tipo lingüístico para indagar en el uso de la lengua, los estudios de estilística tradicional se ocupan de esos esos mismos textos con el fin de valorarlos en sí mismos, pues el valor artístico es un aspecto susceptible de análisis que poco tiene que ver con el uso de la lengua en general. La estilística de corpus, como la estilística tradicional, se acerca al estudio de textos literarios en un intento de relacionar la descripción lingüística del texto propia del lingüista con la apreciación artística propia del crítico literario (Leech y Short 2007 [1981], 11).

Conviene incidir en que la relación entre la descripción lingüística y la apreciación artística es bidireccional, pues cada una ofrece una explicación y una motivación a la otra. Spitzer (1948) se refirió a esta relación con un esquema que bautizó como el «círculo filológico», que ha resultado muy útil para entender el componente interpretativo basado en cuestiones de tipo lingüístico de la estilística tradicional. Efectivamente, las valoraciones estilísticas no se caracterizan por priorizar la apreciación artística sobre la descripción lingüística –o al revés–. Por el contrario, ambos componentes se encuentran indisolublemente unidos y cada uno de ellos se relaciona con el otro en una suerte de relación

simbiótica de la que surge el análisis del texto literario. La estilística de corpus se construye sobre este principio interpretativo, pero añade un tercer elemento: las metodologías propias de la lingüística de corpus comentadas más arriba. Este tercer elemento hace posible lo que Mahlberg (2013, 12) ha llamado «círculo de la estilística de corpus», en clara alusión al círculo filológico de Spitzer (1948). Lejos de ser una simple reformulación de la teoría *spitzeriana*, el elemento metodológico del círculo de la estilística de corpus plantea un nuevo modo de hacer estilística. Como asegura la propia Mahlberg (2013, 12), las metodologías de corpus han desempeñado un papel crucial «in changing linguistic observation», pues «the language looks rather different when you look at a lot of it at once» (Sinclair 1991, 100). Efectivamente, la identificación de patrones puede ayudar a alcanzar un mayor nivel de comprensión en el plano de la descripción lingüística, lo que contribuye (o al menos tiene el potencial de hacerlo) a una valoración artística más precisa.

3.1.2. Diferencias teóricas y metodológicas con la lingüística de corpus y la estilística tradicional

Del mismo modo que la estilística de corpus se construye sobre preceptos teóricos y procedimentales de la lingüística de corpus y la estilística tradicional, en su conceptualización también encontramos grandes diferencias con ambas disciplinas. Estas diferencias son las que hacen que la estilística de corpus sea una disciplina de estudio autónoma, y no una rama subsidiaria de la lingüística de corpus o de la estilística tradicional. En primer lugar, una de las mayores diferencias entre la lingüística de corpus y la estilística de corpus reside en la prominencia del componente cuantitativo que las caracteriza. De un lado, el foco de interés de la lingüística de corpus reside en realizar generalizaciones sobre fenómenos lingüísticos basados en la identificación de patrones localizados en los distintos textos que se analizan. Este es su objetivo, que por definición descarta valoraciones artísticas de la lengua en el texto literario. Esta es la razón fundamental por la que lingüistas que emplean enfoques de corpus en sus análisis consideran que «a distinctive literary text is just not worth including in a general corpus because it will disappear below the waves since its textual patterns are not echoed in other texts» (Sinclair 2007, 3). Efectivamente, las particularidades de un texto literario se diluyen en el conjunto de textos de cualquier corpus desde un punto de vista analítico, pues la lingüística de corpus se encarga de llevar a cabo generalizaciones

sobre el uso de la lengua. Los estudios de estilística de corpus, por su parte, además de en cuestiones estilísticas de tipo general, «are also interested in the meanings of individual texts, the patterns that are relevant to a particular text or even linguistic phenomena that are unique to a text» (Mahlberg 2016, 144). En otras palabras, la estilística de corpus está interesada, en sus estudios sobre el valor artístico de la lengua, en usos específicos que pueden encontrarse en textos concretos. El componente cualitativo, por tanto, no es solo un rasgo distintivo de la estilística de corpus, sino una característica que la diferencia de la lingüística de corpus.

La otra gran diferencia de la estilística de corpus con la lingüística de corpus tiene que ver con sus fines analíticos, pues ambas incluyen textos literarios en sus corpus de estudio. Como decíamos más arriba, la lingüística de corpus se encarga de analizar cuestiones relacionadas con el funcionamiento de la lengua. En estos estudios, las características lingüísticas se exploran en el marco de un corpus de textos que suele incluir textos que pertenecen a distintos registros para cubrir un espectro lo más amplio posible. Esta es la razón principal por la que la mayoría de los corpus que se emplean en estudios de lingüística incluyen textos literarios (Sinclair 2004, 51), tanto a nivel internacional (el British National Corpus en inglés, por ejemplo) como en el ámbito hispánico (el Corpus de Referencia del Español Actual de la RAE, por ejemplo). El hecho de incluir textos literarios en sus estudios, sin embargo, no debe confundirnos, pues la lingüística de corpus no se encarga de su análisis crítico. Su propósito es identificar y estudiar patrones lingüísticos para analizar la lengua en general. Y la literatura es otro ejemplo más. La estilística de corpus, por su parte, tiene motivaciones teóricas similares a las de la estilística tradicional, como decíamos en el apartado anterior, pues se acerca al texto literario como una creación artística susceptible de análisis en sí misma.

Esta motivación teórica compartida entre los estudios de estilística tradicional y los estudios de estilística de corpus no deben ser óbice para encontrar diferencias notables entre ambas disciplinas. Estas diferencias son fundamentalmente de tipo metodológico, pues en esta dimensión la estilística de corpus es apoya en la lingüística de corpus, como también comentábamos en el apartado 3.1.1. En concreto, la estilística de corpus se erige sobre un pilar metodológico en el que el plano textual tiene un peso mayor que en la estilística tradicional. Esta prominencia del plano textual en ocasiones se materializa en que haya quienes pongan en duda el valor (excesivo, a su juicio) concedido al plano cuantitativo en detrimento de aspectos más recónditos que puedan resultar igualmente significativos. Sin embargo, debemos insistir en que la

estilística de corpus no solo no desdeña la parte cualitativa –ya hemos dicho que el interés de un estudio de estilística de corpus puede residir en un texto concreto– sino que «provides the tools to systematically aim to describe the relationships between textual cues and literary effects» (Mahlberg 2016, 140). Las herramientas que permiten desarrollar metodologías de corpus no hacen sino reforzar esa sistematicidad (cf. Stubbs 2005; O'Halloran 2007), lo que en cierto modo contribuye a dotar de una mayor solidez a las interpretaciones que se llevan a cabo, incluso cuando estas se llevan a cabo sobre un solo texto –como mostramos en nuestro análisis de *La Fontana de Oro* en el capítulo 4, por ejemplo–.

En definitiva, de todo lo dicho en estos dos subapartados podemos extraer dos conclusiones. Por un lado, aunque la estilística de corpus difiera de los presupuestos teóricos fundamentales de la lingüística de corpus, otros –los de tipo metodológico– los abraza. Por otro lado, aunque utilice procedimientos diferentes a la estilística tradicional, sus objetivos son similares: la realización de una valoración literaria del texto mediante la observación de tipo lingüístico. Por ello, consideramos que la mejor manera de entender la estilística de corpus es como una disciplina en el marco de los estudios literarios donde convergen la lingüística de corpus y la estilística tradicional, sin que ninguna de ellas deba considerarse la matriz desde la que surge. Una vez hechas estas (necesarias) matizaciones, podemos descender a ejemplos concretos de análisis que han contribuido a la conceptualización de la estilística de corpus como disciplina de estudio autónoma.

