ANUARIO Lope de Vega texto · literatura · cultura

RESEÑA

Joan Torruella Casañas, *Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación científica*, Peter Lang, Frankfurt am Main, 2017, 279 pp. ISBN: 9783631717189.

Assumpció Rost Bagudanch (Universitat de les Illes Balears)

DOI: <https://doi.org/10.5565/rev/anuariolopedevega.324>

En ocasiones, cuando uno busca un libro que sirva de manual y de referencia básica, precisa y clara para un determinado tema, o no lo encuentra o se da de bruces con obras de una complejidad excesiva y gratuita. Este no es el caso del trabajo de Joan Torruella, un volumen útil y accesible para todo aquel que desee acercarse por primera vez a la lingüística de corpus o que quiera hallar una síntesis práctica de los aspectos fundamentales de esta disciplina y que esté particularmente interesado en los corpus lingüísticos de tipo histórico.

Lingüística de corpus: génesis y bases metodológicas de los corpus (históricos) para la investigación científica presenta una estructura clara, bien establecida y plenamente justificada: tras unas páginas de presentación (pp. 15-19), la primera parte (pp. 21-60) funciona como una introducción a las disciplinas de base, la segunda (pp. 61-233) supone un conciso pero completo manual para la confección de corpus lingüísticos y la tercera (pp. 235-259) quiere dar cuenta de los aspectos relativos al análisis de los datos que se pueden extraer de los corpus. Cada una de ellas está constituida por cuatro capítulos (más un epílogo, en el caso de la tercera) dedicados a desarrollar aspectos específicos dentro de la temática general de la sección.

Así, el primer capítulo («La lingüística de corpus», pp. 23-29) presenta la lingüística de corpus como disciplina, traza brevemente su historia y esboza la perspectiva de diversos autores relevantes en la materia. En él se pone de relieve la importancia de los corpus para llegar a resultados y conclusiones fiables al abordar el estudio de los hechos lingüísticos, aunque sin dejar de lado el necesario trabajo

filológico de tipo más tradicional, que también se pone en valor y es reconocido como parte esencial de la investigación. Este es un acertado punto focal en todo el trabajo: no hay un alegato exclusivo a favor de lo nuevo en detrimento de los métodos "antiguos", sino que se enfatiza cómo las innovaciones de la tecnología suponen una ayuda más que una sustitución. Otra idea importante es que los corpus son útiles en el estudio de todos los niveles de la lengua y que su uso es pertinente desde cualquier postulado teórico.

El segundo capítulo («Corpus textuales», pp. 31-39) trata específicamente de los corpus textuales, centrándose de forma más concreta en los de tipo escrito. Se hace notar su relevancia en la lingüística actual, que hoy en día no se puede obviar la necesidad de partir de los datos reales y no exclusivamente de la intuición lingüística. En este sentido, se contrapone esta manera de investigar con la perspectiva generativista más tradicional, de tipo racionalista y menos dada a la exploración de los datos reales (al menos en sus inicios). Así, se reflexiona sobre la revolución que ha supuesto el empleo de los corpus, ya que ha significado poder llegar a conclusiones empíricas y más objetivas sobre el funcionamiento de la lengua en todos sus niveles. Es interesante, sin embargo, que se haga notar que los principales campos que han acudido a la lingüística de corpus han sido el léxico, la morfología y la sintaxis, mientras que se deja más de lado los avances que se están dando en este terreno en otros, como la fonética (baste señalar la existencia de corpus orales como los del proyecto PRESEEA o Val.Es.Co en el ámbito del español, o, ya preparados para uso específico en fonética, el corpus Ahumada o el Albayzín, así como los corpus EUROM o SpeechDat en el ámbito europeo).

El capítulo 3 («Parámetros clasificatorios de los corpus», pp. 41-57) revisa los diversos parámetros que hay que tener en cuenta a la hora de confeccionar un corpus y definir sus características básicas; dicho de otro modo, se ocupa de cuáles son los factores que determinan las propiedades de todo corpus: su modalidad, su temática, la época que abarca, su temporalidad, su magnitud en términos de volumen de palabras, su evolución, la distribución de los documentos que lo componen respecto a las variables que se han de considerar, el número de ediciones de un mismo texto que se admitirán, el de las lenguas que podrán entrar en él, el tipo de ediciones que se van a seleccionar, las muestras y el tipo de marcaje que se pretende llevar a cabo. Como puede observarse, se trata de parámetros dirigidos a corpus formados a partir de textos escritos, primordialmente. A lo largo del capítulo, se caracteriza cada

uno de ellos y se identifican las posibles dificultades a la hora de establecer los criterios que condicionarán los documentos que han de formar parte del futuro corpus.

