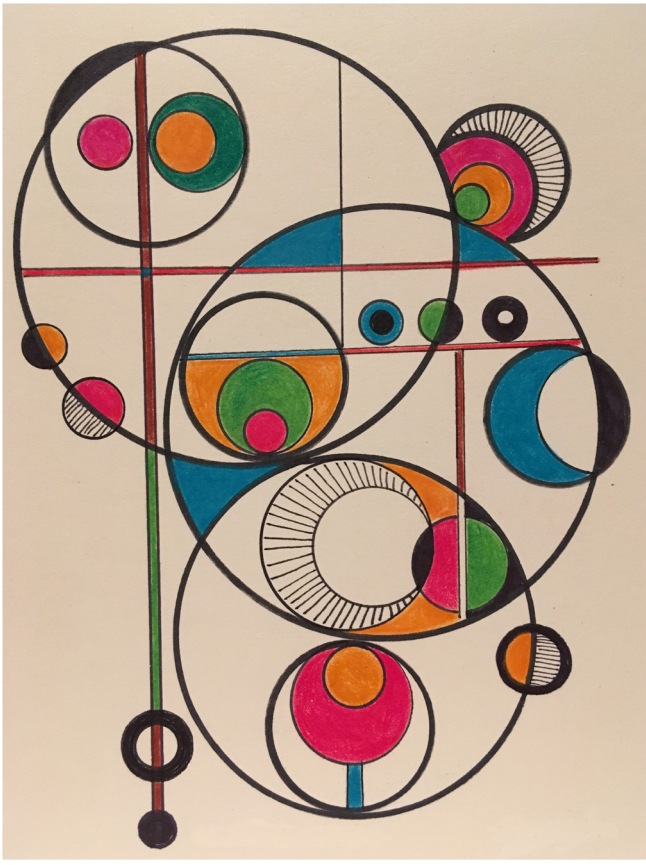
LA CONSTITUCIÓN DEL CORPUS

Algunas reflexiones teórico prácticas para investigaciones con análisis de contenido o de corpus

Melanie
Salgado López



BREVIARIOS de Lingüística

UNAM POSGRADO Lingüística

Dr. Leonardo Lomelí Vanegas
Rector

Dra. Patricia Dolores Dávila Aranda
Secretaria General

Dra. Diana Tamara Martínez Ruiz
Secretaria de Desarrollo Institucional

Dra. Cecilia Silva Gutiérrez
Coordinadora General de Estudios de Posgrado

Dra. María del Carmen Curcó Cobos
Coordinadora del Programa de Maestría y Doctorado en Lingüística

LA CONSTITUCIÓN DEL CORPUS

BREVIARIOS de Lingüística

La Colección **Breviarios de Lingüística** publica textos breves sobre temas selectos de lingüística, así como estudios específicos recientes con una dimensión didáctica. Se propone difundir propuestas académicas generadas en nuestro programa, pero también recibe trabajos externos.

COORDINACIÓN GENERAL DE ESTUDIOS DE POSGRADO
Programa de Maestría y Doctorado en Lingüística

LA CONSTITUCIÓN DEL CORPUS

Algunas reflexiones teórico prácticas para investigaciones con análisis de contenido o de corpus

Melanie Salgado López

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
México, 2023

Catalogación en la publicación UNAM. Dirección General de Bibliotecas y Servicios Digitales de Información

Nombres: Salgado López, Melanie, autor.

Título: La constitución del corpus : algunas reflexiones teórico prácticas para investigaciones con análisis de contenido o de corpus / Melanie Salgado López.

Descripción: Primera edición. | México : Universidad Nacional Autónoma de México, 2023. | Serie: Breviarios de Lingüística.

Identificadores: LIBRUNAM 2223386 | ISBN 978-607-30-8543-4.

Temas: Corpus lingüístico. | Análisis lingüístico. | Análisis del discurso. | Investigación -- Metodología.

Clasificación: LCC P128.C68.S35 2023 | DDC 415.0285—dc23

Esta publicación fue sometida a un proceso de dictaminación
a doble ciego por pares académicos.

Ilustración de la portada:
Circle Games # 3, Rosa María C. Dies

Diseño de portada: Diego García del Gállego

Primera edición: noviembre de 2023

Coordinación General de Estudios de Posgrado
Programa de Maestría y Doctorado en Lingüística

ISBN 978-607-30-8543-4

Hecho en México

doi: https://doi.org/10.22201/cgep.9786073085434e.2023

Para Noëlle Groult,
porque no tengo espacio para
acomodar tanto vacío.

Tabla de contenido

Resumen 10

Introducción 12

1.Definiciones básicas 17
1.1 Introducción 17
1.2 ¿Qué es un corpus? Su definición laxa 18
1.3 Corpus: su definición especializada 18
1.3.1 *El corpus: una manera distinta de acercarse a los datos y de construir objetos de estudio* 25
1.3.2 *El corpus es una construcción metodológica (teórico- metodológica) de la investigación misma* 28
1.3.3 *Una pertinencia y representatividad diferentes* 30
1.3.4 *El rechazo a la neutralidad* 31
1.4 Recuento general 33
1.5 Reflexiones recomendadas 35
1.6 Para leer más acerca de la definición específica del corpus 36

2. Tipos de corpus 38
2.1 Introducción 38
2.2 Tipos de clasificaciones y corpus 39
2.2.1 *Por el tipo de soporte en el que se presenta el contenido o material del corpus* 39
2.2.2 *Por el periodo de los datos que reúne* 42
2.2.3 *Por el tipo de unidades que se reúnen o estudian* 43
2.2.4 *Por el grado de representatividad con que se puede observar el fenómeno de estudio en el material que se reúne* 44
2.2.5 *Por el grado de análisis o trabajo de etiquetado que contiene o no* 46
2.2.6 *De acuerdo a si reportan o no su procedencia y criterios* 46
2.2.7 *Por la función que cumple en la investigación* 47
2.2.8 *Por el procedimiento que nos lleva a él* 48

2.3 Reflexiones recomendadas. 50
2.4 Para leer más acerca de los tipos de corpus. 51

3. Universo, acervo y corpus. 53
3.1 Introducción 53
3.2 Definiciones básicas 53
3.2 Del universo al corpus 57
3.3 Facetas en el camino para llegar al corpus 58
3.3.1 *Para constituir un corpus* 58
3.3.2 *Para recolectar o crear un corpus* 59
3.3.3 *Para quienes van a replica un corpus o una metodología de corpus*. 59
3.4 Reflexiones recomendadas. 60
3.5 Ejemplos de trabajos en donde se reporta el tránsito del universo al corpus 61
3.6 Para leer más sobre universo, acervo y corpus 62

4. Los criterios del corpus 63
4.1 Introducción 63
4.2 Variables y criterios operativos 63
4.3 Los criterios que permiten llegar al corpus: Homogeneidad/heterogeneidad; internos y externos 64
4.4 Los criterios en distintos tipos de corpus 66
4.5 ¿Es mi corpus comparable?: La normalización del corpus. 68
4.6 La subjetividad y neutralidad en el corpus. 69
4.7 Para leer más sobre los criterios y la normalización del corpus 70

5. La presentación argumentada del corpus 71
5.1 Introducción 71
5.2 Descripción y argumentación del proceso y los criterios metodológicos y su pertinencia 71
5.3 El acceso al corpus 73
5.4 Para leer ejemplos de reportes de metodología en la constitución de corpus 74
5.4.1 *Sobre constitución de corpus en investigación de Análisis del discurso* 74
5.4.2 *Sobre la constitución y retomado de corpus para análisis lingüístico*. 74
5.4.3 *Sobre la constitución de corpus para obtención de datos en enfoques experimentales*. .75
5.4.4 *Sobre el trabajo con corpus multimodales* 75

Bibliografía 76

Resumen

El objetivo de este trabajo es ofrecer un material, introductorio y básico, acerca de las reflexiones teórico-prácticas que son fundamentales en el proceso de selección o diseño de un corpus (sin importar si este es de manifestaciones del lenguaje verbal escrito u oral o de otra naturaleza multimodal o semiótica) para una investigación. Del mismo modo, el lector podrá acercarse a una revisión panorámica y general tanto de los enfoques teóricos como de algunos ejemplos que se relacionan con la selección del corpus. En todos los casos se hace énfasis en las correspondencias que este proceso entabla con el diseño de la investigación misma.