3.2. Antecedentes teóricos

La conceptualización definitiva de la estilística de corpus como una disciplina de estudio autónoma en el campo de los estudios literarios tiene lugar hacia la primera mitad de la década del 2000, cuando hay «an increasing number of studies that show how a range of corpus linguistic tools and concepts can usefully be employed to support the analysis of literary texts» (Mahlberg, Smith y Preston 2013, 35). Hasta entonces, los enfoques cuantitativos aplicados al estudio de textos literarios carecían de la dimensión cualitativa que muchos trabajos comenzaron a mostrar hace aproximadamente dos décadas. A partir de ese momento, asistimos a la publicación de trabajos que marcan una nueva tendencia, con un nuevo enfoque que «draws on corpus methodology but at the same time, it emphasises the link that literary stylistics provides to

literary criticism» (Mahlberg 2013, 2). La publicación de numerosos trabajos sobre autores canónicos ingleses en un corto espacio de tiempo sin duda contribuyó a la consolidación del campo. Figuras de referencia como Virginia Woolf (Adolphs y Carter 2002; Balossi, 2014), Joseph Conrad (Stubbs 2005), William Shakespeare (Culpeper 2009), Jane Austen (Fischer-Starcke 2010) o Charles Dickens (Mahlberg 2013), entre otras, han sido analizadas utilizando enfoques de estilística de corpus. Dickens es tal vez el autor más analizado en este sentido. La publicación del libro *Dickens and Corpus Stylistics* (Mahlberg 2013) supuso un hito en la conceptualización de la estilística de corpus como una disciplina de estudio plenamente consolidada, pues se publicaba, por vez primera, un trabajo a gran escala capaz de conjugar la dimensión cuantitativa de las metodologías de corpus con el componente cualitativo de enfoques literarios. De un lado, desde un punto de vista cuantitativo, en este libro la autora es capaz de abordar toda la producción de Dickens, comparándola con un corpus de novela decimonónica inglesa. De otro lado, además, desde un punto de vista cualitativo, la autora es capaz de analizar con éxito «the relationship between linguistic units and the contributions they might make to the effects that texts have on readers» (Mahlberg 2013, 6). Cabe destacar que *Dickens and Corpus Stylistics* es la culminación de una serie de publicaciones de la autora sobre el propio Dickens, también de estilística de corpus (Mahlberg 2007a, 2007b, 2007c, 2009, 2010, 2012a, 2012b, 2013), que ya venían mostrando un carácter mucho más refinado y con una dimensión cualitativa mucho mayor que otros trabajos sobre la figura del autor victoriano realizados con anterioridad por Tabata (1991, 1993, 1994, 1995, 2002) y Hori (1993, 1999, 2002, 2004), también sobre Dickens y que podríamos considerar en una posición intermedia entre la estilística computacional y la estilística de corpus. Los trabajos de Mahlberg sobre Dickens sirvieron para consolidar la estilística de corpus como disciplina de estudio autónoma y han sido precursores de otros trabajos de estilística de corpus sobre el autor victoriano cada vez más refinados, como los de Mahlberg y Smith (2012), Mahlberg, Smith y Preston (2013) o Stockwell y Mahlberg (2015), Ruano San Segundo (2016, 2018a) o Mahlberg *et al.* (2019).

El hecho de que la estilística de corpus sea una disciplina consolidada se demuestra también en la creciente popularidad de los enfoques de corpus en las publicaciones de estilística en general. Sirva como ejemplo la revista científica *Language and Literature*. Como señalaba su editor hace apenas cinco años, el número de artículos recibidos y publicados por la revista que adoptan enfoques de estilística de corpus ha crecido de manera exponencial en la

última década (McIntyre 2017). Lo mismo ocurre con los manuales de introducción a la estilística que incluyen capítulos específicos sobre este enfoque (por ejemplo, Jeffries y McIntyre 2010; Leech y Short 2007 [1981]; McIntyre y Busse 2010; Burke 2014; Sotirova 2016; Mastropierro y Ruano San Segundo 2022) o los manuales de lingüística de corpus que han comenzado a prestar atención al análisis de textos literarios utilizando este tipo de enfoques (Chapelle 2012; Flowerdew 2012; Lindquist 2009; O'Keeffe y McCarthy 2010). Además, también son dignos de mención los proyectos que adoptan la estilística de corpus como eje vertebrador de su aparato analítico, como el CLiC Dickens, sobre la novela decimonónica inglesa en general y la obra de Charles Dickens en particular (Mahlberg *et al.* 2016), o el Shakespeare Encyclopaedia of Shakespeare's Language Project (Shakespearelang, s.f.), sobre el lenguaje literario de William Shakespeare. Finalmente, la consolidación de la estilística de corpus ha culminado con la publicación del primer manual de estilística de corpus: *Corpus Stylistics: Theory and Practice* (McIntyre y Walker 2019). Este es el primer libro de carácter monográfico sobre esta disciplina, en el que se abordan cuestiones teóricas y de tipo práctico para ofrecer al estudioso del texto literario sin conocimientos previos en la disciplina opciones de análisis que pueda poner en práctica.

3.2.1. El ámbito hispánico: una disciplina aún por explotar

Desgraciadamente, la consolidación de la estilística de corpus en el ámbito internacional no se ha materializado en el desarrollo de esta disciplina en el mundo hispánico, donde apenas encontramos trabajos que adopten este tipo de enfoques. De hecho, hasta hace pocos años apenas podían mencionarse unos pocos estudios computacionales situados más cerca del análisis estadístico de textos o la estilometría que de la estilística. Un ejemplo representativo, que además aborda la figura de Galdós, es el de Irizarry (1997). Su trabajo puede situarse en la línea de los estudios previos a la estilística de corpus como tal, como los de Tabata (1991, 1993, 1994, 1995, 2002) y Hori (1993, 1999, 2002, 2004) realizados sobre Dickens, comentados en el apartado anterior. En su estudio, la autora desarrolla procedimientos metodológicos y opciones de análisis que sirven para calibrar aspectos estilísticos de distintos autores. Así, además de sobre Galdós, su investigación cuenta con breves análisis sobre autores como Octavio Paz y Rosario Castellanos, Rodríguez-Juliá, Rafael Dieste, Valle-Inclán, Eugenio Fernández Granell, Buero

Vallejo o Bécquer. No obstante, en estos estudios la autora apenas pasa de puntillas por valoraciones de tipo literario sobre la masa de texto que analiza, pues apenas desciende a cada obra en tanto que creación artística. Naturalmente, esta carencia es fruto del tipo de análisis que realiza y no el resultado de una investigación pobre en términos exegéticos. Sirva como ejemplo el caso de Galdós, en el que nos detendremos brevemente por ser el autor cuya producción narrativa analizamos en este libro. En su estudio, Irizarry investiga (o tal vez sería más acertado decir que cuantifica) el uso que el autor canario hace de las palabras, las oraciones y la segmentación discursiva para explorar su evolución como novelista. Este acercamiento a lo que ella llama «estructura estilística "profunda"» (Irizarry 1997, 121) consiste en analizar muestras de cinco mil palabras de veintisiete novelas para indagar en el uso que el autor canario hace del diálogo. Desde luego, la masa de texto analizada, de apenas cinco mil palabras, constituye un problema a la hora de realizar cualquier tipo de valoración –en total, su estudio se basa en 135 000 palabras de texto, por las más de seis millones que se analizan en este libro, por ejemplo–. Además, las cinco mil palabras analizadas son las cinco mil palabras iniciales de cada novela. Si lo que se pretende analizar es el diálogo, como apuntábamos hace un momento, esta muestra de texto no parece ser la mejor opción, pues la muestra de diálogo que puede haber en las cinco mil palabras iniciales de una novela, donde se introduce el universo ficticio al lector, puede tener poco que ver con la cantidad real de diálogo de la novela en conjunto. Tal vez por ello la autora descubre que

> El porcentaje de diálogo varía enormemente en las muestras, desde un porcentaje minúsculo del 4 % en *La Fontana de Oro* hasta constituir el 90 % del texto en *La sombra* […]. El diálogo alcanza un promedio del 41 % de las muestras en general, y varias novelas contiguas coinciden alrededor del 27 % (Irizarry 1997, 123).