Para terminar la primera parte, el capítulo 4 («Corpus de lectura», pp. 59-60) define la naturaleza de los corpus de lectura y subraya su importancia capital como complemento indispensable de los corpus informatizados. De hecho, se observa que, a la hora de abordar cualquier investigación con un mínimo de rigor y de dominio de los materiales, es recomendable dominar el tipo de textos con los que se deberá trabajar con el fin de realizar calas previas y familiarizarse con su tipología y características. Así pues, la labor de lectura en sentido más tradicional no es algo que deba pasar a la historia con la aparición de los corpus informatizados, sino que cobra un sentido diferente, como trabajo preliminar ineludible para llevar a cabo análisis rigurosos e interpretaciones correctas de los datos que se van a extraer de los repertorios informatizados.

La segunda parte comienza con un breve capítulo, el 5 («Fases en la construcción de un corpus», pp. 63-65), que funciona como unas consideraciones iniciales acerca del proceso de constitución de un corpus. En él se enumeran las fases que habrán de sucederse en ese proceso y se ofrecen una serie de consejos y observaciones generales de tipo práctico para encarar con solvencia tal empresa. Se trata de unas páginas de lectura agradable, de contenido riguroso, que anticipan lo que se va a explicar con mayor detalle en los capítulos siguientes, cada uno de ellos dedicado a una o varias de estas fases.

En efecto, el capítulo 6 («Estructura y ejes principales», pp. 67-128) describe la estructura y los ejes que actúan como pilares de un corpus y supone una revisión de las bases teóricas que han de sustentar los proyectos de esta clase. De hecho, merece la pena resaltar un aspecto que se cree muy enriquecedor de esta sección: la relación que establece el autor entre la estructura de un corpus y los aspectos teóricos de la variación lingüística, así como los de la pragmática y la lingüística del texto. Se vincula la creación del corpus con los aspectos de lingüística teórica que han de permitir analizar y explicar satisfactoriamente la evolución lingüística, algo que a veces se pierde de vista cuando se tratan cuestiones técnicas ligadas a corpus. Llegados a este punto, debe mencionarse que se restringe la explicación a los corpus de tipo histórico.

Los tres ámbitos que centran la atención del autor son el temporal, el diatópico y el tipológico. El primero resulta fundamental por cuanto ha de permitir establecer las épocas en que se inscriben los textos contenidos en el corpus y su periodización en

intervalos de tiempo previamente acordados. El segundo atañe a la relación de los datos lingüísticos con la geografía lingüística (la variación dialectal), algo importantísimo si se tiene en cuenta que esta puede jugar un papel relevante en la reconstrucción histórica y en el estudio de la variación en sí. El tercero, a su vez, entronca con la teoría del discurso y de la comunicación en general, esencial para entender la progresión del cambio, ya que la tipología lingüística se hace eco de situaciones comunicativas concretas y de la variación social y contextual. Es, por lo tanto, el indicador de la variación diafásica, un factor crucial en la explicación del cambio. En los tres casos se presentan diversas opciones de clasificación de los textos, los problemas que estas entrañan y las ventajas que conlleva cada una. En cualquier caso, se advierte que cada alternativa resultará más o menos adecuada en función de la finalidad para la que se crea el corpus. Además, se propone siempre una solución concreta justificando certeramente su elección en términos prácticos, algo que es muy de agradecer para el público inexperto. Otro aspecto que cabe resaltar es que se ofrecen ejemplos de cómo se ha tratado cada uno de estos tres ejes en diversos corpus históricos del ámbito iberorrománico, lo que clarifica sobremanera las explicaciones efectuadas en cada punto. El capítulo se cierra con una serie de consideraciones referentes al tratamiento de los textos traducidos (si han de incluirse o no, pros y contras de hacerlo), pese a que en este caso el autor no ofrece una propuesta clara de actuación.

