REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
 54(107) 919-941

Notas epistemológicas sobre variación para una lingüística de corpus

Epistemological notes on variation for corpus linguistics

Francisco Moreno Fernández
UNIVERSITÄT HEIDELBERG
ALEMANIA
UNIVERSIDAD DE ALCALÁ
ESPAÑA
francisco.moreno@uni-heidelberg.de

Recibido: 14-VI-2021 **/ Aceptado:** 15-X-2021
DOI: 10.4067/S0718-09342021000300919

Resumen

Este estudio ofrece reflexiones y comentarios sobre el modo en que la variación se recoge y refleja en los corpus lingüísticos. En unas ocasiones, los corpus se construyen prestando atención específica a los fenómenos variables de las lenguas y sus modalidades; en otras, los corpus no se interesan específicamente por la variación de la lengua, pero ello no significa que no tengan capacidad de reflejarla en sus distintas manifestaciones. En estas páginas se presta atención a la dimensión dialectal de los corpus lingüísticos, atendiendo especialmente al modo en que la configuración geográfica se refleja en ellos y a la representatividad de las áreas lingüísticas. Asimismo, se atiende a las dimensiones social y estilística de los corpus. Para ello se explica el tratamiento dado a estas dimensiones tanto en los corpus especializados, como en los corpus generales o de referencia. Finalmente, se proponen unas reflexiones sobre cuestiones de fondo concernientes a la metodología de la elaboración de corpus, como la representación de la diversidad dialectal dentro de espacios geográficos de mayor amplitud o su relación con el tamaño de los corpus.

Palabras Clave: Corpus, dialectología, sociolingüística, variación, metodología.

Abstract

This study offers reflections and comments on the way in which variation is captured and reflected in linguistic corpora. On some occasions, corpora are built with specific attention to the variable phenomena of languages and their modalities; on other occasions, corpora are not specifically interested in language variation, but this does not mean that they do not have the capacity to reflect it in its different manifestations. These analyze the dialectal dimension of the linguistic corpora, paying special attention to the way in which the geographical configuration is reflected in them and to the

representativeness of the linguistic areas. Attention is also paid to the social and stylistic dimensions of the corpora. To this end, the treatment given to these dimensions in both specialized corpora and general or reference corpora is explained. Finally, some reflections are proposed on fundamental questions concerning the methodology of corpus elaboration, such as the representation of dialectal diversity within larger geographical spaces or its relation to the size of the corpora.

Key Words: Corpus, dialectology, sociolinguistics, variation, methodology.

INTRODUCCIÓN

En el artículo "La variación geográfica y social en los corpus hispánicos" (Moreno Fernández, 2021) presenté una reflexión sobre los corpus del español más relevantes para el estudio de la variación geográfica y social. Estos corpus se describían por sus características y objetivos generales, pero detallando los fundamentos metodológicos sobre los que se había construido su configuración. En ese estudio se explicaba que la elaboración de corpus con aplicaciones dialectológicas y sociolingüísticas implica el tratamiento de cuestiones teóricas y metodológicas relevantes (Anderwald, 2009; Hansen, 2018). Por esta razón, el desarrollo futuro de tales corpus estaría ligado en gran medida a su evolución metodológica, incluidas las posibilidades que pueda ofrecer la tecnología.

El mencionado texto se elaboró con una intención comprehensiva y abarcadora, como puede esperarse del manual universitario en el que se incluía como capítulo (Parodi, Cantos & Howe, 2021). Sin embargo, las cuestiones que ese formato obligaba a soslayar o a tratar de pasada eran muchas. Estas páginas servirán en gran parte para paliar algunos de los aspectos entonces soslayados, así como para introducir algunas reflexiones complementarias de orden epistemológico. La lectura de este texto adquiere pleno sentido si se conecta con la del capítulo mencionado. No obstante, aquí se ofrece un texto autónomo, que no exige mirar de reojo ni de modo constante al capítulo primigenio.

La intención de estas páginas es, pues, abordar cuestiones relativas a la variación dialectal y sociolingüística en los corpus lingüísticos, prestando especial atención a aspectos teóricos y metodológicos referidos tanto a los criterios seguidos actualmente, como a sus desarrollos potenciales.

1. La dimensión dialectal en los corpus lingüísticos

Cualquier reflexión detenida sobre las condiciones en que la variación dialectal o geolingüística se hace presente en los corpus lingüísticos conduce indefectiblemente a cuestiones de fondo sobre la naturaleza misma de la lengua hablada. En este sentido, cabe distinguir entre los aspectos geosociolingüísticos que afectan a todo corpus y aquellos que se manejan específicamente para su construcción. Esta distinción resulta

apropiada tanto para la lengua de los corpus, como para cualquier otra manifestación o producto lingüístico.

Efectivamente, podría decirse que toda muestra de lengua, en cualquiera de sus modalidades, ofrece un perfil dialectal y sociolingüístico, en sentido amplio, que debería tenerse en cuenta para su comprensión y análisis. Y esto sería así independientemente de su modo, campo y tenor, siguiendo la distinción establecida por Halliday (1979) a propósito de los registros. Como prueba de ello podría servir la experiencia de De Kock, quien utilizó como base, para el análisis gramatical del español, dos corpus de lengua escrita, en su mayoría de obras firmadas por autores de primer nivel. El objetivo principal de De Kock era llegar a conclusiones válidas para la comprensión de la gramática del español y, muy especialmente, para su enseñanza, a partir de autores y textos representativos. Con este fin, partió precisamente de consideraciones que afectan a la variación lingüística; por un lado, referidas al estilo; por otro, referidas a la procedencia geográfica de los autores.

En lo que afecta a la procedencia geográfica, los textos elegidos por De Kock (1990) para su primer corpus (19) estaban firmados por Dámaso Alonso, Francisco Ayala, Américo Castro, Camilo José Cela, Rafael Sánchez Ferlosio, José Gaos, Juan Goytisolo, Antonio Machado, Juan Ramón Jiménez, José Ortega y Gasset, Ramón Pérez de Ayala, Pedro Salinas, Miguel de Unamuno (españoles), Mario Benedetti (uruguayo), Alejo Carpentier (cubano), Rómulo Gallegos (venezolano), Ezequiel Martínez Estrada (argentino), Alfonso Reyes (mexicano) y Mario Vargas Llosa (peruano). El segundo corpus de De Kock (1992) (20 textos) incluyó escritos de filólogos, pensadores y escritores distribuidos por su origen geográfico entre España (12 autores), Argentina (3), México (2), Cuba (1), Colombia (1) y República Dominicana (1).

A pesar de que la búsqueda del 'pluralismo geográfico', como hemos visto, fue decisiva en la construcción del corpus de De Kock, lo cierto es que el gramático belga le quitó trascendencia lingüística, singularmente gramatical, como se aprecia en la siguiente cita (De Kock, 1990: 130):

> "Para algunos la dosificación de los autores según su procedencia geográfica es un factor decisivo de representatividad. La importancia que se le otorga parte en muchos casos de una reacción nacionalista o política, consciente o inconsciente. Aunque nos hemos esforzado en mantener el pluralismo geográfico, la adecuación del texto al propósito metodológico ha sido siempre nuestro primer criterio de selección antes que el lugar de nacimiento de su autor".

Es significativo que tan relevante información se aporte en nota a pie de página. De Kock (1990) aducía que los españoles a menudo vieron transcurrir su carrera en otro país, del mismo modo que los latinoamericanos pasaron su vida fuera de sus países respectivos, hecho que restaba relevancia a la estricta procedencia geográfica. Sin

embargo, la cuestión era insoslayable y De Kock (1990: nota 22) concluía su nota de modo claro: "las diferencias o semejanzas entre el español de España y de América serán señaladas".