Desde luego, la diferencia que plantea la autora está directamente relacionada con la elección de su muestra de texto. El hecho de no descender al texto, como decíamos más arriba, para realizar una valoración cualitativa hace que las conclusiones se basen únicamente en los datos obtenidos y no en un análisis crítico de estos –lo que desde luego resulta peligroso a la luz de cuáles son algunos de los resultados obtenidos, como acabamos de ver–. En cualquier caso, insistimos, los problemas que plantea este estudio no son responsabilidad de un descuido de la autora, sino propio de las característi-

cas de los estudios de estilometría. De hecho, a pesar de sus limitaciones, es de justicia reconocer el valor del trabajo de Irizarry. Su carácter pionero lo convierte en una suerte de punto de partida de trabajos venideros. La propia autora, consciente de las limitaciones de su estudio, se muestra confiada en que esta nueva vía de análisis hasta entonces inédita en el ámbito hispánico abriera nuevos acercamientos cuantitativos al análisis de textos literarios, que actuarían como «un auxilio vital para ensanchar el pensamiento y las metodologías de la crítica literaria» (Irizarry 1997, 9). Nosotros confiamos en que el enfoque de estilística de corpus que presentamos en este libro pueda ser un buen ejemplo de ello.

3.2.2. Galdós y la estilística de corpus

Como comentábamos, la estilística de corpus no es una disciplina consolidada en el ámbito hispánico. Eso no quiere decir, sin embargo, que no existan estudios que en los últimos años hayan adoptado este tipo de enfoques, en lo que ojalá esté siendo el germen de la consolidación de este tipo de análisis. Como suele ocurrir en cualquier metodología de estudio de carácter pionero, los primeros trabajos en los que observamos un enfoque de estilística de corpus se acercan al texto en combinación con alguna otra disciplina donde ya han sido empleadas con éxito en el pasado, en una suerte de intento de justificar su validez. Eso no los convierte en trabajos menos valiosos, desde luego, pero sí en investigaciones donde algunos de los presupuestos teóricos de tipo estilístico no están plenamente consolidados. Aun así, son trabajos dignos de reseña, por su carácter pionero y por su contribución a la dimensión exegética de los textos que analizan. Algunos de los primeros trabajos sobre textos literarios en los que observamos enfoques de estilística de corpus tienen que ver con enfoques como la traducción o la intertextualidad. Y de nuevo Galdós aparece como figura habitualmente analizada.

En el caso de la traducción, por ejemplo, encontramos el trabajo de estilística de corpus de Ruano San Segundo (2015) sobre la versión española de Galdós de *Pickwick Papers* (Ramoneda 1989), de Charles Dickens, única traducción que el autor realizaría a lo largo de su vida. En concreto, en este trabajo se explora el valor caracterizador de los verbos de habla en la obra del autor inglés, así como la forma de traducirlos al español por parte de dos traductores: José María Valverde y el propio Pérez Galdós. Gracias al enfo-

que de corpus empleado, Ruano San Segundo es capaz de detectar el valor caracterizador de los verbos de habla en el texto de partida, así como la importancia de mantener este valor en su traducción al español. Mediante el uso de un corpus paralelo formado por la novela inglesa y dos traducciones (la del propio Galdós y la traducción de José María Valverde (2004)), es posible establecer una comparación metódica de la traducción de los verbos objeto de análisis. En el caso de Galdós, el estudio demuestra una pérdida sistemática de matices en su traducción de la novela. En concreto, la omisión, neutralización o modificación por parte de Galdós de los verbos de habla del texto original anula un aspecto esencial del estilo caricaturesco del autor victoriano. Como asegura Ruano San Segundo, este análisis no pretende menoscabar la figura del autor canario, sino demostrar la importancia que tiene reflejar en la traducción un elemento que, debido a su carácter disperso, no siempre resulta perceptible a simple vista. Gracias a la metodología computacional empleada es posible localizarlo y aislarlo en el texto de partida, lo que permite realizar un análisis sistemático de su traslado al español y evaluar hasta qué punto los traductores son capaces de preservar su valor estilístico.

En el caso de la intertextualidad, Nieto Caballero (2019a) analiza la influencia que el propio Charles Dickens ejerció en el estilo de Benito Pérez Galdós utilizando un enfoque de estilística de corpus, centrándose en el lenguaje gestual de los personajes de las novelas de ambos autores. El análisis consiste en un estudio de *clusters* de al menos cinco palabras empleados de forma sistemática por ambos novelistas que hacen referencia a alguna de las siguientes partes del cuerpo: los ojos, la cabeza, las manos o los hombros. El lenguaje gestual se erige en un sistema autónomo en la construcción del universo ficticio en el género novelesco en general (Korte 1997, 4) y uno de los rasgos distintivos del estilo de Dickens en particular (Mahlberg 2013, 100-127). Los ejemplos identificados revelan un claro paralelismo en el estilo de ambos autores, lo que ayuda a demostrar la influencia del primero sobre el segundo desde un punto de vista estilístico. Por citar un ejemplo, en este estudio se muestra como en el uso de *her handkerchief to her eyes* y *el pañuelo a los ojos* se advierte un uso sorprendentemente similar por parte de ambos novelistas. Ambos se asocian al habla de personajes femeninos y, además, son empleados en momentos dialógicos para reforzar la tristeza de estos. A la luz de la comparación de estos y otros ejemplos, parece claro que el autor canario pudo imitar esta estrategia para realzar la tristeza de varios de sus personajes femeninos en sus novelas. El enfoque de corpus adoptado permite establecer un paralelismo en el estilo de ambos autores que refuerza la influencia del

autor victoriano en Galdós desde un punto de vista estilístico. Gracias a la metodología de corpus se identifican segmentos de texto en la producción de ambos autores relacionados con la configuración del lenguaje gestual de los personajes que, a la luz de las similitudes tanto en términos formales como funcionales, revelan una concomitancia entre Dickens y Galdós que hasta ahora ha pasado desapercibida, lo que refuerza una influencia que, si bien está fuera de toda duda, habitualmente no se sustenta sobre una base lo suficientemente sólida desde una perspectiva puramente estilística.