El capítulo 7 («Composición del corpus», pp. 129-163) se concentra en la fase práctica de la elaboración de los corpus históricos: atendiendo a los criterios de selección de las obras explicados en el capítulo anterior, se ha de elegir el/los documento/s más adecuado/s para cada una de ellas con el fin de cubrir las variedades lingüísticas contempladas y previamente establecidas. Para ello, se hace hincapié en una serie de conceptos fundamentales. Por una parte, y de forma más exhaustiva, se exponen las cuestiones de la representatividad de las muestras y de su equilibrio. Por otra, se trata la naturaleza provisional o definitiva de los corpus, los criterios de selección de las obras, los que afectan a la selección de los documentos para cada obra y los problemas relacionados con la filiación de los mismos (lo que técnicamente se conoce como metadatos). Como síntesis, se reproducen los diez principios básicos de Sinclair (2005:1-14)[1] para la construcción de corpus (p. 163).

1. Véase John Sinclair, «Corpus and text — Basic principles», en Martin Wynne, ed., *Developing Linguistic Corpora: A Guide to Good Practice*, Oxbow Books, Oxford, 2005, pp. 1-16.

Tal y como ya ocurriera en el capítulo 6, se muestran las dificultades más habituales en cada uno de estos aspectos y se ofrecen las soluciones consideradas más pertinentes para cada uno.

Finalmente, la segunda parte se cierra con una sección dedicada a la preparación de los documentos para su inclusión en los corpus (capítulo 8 «Preparación de los textos», pp. 165-233). Se trata de unas páginas que resumen de forma clara y concisa los principios básicos que rigen el tratamiento informático de la documentación que ha de constituir un corpus. Como saben quienes han colaborado alguna vez en tareas de este tipo, es un conocimiento que no es fácilmente accesible y que hay que buscar o bien en opúsculos muy concretos, o bien en obras excesivamente generales y extensas, o bien a través de resúmenes propios de cada grupo/proyecto de investigación, que no siempre aparecen sintetizados y sistematizados de manera clara y eficiente. Este apartado, pues, supone una aportación importante. En él se distingue entre los conceptos y las características de la edición textual, la edición filológica digital y la edición lingüística. Para cada una se ofrece una definición y descripción de su alcance, los problemas más frecuentes y se aportan soluciones para la homogeneización de criterios, así como consejos prácticos. Asimismo, se presentan ejemplos que ilustran la explicación y contribuyen a una comprensión más efectiva. Se trata de una sección altamente interesante porque hace asequible la entrada al lenguaje de codificación informática propio de cada uno de estos estadios de edición, sea cual sea el nivel de conocimientos previos del lector.

La tercera parte es la más breve. Como hemos dicho ya, se divide en cuatro capítulos seguidos de un quinto, titulado «Epílogo», que funciona como *petitio benevolentiae* de cara a justificar las posibles lagunas en cuanto a contenidos en los cuatro precedentes. Estos se fijan en la fase posterior a la elaboración de un corpus: la de su explotación, es decir, la extracción de los datos y su tratamiento para lograr explicaciones científicas.

El capítulo 9 («Elementos base en la investigación científica», pp. 237-242) viene a ser una introducción al método científico aplicado a los estudios de lingüística, especialmente en su vertiente histórica. Define sus bases más generales estableciendo una distinción clara entre método experimental y método observacional, también denominado comparativo (que sería el empleado en el ámbito de la diacronía lingüística). Resulta una buena descripción de síntesis, aunque le

faltaría incidir en cuestiones como la formulación de objetivos de trabajo a partir de las preguntas de investigación, que deberían ser previos al planteamiento de las hipótesis.

El siguiente paso, que corresponde al capítulo 10 («Método comparativo», pp. 243-246), quiere ser una descripción de los fundamentos del método comparativo. El autor enfatiza las ventajas de su empleo en estudios de lingüística histórica por la idoneidad de sus postulados básicos: como no es posible la experimentación en este ámbito, hay que trabajar a partir de la observación de los testimonios documentales disponibles. Así pues, la observación y el análisis numérico de los fenómenos lingüísticos detectados en la documentación de los corpus es la que ha de proporcionar resultados que permitan establecer la existencia de leyes de cambio (y, si es el caso, de teorías). Se insiste, con mucha razón, en que la vía más objetiva para el estudio histórico de la lengua es precisamente la que marca el método científico y, en concreto, el observacional derivado de él. Como ya sucedía en el capítulo anterior, se lleva a cabo una síntesis de sus bases, sin entrar en profundidades técnicas ni en especificidades de más calado, como podría ser una explicación detallada de las fases que hay que recorrer para estudios de esta clase.