A la luz de la experiencia de De Kock y de su proyecto sobre "Gramática española: Enseñanza e investigación" queda de manifiesto que la finalidad última de los corpus determina la importancia que se concede a los factores de variación lingüística, incluida la geográfica o dialectal, en la selección de las muestras textuales que los conforman. Sin embargo, queda también en evidencia que se trata de un factor ineludible. La lengua escrita y publicada en la prensa, muy presente en los corpus lingüísticos de referencia, a menudo pasa por representativa de un nivel promedio o 'estándar' de lengua, con capacidad de minimizar el impacto de la diversidad dialectal. Sin embargo, tal planteamiento no por recurrente se convierte en verdadero. Basta con pensar que hay estudios dialectológicos, cartografía incluida, realizados, por ejemplo, para la lengua alemana, que se elaboran a partir materiales publicados en prensa (Adarve, 2020).

En relación con la inevitable presencia del factor dialectal en los corpus lingüísticos, podría pensarse en una excepción que, hasta donde sabemos, no se ha convertido por ahora en realidad. Se trataría de un corpus conformado por muestras de la variedad denominada 'español neutro' (Gómez Font, 2012; Guevara, 2013), también conocido como 'español internacional', 'español globalizado' o incluso como 'español latinoamericano'. Esta 'variedad' se ha construido en las producciones audiovisuales de las grandes empresas de comunicación: es el español de productos culturales, informativos y de entretenimiento, comprensibles para cualquier hispanohablante sin que las diferencias se aprecien como significativamente extrañas; es el español de las cadenas de televisión que emiten sus noticieros para que sean seguidos en todo el mundo, y lo es también de numerosas producciones documentales creadas para la televisión. Este 'español neutro', como producto creado, enseñado, entrenado y difundido es consecuencia de la globalización y el llamado 'español latinoamericano', ofrecido como alternativa en muchos productos audiovisuales, es una de sus manifestaciones. Esta modalidad no existe como variedad natural, sino que emerge como un interdialecto artificioso y aprendido que ha conseguido ser ampliamente aceptado en calidad de registro característico del medio audiovisual, especialmente en América.

Siendo así, un corpus formado por muestras de este 'español neutro' o 'español latinoamericano' (escrito y hablado) parecería escapar a cualquier catalogación dialectal. En la práctica, sin embargo, no ocurriría así realmente, como puede deducirse del uso mismo de la etiqueta 'latinoamericano', que supone una filiación geolingüística, amplia, aunque concreta. Además, este español tendría una base compatible con una variedad específica: el español mexicano. En el plano léxico, esta modalidad propone soluciones de extensa implantación, pero generalmente propias también de México, lo que a

menudo supone excluir alternativas de otras áreas dialectales, como el Cono Sur, por no mencionar las áreas europeas o africanas.

2. Corpus y configuración de espacios lingüísticos

Todo ello nos obliga a considerar que la forma en que la variación dialectal aflora en los corpus se ve afectada por factores diversos. Uno de ellos, fundamental, es la propia configuración dialectal y social de cada espacio lingüístico. Así, es bien conocida la propuesta de Trudgill (1974), para el inglés británico, de una pirámide dentro de la cual las diferencias interregionales pueden observarse claramente en su base, donde aparecen los estratos socioculturales más bajos de una comunidad. Conforme los hablantes se ubican en estratos más altos, su forma de hablar va reduciendo sus posibilidades de variación dialectal, hasta llegar al vértice de la pirámide, alcanzado por las clases más altas, tras el sometimiento a un proceso educativo. Al inglés utilizado en ese vértice, por los hablantes más cultos, acomodados y prestigiosos de la comunidad británica, se le da el nombre de 'pronunciación recibida' (*Received Pronuntiation*), inglés de la Reina / Rey o inglés de la BBC. Como consecuencia de esta configuración, cabe pensar que los corpus británicos reflejan menores diferencias regionales o dialectales conforme su origen social y estilístico es más elevado, de modo que el inglés estándar, como el que se reúne en muchos corpus de referencia, apenas deja espacio para la variación dialectal. Ahora bien, la consideración conjunta de las principales variedades del inglés (británico, estadounidense, australiano…), incluidas sus propias variantes, alejaría la posibilidad de un corpus despojado absolutamente de elementos marcados dialectalmente.

El caso del español sería diferente, por cuanto, aceptando la representación piramidal, en la cima no aparecería 'una pronunciación recibida'. De hecho, no existe un vértice o cumbre como tal, sino una sucesión de cumbres de una serie de pirámides, que representarían las normas cultas de cada región hispánica. Cada pirámide correspondería a la modalidad de un ámbito geográfico determinado. En su base estarían los usos populares regionales, más visibles en los hablantes de niveles socioculturales bajos, y los usos populares o vulgares que son comunes a todo el dominio hispanohablante, ya que responden a tendencias generales de la lengua. La consecuencia lógica de ello es que los corpus de la lengua española, especialmente los de referencia, no pueden eludir la filiación geográfica o dialectal de sus muestras, que forzosamente han de adscribirse a una u otra de las pirámides regionales.

Las lenguas alemana e italiana ofrecen una configuración igualmente particular. Así, la situación actual del alemán en Alemania también distribuye los usos lingüísticos a lo largo de un eje estándar-dialectal, que opone una base con diferencias lingüísticas en el espacio geográfico a una zona de prestigio asociado a un 'alemán estándar', muy vinculado a la lengua escrita por razones históricas bien conocidas (Lameli, 2010). Si bien la situación dialectal no es idéntica en el norte y en el sur de Alemania, puede

hablarse de una permeabilidad creciente entre el alemán estándar escrito y las variedades de alemán habladas cotidiana y regionalmente, de modo que la brecha entre uno y otras se va cerrando hasta establecerse un continuo por el que los hablantes se mueven en función de sus necesidades comunicativas. Esta permeabilidad, más perceptible en el sur (alto alemán y alemán central) que en el norte (bajo alemán), también se produce en la lengua escrita, como la utilizada en la prensa. Por eso, los corpus construidos sobre materiales periodísticos ofrecen también una valiosa información sobre variación marcada geográficamente, incluida la variación gramatical (Adarve, 2020).

En cuanto a la lengua italiana, encontramos algunos paralelismos con el caso del alemán (Berruto, 2018), pero presenta sus peculiaridades. La relación entre el italiano, entendida como lengua estándar, y los dialectos o lenguas regionales se ha representado como una doble pirámide (también un doble cono), cuyo elemento superior ofrecería un vértice con el estándar (escrito y hablado) y una base con los estándares de cada uno de los dialectos, que, a su vez, son vértice del segundo elemento, en cuya base estarían las hablas locales y rurales (Auer, 2011). Esta situación da lugar a una constelación jerarquizada de variedades que, en la sociolingüística estadounidense podría ser catalogada como 'diglosia de esquema doble'. Aquí, lo interesante es precisar que el vértice de la pirámide o cono superior no es tal, sino que tiene forma truncada, ya que el italiano estándar propiamente dicho no coincide con ninguna variedad hablada realmente en una región determinada o por una clase o grupo social determinado De hecho, nadie en Italia puede considerarse un verdadero hablante nativo del italiano estándar. En cambio, existen variedades regionales estandarizadas de italiano, que son 'estándares por el uso' de acuerdo con el concepto propuesto por Ammon (2003). Esta compleja situación geosociolingüística ha determinado en buena medida los objetivos de varios corpus del italiano. Así, por ejemplo, el corpus elaborado para el proyecto *Lessico di frequenza dell'italiano parlato* (De Mauro, Mancini, Vedovelli & Voghera, 1993) incluía entre sus objetivos la observación del mayor o menor grado de alejamiento de las características puramente locales del habla, que, en general, incluían rasgos dialectales y regionales.