Además de estos estudios de estilística de corpus relacionados con aspectos como la traducción o la intertextualidad, también se han publicado, en los últimos cuatro años, algunos estudios que adoptan un enfoque de estilística de corpus para analizar la producción narrativa del autor canario de forma concreta. El trabajo de Nieto Caballero (2018) sobre patrones léxico-gramaticales identificados con WordSmith Tools 6 (Scott 2016) relacionados con la proyección del discurso de los personajes del universo galdosiano por parte de los narradores de las distintas novelas es un buen ejemplo. En este estudio se compara la producción narrativa de Galdós con un corpus de novela realista española, como hacemos en este libro. Además de para calibrar la preponderancia de la representación del discurso en la obra de Galdós, esta comparación sirve para identificar bloques textuales típicamente galdosianos que contribuyen a la construcción de los mundos textuales del autor. Este es uno de los objetivos habituales de la estilística de corpus, como comentamos en el apartado 3.3. En concreto, en este trabajo se explica cómo el lenguaje gestual aparece frecuentemente unido a la representación del discurso verbal, sobre todo en estilo directo (Korte 1997, 94 y ss.). Esta combinación facilita la construcción del universo ficticio que se presenta al lector a través de la alternancia de la voz del narrador y el discurso de los personajes. En el artículo se ofrece un análisis de la fórmula *dijo abrazándo(le)(se)(me)*, de la que mostramos un ejemplo en (1). Esta fórmula representa un patrón formal que se materializa en una suerte de estilema en la representación del discurso de los personajes: contiene una proposición adverbial que completa la representación del discurso mediante una descripción del movimiento del personaje (lenguaje gestual) y se sitúa, además, interrumpiendo la proposición proyectada, partiendo esta en dos. Gracias a esta posición intermedia de la proposición proyectora, conocida como «suspended quotation» o «suspension» (Lambert 1981) se crea un efecto de sincronía entre las palabras del personaje y su movimiento, pues «suspensions can create an impression of simultaneity between the speech and the contextual information described by

the narrator, which in turn can suggest similarities to the simultaneous occurrence of speech and body language in real life» (Mahlberg, Smith y Preston 2013, 40). El enfoque de corpus permite identificar estos casos y analizarlos sistemáticamente, descubriendo un aspecto del estilo de Galdós habitualmente desapercibido en las valoraciones críticas de su estilo.

(1) –Somos huérfanas –*le dijo, abrazándose las dos estrechamente*–; somos huérfanas, Dios no ha querido que entremos en casa con nuestro padre (*De Oñate a la granja*, capítulo 28)[1].

En un trabajo posterior, y de corte analítico más concreto, Nieto Caballero investiga la narración de los silencios en la producción narrativa de Galdós (Nieto Caballero 2019c). Partiendo de la premisa de Caldas-Coulthard de que en la representación del discurso en el género narrativo «if silence is reported, then it is because it is super-significant» (Caldas-Coulthard 1987, 164), en este artículo se explora el modo en el que los narradores de las novelas de Galdós informan de los silencios que rodean a la representación del discurso de los personajes. En este análisis se muestra cómo los silencios se encuentran frecuentemente narrados de forma retrospectiva, como se muestra en (2). Como se puede advertir en el ejemplo, a los lectores se nos informa de las palabras del personaje y luego de que estas tienen lugar tras una pausa. Esto comporta una serie de implicaciones estilísticas que, tal vez debido a la falta de herramientas adecuadas, no han podido ser identificadas con anterioridad –o al menos no han podido ser analizadas de forma metódica– y que, en cierto modo, contribuyen a definir su estilo. En concreto, estas pausas suelen aparecer de nuevo en suspensiones (*suspensions*). Así pues, no se representa todo el acto de habla antes de informar de la pausa que tuvo lugar antes de este, sino que el acto de habla se introduce parcialmente, para interrumpirlo e informar de la pausa que en realidad tuvo lugar antes de su pronunciación, y luego continuar con el resto del acto de habla. Este recurso sirve para provocar un momento de tensión en el acto de lectura que sirve para reflejar la tensión de la situación que tiene lugar en el propio diálogo de la historia. Resulta interesante que este rasgo sea un hábito que domine la producción de Charles Dickens y caracterice su estilo, pues ayuda a reforzar el eco *dickensiano* en

[1] Todos los ejemplos citados en el libro han sido extraídos de una versión electrónica de los textos (véase apartado 3.4.1). Por lo tanto, en lugar de la paginación se ofrece como referencia el capítulo en el que aparecen.

la obra de Galdós que ya se estudiaba en el trabajo señalado anteriormente –y del que en este libro nos ocupamos en el capítulo 9–.

(2) –No me será muy difícil creer *–dijo después de una larga pausa–* que no estoy delante de un ladrón, bandolero, o asesino. Bien veo por su lenguaje que no pertenece usted a esa pobre clase plebeya de la cual salen todos los malvados (...). (*Un voluntario realista*, capítulo 20)

Finalmente, la evolución de Galdós a través de la frecuencia y el uso de distintas estrategias de representación del discurso también ha sido analizada utilizando un enfoque de estilística de corpus (Nieto Caballero 2019d). Gracias al procesamiento de las novelas del autor canario con un *software* de concordancias, en este análisis se muestra una opción de análisis mediante el uso de la forma verbal *dijo* que permite identificar aspectos estilísticos relevantes de la producción del autor, así como calibrarlos a lo largo de su trayectoria. Este análisis permite, por ejemplo, distinguir una fase inicial en la producción de Galdós bien diferenciada del resto de su trayectoria, al menos en lo que a representación del discurso verbal se refiere. En esta fase inicial el autor utiliza la forma *dijo*, de media, aproximadamente el doble (0,37 %) que en el resto de sus obras (0,19 %). La distinción entre ambos estadios de la producción del autor canario la marca *La desheredada*. La crítica ya se había referido a esta obra como el texto que marca un cambio de paradigma en la técnica narrativa de Galdós, ya que es considerada por muchos la primera novela naturalista. Gracias a una metodología de corpus, este cambio de paradigma puede calibrarse con un alto grado de precisión e identificar cómo el autor comienza a utilizar técnicas narrativas como el monólogo interior, el estilo indirecto libre o el multiperspectivismo, entre otras. Este estilo, que se articula en gran medida en torno al plano psicológico de los personajes, supone un cambio con respecto a la etapa inicial y se traduce en una pérdida de protagonismo de la voz del narrador. La metodología de corpus como la que se emplea en este estudio permite explicar, con la precisión que ofrecen los datos, el cambio de estilo que algunos críticos literarios ya habían vislumbrado con buen criterio en el pasado.

Sin duda, este libro sigue la estela marcada por estos últimos trabajos que acabamos de mencionar más que la de los multidisciplinares apuntados más arriba sobre traducción (Ruano San Segundo 2015) e intertextualidad (Nieto Caballero 2019a). Sin embargo, en este libro lo hacemos desde una perspectiva concreta. Así, en el análisis que planteamos nos ocuparemos de cuestiones

relacionadas fundamentalmente con la creación de los mundos textuales, con el fin de demostrar cómo los universos ficticios que el autor nos plantea son el resultado de la utilización hábitos estilísticos hasta ahora desapercibidos en la exégesis de su obra. Gracias a un enfoque de corpus podremos calibrar con precisión la sistematicidad de estos hábitos y su relevancia estilística. El hecho de centrarnos en la creación de los mundos textuales no es casualidad, pues se trata de uno de los puntos de partida de análisis más habituales de los estudios de estilística de corpus, como se explica a continuación.