El capítulo 11 («Bases estadísticas en la investigación con corpus», pp. 247-253) entra en las cuestiones del tratamiento estadístico de los datos, precisamente porque una de las posibilidades que ofrecen los corpus es la cuantificación de los mismos. Esto significa que pueden aplicarse técnicas empíricas, matemáticas, de análisis. Pese a esta innegable ventaja, se pone acertadamente el acento en que la estadística no ha de ser un fin en sí misma, ni la varita mágica que aporte soluciones: es labor del investigador, con sus conocimientos y su capacidad de estudio filológico de interpretación fina, quien ha de poder extraer conclusiones a partir de los resultados de las pruebas estadísticas. Estas orientan, indican tendencias, la ausencia o presencia de relaciones entre las variables, pero no explican los hechos por sí mismas. En cuanto a las técnicas de análisis, se realiza la distinción entre la vía cualitativa y la cuantitativa de análisis de datos, pero no se detalla mucho más. De hecho, respecto a los límites entre las dos, no estaría de más señalar que "cuantitativo" no implica únicamente el recuento de casos (teniendo en cuenta, de hecho, que esta es la opción más habitual en estudios como los que se describen aquí) y que la vía cualitativa no supone solamente una descripción somera. Tampoco estaría de

más relacionarlas con la tipología de las variables (dependientes e independientes) que se barajan en una investigación y con el tipo de pruebas estadísticas que pueden aplicarse.[2]

El penúltimo capítulo («El valor de la estadística», pp. 255-258), el número 12, trata el tema de la representatividad de los resultados estadísticos en términos de investigación en diacronía: cómo se puede asegurar que los resultados son fiables y realmente significativos. Para ello se centra en los conceptos de margen de error y de nivel de confianza. No obstante, se reconoce que para poder hablar de solidez de las pruebas y de relevancia estadística, hay que tener en cuenta otros conceptos que ya han ido apareciendo en apartados anteriores: los de población y muestra. Es indispensable poder establecer cuál es la población de sujetos susceptibles de entrar en un estudio y, a partir de ello, seleccionar una muestra proporcional y representativa. En el caso de los estudios en lingüística histórica, se parte de un problema de base insalvable: es imposible determinar la población total posible, entre otras razones porque se ha perdido una cantidad importantísima de documentación. Así pues, se concluye que los resultados de cualquier análisis han de ser necesariamente cautos, puesto que podrán revelar tendencias en el comportamiento lingüístico, pero no podrán dar cuenta de hechos irrefutables por la misma naturaleza de la disciplina.

Hay varios aspectos que merecen resaltarse en esta obra. El primero de ellos es la modestia del autor, que insiste en varios pasajes en que su pretensión no es hacer un volumen exhaustivo que compendie todos los conocimientos existentes sobre la lingüística de corpus, la creación de los mismos y su explotación, sino proporcionar una buena composición de lugar sobre la materia. Debe decirse que si este es el objetivo (que no es poco, ni sencillo), se ha conseguido con creces. Se ha compuesto un libro que puede convertirse en un muy buen manual, como se indicaba al inicio de esta reseña, muy recomendable para aquellos que quieran acercarse *ex novo* a las cuestiones que se tratan en él y muy aconsejable para los que necesiten consultar aspectos prácticos de la elaboración de corpus. Su lectura es amena; incluye ejemplos pertinentes e interesantes, se ilustran todos los elementos complejos de forma que se logra algo al alcance de pocos: hacer parecer fácil lo que, en realidad, es francamente complicado.