En ese sentido, el libro pretende funcionar como un material de acercamiento que guíe a estudiantes, profesores e investigadores para identificar los retos fundamentales de diseño metodológico de las investigaciones que apuestan por aproximarse a un fenómeno mediante el análisis de contenido o de corpus (se sitúen o no dentro de la Lingüística), el tipo de congruencia y concordancia que hay que cuidar en estos enfoques y algunas definiciones, posturas teóricas y ejemplos que ayudan para la comprensión y reflexión de este trabajo. Por ello mismo, este material permite conocer el panorama necesario para el diseño de un corpus, así como para desarrollar una postura crítica, alimentada por el conocimiento de fuentes reconocidas, ante las tareas que el investigador o estudiante debe enfrentar. Con el objetivo de que tal cometido se cumpla, en el trabajo se encontrarán nociones, definiciones, ejemplos, reflexiones y un listado de fuentes en el que se puede profundizar en los aspectos abordados.

El texto se ha organizado de la siguiente manera. Luego de una introducción, aparece el primer capítulo donde se exponen las definiciones fundamentales con respecto al corpus, así como un panorama general acerca de las distintas (y más importantes o destacadas) posturas teóricas que se han desarrollado con respecto al tema, con el objetivo de partir de un marco teórico o estado de la cuestión mínimo que destaca el potencial que el trabajo con corpus y su diseño ofrece para los investigadores; enseguida se habla sobre las distintas tipologías y clasificaciones que se han construido con respecto a los tipos de corpus; luego se analizan tres conceptos fundamentales: el universo, el acervo y el corpus. Ya en el capítulo 4 se reflexiona acerca de algunas cuestiones y retos fundamentales en el camino hacia la selección del corpus y, finalmente, en el último, se plantean las cuestiones metodológicas que es deseable reportar con respecto al proceso de diseño de corpus en una investigación. En cada apartado se postula un diálogo con los autores fundamentales que han abordado estos constituyentes que giran en torno al corpus de una investigación; sin embargo, hay que advertir que la división y el orden que estos presentan no supone una correspondencia

con un orden que se deba seguir ni con un grado de importancia, simplemente suponen diversas dimensiones del trabajo de diseño de un corpus.

Esperamos que el lector encuentre aquí un incentivo estimulante, un acompañamiento que lo reatroalimente en el trabajo de constitución, obtención o recolección de un corpus, y que le permita repensar críticamente su propia práctica y hacerse consciente de los aspectos más fundamentales en esta ardua y primordial labor.

Introducción

Antes de exponer los asuntos fundamentales que deseamos dirigir, a manera de introducción, a quien se acerca a esta publicación, es necesario hacer algunas aclaraciones importantes.

En primer lugar, el lector debe de saber que, a lo largo de este trabajo, se ha decidido conscientemente utilizar corpus cuya forma es invariable en el plural, y no usar el término *corpora* que, por influjo del inglés, se utiliza a veces replicando la forma de pluralización latina. Tampoco se usarán las cursivas cuando aparezca la palabra *corpus*, pues tal término forma parte del inventario léxico del español. Más adelante, se definirá claramente el término especializado en el sentido que interesa para este trabajo. En segundo lugar, es importante explicar que, pese a que en la tradición lingüística se suele utilizar el término *elicitar* para hacer referencia a la acción de obtener, a partir de ciertos materiales y tareas, los datos de una investigación, nos hemos decantado por usar el verbo obtener y sus conjugaciones, debido a que el primer verbo está indicado como un calco del inglés, sin embargo, como bien puede notar el lector, sobre todo quien provenga de la tradición de estudios lingüísticos, nos estamos refiriendo a lo mismo que en ciertos textos y referencias aparecerá como elicitación o elicitar.

Del mismo modo, es importante advertir que, a lo largo del libro, se utilizarán las palabras selección y/o diseño para hacer referencia general al proceso global de la construcción de un corpus. Con las formas de los lemas de los verbos seleccionar o diseñar, por lo tanto, se hace una mención general y laxa de tal proceso, sin poner énfasis en el tipo de método utilizado para tal fin. El lector encontrará en el Capítulo 2 las definiciones más precisas que se propone utilizar para referirse específicamente a distintos procedimientos, métodos o metodologías para diseñar un corpus. De ahí también que sea importante advertir desde ahora que, a lo largo de este trabajo, se parte del supuesto de que existe una diferencia clara entre constituir y construir un corpus. En el primer caso se hace referencia específica al procedimiento en el que el investigador selecciona un conjunto de manifestaciones o unidades que se han generado en el proceso de la comunicación social (con sus propios fines y contextos), y que el investigador elige (bajo principios, criterios y metodologías claras) para la construcción y análisis de su objeto de estudio; mientras que obtener un corpus señala específicamente el método en el que el investigador ha de diseñar una serie de instrumentos que le permitan obtener el conjunto de materiales con los que trabajará como la fuente de datos de su investigación. La diferencia entre seleccionar manifestaciones que no fueron creadas para nuestra investigación y usar instrumentos para obtener los elementos

que formarán parte de nuestro corpus no es menor y se abordará con cuidado en el capítulo ya mencionado.

En tercer lugar, y como última aclaración, queremos insistir en que es cierto que este texto se concibió para que formara parte de un proyecto de publicaciones, del Programa de Maestría y Doctorado en Lingüística, de acercamiento general a temas importantes lingüísticos, de ahí que se apele constantemente a investigaciones en este campo. Sin embargo, lo dicho a lo largo del trabajo aplica para investigaciones de distintas disciplinas de las Ciencias Sociales que deseen hacer un análisis a partir de un corpus. Por ello mismo, en cada apartado, así como en la bibliografía final y en algunos ejemplos, se ofrece información útil para quien está trabajando con fenómenos que suponen no solo unidades de estudio del lenguaje verbal, escrito u oral, sino para quienes trabajan con materiales semióticos o multimodales. Entendemos la multimodalidad en el sentido en que lo ha postulado O'Halloran (2012) al especificar que las unidades de análisis de los procesos de comunicación y semiótica social no se construyen ni generan solo con los elementos lingüístico verbales (escritos u orales), sino por medio de un sinfín de elementos como las imágenes, colores, músicas, gestos, tamaños, etc., que desempeñan un papel importante en el proceso de significación social.

Por tanto, es importante entender que aunque la creación de este libro nació de las expectativas de fortalecer el rigor metodológico de estudios lingüísticos de corpus (incluidos los del análisis del discurso) que se realizan actualmente, el lector notará muy pronto que los aspectos abordados aplican para muchas otras disciplinas; no obstante, es la naturaleza de este proyecto y de sus objetivos mismos lo que nos ha llevado a decidirnos por poner más énfasis en los ejemplos y materiales lingüísticos sin que eso quiera decir que ignoremos que existen muchos otros ejemplos y metodologías para trabajar con unidades que no son lingüísticas y que, aunque con menos insistencia, se han abordado en cada uno de los capítulos.