3.3. La dimensión exegética de la estilística de corpus

El objetivo de un análisis de estilística de corpus es, ante todo, explicar el porqué de un fenómeno lingüístico desde un punto de vista literario. Esta fundamentación teórica sitúa a la estilística de corpus en el marco de los estudios explicativos (*explanatory*) –y no en los descriptivos (*descriptive*)– a los que Leech (2008) se refiere cuando establece sus dos modelos de estilística. Del mismo modo, esta fundamentación teórica también establece puntos de partida de análisis muy concretos, pues, además de la producción de un autor, el interés de un estudio de estilística de corpus puede residir en un solo texto o, por qué no, en una palabra o segmento concreto –la fórmula *dijo abrazándo(le)(se)(me)* comentada en el apartado 3.2.2 en el estudio de Nieto Caballero (2018), por ejemplo–. Cabe destacar, sin embargo, que saber qué queremos analizar dentro de la producción de un autor (o en una de sus obras) o la manera de enfocarlo no siempre es una tarea sencilla, pues, como acertadamente apunta Wales (2001, 237), «language is literary not in a special sense, but only in the sense that it belongs to a work regarded generically as literature, as opposed to a newspaper or recipe». Un enfoque de estilística de corpus puede ayudarnos a acotar el tipo de análisis que queremos llevar a cabo, pues es capaz de detectar aspectos relevantes desde un punto de vista lingüístico que pueden comportar un ejercicio de *literariedad* (*literariness*) (Carter 2004, 69) susceptible de análisis. En este sentido, conviene mencionar dos aspectos clave de los estudios de estilística de corpus que nos ayudarán a comprender su dimensión exegética: la comparación y el concepto de desviación lingüística.

La comparación, de un lado, constituye una de las bases metodológicas de los estudios de estilística de corpus. Sea cual sea la masa de texto que se analice –desde la producción de un autor hasta el uso de una construcción concreta, como mencionábamos hace un momento–, esta se compara con otra

masa de texto similar en un corpus de referencia. Esta comparación juega un papel esencial en el análisis que se realiza, pues no solo se utiliza para calibrar la presencia de un fenómeno lingüístico en un corpus (el de estudio) frente a otro (el de referencia) para calibrar su representatividad –después de todo, describir el estilo de un texto no es sino analizar los rasgos que lo hacen distinto de otros–, sino que se convierte en un recurso con el que poner a prueba cuestiones enraizadas en teorías literarias (cf. Semino y Short 2004; Toolan 2009).

De otro lado, la desviación lingüística (en inglés, *foregrounding*) tiene que ver con la desviación de la norma como elemento susceptible de análisis por su relación con la representatividad literaria. Se trata de un componente esencial para entender cómo se construyen las metodologías de corpus. Además de ser uno de los aspectos sobre los que tradicionalmente se han construido las interpretaciones de textos literarios (cf. Culpeper 2001, 129 y ss.), en los estudios de estilística de corpus nos sirve para medir la literariedad a la que nos referíamos más arriba. Por definición, la desviación lingüística o *foregrounding* se relaciona directamente con la «unexpectedness, unusualness, and uniqueness» (Mukarovsky 1970, 53-54), elementos todos ellos susceptibles de análisis en el plano literario. Además, cabe destacar que un autor no tiene por qué ser consciente de esa desviación (Short 1996, 16). En esta línea, las metodologías de corpus resultan enormemente útiles frente a enfoques tradicionales. Así, aunque es cierto que un enfoque de corpus «cannot provide direct insights into the process of understanding and interpreting a text», sus metodologías «can reveal patterns of language that speakers of a language may not be aware of» (Mahlberg 2013, 11). Y eso incluye al autor de una obra literaria, naturalmente. Dicho de otro modo, un estudio de estilística de corpus nos puede ayudar a detectar casos de desviación lingüística que ayuden a identificar ejemplos de literariedad en un texto, que resulten significativos desde un punto de vista estilístico para la interpretación de dicho texto (van Peer 1986).

Sobre el concepto de desviación lingüística se erigen dos de los planos fundamentales de los análisis de estilística de corpus: la caracterización de los personajes que pueblan los universos ficticios de las historias y la creación de los mundos textuales. El concepto de caracterización tiene que ver con «how we form impressions of characters in our minds –not just characters themselves or their personalities–» (Culpeper 2001, 2). La relación del estudio de la caracterización con la estilística de corpus es fácil de intuir: un enfoque de corpus nos puede ayudar a identificar patrones de los que *a priori* no somos conscientes como lectores (o incluso como estudiosos de un texto) pero

que forman parte del conjunto de mecanismos literarios empleados por un autor para construir a esos personajes, que constituyen la base textual sobre la que realizamos abstracciones que nos permiten transformar ese texto en criaturas concretas dentro de los universos ficticios (véase Nieto Caballero y Ruano San Segundo 2020, 55-58). Este es uno de los puntos de partida más habituales de los estudios de estilística de corpus. El otro tiene que ver con la creación de mundos textuales, del que nos ocupamos por separado en el siguiente apartado por ser un pilar fundamental del bloque analítico de este libro (capítulos 4, 5 y 6).

3.3.1. La creación de mundos textuales

Los mundos textuales se definen como el conjunto de «mental representations constructed in the process of discourse comprehension» (Mahlberg 2013, 34). La estilística de corpus no es la disciplina de estudio que los analiza de manera específica –de ello se ocupa la teoría de los mundos textuales (*Text-World Theory*) (Werth 1999; Gavins 2007)– pero sus características (textuales) los convierten en el marco sobre el que suelen plantearse numerosos estudios de este campo. Este libro no es una excepción. Los mundos textuales –a los que también nos referiremos como universos ficticios– definen las circunstancias físicas y el contexto social que hacen posible la existencia de los personajes que pueblan la historia. Desde un punto de vista textual, los universos ficticios se configuran mediante la combinación de bloques textuales que se materializan en funciones literarias determinadas. Aunque pueda (y deba) resultar una obviedad incidir en ello, conviene mencionar que un mundo textual no existe fuera del texto literario, incluso aunque represente una realidad que sí exista (una ciudad, por ejemplo). En otras palabras, aunque la historia de una novela transcurra en una ciudad concreta (la Barcelona de Zafón, por ejemplo), la representación de esta es fruto de una construcción literaria que solo existe en el plano textual, al igual que el resto de detalles que hay en ella (los personajes que pueblan ese mundo textual, por ejemplo). Si todo ese mundo textual existe únicamente en el plano literario, parece claro que su construcción dependerá ante todo de las elecciones del autor para darle forma en el plano textual. Es por ello que todos los segmentos de texto empleados para configurar los mundos textuales son susceptibles de análisis, pues, como apunta Page, son parte de un texto finito y, por tanto, potencialmente relevantes desde un punto de vista estilístico:

> Detail in a work of fiction, whether of action, description or speech, and however apparently fortuitous or excessive, can hardly be dismissed as irrelevant, since it belongs to the strictly finite amount of material laid at our disposal by the writer, as distinct from the unselective and virtually unlimited offering made by «reality» (Page 1973, 2).