---

2. Al estilo de, por ejemplo, Keith Johnson, *Quantitative Methods in Linguistics*, Blackwell, Oxford, 2008.

Esto se explica, en parte, por la vasta experiencia del autor en el ámbito de la lingüística de corpus. Su trayectoria se trasluce en los contenidos del libro y en la manera de explicarlos: su buen hacer como investigador y miembro de proyectos que se han centrado en la creación de corpus, así como en la aplicación de las nuevas tecnologías a la filología (*Corpus Informatitzat del Català Antic*, *Portal del Léxico Hispánico* o la informatización del *Diccionario Crítico Etimológico Castellano e Hispánico* de J. Corominas y J.A. Pascual, por citar unos ejemplos)[3] se traduce en una transmisión ordenada, adecuada y agradable de sus conocimientos, los cuales se fundamentan y se complementan en un amplio dominio de la bibliografía, pertinente, significativa y muy bien seleccionada.

Si hay que poner algunas objeciones a la obra, serían relativas al título y a la tercera parte. Respecto al primero, el adjetivo "históricos" entre paréntesis acaba resultando algo dubitativo. A nuestro entender, el volumen se concentra de forma clara en los corpus pensados para el análisis histórico de la lengua, salvo por los primeros capítulos de corte más general. No debería haber complejos, si se trata de abordar la lingüística de corpus dedicada a la lingüística histórica. De hecho, esta es una de las contribuciones del volumen, puesto que existen otros dedicados a la lingüística de corpus,[4] pero no es frecuente encontrarlos sobre esta vertiente específica, por lo que vale la pena destacarlo.

En cuanto a lo segundo, la tercera parte resulta prometedora en su planteamiento, pero queda algo descompensada en relación al resto del libro. Como se ha explicado ya, en el capítulo 13 se indica expresamente que no se quiere hacer una relación exhaustiva de las cuestiones referentes al análisis de los datos; sin embargo, hubiera sido una buena oportunidad para reunir las nociones indispensables para llevar a cabo un análisis estadístico en condiciones, tanto desde el punto de vista descriptivo como comparativo (de relación entre variables), algo que sería de mucha utilidad para los lectores. En efecto, del mismo modo que se puede seguir sin

3. Joan Torruella, Manel Pérez Saldanya, Josep Martines, dirs., *Corpus Informatitzat del Català Antic*, http://www.cica.cat/; Seminario de Filología e Informática, *Portal del Léxico Hispánico*, http://portaldelexico.es/; Joan Corominas, José Antonio Pascual, *Diccionario Crítico Etimológico Castellano e Hispánico. Edición en CD-Rom*, Gredos, Madrid, 2012.

4. Véanse, por ejemplo, Douglas Biber, Susan Conrad, Randi Reppen, *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge University Press, Cambridge, 1998; Giovanni Parodi, *Lingüística de corpus: de la teoría a la empiria*, Iberoamericana, Frankfurt, 2010; Graeme Kennedy, *An Introduction to Corpus Linguistics*, Routledge, Londres, 2014, por citar referencias recientes, algunas también mencionadas por el autor.

dificultad el proceso de confección de un corpus y qué criterios hay que tener en cuenta para ello, hubiera sido muy adecuado proporcionar información sobre los pasos que constituyen el método científico y las técnicas de análisis estadístico básicas para un primer procesamiento de los datos. Hoy en día, este tipo de conocimientos hay que ir a buscarlos en obras que no suelen tener en cuenta el perfil del lingüista y, menos aún, las necesidades del especialista en lingüística histórica.

No se quiere acabar sin recalcar que se considera muy pertinente e interesante que no se haya enfocado la obra únicamente como una serie de procedimientos para la confección de un corpus y en cómo pueden analizarse los datos. Es importantísimo que se haya relacionado continuamente este aspecto más procedimental con sus bases teóricas de fondo, que entroncan directamente con disciplinas tan variadas como la sociolingüística, el análisis del discurso, la pragmática o la geografía lingüística. El hecho de no perder de vista que el corpus no es simplemente un conjunto de textos que se han recopilado para tener una colección de palabras susceptibles de ser analizadas asépticamente, sino que existe un vínculo estrechísimo entre su selección cuidadosa y la naturaleza poliédrica del cambio es fundamental: es una obra que contiene una profunda reflexión, más o menos obvia, más o menos latente, pero continuamente presente, sobre el cambio lingüístico en prácticamente todos sus aspectos. Y se pone mucha atención en que quede claro en cada uno de los estadios de la creación de corpus que se desgranan. Esto, a juicio de esta reseñadora, marca la diferencia y supone un ingrediente de excelencia.