Pese a que es posible que consideremos que el diseño de un corpus se hace necesario tan solo cuando uno quiere hacer un análisis discursivo o un análisis lingüístico, hay muchas otras disciplinas en las que los trabajos e investigaciones se basan en los datos obtenidos con un corpus (recorte del fenómeno total o muestra de él) que el mismo investigador ha construido, constituido, definido u obtenido para tales fines. Las diversas disciplinas que, por distintas razones, abordan una investigación que tiene por objeto de estudio las manifestaciones del lenguaje en uso (o de cualquier otra manifestación semiótica), apuestan, al decidirse a trabajar con un corpus por un enfoque metodológico específico y, por eso mismo, requieren abordar este aspecto con rigor científico. Si bien es cierto que el diseño de un corpus no es una labor específica de la Lingüística, sí es una que los lingüistas utilizan ya sea para construir un repertorio en el que se pueda observar el lenguaje (en uso o desuso, de una o muchas regiones, de

un tipo u otro de hablantes) o bien para reunir un conjunto de elementos que serán analizados para la investigación y que resultan pertinentes para contestar las preguntas planteadas y abordar el objeto de estudio.

Hoy en día se pueden encontrar con mucha frecuencia artículos e investigaciones que, conscientemente o no, obvian precisamente su enfoque metodológico de acercamiento mediante el corpus: los trabajos condenan al silencio o esconden (en el peor de los casos) la descripción y argumentación del diseño de ese corpus, de su pertinencia y la relación con los objetivos de la investigación y con la construcción del objeto de estudio y, por lo tanto, omiten hablar de los límites que su selección impone a la lectura de los resultados del trabajo y sus posibilidades de generalización.

Actualmente, aunque es poco frecuente, aún podemos encontrar ejemplos en los que se habla del corpus como si este fuese tan solo un accidente, una circunstancia secundaria de la investigación. La reserva que debería causar este tipo de afirmaciones es similar a la que todos sentiríamos si un equipo de microbiólogos nos dice que ha creado una vacuna efectiva que no genera ningún tipo de consecuencia negativa ni efecto secundario en los seres humanos y que, luego de ser cuestionados acerca de la muestra de la población con la que se elaboró la prueba que les permite sustentar esto, nos dijeran muy confiados "eso no es importante, porque la muestra del estudio es lo de menos". Una consideración semejante nos causaría leer en un artículo especializado la conclusión de que la vacuna estudiada no causa ningún efecto secundario ni reacción en los seres humanos, luego de haber leído en la descripción de los estudios que se trabajó tan solo con 25 hombres y mujeres con edades promedio de 35 años y pertenecientes al grupo poblacional de procedencia europea. Por supuesto que todos notaremos (quizá con alarma) que es imposible que, a partir de tal muestra, se pueda extender como generalización verdadera lo que se afirma. ¿Por qué no nos causa la misma inquietud encontrar trabajos en donde se afirma que un fenómeno se comporta o es de *x* modo luego de haber observado tan solo una, quizá tres manifestaciones de ese fenómeno en realizaciones muy concretas de la lengua en uso?

El lector avispado estará ya preparando su respuesta a este cuestionamiento: los trabajos que por medio de un corpus ponen en el centro fenómenos del habla, del lenguaje en uso o de la palabra en términos bajtinianos (Bajtín 2005) no se interesan por un sistema, sino por un fenómeno vivo, sin embargo, no se puede encontrar una argumentación sólida que vincule esto con la falta de rigor en el diseño del corpus y con las reflexiones acerca de ello. Tampoco anula este cuestionamiento el hecho de que existen enfoques y metodologías ampliamente reconocidas, sustentadas, aceptadas y difundidas que trabajan con muestras pequeñas a profundidad, como el estudio de caso (Blásquez Martínez y López Moreno 2016). Pero queremos insistir en que en ningún fundamento de estos enfoques se anula la

importancia con la que debe argumentarse la pertinencia, solidez y adecuación justamente del caso que ha de revisarse a profundidad.

Si bien es cierto que es imposible construir un modelo o hablar del proceso de la selección del corpus sin cometer injusticias contra muchas especificidades y particularidades, la ventaja que trae la sistematización y abstracción consiste también en que permite estimular un modo de hacer que se valida socialmente y que, desde nuestro punto de vista, haría muy bien dentro de un contexto en el que abundan publicaciones y análisis de muy dudosa procedencia; en pocas palabras: retomar lo que es fundamental en el diseño metodológico con respecto al corpus y la construcción del objeto de estudio aportaría en la construcción de la validez, rigurosidad y cientificidad de nuestros estudios.

El corpus no es lo de menos y es, como todo aspecto metodológico, uno de los pilares que sostiene una investigación honesta, rigurosa y sólida. De la misma manera en que es importante que la muestra sea suficientemente similar y representativa a la población que se quiere estudiar y que sea pertinente para responder a los objetivos y las preguntas del diseño metodológico de la investigación, es importante que el investigador dedique el tiempo suficiente a analizar las implicaciones metodológicas del corpus con el que ha decidido trabajar, y la pertinencia o no de los criterios que lo llevaron a escoger ese y no otro corpus.

El objetivo de este libro es insistir en la trascendencia metodológica que supone, no solo la elección de una metodología de trabajo por medio de corpus, sino el diseño de un corpus de estudio; aclarar conceptos y nociones que muchas veces parecen vagos y, sobre todo, insistir en que es necesario, ante el evidente hecho de que no existe un modo único de diseñar un corpus ni existe uno que resulte pertinente para toda investigación, seguir (a veces se hace necesario construir) un camino y reportarlo, pues esto dota de solidez y fuerza a nuestras investigaciones, además de que las hace replicables y, por lo tanto, refutables. Al mismo tiempo se busca insistir y hacer explícita la estrecha relación metodológica que se establece entre la construcción del corpus y la construcción misma del objeto de estudio de una investigación. Por lo que la rigurosidad en lo primero necesariamente impacta en el segundo aspecto.

El acento y el hilo conductor de este trabajo no es teórico, sino fundamentalmente práctico. En ese sentido, se ha tratado de sistematizar aquello que hemos aprendido al lado de colegas, alumnos, asesorados, asesores y amigos en experiencias académicas con respecto a lo que debe considerarse cuando uno se enfrenta a la labor de seleccionar un corpus. El lector encontrará, por tanto, definiciones cortas que enseguida se problematizan con el objetivo de insistir en los elementos que deben ocupar nuestras reflexiones y que, desde nuestro punto de vista, son básicos para la conformación del corpus. Es decir, no encontrarán en este libro todo lo que se debe o puede saber acerca del estado de la cuestión de los conceptos

del corpus, sino más bien los aspectos básicos, panorámicos e introductorios, junto con una serie de reflexiones, ejercicios y ejemplos que ayudan a cualquiera a acercarse al proceso de diseño de corpus de una manera más consciente y reflexiva.

1. Definiciones básicas

1.1 Introducción

Se ha vuelto tan natural leer, escuchar y plantear análisis que suponen el trabajo con un corpus, que se suele recibir este concepto con suma naturalidad y ha quedado en segundo plano que este tipo de acercamientos metodológicos tiene detrás muy profundas e importantes discusiones. En ellas, lo primero que se destaca es la complejidad para poder establecer cuándo deja uno de definir el corpus y comienza a hablar de postulados metodológicos y de diseño de investigación. Acercarnos a los trabajos que han intentado profundizar en la definición del concepto corpus en un sentido especializado, como metodología y diseño de una investigación, se vuelve necesario para tener claridad suficiente acerca de todas las implicaciones que el investigador está asumiendo al apostar por una investigación cuyos datos provienen del diseño de un corpus.