A propósito de las situaciones lingüísticas descritas, podría decirse que se observa una tendencia a la aproximación de los usos considerados como estándares y los usos marcados regional o localmente; o, al menos, un progresivo difuminado de sus fronteras. Si estas situaciones se caracterizaran como diglósicas, podría decirse que los límites entre las variedades altas y bajas tienden a hacerse más porosos, permitiendo la aparición de elementos estándares en las hablas dialectales, pero también elementos dialectales en la modalidad estándar. Esta tendencia podría adscribirse a la corriente de 'realismo lingüístico' que, alineada con una ideología de la autenticidad (Woolard, 1998), promueve la inclusión de elementos vernáculos populares en todo tipo de manifestaciones lingüísticas, habladas y escritas. Siendo así, los corpus lingüísticos han

de reflejar la realidad dialectal que los usos lingüísticos muestran en todo tipo de registros.

Desde una perspectiva más teórica, las dificultades para el establecimiento de límites o fronteras entre variedades serían trasunto de las existentes para delimitar lenguas entre sí y estarían en el foco de los planteamientos que reciben la etiqueta general de *translanguaging*. El *translanguaging* o translingüismo supone un proceso mediante el cual los hablantes utilizan sus lenguas y variedades como un sistema de comunicación integrado. Esto lleva a que, en el uso de la lengua, estén implicados tanto la mera producción lingüística, como la comunicación efectiva, las funciones del lenguaje y los procesos de pensamiento. El translingüismo propone la integración no intencional de múltiples sistemas lingüísticos, convirtiéndose en una alternativa interpretativa al concepto de bilingüismo (García & Wei, 2014; Otheguy, García & Reid, 2015). El prefijo 'trans' enfatiza que se trata de prácticas fluidas, que trascienden los sistemas y estructuras de lenguaje socialmente construido, para involucrar diversos sistemas generadores de significados. Una vez más, esta concepción teórica viene a justificar la inevitabilidad de la dimensión dialectal de los corpus lingüísticos, relegando a un plano secundario la adjudicación de etiquetas geolingüísticas para los elementos que en ellos puedan aparecer.

3. Corpus y áreas lingüísticas

Efectivamente, el establecimiento de límites entre variedades lingüísticas supone una dificultad que se traslada a la selección de las variedades que han de quedar representadas en un corpus. Las interpretaciones que se hagan sobre las relaciones entre variedades o sobre la relación de una lengua con otras pueden llevar a decisiones dispares, como la exclusión o la inclusión en los corpus de muestras bilingües de uso, de mezclas de código o de variedades criollas. Por otro lado, como ya dejamos de manifiesto anteriormente (Moreno Fernández, 2021), la decisión sobre qué áreas de una lengua han de quedar representadas en un corpus y con qué proporciones es un asunto polémico, sin trazas de ser resuelto desde la lingüística de corpus, y así se ha comprobado a propósito de la zonificación del español (Alba, 1992; Moreno Fernández, 1993; Ueda 1995; Quesada Pacheco, 2014).

Desde una perspectiva geográfica, la zonificación del dominio de una lengua no puede resolverse desde una sola posición teórica. En el caso de la lengua española, por ejemplo, han sido numerosas las propuestas de división dialectal manejadas durante el último medio siglo: desde las que plantean una ficticia distinción entre español de España y español de América, hasta las que proponen decenas de zonas según las denominaciones dadas a determinados objetos, acciones o conceptos. Esta divergencia a la hora de entender la división dialectal de una lengua tiene su reflejo en los distintos modos en que tal división se ha manejado para la arquitectura de los corpus. Veamos

algunos ejemplos a propósito de la lengua española, si bien podrían encontrarse equivalentes a propósito de otras realidades lingüísticas.

Cuando De Kock (1990) decide construir una gramática cuya enseñanza se organice a partir de las conclusiones obtenidas de una lingüística de corpus, considera ineludible, aunque no lo califique así, reunir materiales procedentes de autores de distinta procedencia geográfica. Sin embargo, en su selección de textos, ya comentada, se observan dos hechos con claridad: en primer lugar, la preeminencia concedida a los autores procedentes de España (13 en el primer corpus; 12 en el segundo) frente a una minoría de autores americanos; en segundo lugar, la aparente falta de criterio para elegir autores de un país americano y no de otro, excepto el de evitar la repetición. Podría decirse, pues, que en este caso existe una conciencia tanto de la diversidad dialectal del español, como de su zonificación, solo que esa conciencia no se convierte en un criterio para la representación de las distintas áreas del español. Transcurridos 30 años desde su propuesta de lingüística de corpus y dada la relevancia que han ido adquiriendo las cuestiones de identidad y diversidad, hoy De Kock probablemente habría adoptado otro criterio sin traicionar a sus principios metodológicos.

Un segundo ejemplo de corpus general en el que, de un modo u otro, se atiende a la variación geolingüística es el corpus CUMBRE (Sánchez, 1995), al que ya le prestamos atención en su momento (Moreno Fernández, 2021). Basta ahora tener en cuenta que este corpus distinguió la variedad lingüística de España (65% del total de textos) de la variedad lingüística de los países de habla hispana del continente americano (35% del total de textos). Curiosamente, el corpus daba un 10% más de peso a los textos orales sobre los escritos en América, esgrimiendo como argumento el deseo de 'equilibrar la balanza', aunque esa proporción no compensara ni de cerca el desequilibrio representativo. El corpus CUMBRE identificaba cinco grandes zonas americanas: América Central, México, Venezuela-Colombia-Ecuador, América Andina (Perú-Bolivia-Chile), Argentina-Sur (Argentina-Paraguay-Uruguay). Sin embargo, no ofrecía razones lingüísticas que justificaran tal división dialectal, aunque puedan intuirse motivos metodológicos y prácticos.

Más reciente en el tiempo es la propuesta de la Real Academia Española para la elaboración de sus corpus lingüísticos: concretamente el CREA (*Corpus de Referencia del Español Actual*) (1997) y el CORPES XXI (*Corpus del Español del siglo XXI*) (versión 0.93. 2021). Desde prácticamente el inicio de su política panhispánica, la Asociación de Academias de la Lengua Española ha considerado la existencia de ocho áreas lingüísticas, consideradas teóricamente en pie de igualdad: Chile; Río de la Plata; área andina; Caribe Continental; México y Centroamérica (en un principio separadas y después unidas); Antillas; Estados Unidos y Filipinas; y España (RAE y ASALE, 2010). En esta división se combinan razones geolingüísticas con criterios operativos y organizativos propios de la Asociación. Ahora bien, en la construcción de los corpus académicos, si bien se atiende a las áreas lingüísticas mencionadas, a las que se atribuye

el calificativo de 'tradicionales', el peso de cada una de ellas no es equivalente. El CREA partía de una arquitectura dialectal en que los materiales procedentes de España suponían el 50 % y el resto correspondía a todos los países americanos hispanohablantes, excluidos los Estados Unidos. El peso relativo de cada uno de estos países no se ha especificado. El CORPES XXI, por su parte, corrige las proporciones, dejando en un 30 % los materiales procedentes de España y un 70 % los materiales americanos, esta vez con los Estados Unidos incluidos, así como testimonios de Guinea Ecuatorial y Filipinas, si bien Davies (2016) pone en duda la validez de los materiales ecuatoguineanos (640.000 palabras) y filipinos (100.000 palabras) incluidos en el CORPES XXI, por considerarlos muy escasos para un análisis significativo. Este nuevo corpus académico sí explica y especifica la proporción interna de los materiales americanos, aduciendo un cruce de criterios diferentes: la población, el volumen de publicaciones, el número de ediciones digitales de periódicos y revistas. La relación se cierra con un 'etc.' que se corresponde con el desconocimiento del algoritmo utilizado para obtener las proporciones.

Tabla 1. Distribución general de materiales del CORPES XXI por áreas lingüísticas.