En general, los estudios de estilística (Fowler 1986; Leech 1969, 1985; Leech y Short 2007 [1981]; Semino 1997; Short 1996; Simpson 1993, 1996; Toolan 1988, entre otros) comparten una suerte de objetivo común independientemente de cuál sea el objetivo específico que plantean, pues de un modo u otro todos analizan textos con el fin de «arrive at a detailed account of how readers understand particular texts in the ways they do» (Short y Semino 2008, 117). La estilística de corpus tiene una motivación teórica similar, con el matiz de que las metodologías de estudio permiten identificar patrones estilísticamente relevantes que tradicionalmente han pasado desapercibidos en el análisis de esos mismos textos. En el caso de los mundos textuales, la sistematicidad propia de las metodologías de corpus permite identificar hábitos literarios que pueden resultar relevantes en la construcción de los universos ficticios y tienen una incidencia directa en cómo los percibimos. Dos conceptos son especialmente significativos en este sentido: los bloques textuales (*textual building blocks*) (Mahlberg 2013, 26)[2] y las funciones textuales locales (*local textual functions*) (Mahlberg 2007a, 4). Los bloques textuales son unidades lingüísticas (pueden ser palabras o segmentos de textos concretos) que contribuyen a dar forma a la realidad literaria que se nos presenta, mientras que las funciones textuales locales tienen que ver con el valor de los mencionados bloques textuales en un corpus de estudio concreto –la función de las *calles* como elemento que ayuda a configurar el espacio en la producción narrativa de Galdós, aspecto del que nos ocupamos en el capítulo 6, por ejemplo–. Las características de las metodologías propias de los enfoques de corpus permiten localizar tanto bloques textuales como sus funciones textuales locales, lo que hace posible identificar hábitos estilísticos que tradicionalmente han pasado desapercibidos en la exégesis de un autor pero que atesoran un valor literario claro. Esa es la razón por la que la creación de los mundos textuales constituye un punto de partida habitual en los estudios de estilística de corpus, como ocurre en este libro.

[2] Los bloques textuales (*textual building blocks*) también reciben otros nombres en la literatura científica en inglés como *basic building-blocks* o *world-building elements* (Gavins 2007, 36).

3.3.2. El peligro de la simplificación

Como en cualquier disciplina, un análisis de estilística de corpus ha de estar debidamente motivado. Esto es, la disponibilidad de herramientas y su capacidad para localizar datos que no podrían identificarse con una lectura atenta no deben ser las únicas razones que guíen un análisis estilístico en el que se adopte una metodología de corpus. De hecho, ni siquiera deberíamos considerarlo un análisis. Incidimos en esto porque, en ocasiones, la accesibilidad de los textos en formato electrónico y la posibilidad de procesarlos con un *software* parecen equipararse a la realización de un análisis estilístico. Sin embargo, como señala Mahlberg (2014, 388-389),

> [r]unning a concordance or generating key words is not in itself a useful research method if it is not applied to address a particular research question. The application of corpus procedures without searching for theoretically grounded links to interpretative concerns bears the danger of uninsightful and naive observations on a text.

Este es un problema habitual de quienes se acercan por primera vez a la estilística de corpus. Efectivamente, las metodologías invitan a presentar una gran cantidad de información. Y dado que esa información no es localizable con una lectura de los textos, resulta tentador presentar los resultados obtenidos como si fueran un análisis. Sin embargo, que una palabra sea más frecuente comparativamente en la obra de un autor o que su uso se concentre en una época determinada no es en sí mismo un análisis estilístico, por citar un par de ejemplos fáciles de entender. El análisis estilístico debe tener una motivación teórica, con un punto de partida y unas tesis que deben establecerse *a priori*. Es decir, los datos deben ser el material que nos permita realizar una interpretación cabal en combinación con la literatura científica sobre un asunto. Ho (2011, 11) lo sintetiza muy bien cuando dice que si un investigador «is unable to incorporate or weave his/her quantifications into a synthetic discussion of literature, the findings will always remain cold numbers, alien to the field of the humanities». En definitiva, un análisis estilístico que adopte un enfoque de corpus debe estar debidamente motivado y construido sobre un marco teórico bien definido. La simple identificación de datos (aunque puedan resultar inéditos en la dimensión exegética de un autor) no constituye, por sí misma, un análisis de estilística de corpus.

Además, cabe destacar que, aunque a veces exista una clara motivación teórica y un marco teórico bien definido, debemos ser igualmente conscientes de que «[n]ot every research question can be answered with such methods» (Mahlberg 2016, 153), pues no todos los análisis estilísticos se adaptan a las características propias de un enfoque de corpus. Normalmente, el conflicto suele plantearse en relación con el texto –o los textos– que se analizan. O, mejor dicho, con su extensión. Así, un poema no suele aceptar una metodología de corpus para investigar cuestiones de tipo estilístico, salvo que el análisis se construya en torno a la comparación del poema con un corpus de referencia y esté debidamente motivado –analizar la prosodia semántica de alguna palabra o construcción empleada en el poema, por ejemplo–. Incluso la prosa narrativa, que *a priori* resulta ideal para implementar metodologías de corpus, puede plantear dificultades. Como señala Mahlberg (2014, 387), por ejemplo,

> [a] short story may not yield a sufficient number of 5-word clusters for a useful analysis, or a word that appears interesting in a text extract from a novel may be insufficiently frequent to show patterns in the form of a concordance. Even if there were a reasonable number of key words for a text, the analysis of these key words might not provide observations that have much to contribute to the literary stylistic analysis.

En suma, para adoptar un enfoque de estilística de corpus es necesario que se den ciertas condiciones que favorezcan la implementación de una metodología de estas características. Consideramos pertinente detenernos en esta suerte de nota de precaución dentro de este apartado sobre la dimensión exegética de la estilística de corpus porque no todos los textos (ni todos los análisis estilísticos) son igual de susceptibles al empleo de metodologías de corpus. En el caso de este libro, el estudio de la obra narrativa de Pérez Galdós (6,5 millones de palabras aproximadamente) encuentra en la estilística de corpus un enfoque idóneo para procesar tal cantidad de texto. De hecho, esta vasta extensión es, en sí misma, una motivación para el análisis, pues la cuantificación –y el posterior análisis– de determinados hábitos estilísticos que se repiten a lo largo de sus novelas resulta prácticamente imposible de identificar y de relacionar con la literatura científica disponible mediante una lectura atenta de los textos.

3.4. Creación de un corpus

En este apartado, por último, nos ocupamos de los principios que deben guiar la compilación de un corpus. Para ello estableceremos una distinción. En primer lugar, nos ocuparemos de las características específicas que han de guiar la compilación de un corpus de estudio (apartado 3.4.1). Además de detallar los aspectos más importantes desde un punto de vista teórico, también explicaremos qué textos forman nuestro corpus de estudio de la producción narrativa de Pérez Galdós. En segundo lugar, nos ocuparemos de las características más importantes de los corpus de referencia. Igualmente, además de explicar los aspectos más importantes que han de guiar la elaboración de este tipo de corpus, detallaremos qué textos hemos incluido en nuestro corpus de referencia de novela decimonónica española (apartado 3.4.2) que utilizaremos como base de comparación en nuestro análisis. Antes de ello, sin embargo, cabe mencionar que los principios que rigen la construcción de un corpus en los trabajos de estilística de corpus son los mismos que se aplican, salvo matices de los que nos ocupamos a continuación, en los trabajos de lingüística de corpus en general. Así pues, las muestras de texto deben ser completas, salvo que alguna limitación no lo haga posible (la disponibilidad de los textos objeto de estudio, por ejemplo); la selección de textos debe explicarse, de modo que su pertinencia para el tipo de análisis que se lleve a cabo quede justificada; y si el texto contiene algún etiquetado o anotación concretas, este debe almacenarse por separado, de modo que no sea cuantificado como texto[3]. Más allá de estas generalidades, que como decíamos resultan de aplicación a la construcción de cualquier corpus, ya sea en estudios de lingüística o de estilística, en la elaboración de los corpus de estudio y de referencia de los trabajos de estilística de corpus se aplican principios que son propios y que conviene abordar con detenimiento, pues resultan decisivos a la hora de definir una base textual que permita un análisis estilístico adecuado.