Debemos advertir que la organización de este apartado ha sido una de las tareas más difíciles en la construcción de este trabajo, no es sencillo abordar las aportaciones y discusiones con respecto al tema: podíamos haber optado por un hilo conductor cronológico, o bien por uno basado en corrientes o posturas, sin embargo, luego de mucho corregir, reorganizar y borrar, hemos decidido ordenar este apartado a partir de algunos de los elementos centrales que se han discutido desde la definición especializada de corpus, y que revelan su complejidad y relación con el diseño completo de la investigación. Es decir, nos hemos decantado por una organización temática que, fundamentalmente, intenta poner énfasis en las dimensiones metodológicas y teóricas que muestran la profunda imbricación que existe entre el corpus y la investigación misma. Habría sido imposible dar cuenta de tal estado de la cuestión de manera exhaustiva en un espacio tan pequeño, pues los objetivos y la extensión de este esfuerzo nos impiden hacer justicia a las contribuciones que se han hecho en este sentido. Debido a ello hemos decidido tan solo enfatizar las aportaciones más importantes que nos permitan complejizar y destacar la dimensión del corpus como parte del proceso de la investigación y no como una tarea aparte y separada de ella y el diseño metodológico. Por supuesto, no pretendemos afirmar que las que aparecen mencionadas sean todas, sin embargo, consideramos que hemos planteado, al menos, las más significativas.

Esto nos permitirá cumplir con los dos objetivos centrales de este primer capítulo: presentar al lector la definición especializada de corpus y exponer un muy sucinto resumen y diálogo con respecto a las reflexiones y discusiones metodológicas que la elección de análisis de corpus supone.

1.2 ¿Qué es un corpus? Su definición laxa

En el uso común y no especializado, el corpus es un conjunto de textos, de datos o muestras que se analizarán en una investigación. Esta es la definición que podemos encontrar en un diccionario común, como el *Diccionario de la lengua española* (DLE). Por otro lado, Charaudeau y Maingueneau afirman que el corpus es "una compilación vasta, y a veces exhaustiva, de documentos o datos" (2005: 136). Sin embargo, en ninguna de estas definiciones están plasmadas las profundas (y a veces acaloradas) discusiones y reflexiones que se han suscitado alrededor de las investigaciones de análisis de corpus. Desafortunadamente la poca profundidad con la que en muchas fuentes especializadas se aborda la definición de corpus han permitido que su acepción laxa tenga mayor difusión, de hecho, autores como Briz y Albelda (2009) ya han advertido que la definición laxa de corpus se presta para confundirlo con una base de datos. De ahí que sea necesario acercarnos a la discusión y a las posturas que distintos autores han aportado para entender las especificidades que hacen que la constitución de un corpus adquiera las características de pertinencia y adecuación de acuerdo con los objetivos perseguidos y con la metodología seleccionada para una investigación. Para poder profundizar en las reflexiones que nos interesan, es necesario acercarnos a definiciones especializadas del término corpus.

1.3 Corpus: su definición especializada

La introspección (entendida como una metodología en la que el investigador recupera ejemplos y datos a partir de sus propias reflexiones) fue, por un largo periodo de tiempo, una de las metodologías preferidas en el análisis lingüístico y el de otras disciplinas que se sustentaban en algunos ejemplos que se retomaban o se construían para ilustrar aquello que se reflexionaba. No obstante, es evidente que las reflexiones y análisis así sustentados, pese a que tienen una gran validez y obedecen a principios que son reconocidos como rigurosos en su época, permitieron que otros investigadores se acercaran a los mismos fenómenos luego de haber encontrado ejemplos y contraejemplos que cuestionaban, ponían en duda o mostraban la necesidad de matizar las primeras aproximaciones. En este camino se fue haciendo evidente una verdad innegable: los resultados de ciertas investigaciones dependen en gran medida de los datos revisados. El reconocimiento de este hecho no hizo sino ahondarse cuando, en los años sesenta, la crisis experimentada en los estudios de las Ciencias de la Comunicación postuló, entre muchas otras cosas, un profundo rechazo a concentrarse solo en el sistema y no en el habla, un fuerte cuestionamiento a la afirmación de que las oraciones eran la unidad máxima de estudio de la Lingüística, pues la atención sobre el uso mostró que en realidad, los seres humanos nos comunicamos en unidades superiores a las que Harris (1952) llamaría discursos (y en otras tradiciones nombraron como textos), y el

cuestionamiento a que la lógica veritativo-condicional fuera suficiente para explicar todos los fenómenos del uso y significado del lenguaje. [1]

Entre las muchas consecuencias que tal crisis trajo, está la focalización de la importancia de establecer con claridad las implicaciones que la constitución de un corpus supone en el campo de los estudios de la Lingüística. De hecho, el conjunto de estos cuestionamientos y reflexiones dio origen a la Lingüística de corpus que es "una rama de la lingüística que basa sus investigaciones en datos obtenidos a partir de corpus" (Martín Peris *et al.* 2004: s/p).[2] Lo esencial es entender que la Lingüística de corpus constituye una disciplina que se ocupa de la manera como se puede llegar a resultados científicos, válidos y sólidos acerca del lenguaje en uso utilizando un conjunto parcial de esas manifestaciones para observar el comportamiento de un fenómeno en específico. Es decir, se constituye como una línea que se concentra en la reflexión de los principios metodológicos más pertinentes para la construcción de corpus que están a disposición de los lingüistas para poder acercarse a los fenómenos que les interesan de manera que tales materiales cumplan con una serie de criterios y requisitos que permitan comprender la validez de las afirmaciones extraídas de los datos encontrados.

En este sentido, la Lingüística de corpus ha aportado muchísimo no solo con respecto a algunas definiciones, sino también en la construcción de análisis y reflexiones acerca de las especificidades que supone, metodológicamente, el uso de esta forma de acercamiento y estudio de un fenómeno y, por lo tanto, de los aspectos que debemos cuidar, reflexionar seriamente y explicitar en nuestra investigación con el objetivo de que los datos examinados, a partir de una muestra parcial, puedan ser discutidos y comprendidos.

Desde este punto de vista no hay análisis del lenguaje en uso que no se haga con base en un corpus de estudio (aunque la naturaleza de cada corpus pueda variar muchísimo). Es cierto que muchos de los resultados de los postulados de la Lingüística de corpus buscaron atender el cuestionamiento acerca de los estudios en donde el investigador recuperaba de manera introspectiva y un poco a modo los ejemplos que utilizaba. El objetivo era construir corpus exhaustivos en los que se pudiera encontrar un conjunto muy amplio de materiales que permitiera localizar los ejemplos en los que aparecía el fenómeno que interesaba al investigador con el presupuesto de que estos eran más ampliamente representativos que aquellos que hubiese podido reunir simplemente por medio de lo que el investigador había escuchado o recordaba. Surgieron así muchos de los grandes materiales que actualmente están a disposición de quienes trabajan con el lenguaje y que suponen diferentes periodos,

[1] Evidentemente, hemos mencionado solo algunos de los postulados de tal crisis, si el lector quiere acercarse más a este tema, recomendamos ampliamente leer el trabajo de Gutiérrez, Guzmán y Sefchovich (1988: Cap. IX).
[2] Tampoco vamos a profundizar en esta rama específica. Si el lector quiere ahondar en ella, recomendamos consultar a los siguientes autores: Bernal (1985), Murgunova (2013) y Sinclair (1991, 1996) y Sinclair *et al.* (1996).

fuentes y criterios de datos reunidos. Evidentemente, los estudios de Gramática histórica y sincrónica construyeron, a partir de entonces, una larga tradición de consulta y uso de estos corpus exhaustivos en los análisis que se realizaban.