ZONA	%
Andina	9
Antillas (Caribeña)	12
Caribe continental	9
Chilena	4
Estados Unidos	4
México y Centroamérica	19
Río de la Plata	13
TOTAL América	70
España	30
TOTAL	100

Aún cabe una cuarta muestra de la diversidad de criterios manejados para la construcción de corpus lingüísticos en materia de zonificación dialectal. Se trata del *Corpus del Español: Web / Dialects*, elaborado por Mark Davies desde la Universidad Brigham Young e integrado por 2.000 millones de palabras tomadas de 2 millones de páginas *web* (Davies, 2016). En lo que se refiere a las zonas geográficas propiamente dichas, Davies (2016) no entra en consideraciones lingüísticas, puesto que se limita a clasificar los materiales por su país de procedencia: 21 países en total, desde los Estados Unidos a Argentina. En cuanto a la proporción de los materiales americanos y españoles, *El Corpus del Español* incluye un 78% de americanos y un 22% de españoles y lo justifica explicando que esta proporción representa mejor el equilibrio poblacional real de estas dos zonas, dado que el 90% del mundo hispanohablante es latinoamericano, a la vez que achaca al origen español de la RAE la proporción dada en su corpus a los datos de España. Sin embargo, el resultado de esa búsqueda de equilibrio en el corpus de Davies no es otro que España es el país más representado, en términos

absolutos (440 millones de palabras) y relativos (22%), lejos del 10% que exigiría el anhelado equilibrio poblacional. En cualquier caso, los criterios manejados poco tienen que ver con la estricta realidad dialectal, dado que, junto a los porcentajes poblacionales, se maneja el criterio de los países. La razón de esto último es muy sencilla: la búsqueda de los textos se ha hecho sobre la base de las extensiones de los diferentes países en internet: .es; .mx; .cl...

Hasta aquí hemos podido observar, por tanto, cómo los corpus generales o de referencia, en el caso del español, si bien prestan atención a la variación dialectal de las muestras, a menudo se construyen sobre argumentos o criterios de naturaleza operativa o técnica, lejanos a la dialectología propiamente dicha. De los tres corpus comentados hasta aquí, el corpus CUMBRE y el de la RAE-ASALE son los que más atención prestan a la diversidad dialectal desde una perspectiva lingüística, aunque tampoco sean ajenos a condicionamientos prácticos o metodológicos. La alternativa natural a las limitaciones que la elaboración de un corpus impone, en lo que a la variación dialectal se refiere, es la de trabajar sobre materiales predeterminados o seleccionados con criterios geolingüísticos. Este es el caso, por ejemplo, del *Corpus del Español Mexicano Contemporáneo* (CEMC) (Lara, 1987, 2019).

El CEMC es un corpus construido con fines lexicográficos, como también fue el caso de CUMBRE, desde El Colegio de México. EL CEMC incluye alrededor de 1.000 muestras de unas 2.000 palabras gráficas cada una, procedentes de textos escritos y orales producidos en México entre 1921 y 1974. La geografía, en este caso, está bien delimitada, lo que permitió atender más minuciosamente a otras variables, como las sociales o las etnográficas. No obstante, la consideración de estos factores no se hizo desde un diseño estructural *ex novo*, sino incorporando muestras del *Atlas Lingüístico de México* (Lope Blanch, 1970, 1990) y de otros recursos lingüístico-etnográficos de El Colegio de México. Del millar de textos o documentos incluidos, 211 se clasifican como textos de 'lengua no estándar', entre los que se distinguen 130 textos dialectales. Más allá de la etiqueta no hubo un criterio que distinguiera áreas o variantes dentro del mosaico dialectal mexicano.

Junto al CEMC, otros corpus se han creado con la intención de reunir muestras del español de una región, área o país determinado. Aquí podrían incluirse los corpus dirigidos por Francisco Marcos Marín en los años noventa y que reúnen alrededor de dos millones de formas cada uno: el *Corpus lingüístico de referencia de la lengua española en Argentina* (Marcos Marín, 1992a) y el *Corpus lingüístico de referencia de la lengua española en Chile* (Marcos Marín, 1994). Más reciente es el *Corpus dinámico del castellano de Chile* (CODICACH), desarrollado por Sadowsky (2006) desde la Pontificia Universidad Católica de Chile y que consta de 800 millones de palabras procedentes de lengua escrita, principalmente en prensa, aunque solo ofrece acceso abierto a las listas de frecuencias léxicas derivadas. Todos estos corpus, junto a otros, se interesan por la variación dialectal en la medida que permiten fijar los límites externos de su alcance. Cuestión

distinta es la naturaleza variable de las realidades lingüísticas que quedan incluidas dentro de estas obras nacionales.

En lo que se refiere al análisis de los materiales reunidos en los corpus, es evidente el interés de adoptar una perspectiva comparada para entender y valorar los datos procedentes de unas regiones y de otras. En buena medida, la sociolingüística comparada (Tagliamonte, 2013) maneja datos de ese tipo para proceder a sus investigaciones: el contraste de información sociolingüística procedente de comunidades o áreas geográficas diferentes. Pero las posibilidades del análisis también alcanzan a otros medios y métodos de la dialectología, de aplicación directa sobre los materiales de los corpus. Uno de estos medios es el cartografiado de datos recuperados a partir de corpus lingüísticos. Veamos brevemente algunos ejemplos.

La confección del *Atlas Lingüístico y Etnográfico de Andalucía* (ALEA) (Alvar, Llorente & Salvador, 1961-1973) incorporó la grabación de conversaciones, relatos y otros tipos de textos orales que conformaron un corpus de la lengua hablada en Andalucía (Alvar, Llorente & Salvador, 1995), antes de que existiera una lingüística de corpus como tal y, en consecuencia, sin las posibilidades técnicas de almacenamiento y recuperación de datos que ofrecen los corpus actuales. Aun así, los textos reunidos en ese corpus pionero permitieron elaborar mapas como el que representa los rangos de probabilidad de realización [ø] del fonema /s/ en posición final de sílaba (Figura 1). El total de casos de -/s/ analizados a partir de ese corpus oral en transcripción fonética fue de 2954 (Moreno Fernández 1996-1997).

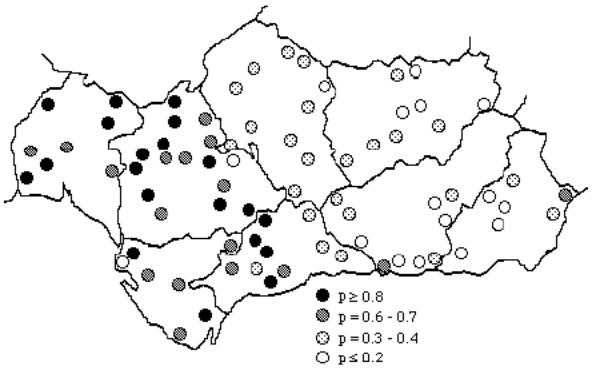

Figura 1. Mapa de la variante [ø] de /s/ implosiva en Andalucía, con indicación de rangos de probabilidad, a partir del corpus oral del ALEA. Fuente: Moreno Fernández (1993).

Otro ejemplo de cartografiado elaborado a partir de materiales de corpus nos lo proporciona el proyecto *Varianten Grammatik* desarrollado desde las universidades de Zúrich y Salzburgo, junto al *Leibniz-Institut für Deutsche Sprache*. La base de este corpus está formada por artículos de 68 periódicos en línea de todas las áreas de lengua alemana. El hecho de contar con materiales digitales bien almacenados permite el cartografiado

automático de mapas como el de la alternancia léxica entre *Bussgeld* y *Geldbusse* 'multa' (Figura 2).

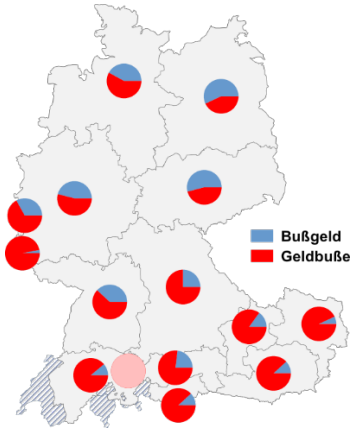

Figura 2. Frecuencia relativa de *Bussgeld* y *Geldbusse* 'multa' en las áreas germanohablantes, a partir del corpus *Varianten Grammatik* Fuente: Adarve (2020).