3.4.1. Características de un corpus de estudio

En cuanto a las características de los corpus de estudio que son propias de los análisis estilísticos, cabe destacar que las muestras de análisis suelen ser

[3] Para una información más detallada sobre en qué consiste un etiquetado en los corpus de estilística de corpus en general, véase Nieto Caballero y Ruano San Segundo (2020, 117-122).

menos extensas que en los análisis de lingüística de corpus. La razón para que esto es así parece lógica: el nivel de especificidad de los estudios de estilística de corpus, que como ya hemos dicho, puede estar interesado en textos concretos (una novela, por ejemplo) o incluso en fragmentos de textos, hace que los corpus en ocasiones sean más reducidos de lo habitual, pues deben posibilitar «not just the quantitative work normally associated with corpus linguistics but also the qualitative study of individual texts from the corpus, a type of analysis which is typical of stylistics» (Semino y Short 2004, 201). Para hacer posible la dimensión cualitativa y el grado de concreción analítica propios de los estudios de estilística de corpus, la muestra de texto es habitualmente más específica y de menor tamaño que en un análisis de lingüística de corpus.

Además, la muestra de texto que forme un corpus de estudio debe ser coherente con el tipo de análisis que se quiera llevar a cabo (cf. Semino y Short 2004, 39). En otras palabras, si un autor escribe más de un género (novela y teatro, por ejemplo) y queremos analizar su producción narrativa, nuestro corpus de estudio no deberá incluir ninguna obra dramática, a pesar de que ello supusiera una masa de texto mayor. Hacerlo distorsionaría el análisis que pudiéramos realizar, pues trabajaríamos con resultados que no guardarían relación con el tipo de análisis planteado. En el caso de Galdós del que aquí nos ocupamos, por ejemplo, estamos interesados en analizar aspectos de su producción narrativa. Por lo tanto, nuestro corpus de estudio no contendrá ningún texto correspondiente a su producción teatral, por escasa que esta sea. En otras palabras, y aunque pueda parecer una obviedad incidir en ello, los textos que se incluyan en un corpus de estudio habrán de estar directamente relacionados con la hipótesis que se pretenda investigar.

Finalmente, también podemos mencionar algunas pautas relacionadas con la siempre controvertida cuestión del número de textos que debemos incluir en nuestro corpus para construir una muestra de estudio de forma adecuada, si bien es cierto que no existe una fórmula que se pueda aplicar a todos los casos. En primer lugar, la selección ha de estar guiada por la hipótesis de análisis. Así, si lo que pretendemos es analizar la obra completa de un autor, el corpus debe estar formado, si las circunstancias lo permiten, por toda su producción. En estos casos, naturalmente, aquellos autores con una producción narrativa más vasta (el caso de Galdós) darán como resultado un corpus de estudio más extenso que en el caso de autores con una producción narrativa menor (Clarín, por ejemplo). Por el contrario, si nuestra hipótesis se constru-

ye sobre la obra de distintos autores (en el caso de querer analizar aspectos estilísticos de un movimiento literario, por ejemplo), intentaremos mantener un equilibrio en la representatividad de estos en muestro corpus. Esta representatividad se puede medir en número de obras o en palabras totales. La segunda opción es la más adecuada, pues la extensión de las novelas puede generar desequilibrios significativos –tres novelas de un autor pueden ser menos extensas que una novela de otro–. De nuevo, en estos casos entrará en juego la disponibilidad de los textos, que influirán en las decisiones que tomemos. En cualquier caso, nuestra selección siempre deberá atender a la hipótesis de partida, de tal suerte que quede demostrado que se trata de la mejor elección y no de una selección de textos que pretenden satisfacer una hipótesis determinada. Precisamente por eso, resulta necesario detenerse a explicar las razones y las decisiones detrás de la elección del corpus de estudio que empleamos en este libro.

La razón principal por la que hemos escogido a Galdós como autor al que analizar en este libro tiene que ver, más allá de las motivaciones estilísticas concretas que iremos desgranando en los capítulos 4, 5 y 6, con que se trata de uno de los novelistas españoles más laureados de siempre –quizá solo por detrás de Cervantes–. Además, es un escritor con una producción narrativa muy extensa, que publicó más de ochenta novelas a lo largo de su vida. Esta prolijidad lo convierte en un autor ideal atendiendo a la dimensión cuantitativa que caracteriza a los enfoques de corpus, pues el procesamiento de sus novelas con las herramientas adecuadas puede hacer que localicemos hábitos estilísticos a lo largo de su trayectoria que no han sido identificados con anterioridad. Se trata, además, de un autor perteneciente a una de las épocas doradas de la novela española: el realismo. Este hecho, que para la construcción del corpus de estudio no tiene mayor trascendencia, sí que resulta importante a la hora de comparar su producción con la de otros autores con los que compartió plana, pues nos permitirá comparar su producción con la de un corpus de referencia y calibrar hasta qué punto los hábitos que identifiquemos pueden ser considerados idiosincrásicos del autor canario.

Sobre las novelas incluidas en el corpus de estudio de Pérez Galdós (al que también nos referiremos a partir de ahora como CorBPG), cabe señalar que han sido descargadas en su totalidad del repositorio digital de la Biblioteca Virtual Miguel de Cervantes (s.f.), al que ya nos hemos referido en el apartado 2.2.1. Toda la producción literaria del autor canario se encuentra almacenada en este repositorio. En concreto, para llevar a cabo este análisis hemos seleccionado su obra narrativa, formada por 89 novelas y con una extensión

de aproximadamente seis millones de palabras. En la Tabla 1 se muestra un resumen de las novelas incluidas en CorBPG[4].

Tabla 1. Resumen de novelas incluidas en CorBPG

Novelas	Palabras
Episodios nacionales. Primera serie	967 310
Episodios nacionales. Segunda serie	640 865
Episodios nacionales. Tercera serie	759 545
Episodios nacionales. Cuarta serie	774 255
Episodios nacionales. Quinta serie	433 611
Novelas serie primera época	683 376
Novelas serie contemporánea	2 184 747
Total palabras	6 443 709

Fuente: elaboración propia.

Los motivos por los que se han escogido sus 89 novelas tienen que ver con el propósito del estudio: dado que en nuestro análisis investigamos hábitos estilísticos de Galdós como novelista, sin duda lo más adecuado parece ser incluir toda su producción narrativa, como se apuntaba en el apartado anterior. Es decir, hemos dejado fuera su (escasa) producción teatral. Para ser lo más coherentes posible, además, hemos optado por no incluir su narrativa breve, por no tratarse de novelas propiamente dichas. Huelga decir que el corpus está formado por novelas completas, que nos ayudará a realizar un análisis más sólido que si hubiéramos escogido una muestra parcial de estas, como hace por ejemplo Irizarry (1997) en el estudio al que hacíamos referencia en el apartado 3.2.1. En definitiva, CorBPG se ajusta a los principios de compilación de un corpus de estudios a los que nos referíamos en el apartado anterior. Esta selección de textos permitirá realizar un análisis cabal del estilo de Galdós como novelista. Como confiamos en ser capaces de demostrar en los siguientes capítulos, aplicar un enfoque de corpus sobre las 89 novelas de Galdós nos ayudará a descubrir aspectos del estilo del autor canario que, o bien han pasado desapercibidos para la crítica, o bien no han podido ser analizados de manera sistemática.