La Lingüística de corpus hizo explícita la necesidad de entender en qué medida un corpus es una muestra representativa del discurso que se estudia y, por lo tanto, de la obligación de tener conciencia precisa de aquello que sí permite capturar y aquello que no, de qué generalizaciones sí pueden hacerse a partir de su análisis y cuáles no. Pese a que el investigador haya optado por una aproximación cualitativa, este punto no puede desatenderse (Biber 1993; Kennedy 1998; Kock 2001; McEnery y Wilson 1996 y Sinclair 1987).

Este primer momento de la discusión teórica quedó invisibilizado debido a que llegaría otro elemento central a provocar una segunda discusión, ya que el aumento en el interés en los textos o discursos provocó que se hablara acerca de que, para este tipo de enfoques, el objeto de estudio no aparecía representado en los primeros esfuerzos de construcción de corpus exhaustivos y el fuerte interés en los fenómenos de análisis de la lengua en uso llevó a los analistas (inicialmente a los de contenido de la Escuela francesa) a asumir que para poder trabajar con aquello que les interesaba no quedaba de otra que construir, como parte de su investigación, su propio corpus. Uno menos exhaustivo y delimitado, en este tipo de casos, por criterios que obedecían más al tipo de dispositivo discursivo (Foucault 1975), texto o tradición discursiva que interesaba.

Fue el surgimiento de esta metodología de estudio (ya sea del lenguaje verbal en uso u otro aspecto) la que mostró que no todo estaba resuelto con la construcción de corpus exhaustivos pues había investigaciones que los requerían de otro tipo.

Si quisiéramos extrapolarlo con ejemplos de análisis científicos, el corpus se parece un poco a la población muestra o población con la que se trabaja. Si soy un investigador científico y quiero analizar los efectos adversos del consumo de cierto colorante (sea por caso el amarillo número 6) en seres vivos, es posible que comience estudiando los efectos del colorante en ratones y/o peces. Ahora bien, si uno de estos grupos (población) presenta deformaciones y tumores luego de la exposición o del consumo del amarillo 6, pero no tengo una población control (es decir una no expuesta), puede ser que descubra tardíamente que las tumoraciones descritas se presentan también en la población de ratones o peces que no se expuso al colorante. Si de mis estudios con una población de ratones expuestos extraigo la conclusión de que el colorante generará siempre tumoraciones en los seres vivos, es evidente que estoy afirmando algo mucho más fuerte que lo que realmente pude observar en mi estudio. En pocas palabras: mi generalización es inadecuada.

En las ciencias (como la Lingüística), disciplinas e investigaciones en las que se trabaja con un corpus, este supone, al igual que la muestra o la población, un conjunto de fuentes

de datos o datos cerrados y parciales para estudiar un fenómeno que ocurre y es mucho más extenso que esa selección.

Todo análisis opera en un campo específico, todo estudio nos lleva a descripciones (y, en el mejor de los casos, a explicaciones), mas no debemos olvidar que las relaciones causales realmente son muy raras y para serlo deben cumplir con criterios específicos y deben estar condicionadas por determinados elementos. A veces creemos que el estudio concreto de cierto corpus (como manifestaciones parciales) nos permite analizar todos los usos del lenguaje o de algún otro mecanismo de comunicación, lo cual es imposible y, más aún, pensamos que la constitución del corpus de donde vamos a extraer los datos es cosa menor, pero no hay nada más falso.[3] El corpus lo es todo, se debe poner tanto cuidado en su constitución como el científico cuida el control de variables en el ambiente experimental para evitar que en la población existan otros factores que no fueron controlados y que pudieran estar influyendo o correlacionándose con lo que queremos observar.

Teresa Carbó (2001c: 22-23) ha recuperado parte de las discusiones que se desataron en este segundo momento a partir de los postulados de los análisis del corpus. La autora data una de esas profundas discusiones formuladas en 1964 durante el Coloquio Internacional de Sociología de la Literatura, donde justamente se discutió la validez o no del método empírico que supondría el acercamiento por medio de corpus.

Los dos procesos de los que hemos hablado contribuyeron al surgimiento de un cambio radical con respecto al método de la introspección que por mucho tiempo se utilizó, y mantiene, también, diferencias con respecto a los procedimientos experimentales. Todas las disciplinas que, por distintas razones, abordan una investigación que tiene por objeto de estudio el lenguaje en uso o fenómenos de la comunicación humana y semiosis social, requieren de la constitución de un corpus y exigen abordar este aspecto con rigor científico. Eso provocó el interés por abordar la ardua tarea de la definición del corpus de un modo más especializado y profundo poniendo en evidencia que hablar de corpus no suponía solo problemas de representatividad sino de posicionamiento y construcción metodológica.

Según Sinclair, uno de los grandes especialistas en el campo de los corpus modernos, el corpus es "A colection o pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair 1996: 7), mientras que el corpus lingüístico informatizado es "a corpus wich is encoded in a stand-

[3] También es posible encontrar las posturas que, en tanto que se afilian al extremo más radical de las posturas cualitativas, suponen que lo que les interesa no es comparar ni estudiar el fenómeno más que como manifestación específica, aislada, peculiar y única, pero incluso en estos posicionamientos hay cuestiones que abordar con respecto al corpus seleccionado.

ardized and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance" (Sinclair 1996: 7).

La relación entre el corpus y el objeto de estudio establece nexos diferentes a los que se instituyen entre una muestra o población en un análisis experimental, pues el objeto de estudio en las Ciencias Sociales se construye al mismo tiempo que el corpus, es el investigador quien, al ir definiendo los criterios para la constitución de ese conjunto parcial de datos, va construyendo el objeto de estudio y sus límites justo a partir de la problemática que establece el propio investigador, quienes los presenta como homogéneos en cierto modo (comparables, parecidos) y heterogéneos en otros (diferentes contrastables).

Hay que mencionar, además, que en algunos, por no decir muchos de los casos, en los que se trabaja con corpus (específicamente aquellos en donde se trata de un conjunto de textos que han sido reunidos y seleccionados a partir de los intereses y criterios del investigador) no hay que perder de vista que cada uno de los elementos que lo componen no ha sido creado o no ha surgido con el fin de que realicemos nuestro análisis, muy por el contrario, es resultado de necesidades e intereses comunicativos para actuar en ciertas esferas de la vida humana.[4]

Esta diferencia no es menor: el científico reúne en su población a un conjunto de sujetos o seres vivos y será él quien los exponga a ciertas condiciones o manipule variables concretas para obtener sus datos. A diferencia de esto, en los casos en los que se constituye un corpus con textos (orales o escritos), es justamente la manera en la que se construye lo que modifica cosas, pues el texto es el producto de un evento de la comunicación humana que el investigador no controló.

Así pues, el corpus es mucho más que solo el conjunto de textos o datos que se analizan en una investigación, ya que, como hemos visto, de él dependen elementos metodológicos fundamentales. El corpus forma también parte de la construcción de nuestro objeto de estudio. Uno que debe ser construido bajos criterios precisos, pertinentes y defendibes. Así que, primero que nada, si queremos hacer un corpus para nuestra investigación debemos darnos cuenta que no es tan sencillo ni poco importante como parece, de ahí que incluso definiciones más acabadas (p.e. Sinclair, Payne y Pérez, 1996) no alcanzan para visibilizar la complejidad del planteamiento de una investigación que trabaja con corpus.