Un último ejemplo de cartografiado dialectal, aunque el proyecto original tenga una raíz sociolingüística, como podrá verse seguidamente, es el mapa que refleja la probabilidad de omisión del pronombre objeto directo en la legua hablada con verbos de comunicación, calculado a partir del corpus de lengua hablada PRESEEA (Moreno Fernández, 2005) (Figura 3).

Figura 3. Probabilidad de omisión del pronombre objeto directo en la lengua hablada con verbos de comunicación (datos de corpus PRESEEA).
Leyenda: Círculo doble: .5. Círculo dividido: .4. Círculo blanco: .3.
Fuente: Moreno Fernández (2019).

Vemos, pues, que la calidad y la cantidad de los materiales que aportan los corpus lingüísticos ofrecen unas magníficas posibilidades para el análisis y la representación de sus resultados, especialmente si los corpus son elaborados con fines dialectológicos, aunque no necesariamente.

4. La dimensión social en los corpus lingüísticos

En relación con la dimensión social de los usos lingüísticos, los corpus son también un reflejo de ella, en cualquiera de sus expresiones, pero las maneras en que ello se produce son muy variadas. Una de ellas es la de atender a criterios sociolingüísticos para configurar la arquitectura del corpus. Esto fue lo que se hizo en el proyecto para el *Estudio coordinado de la norma lingüística culta de las principales ciudades de Iberoamérica y de la Península Ibérica*, conocido como PILEI (Lope Blanch, 1967, 1986; Rabanales, 1992), que reunió materiales orales del español para abordar el estudio de las variedades de una docena de ciudades hispanohablantes a partir del discurso oral de hablantes cultos, con formación universitaria, teniendo en cuenta además la edad y el sexo de los hablantes. El conjunto de los materiales del proyecto fue publicado como *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo* hispánico (Samper, Hernández & Troya, 1998).

Con una intención dialectal semejante, aunque limitada a la península ibérica, se ha elaborado también el *Corpus Oral y Sonoro del Español Rura*l (COSER), dirigido por Inés Fernández Ordóñez (2005) desde la Universidad Autónoma de Madrid. Se trata de un corpus que reúne entrevistas realizadas a informantes procedentes de zonas rurales en más de 600 localidades. En este caso son los usos populares los que reciben atención prioritaria, restringiendo, pues, su alcance sociolingüístico. Asimismo, el corpus reunido dentro del proyecto *Difusión internacional del español* (DIES) (Ávila, 2004), si bien reúne muestras de diferentes áreas hispanohablantes, ciñe su interés a las manifestaciones lingüística de la radio, la prensa y la televisión.

Más amplitud de miras sociolingüísticas presenta el macrocorpus creado dentro del *Proyecto para el estudio sociolingüístico del español de España y de América* (PRESEEA), integrado por entrevistas semidirigidas realizadas en más de 40 comunidades hispanohablantes de Europa y América. Los materiales disponibles son aproximadamente un tercio del total de materiales con los que se está trabajando, lo que supone, en 2019, contar con 1.200 horas de grabación, en las que aparecen 11,5 millones de formas procedentes de 1.300 informantes de 17 ciudades. Estos datos nos permiten afirmar que, por el número de informantes, se trata del mayor corpus reunido hasta el momento de acuerdo con criterios sociolingüísticos y, a la vez, dialectales. Asimismo, por número de palabras, el tamaño del corpus PRESEEA supera ampliamente al componente oral de cualquier corpus de referencia (Moreno Fernández, 2005, 2021).

Los grandes corpus de lengua hablada que acaban de mencionarse muestran con claridad la existencia de vínculos sustanciales entre la variación dialectal y la social, de modo que la relevancia concedida a la dimensión sociolingüística se complementa con la otorgada a la dimensión dialectal, de igual forma que esta se trata correlacionada directamente con la primera. De hecho, existen otros corpus compatibles con estudios de esta naturaleza, sin ser propiamente dialectales o sociolingüísticos, como el *Corpus oral de referencia de la lengua española contemporánea* (CORLEC) (Marcos Marín, 1992b) o el corpus oral C-ORAL (Moreno-Sandoval, de la Madrid, Alcántara, González, Guirao & De la Torre, 2005).

Algo semejante ocurre en otros corpus elaborados con una guía sociolingüística, pero centrados en la lengua hablada en comunidades concretas. Estos proyectos se han venido completando prácticamente desde los años setenta, con los estudios pioneros sobre Las Palmas de Gran Canaria (Alvar, 1972) y San Juan de. Puerto Rico (López Morales, 1983), hasta la actualidad. En esta larga serie se inscriben el corpus para el estudio de las lenguas en contacto en Los Ángeles (Silva-Corvalán, 1994), el corpus *Alicante Corpus del Español* (ALCORE), dirigido por Dolores Azorín (2005), el corpus *Vernáculo Urbano Malagueño*, dirigido por Juan Villena Ponsoda (Lasarte, Sánchez, Ávila & Villena, 2008), el corpus sociolingüístico de Castellón de la Plana y su área metropolitana (Blas Arroyo, 2009) o el corpus Otheguy-Zentella del español de Nueva York (Otheguy & Zentella, 2012), entre otros muchos.

Otra colección de corpus con valor sociolingüístico es la formada por muestras procedentes de grupos sociales específicos, entre los que sobresale el de los jóvenes. En este apartado pueden inscribirse el *Corpus oral para el estudio del lenguaje juvenil en Alicante* (COVJA), dirigido por Dolores Azorín (2005), o el *Corpus oral de lenguaje adolescente* (COLA), dirigido por Annette Myre Jörgensen (2009) desde la Universidad de Bergen. Este último añade a su valor sociolingüístico una dimensión geográfica, dado que incluye muestras reunidas en Madrid, Santiago de Chile, Buenos Aires y Ciudad de Guatemala.

Pero si el vínculo entre lo social y lo geográfico queda patente en los corpus hasta aquí mencionados, no menos evidente resulta la esencial relación entre lo sociolingüístico y lo estilístico o lo pragmático. De hecho, en los corpus citados se observa este vínculo con claridad cuando se explica que los materiales de lengua hablada fueron reunidos mediante entrevistas semidirigidas o desde un medio de comunicación, por ejemplo. El propio corpus mexicano CEMC presta atención a factores socioestilísticos cuando incluye entre sus textos integrantes muestras de habla culta y popular (discursos, conversaciones), así como jergas y etnotextos; esto es, materiales marcados sociolingüísticamente.

En realidad, la dimensión social y estilística puede escudriñarse incluso en los corpus considerados de 'referencia', en los que las reflexiones sobre lengua estándar y lengua

vernácula o entre lengua escrita y lengua hablada resultan obligadas. Estas reflexiones se han producido en relación con los corpus prácticamente de todas las lenguas. En el caso de las lenguas alemana e italiana, existen corpus de lengua cotidiana diferenciados de los corpus de lengua estándar o lengua escrita, con todos los matices técnicos y teóricos que ello supone. A propósito del español, el estilo fue un factor fundamental en la selección de los textos integrantes del corpus de De Kock, quien rechazó como base del análisis y la enseñanza de la gramática tanto los textos técnicos (por estereotipados y empobrecidos), como los textos académicos (por forzados) y los literarios (por personales y marginales) (De Kock, 1990: 35), y apostó por textos "no literarios o vulgarizantes" de escritores modernos: ensayos, artículos de revistas o periódicos, conferencias, discursos… Esta decisión se tomó por "razones didácticas" (De Kock, 1990: 130).