[4] Para una información detallada de las novelas incluidas en CorBPG véase el Apéndice 1.

3.4.2. Características de un corpus de referencia

Por su parte, los principios que guían la construcción de un corpus de referencia son ligeramente distintos a los que se aplican a la construcción de un corpus de estudio que acabamos de mencionar. Esto se debe a la función que este corpus desempeña en el análisis. Un corpus de referencia es una muestra de texto que nos permite calibrar la representatividad de lo que hemos encontrado en nuestro corpus de estudio mediante una comparación con una serie de textos que, por su naturaleza, admiten un tratamiento similar a los del corpus de estudio. En otras palabras, un corpus de referencia sirve, ante todo, para comparar y dar validez a las conclusiones que se extraen del procesamiento del corpus de estudio. Como ya decíamos en el apartado 3.3, la comparación es un elemento clave en los estudios de estilística de corpus. Es por ello que la compilación de un corpus de referencia constituye un aspecto de importancia cardinal en este tipo de análisis. Su diseño, sin embargo, no sigue unas pautas específicas que se apliquen a cualquier corpus de referencia que queramos construir. De hecho, como admite Culpeper «[t]here is no magic formula» (Culpeper 2002, 15) para seleccionar una muestra de texto adecuada. En la misma línea se expresa Mahlberg (2013, 9), quien asegura que «[f]inding an appropriate general purpose corpus […] is not a straightforward matter». Lo que sí parece claro es que un mal corpus de referencia puede invalidar las conclusiones que se extraigan de un análisis. Es por ello que si bien no existe una fórmula mágica (por utilizar las palabras de Culpeper) para diseñar un corpus de referencia, sí que podemos una clave que puede ayudarnos a realizar nuestra selección de textos: la representatividad de la muestra de texto.

El concepto de representatividad está estrechamente relacionado con la dimensión cuantitativa propia de los enfoques de corpus en general. En concreto, un corpus de referencia debe ser lo suficientemente representativo (o, si se quiere, extenso) como para admitir una comparación con el corpus de estudio. Biber lo sintetiza muy bien cuando dice que todo corpus de referencia debe diseñarse con el fin de «represent “typical” patterns of use, making it possible to empirically identify distinctive linguistic patterns in the target corpus that depart from those typical patterns» (Biber 2011, 16). En efecto, un corpus de referencia formado por una cantidad de texto escasa difícilmente admitirá una comparación con un corpus de estudio, pues no será una masa de texto representativa. Es cierto que el concepto de representatividad es relativo, pues «corpora cannot provide “absolute” norms, but provide “relative” norms instead» (Mahlberg 2013, 9-10). Sin

embargo, sí que podemos asegurar, sin miedo a equivocarnos, que para que un corpus de referencia sea representativo deberá tener, al menos, una extensión similar a la de nuestro corpus de estudio. De este modo, el corpus de referencia será, al menos teóricamente, una muestra de texto susceptible de arrojar resultados similares a los obtenidos en el corpus de estudio –en términos cuantitativos, lógicamente–. Una muestra de texto representativa garantizará que la comparación a la que sometamos los resultados obtenidos con nuestro corpus de análisis no esté desvirtuada. Y eso, en sí mismo, es un factor clave desde un punto de vista procedimental. En el caso de este libro, hemos intentado respetar la cuestión de la representatividad a la hora de seleccionar los textos de nuestro corpus de referencia.

En concreto, nuestro corpus de referencia (al que en adelante nos referiremos también como CorXIX) se nutre fundamentalmente de novela decimonónica española. Al igual que en el caso de CorBPG, CorXIX está formado por novelas completas. En concreto, nuestro corpus de referencia lo componen 79 novelas de ocho autores canónicos: Pedro Antonio de Alarcón, Vicente Blasco Ibáñez, Leopoldo Alas «Clarín», Luis Coloma, Armando Palacio Valdés, Emilia Pardo Bazán, José María de Pereda y Juan Valera. Estas 79 novelas dan como resultado una masa de texto de más de seis millones de palabras, lo que supone una extensión similar a la de CorBPG. En la Tabla 2 se muestra un resumen de la presencia de cada autor en el corpus[5].

Tabla 2. Resumen de novelas incluidas en CorXIX

Autor	Núm. novelas incluidas	Palabras
Pedro Antonio de Alarcón	4	186 039
Vicente Blasco Ibáñez	12	1 191 131
Leopoldo Alas «Clarín»	2	399 765
Luis Coloma	7	526 070
Armando Palacio Valdés	4	375 981
Emilia Pardo Bazán	20	1 294 616
José María de Pereda	19	1 550 494
Juan Valera	11	660 769
Total	79	6 184 865

Fuente: elaboración propia.

5 Para una información detallada de las novelas incluidas en CorXIX véase el Apéndice 2.

La selección de textos incluidos en CorXIX se ha llevado a cabo teniendo en cuenta la noción de representatividad comentada en el apartado anterior. En efecto, la masa de texto de seis millones de palabras admite, cuantitativamente, una comparación cabal con la masa de texto de CorBPG. Se trata, además, de un corpus formado por (i) obras literarias pertenecientes al género narrativo, (ii) son novelas completas y (iii) fueron publicadas en un espacio de tiempo similar al de las novelas de Galdós, por lo que resulta una selección que cumple con los criterios habituales de compilación de un corpus y admite una comparación con CorBPG también desde un punto de vista cualitativo –la comparación de un fenómeno lingüístico concreto en ambos corpus, por ejemplo–.

Conviene detenerse, aunque sea brevemente, en el claro desequilibrio que se advierte en lo que a la presencia de cada autor en CorXIX (medida en número de palabras) se refiere. Esta presencia desigual de cada autor no tiene que ver con un descuido a la hora de incluir más o menos obras de distintos novelistas en el corpus, sino con el grado de prolijidad de estos. Así, autores como José María Pereda, con una vasta producción, tienen una presencia mayor en el corpus que otros como Clarín, con una reducida producción narrativa. Además, cabe destacar que en el caso de aquellos autores más prolijos (el propio Pereda) se ha llevado a cabo una selección de como mucho veinte títulos de su bibliografía, de tal suerte que las diferencias entre los distintos autores no sean excesivamente acusadas. Además de por cuestiones literarias o de extensión de las novelas, esta selección está directamente influida por la disponibilidad de los textos en la Biblioteca Virtual Miguel de Cervantes. Así, algunas obras de los autores que aparecen en CorXIX no han sido necesariamente descartadas de forma deliberada, sino que no han podido ser integradas en el corpus de referencia por no encontrarse digitalizadas en el repositorio que hemos utilizado (el de la Biblioteca Virtual Miguel de Cervantes). En cualquier caso, y dejando a un lado estas (poco relevantes) limitaciones de la Biblioteca Virtual Miguel de Cervantes, creemos que CorXIX se adapta a la noción de representatividad que debe guiar la construcción de un corpus de referencia, así como a los principios básicos de selección de cualquier corpus en general. Es por ello que este corpus de 79 novelas de ocho autores canónicos constituye, a nuestro juicio, una muestra de texto fiable para poder realizar comparaciones con las que calibrar los resultados obtenidos en nuestro análisis estilístico de la obra narrativa de Galdós.