De hecho, autores como Torruella y Llisterri (1999) definen al corpus como "un conjunto heterogéneo de muestras de lengua de cualquier tipo (orales, escritos, literarios, coloquiales, etc.) los cuales se toman como un modelo de un estado o nivel de la lengua predeterminado" (52) e incluso explicitan más su definición al aclarar que un subcorpus

[4] Entendemos *texto* como una unidad de la comunicación humana. No vamos a entrar en los muchos debates y recovecos de esto, si eso es de interés para el lector recomendamos consultar autores como Beaugrande y Dressler (1997); Bernárdez (1982); Gregorio De Mac y Rébola de Welti (1992) y Van Dijk (1978, 1980, 1999).

será "una selección estática de textos derivada de un corpus, normalmente más general y complejo, el cual está dividido en grupos de muestras textuales más específicas" (1999: 52).

Existen otros autores que destacan más bien el hecho de que un corpus no es cualquier colección de materiales, pues solo "cuando los materiales que contienen los datos han sido seleccionados y ordenados a partir de criterios que se desprenden de la disciplina que guía la investigación, hablamos ya de un corpus y esos criterios pueden ser de órdenes diversos pueden ser: a) externos o b) internos" (Torruella y Llisterri 1999: 52).

Por medio del escueto recorrido que hemos presentado al inicio de este subapartado y de las definiciones que hemos ido recuperando, interesa destacar la definición especializada del corpus, una que va mucho más allá de la simple reunión de piezas o el conjunto de unidades y que, en nuestra opinión, es la que debe tener presente el investigador al construir su corpus. Como podemos ver, en las definiciones que se han retomado el énfasis no está puesto en el hecho de que un corpus sea un conjunto de datos, sino en que ese conjunto ha sido construido con criterios que se desprenden de las disciplinas, objetivos y enfoques de la investigación y las interrogantes que esta plantea. Esto muestra que, en un término más especializado, el corpus es una parte estructural del objeto de investigación y de la metodología de análisis misma.

Es decir, identificamos tres momentos clave en el debate hacia la definición especializada del corpus y, lamentablemente, son muy pocos autores los que los destacan al ofrecer una acepción especializada. De ninguna manera el apartado anterior supone que se haya agotado la riqueza de las pocas, pero bastante complejas y necesarias contiendas con respecto a las definiciones y fundamentaciones del corpus, sin embargo, hemos intentado destacar al menos las aristas metodológicas y los momentos que han contribuido a la discusión y profundización con respecto al estado de la cuestión en la disputa de la definición especializada de corpus.

Con todo, el concepto especializado que nos compete termina por ser una correspondencia de varios cruces en el proceso metodológico de una investigación, tal y como lo muestran las siguientes reflexiones de Carbó:

> La estructura misma de indagación y escrutinio (lo que los franceses llaman la *grille* de análisis) adopta la forma de una red: una trama de puntos interrelacionados. Por ejemplo, de preguntas asociadas, o que postulan una cierta asociación entre elementos o conjuntos, de modo que, a medida que el trabajo prosigue, la ocurrencia (construcción) de ciertos fenómenos (por ejemplo, algunas respuestas preliminares) ocasionan efectos o reacomodos en otros puntos de la misma red, que es dinámica y en proceso siempre de "re-configuración" (Carbó, 2002: 22)

Es decir, el corpus es el resultado de un proceso previo en el que la investigación ya ha comenzado. Carbó traza paralelos entre el proceso de constitución del corpus y la fotografía,

sugiriendo que constituir un corpus es una actividad parecida a la del enfoque de un lente de *zoom*:

> La analogía con el *close-up* fotográfico conviene a la evocación de este momento metódico en el que la mirada (que lee) encuentra la promesa de una mayor densidad operacional (analítica) en ciertas áreas del material del acervo, a la manera como la ampliación de una imagen hace aparecer en ella texturas, irregularidades, rasgos, elementos, que no eran visibles desde una mayor distancia, aunque allí estuvieran antes, ocultos al ojo desnudo (Carbó, 2002: 23).

En cuanto a las dimensiones del corpus, la misma autora señala:

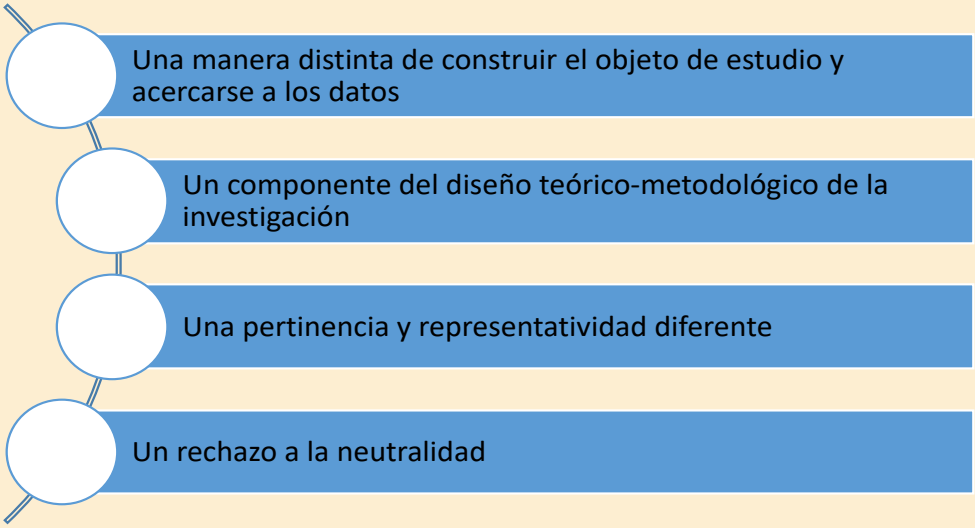
> El tamaño preciso del corpus, su extensión y alcance dentro de la economía energética y temporal de la investigación, o sea, el equilibrio siempre difícil entre ahorro y gasto, se relaciona de manera crucial con esta idea, quizás un poco rara, de un potencial de mayor densidad operatoria en ciertos segmentos del material (Carbó, 2002: 24).

El abandono del abordaje de estos elementos en las investigaciones puede tener muchas explicaciones y causas, sin embargo, es evidente que una de ellas es la complejidad de poder reportar en su justa dimensión la complejidad del corpus entendido de manera especializada, no obstante, al renunciar a tal complejidad, necesariamente condenamos al silencio una de las etapas más complejas de la investigación, de la que depende, en gran medida, la evaluación de la pertinencia y valía de lo encontrado y, al mismo tiempo, perpetuamos ese silencio que ningún bien le hacen a las investigaciones que trabajan con corpus.

Ahora bien, es importante decir y destacar que, pese a la riqueza de las fuentes y autores que hemos mencionado, observamos en la práctica de la investigación a partir de corpus una tendencia al abandono de los supuestos que subyacen al decidir esta metodología. Como otras autoras han destacado hay incluso una tendencia a ni siquiera mencionar los criterios metodológicos del corpus con el que se realiza la investigación. Los aprendizajes y los retos que se presentan en el camino justo de su definición se pierden o quedan relegados a un profundo silencio que tiene implicaciones en cuanto a la honestidad y los alcances de los resultados presentados.

El corpus supone una manera o metodología para reunir de una serie de datos, textos, unidades o materiales a la luz de una disciplina, y también entraña la construcción de un objeto de estudio y del diseño de preguntas de investigación o hipótesis específicas, por lo que su definición especializada despliega las dimensiones metodológico-epistemológicas mostradas en el Esquema 1:

Esquema 1. Dimensiones implicadas en las investigaciones metodológicas que trabajan con corpus



Para ahondar un poco más en las implicaciones de esta definición especializada del corpus, a partir de algunos aportes de distintos autores, abordaremos cada uno de los elementos mencionados en el Esquema 1 en los siguientes apartados.