Con todo, existen corpus elaborados con énfasis específico en el ámbito estilístico y pragmático, y que ofrecen interesantes posibilidades para el conocimiento del español en su dimensión sociopragmática y estilística. El *Corpus conversacional de Alcalá de Henares* (ACUAH), construido por Ana Cestero (1994), y el *Corpus conversacional de Barcelona y su área metropolitana*, proyecto dirigido por Rosa Vila (2001), son dos de ellos. Sin embargo, por su dimensión y repercusión merece una especial mención el corpus ValEsCo (*Valencia Español Coloquial*), coordinado por Antonio Briz (Briz & Albelda, 2009) y, como secuela ampliada de este, el corpus AMERESCO (*América Español Coloquial*) (Albelda & Estellés, 2016), aún en construcción, pero que cubrirá buena parte de las necesidades de estudio de la lengua española coloquial.

5. Consideraciones metodológicas

En el ya citado capítulo sobre la variación en los corpus (Moreno Fernández, 2021), hubo ocasión de reflexionar sobre algunas cuestiones metodológicas fundamentales, tanto en lo que se refiere a las prácticas actuales, como en lo que atañe a las posibilidades futuras. Sin embargo, la relación de cuestiones relevantes en este campo no puede quedar agotada en unas pocas líneas. Por ese motivo, se presentan a continuación algunas consideraciones más, relativas principalmente a las técnicas de confección de los corpus.

En primer lugar, las cuestiones teóricas con consecuencias inmediatas sobre el proceso de confección de los corpus no se agotan con las comentadas más arriba. Ni mucho menos. Existe, por ejemplo, un hecho de singular relevancia por encima del cual suele pasarse sin mayores reparos. Se trata de la variabilidad o heterogeneidad interna de las áreas dialectales identificadas como relevantes para los corpus a efectos metodológicos. Así, por ejemplo, el corpus CUMBRE maneja como categorías dialectales los conceptos de español de España y de español de Hispanoamérica. Sin embargo, es evidente lo inadecuado de trabajar sobre unas inexistentes 'variedad de

España' y 'variedad americana', ya que se trata de realidades que en sí mismas son muy complejas geolingüísticamente. Como es comprensible, no estamos ante un problema de desconocimiento de la realidad, sino ante una decisión metodológica. El problema se deriva de manejar un número de variables que han de quedar debidamente representadas, todas ellas. Cuantas más variables se manejan, el número de palabras necesario para su proporcionada representación crece exponencialmente, sin que se tenga la seguridad de poder reunir un mismo número de palabras para cada una de las categorías o variables establecidas. Se trata, pues, de una cuestión que afecta de un modo importante a la construcción de cualquier corpus. Asimismo, una vez superado el problema técnico de la limitación del tamaño total, por mucho que el CORPES XXI decida incorporar muestras representativas del español de Filipinas o de Guinea Ecuatorial probablemente se esté topando con la dificultad de encontrar fuentes con un volumen semejante al de otras áreas de uso del español.

Pero el asunto de la diversidad –dialectal, sociolingüística, estilística– en el interior de los corpus no termina aquí. Y es que podemos tener perfectamente localizado un español de Andalucía, del Río de la Plata o de Chile, pero ello no supone la homogeneidad de sus manifestaciones, sea el área que sea, con su correspondiente efecto sobre las muestras o textos que se eligen finalmente para la confección del corpus. Pensemos en el 'español de los Estados Unidos', cuya primera polémica se suscita por el uso mismo de la preposición 'de' o de la preposición 'en' para su simple denominación. Efectivamente, el español estadounidense no es una realidad homogénea, como no lo es la de ningún otro espacio lingüístico, con el agravante de que en los Estados Unidos el uso del español, junto al inglés, está lleno de contactos e hibridaciones, en todos los niveles de uso y en todas las manifestaciones lingüísticas: desde los textos en español de la Casa Blanca, hasta los carteles o avisos más cotidianos. Es la propuesta metodológica de cada corpus la que debe dar respuesta a las dudas que estas cuestiones suscitan.

Esta cuestión está claramente relacionada con la representatividad, y así se deduce de lo explicado. Como es sabido, la representatividad es un concepto de base cualitativa, que ofrece una proyección cuantitativa desde la que se mide la significación estadística, relacionada con el volumen de los materiales del corpus, en su conjunto y en sus partes integrantes. La cuestión es que el tamaño del corpus ha de tener unos límites, puesto que la producción lingüística es inabarcable en su totalidad. Lógicamente, si el tamaño del corpus era cuestión decisiva en los años noventa y caballo de batalla para la validez del corpus, en la actualidad ocupa un lugar secundario en las preocupaciones metodológicas, puesto que ya es posible reunir con cierta facilidad miles de millones de formas. Sigue siendo relevante el número de palabras cuando se trata de corpus pequeños, dado que la representatividad puede resultar más dudosa y las pruebas estadísticas analíticas pueden encontrar más dificultades. Cuando el número de palabras es muy grande, sin embargo, el problema de la representatividad puede hallarse a la hora

de identificar submuestras de valor dialectal o sociolingüístico. Pensemos, por ejemplo, en la práctica imposibilidad de identificar dentro de un megacorpus elaborado con textos descargados de internet a partir de la extensión de internet '.es' cuáles de ellos reflejan una variedad andaluza o canaria; incluso, cuáles una variedad americana, dado que muchos autores hispanohablantes de América publican textos en dominios '.es'. Ni que decir tiene que el problema se vuelve irresoluble cuando se integran textos procedentes de dominios con la extensión '.com' o '.net'.

En cuanto al tamaño de las submuestras que conforman un corpus, cuando estas se han predeterminado e identificado, la lingüística de corpus siempre ha demandado equilibrio y proporcionalidad, para hacer posible la comparabilidad. Desde este punto de vista, un corpus que aspire a representar la diversidad dialectal de un territorio debería incluir muestras equilibradas y proporcionales de cada una de las variedades que en él se reúnen, y lo mismo podría decirse cuando se trata de comparar variedades sociolectales. Siendo esto aconsejable, cuando no imprescindible, la estadística pone a disposición de los analistas recursos que compensan las desigualdades entre muestras o submuestras y que consisten en utilizar unos valores o frecuencias normalizados: frecuencia relativa, frecuencia normalizada por millón de palabras, índice normalizado de dispersión (Rojo, 2010; Molina Salinas & Sierra Martínez, 2015). De esta forma no es imprescindible que todos los componentes de un corpus tengan el mismo tamaño, dado que la comparabilidad vendría garantizada por otros procedimientos cuantitativos.

Finalmente, la fuente de los materiales integrados en los corpus es uno de los aspectos cruciales en todo el proceso metodológico. Recientemente, como ya se ha mencionado, la *web* o internet se ha convertido en un medio, cuando no en una fuente directa, para la recolección de palabras destinadas a los corpus. Un ejemplo de ello es el *Wikicorpus*, que para el inglés, por ejemplo, recoge 600 millones de palabras procedentes de artículos de la *Wikipedia*; para el español, se reúnen 120 millones de palabras; para el catalán, 50 millones. Además, se ofrecen en una versión etiquetada gramaticalmente y en otra sin etiquetar (Reese, Boleda, Cuadros, Padró & Rigau, 2010). El uso de la *web* con fines lingüísticos obliga a plantearse reservas y precauciones, comenzando por su conceptualización. Estas reservas alcanzan su grado máximo si pensamos en fines dialectales y sociolingüísticos. Sinclair (1996) afirma que la *web* no constituye un corpus, sino un simple cúmulo de archivos.

En el terreno de la representatividad, la *web* muestra como una de sus debilidades la visibilidad parcial de sus fondos. De acuerdo con Rojo (2014), la red muestra solamente un tercio de lo que contiene y los buscadores solo encuentran una parte de ese tercio, por no mencionar los sesgos que se producen en sus búsquedas. Este problema parece quedar desdibujado cuando se afirma, como hace Lew (2009), que la *World Wide Web* contiene unos cinco billones (5.000.000.000.000.000) de palabras, lo que la convierte en una colección unas 50.000 veces más grande que el *British National Corpus* (BNC).