1.3.1 El corpus: una manera distinta de acercarse a los datos y de construir objetos de estudio

Existen elementos de la definición especializada del corpus que están estrechamente vinculados con el tipo de enfoque y metodología que elige quien se dispone a trabajar con un corpus. Como ya vimos, la Lingüística, junto con muchas de las disciplinas de las Ciencias Sociales, por mucho tiempo trabajó con esquemas de introspección. Sin embargo, el vuelco, a partir de los años sesenta, hacia el estudio del uso (como reacción a la carga que se había dado al sistema) y el interés en fenómenos de dimensiones supraoracionales supuso la necesidad de poder trabajar con mecanismos que permitieran cambiar las aproximaciones introspectivas (que nunca se han abandonado del todo) por acercamientos más empíricos.

En otras ocasiones ya se ha señalado que el corpus es una forma específica de recuperar los datos:

> En la investigación social empírica, es posible establecer una distinción entre los métodos de extracción y los de evaluación, es decir, entre las formas de obtención de datos (ya sea en el laboratorio o mediante el trabajo de campo) y los procedimientos que han sido desarrollados para

> el análisis de los datos recogidos. Los procedimientos metodológicos para la recogida de datos organizan la observación, mientras que los métodos de evaluación regulan la transformación de los datos en información y la ulterior restricción de las oportunidades abiertas a la inferencia y a la interpretación" (Meyer en Wodack y Meyer 2003: 40-41).

Es decir, los investigadores que se decantan por implementar un método de análisis que trabaje con un corpus apuestan por una organización de los datos y una evaluación muy distinta a la de enfoques introspectivos o experimentales. La preferencia o el vuelco hacia esquemas empíricos no es resultado de un simple capricho ni solo de un desencanto, sino del tipo de retos y necesidades que supone el objeto que se construye como objeto de estudio. Por ejemplo, las investigaciones de análisis del discurso, por mencionar algunas, suponen justamente focalizar el uso del lenguaje por personas condicionadas por elementos contextuales que impactan en ese uso porque precisamente tal situación genera efectos que le interesa capturar y analizar. De hecho, Carbó (una de las autoras que más ha reflexionado con respecto al corpus) insiste en que la constitución de un corpus es siempre una configuración de los datos (véase 2001a, 2001b, 2001c).

La configuración de los datos hace referencia a la disposición, la manera de organizar, acomodar y componer un fenómeno o hecho. Es decir, supone el proceso de darle forma y propiedades a un fenómeno que se va constituyendo como un objeto de estudio a partir, justamente, de la selección del corpus en donde queremos observarlo. Si bien de manera apresurada podría pensarse que esto no aplica para quienes trabajan con corpus exhaustivos como los que se mencionaron antes, tal apreciación es falsa. Incluso en esos grandes esfuerzos de constituir corpus exhaustivos de largos periodos de tiempo o de extensas poblaciones etarias o geográficas, hay que estar conscientes de una cosa: esos corpus exhaustivos están conformados por campos de esferas de la comunicación muy específicos, por ejemplo, obras literarias, jurídicas y ya más recientemente periodísticas. La forma y las propiedades que tienen los resultados descritos en los trabajos que han recurrido a estos corpus exhaustivos están restringidos a la configuración que implica trabajar solo con estas fuentes y eso supone necesariamente reconocer que no dan cuenta del uso total del habla, sino del uso de habla configurado por el tipo de dispositivos, tradiciones y géneros que sí fueron incluidos en este tipo de corpus y solo en el grado de representatividad que tengan.

Quien recurre a estos corpus exhaustivos para sus análisis, acepta ya una configuración que será impuesta o heredada a su trabajo debido justamente a las decisiones metodológicas de reunión de datos que ya está presente en los criterios con los que se construyeron tales corpus y debe ser consciente de ellos. Mientras que, por otro lado, quien decide diseñar su propio corpus a partir de selecciones propias, lo que está haciendo es justamente construir una

configuración de datos y de corte metodológico en términos amplios, es decir está tomando decisiones que le darán forma y ciertas propiedades a sus datos. Es por ello que Fonte afirma que "la construcción de un corpus es una particular interpretación de la realidad" (2008: 66). A pesar de ello, y es este uno de los elementos más importantes, no se debe confundir el hecho de que sea una particular interpretación de la realidad o construcción del objeto de estudio con el hecho de que por ello sea una construcción aleatoria, sin justificaciones ni argumentaciones que muestren la pertinencia de las decisiones que se han tomado.

Una de las características de las investigaciones de quien decide trabajar con análisis de corpus es, por tanto, la decisión (que ha de ser consciente y clara) de mirar y acercarse a los datos de una manera específica, una manera que supone, de hecho, el inicio de la construcción misma tanto del objeto de análisis como de la metodología y soporte teórico de la investigación. De ahí que Strauss (1987) haya afirmado:

> En esta modalidad de procedimiento, la recogida de datos es un elemento que nunca se excluye por completo, y siempre surgen nuevas cuestiones que solo pueden abordarse si se obtienen nuevos datos o si se reexaminan los datos recogidos con anterioridad (1987: 56).

Si bien es posible, como hemos enfatizado, rastrear reflexiones teóricas del trabajo con corpus, nos parece que en ellas (salvo en Carbó, 2001ª, 2001b, 2001c, 2002, 2004 y 2007) no hay una reflexión explícita sobre los matices y tipos de corpus a partir más bien de las estrategias generales que se implementan en su diseño. Por ese motivo, nos gustaría añadir un aspecto más que está implícito en las reflexiones (las que ya hemos mencionado y muchas otras), y que supone un cambio diametral en los acercamientos con análisis de corpus. En la mayoría de estas aproximaciones (pero no en todas), las unidades (de distinta índole) que han de componer el corpus son unidades comunicativas que surgieron como enunciados en el sentido estricto de esta definición a manera de unidad de estudio pragmático (Curcó, 2021:14), es decir son piezas que fueron pensadas y ejecutadas para fines comunicativos específicos con los que se operó en las esferas de la comunicación social. Ninguna de ellas ha sido creada para ser el cuerpo de análisis de una investigación, quizá eso sea lo que nos hechiza para tratar de componer, por medio de una reunión metódica de estas piezas, un conjunto que embonará y basará su pertinencia a partir de la destreza con la que el investigador haga evidente lo que la reunión regida por criterios, objetivos y posturas teóricas permite construir como un lente especializado para mirar ciertos fenómenos o efectos.

La pertinencia de la que hablamos no está presente en estas piezas en tanto que son usos del lenguaje con sus propios contextos e intenciones comunicativas: construir un corpus es

mucho más que reunir elementos que arrojen datos observables, significa construir también la pertinencia y el enlace, *a posteriori* (pues tal pertinencia no forma parte de las condiciones de creación de estas unidades) de tales manifestaciones que quedan representadas a partir del trabajo de diseño de corpus como objeto de estudio para poder asir lo inasible. Evidentemente, construir con destreza esa conjunción de piezas del lenguaje en uso supone un reto que implica mostrar y argumentar tal selección como una que resulta clara y pertinente para observar el terreno en donde se construye, disputa o gesta un fenómeno de la comunicación, el uso del lenguaje o la semiosis social.