Parecería que, con el tiempo, la *web* podría registrar casi la totalidad de los usos escritos y alcanzar el ideal de la cartografía: confeccionar un mapa del tamaño del territorio, como en la historia de Lewis Carrol (*Sylvia y Bruno*, 1893) y la pequeña narración de Jorge Luis Borges (*Del rigor en la Ciencia*, 1946). El tamaño puede solucionar muchos problemas, incluso de representatividad, pero no los soluciona todos; y mucho menos el adecuado tratamiento de la variación dialectal y sociolingüística. Además, los resultados de las búsquedas a menudo llegan sesgados porque la frecuencia de las palabras pertenecientes a algunos campos o contenidos aparece sobredimensionada, lo que limita mucho los análisis lingüísticos, incluidos los dialectales y sociolingüísticos.

Otro plano en que dialectología y sociolingüística podrían encontrarse con la lingüística de corpus es el de las redes sociales. Los materiales de lengua en ellas vertidos son un claro objeto de rastreo, almacenamiento y recuperación, como ya se ha practicado en algunas primeras investigaciones, como la que ha escudriñado anglicismos en un corpus de 15 millones de palabras en español utilizadas en 85.000 mensajes de *Twitter* enviados desde los Estados Unidos en 2016 (Moreno Fernández & Moreno Sandoval, 2018). Sin embargo, los problemas técnicos derivados del uso lingüístico en este tipo de soportes son muy numerosos, puesto que este refleja un alto grado de incuria y espontaneidad por parte de los hablantes. Como apuntamos en 2021, tal vez el *machine learning* pueda solventar en el futuro cuestiones que hoy parecen irresolubles.

CONCLUSIÓN

La variación dialectal y sociolingüística, profundamente vinculadas también a la histórica y la estilística, es una característica esencial de las lenguas que afecta a todos sus niveles y que, por lo tanto, resulta ineludible en los materiales que conforman los corpus lingüísticos. Siendo así, la lingüística de corpus debe dar respuesta a cuestiones como la representatividad dialectal y social de las muestras seleccionadas, tanto si se trata de corpus especializados en la variación, como si se trata de corpus generales o de referencia. Estas cuestiones plantean dificultades epistemológicas que afectan al estudio de las lenguas en general, pero que hallan en la lingüística de corpus derivaciones específicas.

Estas páginas han presentado una serie de reflexiones sobre aspectos teóricos y metodológicos que afectan a la elaboración y manejo de los corpus lingüísticos. Se ha prestado atención a los corpus construidos expresamente con la intención de abordar estudios desde la geolingüística y desde la sociolingüística, pero también se ha evidenciado la realidad de que todo corpus lingüístico incluye, de un modo u otro, una dimensión de variabilidad. Los especialistas en lingüística de corpus han de ser conscientes de la trascendencia de sus decisiones metodológicas porque pueden estar incluyendo sesgos que afecten al análisis e interpretación de la variación en los materiales reunidos.

REFERENCIAS BIBLIOGRÁFICAS

Adarve, G. (Coord.) (2020). *Variantengrammatik des Standarddeutschen*. Universität Zurich – Universität Salzburg [en línea]. Disponible en: http://mediawiki.ids-mannheim.de/VarGra/index.php/Start

Alba, O. (1992). Zonificación dialectal del español de América. En C. Hernández (Ed.), *Historia y presente del español de América* (pp. 63-84). Valladolid: Junta de Castilla y León.

Albelda, M. & Estellés, M. (Coords.) (2016). *Corpus Ameresco*, València: Universitat de València.

Alvar, M. (1972). *Niveles socioculturales en el habla de Las Palmas de Gran Canaria*. Las Palmas: Cabildo Insular.

Alvar, M., Llorente, A. & Salvador, G. (1961-1973). *Atlas Lingüístico y etnográfico de Andalucía* (6 vols). Madrid: CSIC.

Alvar, M., Llorente, A. & Salvador, G. (1995). *Textos andaluces en transcripción fonética*. Madrid: Gredos.

Ammon, U. (2003). On the social forces that determine what is standard in a language and on conditions of successful implementation. *Sociolinguistica,* 17, 1-10.

Anderwald, L. (2009). Corpus linguistics and dialectology. En A. Lüdeling & M. Kytö (Eds.), *Handbook of Linguistics and Communication Science* (pp. 1126-1140). Berlín – Nueva York: de Gruyter.

Auer, P. (2011). Dialect vs standard: A typology of scenarios in Europe. En B. Kortmann & J. van der Auwera (Eds.), *The Languages and Linguistics of Europe. A Comprehensive Guide* (pp. 485-500). Berlín: de Gruyter.

Ávila, R. (2004). *Difusión internacional del español por radio, televisión y prensa: Unidad y diversidad de la lengua (DIES-RTP)*. México: El Colegio de México [en línea]. Disponible en: https://raulavila.colmex.mx/index_archivos/page0003.html

Azorín, D. (2005). Corpus oral para el estudio del lenguaje juvenil y del español hablado en Alicante: El corpus ALCORE y COVJA. *Oralia: Análisis del discurso oral,* 8, 265-288.

Berruto, G. (2018). The languages and dialects of Italy. En W. Ayres-Bennett & J. Carruthers (Eds.), *Manual of Romance Sociolinguistics* (pp. 494-525). Berlín: de Gruyter. DOI:org/10.1515/9783110365955-001

Blas Arroyo, J. L. (Coord.) (2009). *Corpus sociolingüístico de Castellón de la Plana y su área metropolitana*. Castellón: Universitat Jaume I.

Briz, A. & Albelda, M. (2009). Estado actual de los corpus de lengua española hablada y escrita: I+D. *El español en el mundo. Anuario del Instituto Cervantes 2009*. Madrid: Instituto Cervantes [en línea]. Disponible en: https://cvc.cervantes.es/lengua/anuario/anuario_09/briz_albeida/p01.htm

Cestero, A. M. (1994). *Alternancia de turnos de palabra en lengua española*. Alcalá de Henares: Universidad de Alcalá.

Davies, M. (2016). *El Corpus del Español*. Provo: Brigham Young Univesity.

De Kock, J. (1990). *Gramática española. Enseñanza e investigación. I. Gramática. Apuntes metodológicos*. Salamanca: Universidad de Salamanca.

De Kock, J. (1992). *Gramática española. Enseñanza e investigación. III. Textos. 2. 20 textos*. Salamanca: Universidad de Salamanca.

De Mauro, T., Mancini, F., Vedovelli, M. & Voghera, M. (1993). *Lessico di frequenza dell'italiano parlato*. Milano: EtasLibri.

Fernández Ordóñez, I. (Dir.) (2005). *Corpus Oral y Sonoro del Español Rural* (COSER).

García, O. & Wei, L. (2014). *Translanguaging: Language, bilingualism and education*. Nueva York: Palgrave Macmillan.

Gómez Font, A. (2012). *Español neutro o internacional*. Madrid: Fundación del Español.

Guevara, A. (2013). *El español neutro (Realización hablada)*. Buenos Aires: Iberoamericana Comunicación.

Halliday, M.A.K. (1979). *El lenguaje como semiótica social*. México: FCE.

Hansen, B. (2018). *Corpus linguistics and sociolinguistics: A study of varuation and change in the modal systems of World Englishes*. Leiden: Brill.

Lameli, A. (2010). Deutsch in Deutschland. Standard, dialektale und regionale Variation. En H. J. Krumm, Ch. Fandrych, B. Hufeisen & C. Riemer (Eds.), *Handbuch Deutsch als Fremd. Und Zweitspreche* (pp. 385-398). Berlín: de Gruyter.

Lara, L.F. (1987). Características del corpus del español mexicano contemporáneo. En H. López Morales & M. Vaquero (Eds.), *Actas del I Congreso Internacional sobre el Español de América, San Juan, Puerto Rico, del 4 al 9 de octubre de 1982* (pp. 579-586). Madrid: Academia Puertorriqueña de La Lengua Española-La Muralla.

Lara, L. F. (2019). *Diccionario del Español de México. Corpus del Español Mexicano Contemporáneo* (CEMC) [en línea]. Disponible en: <http://www.corpus.unam.mx/cemc>, software AMATE ver. 1.0.

Lasarte M. C., Sánchez J. M, Ávila, A. & Villena, J.A. (Eds.) (2008). *El español hablado en Málaga II. Corpus oral para su estudio sociolingüístico. Nivel de estudios alto*. Málaga: Sarriá.

Lew, R. (2009). The Web as corpus versus traditional corpora. *Computer Science* [en línea]. Disponible en: https://pdfs.semanticscholar.org/146e/710d351628cc1c3f18e771afaa05de70c6b4.pdf?_ga=2.136160049.1484498069.1575741642-2129020191.1572797599

Lope Blanch, J. M. (1967). Proyecto de estudio del habla culta de las principales ciudades de Hispanoamérica. En *El simposio de Bloomington. Agosto de 1964. Actas, informes y comunicaciones* (pp. 255-264). Bogotá: Instituto Caro y Cuervo.

Lope Blanch, J. M. (1970). *Cuestionario para la delimitación de las zonas dialectales de México*. México: El Colegio de México.

Lope Blanch, J. M. (1986). *El estudio del español hablado culto. Historia de un proyecto*. México: UNAM.

Lope Blanch, J. M. (Dir.) (1990). *Atlas lingüístico de México. Tomo I. Fonética. Volumen I*. Estudios de dialectología mexicana (IV). México: El Colegio de México, Fondo de Cultura Económica.

López Morales, H. (1983). *Estratificación social del español de San Juan de Puerto Rico*. México: UNAM.

Marcos Marín, F. (1992a). *Corpus lingüístico de referencia de la lengua española en Argentina* [en línea]. Disponible en: http://www.lllf.uam.es/ESP/Argentina.html

Marcos Marín, F. (1992b). *CORLEC: Corpus Oral de Referencia de la Lengua Española Contempóranea*. Madrid: Universidad Autónoma de Madrid [en línea]. Disponible en: https://web.archive.org/web/20171101122930/http://www.lllf.uam.es/ESP/Corlec.ht…

Marcos Marín, F. (1994). *Corpus lingüístico de referencia de la lengua española en Chile* [en línea]. Disponible en: http://www.lllf.uam.es/ESP/Chile.html

Molina Salinas, C. & Sierra Martínez, G. E. (2015). Hacia una normalización de la frecuencia de los corpus CREA y CORDE. *Revista Signos. Estudios de Lingüñistica*, *48*(89), 307-331. DOI: http://dx.doi.org/10.4067/S0718-09342015000300002.

Moreno Fernández, F. (1993). *La división dialectal del español de América*. Alcalá de Henares: Universidad de Alcalá.

Moreno Fernández, F. (1996-1997). La variación de /s/ implosiva en las hablas andaluzas: Análisis cuantitativo. *Studia Hispanica in honorem Germán de Granda. Anuario de Lingüística Hispánica*, XII-XIII, 939-957.

Moreno Fernández, F. (2005). Corpus para el estudio del español en su variación geográfica y social. El corpus PRESEEA, *Oralia* 8, 123-139.

Moreno Fernández, F. (2019). Macro-regional sociolinguistics: Uses and perceptions on null direct objects in Spanish. *Journal of Linguistic Geography*, *7*(01), 1-15.

Moreno Fernández, F. (2021). La variación geográfica y social y en los corpus lingüísticos. En G. Parodi, P. Cantos & Ch. Howe (Eds.), *The Routledge Handbook of Spanish Corpus Linguistics* (xx-xx). Londres: Routledge.

Moreno Fernández, F. & Moreno Sandoval, A. (2018). Configuración lingüística de anglicismos procedentes de *Twitter* en el español estadounidense. *Revista Signos. Estudios de Lingüística, 51*(98), 382-409.

Moreno-Sandoval, A., de la Madrid, G., Alcántara, M., González, A., Guirao, J. M. & De la Torre, R. (2005). The Spanish Corpus. En E. Cresti & M. Moneglia (Eds.), *C-Oral-Rom. Integrated Reference Corpora for Spoken Romance Languages* (pp. 35-161). Ámsterdam: John Benjamins.

Otheguy, R. & Zentella, A.C. (2012). *Spanish in New York*. Oxford: Oxford University Press.

Otheguy, R., García, O. & Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review, 6*(3), 281-307.

Parodi, G., Cantos, P. & Howe, Ch. (Eds.) (2021). *The routledge handbook of Spanish corpus linguistics*. Londres: Routledge.

Quesada Pacheco, M. A. (2014). División dialectal del español de América. *Boletín de Filología*, *49*(2), 257-309.

Rabanales, A. (1992). Fundamentos teóricos y pragmáticos del 'Proyecto de estudio coordinado de la norma lingüística culta del español hablado en las principales ciudades del mundo hispánico', *BFUCh,* XXXIII, 251-272.

Real Academia Española: Banco de datos (CREA) [en línea]. Disponible en: *Corpus de referencia del español actual.*

Real Academia Española: Banco de datos (CORPES) [en línea]. Disponible en: *Corpus del español del siglo XXI.*

Real Academia Española (2013). *Corpus del español del siglo XXI (CORPES). Descripción del sistema de codificación. Libros y prensa*. Recurso de Internet. Madrid: Real Academia Española.

Real Academia Española y Asociación de Academias de la Lengua Española (2010). *Ortografía de la lengua española*. Madrid: Espasa.

Reese, S., Boleda, G., Cuadros, M., Padró, Ll. & Rigau G. (2010). Wikicorpus: A Word-Sense Disambiguated Multilingual Wikipedia Corpus. En *Proceedings of 7th Language Resources and Evaluation Conference (LREC'10)*, La Valleta, Malta.

Rojo, G. (2010). Sobre codificación y explotación de corpus textuales. Otra comparación del Corpus del Español con el CORPES y el CREA. *Lingüística*, 24, 11-50.

Rojo, G. (2014). Hispanic corpus linguistics. En M. Lacorte (Ed.), *The Routledge Handbook of Hispanic Applied Linguistics* (pp. 371-387). Nueva York: Routledge.

Sadowsky, S. (2006). *Corpus Dinámico del Castellano de Chile (Codicach)*. Base de datos electrónica [en línea]. Disponible en: http://sadowsky.cl/codicach.html

Samper, J. A., Hernández, C. E. & Troya, M. (Eds.) (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*. Edición en CD-ROM. Las Palmas de Gran Canaria: Universidad de las Palmas de Gran Canaria.

Sánchez, A. (1995). *Cumbre. Corpus lingüístico. Del español contemporáneo*. Madrid: S.G.E.L.

Silva-Corvalán, C. (1994). *Language contact and change: Spanish in Los Angeles*. Oxford: Clarendon.

Sinclair, J. (1996). The search for units of meaning.*TEXTUS,* IX, 75-106.

Tagliamonte, S. (2013). Comparative sociolinguistics. En J. K. Chambers & N. Schilling (Eds.), *Handbook of Language Variation and Change* (pp. 729-763). Oxford: Balckwell.

Trudgill, P. (1974). *Sociolinguistics*. Harmondsworth: Penguin.

Ueda, H. (1995). Zonificacion del espanol del mundo. Palabras y cosas de la vida urbana. *Lingüística*, 7, 43-86.

Vila Pujol, M. R. (2001). *Corpus del español conversacional de Barcelona y su área metropolitana.* Barcelona: Universitat de Barcelona.

Woolard, K. (1998). Language Ideologies as a Field of Inquiry. En B. Schieffelin, K. Woolard & V. Kroskrity (Eds.), *Language ideologies: practice and theory* (pp. 3-43). Nueva York: Oxford University Press.