En otros casos se crean o se aplican tareas o instrumentos para obtener un conjunto de materiales: enunciados, entrevistas, historias de vida, etc. En estas situaciones existe una pequeña variación, es cierto, debido a que el investigador diseña y adapta una serie de instrumentos y tareas para obtener unidades comunicativas que sí surgen como resultado de fines comunicativos relacionados con la investigación misma. Aunque si lo pensamos con detenimiento, en la mayoría de los trabajos de este corte el investigador diseña una serie de instrumentos y tareas que suponen, para los sujetos que participan, la simulación de una situación comunicativa que intenta, de algún modo, emular cierta naturalidad para que el sujeto no piense que está diciendo cosas para el corpus o la reunión de datos de una investigación específica, sino para que considere que está haciendo algo comunicativo.

Hemos explicado ya que el uso del corpus en una investigación supone mirar, observar de un modo distinto el fenómeno de análisis, pero no se ha abordado nada acerca del posicionamiento epistemológico ante la construcción misma del objeto de estudio, por lo que este aspecto se abordará con mayor énfasis enseguida.

1.3.2 El corpus es una construcción metodológica (teórico-metodológica) de la investigación misma

El hecho de que al hacer análisis de corpus se configuren los datos de un cierto modo y con una postura particular implica que el investigador considera que el objeto de estudio requiere justamente una mirada afín. Es más, se requiere la construcción de un modo de observar que construya los posicionamientos y planteamientos epistemológicos y metodológicos a la par que se fabrica la ventana que le permitirá observar su objeto de estudio.

Es por ello que, en este mismo texto, la autora insiste en que ha de reportarse tanto el camino de construcción de las preguntas de investigación como el camino andado para la construcción del corpus, por el simple hecho de que no estamos hablando de dos caminos ni de dos momentos distintos de la investigación. Contrario a lo que muchas veces se piensa, el análisis del fenómeno ha comenzado desde el momento mismo en que se ha empezado a pensar en el corpus y viceversa.

La importancia de que se comprenda que en los análisis de corpus no podemos separar el subdiseño del posicionamiento metodológico mismo de todo el proceso de construcción de la investigación y del análisis ha sido enfatizada de otras maneras:

> El foco de estos trabajos así orientados se centra en el tema del corpus, porque en mi propia experiencia de investigación, y en otras, la construcción de ese objeto ineludible se manifestó, desde las etapas metodológicas que pueden considerarse previas al análisis en sentido estricto (o más convencionalmente así reconocido), como una clave, una llave de sorprendente eficacia para la concepción y trazo de un recorrido cognoscitivo que aumenta su capacidad de comprensión y respuesta a las propias preguntas de cada estudio solo a medida que avanza. Es ése un camino que se va haciendo al andar (como dice Machado por voz de Serrat, según el acervo musical de mi generación y entorno) (Carbó 2016: 13).

Huffschmid, en su excelente trabajo "De los cuerpos al corpus", habla de esto mismo cuando advierte que en la construcción de un corpus:

> si somos afortunados, logramos aprovecharlo como portador de sentido(s) y placeres. [...] En el análisis de lo dicho, es también el corpus el que carga y ancla el sentido, una especie de tierra. Sin la tierra (empírica) no hay vuelo (teórico), y viceversa. Pero este cuerpo no nos está dado de antemano, no existe como tampoco existe el discurso, sino que lo construimos. Con cuidado y delicadeza hacemos de lo mucho que nos rodea un algo que nos permite mirar de cerca, descubrir y detectar las marcas del habla como hacer significante (2016: 85).

El aspecto que queremos enfatizar en este subapartado no termina de estar claro si no explicitamos que la definición especializada de corpus no solo implica la construcción misma del fenómeno como un objeto de estudio, sino un posicionamiento:

> Miramos y enfocamos, escogemos ángulo y encuadre, cambiamos el lente, nos acercamos con lupa o nos distanciamos para lograr una vista panorámica. [...] Así que nuestro corpus será un mundo propio y ficticio, en estrecha relación con el mundo exterior pero formado sobre la base de los intereses y saberes de quien lo construye (Huffschmid, 2016: 87).

Ese posicionamiento se irá afinando al tiempo que diseñamos el corpus, como bien advierte Eva Salgado (2016) en su trabajo "Un corpus discursivo para entender el presidencialismo en México" cuando afirma que "es imposible construir los criterios de constitución del corpus sin los objetivos de la investigación, y sin conocer el acervo, luego de tener el acervo y

conocerlo entonces sí viene el salto al corpus, [...] tomar una decisión drástica respecto a cuáles partes serán analizadas" (2016: 156).

La teoría fundamentada ha marcado claramente que, en ciertos procedimientos, que se han validado como métodos de estudio, la fase de la recolección de datos no tiene que ser una fase cerrada y concluida al iniciar el análisis, sino más bien es parte del proceso mismo de la construcción metodológica (Glaser y Strauss, 1967). Las ventajas y desventajas de tal comprensión de un hacer-conocer, así como el cuidado que hay que tener con el establecimiento de conclusiones han sido magistralmente discutidos en Baker (2006).

Evidentemente, los planteamientos y las reflexiones que se han suscitado acerca del trabajo con corpus conducen necesariamente a que los principios de representatividad y pertinencia cambien en este tipo de acercamientos. Eso no quiere decir que no exista, entre la construcción de los planteamientos metodológicos de la investigación y la constitución misma de la forma de mirar el fenómeno a observar y el corpus en el que miraremos, un compromiso y una apuesta epistemológica con la que el corpus también guarda una relación estrecha.

1.3.3 Una pertinencia y representatividad diferentes

¿Qué es aquello que el investigador trata de capturar en la constitución del corpus al mismo tiempo que construye su proyecto? No hay nadie que lo haya dicho con tanta precisión:

> [...] un deseo de armonía y claridad; luego entonces también de precisión, finura, amplitud y detalle [...] no nos exime de la obligación de reflexionar sobre el proceso por medio del cual, luchando contra el sinsentido y la barbarie, el análisis del discurso construye sus datos [...] y con base en ellos sus aseveraciones, sean estas del alcance que sean [...] la confianza que se pueda tener en la capacidad del corpus para exhibir rasgos significativos con respecto al asunto que se analiza [...] que sean significativamente a los de la totalidad mayor [...] la avaricia o la largueza con que ese fragmento de mundo ha sido recortado (Carbó 2001a: 20 y 21).

La relación de pertinencia entre el fenómeno que se quiere observar con los posicionamientos con que se ha construido el objeto de estudio, las preguntas que nos planteamos contestar, aquello de lo que vamos a poder dar cuenta y lo que se nos ha de escapar debido a que no todo puede ser capturado con el mismo tipo de red y a que no todo ha de comportarse como lo ha hecho la muestra que hemos construido para nuestra investigación explica la importancia de que junto con la presentación del corpus se explique su organización, pertinencia y representatividad.

Es vital que en la construcción del corpus, el investigador no pierda de vista el anhelo de tal equilibrio entre la pertinencia y la representatividad que esa investigación exige y que

nunca olvide reportar las maneras, procedimientos y métodos que ha implementado para no perder de vista este horizonte, ni de reportar críticamente qué tanto nos hemos podido acercar a ese anhelo y en qué modos no lo hemos logrado junto con las honestas implicaciones de estas limitaciones.

El corpus supone la construcción (arriesgada, sí) de una reproducción que aspira a la máxima fidelidad posible de las características del fenómeno de estudio que logre capturar al menos lo más importante y representativo de los elementos que constituyen la realidad que nosotros hemos transformado en objeto de estudio, pero como bien advierten los autores de corpus textuales y orales:

> [...] solo una codificación, ordenación y organización de estos datos en la proporción adecuada pueden salvarlo de un naufragio en un mar inmenso de información [...] las pautas para obtener un corpus suficientemente organizado y representativo de la realidad que quiera reflejar [...] Svartvik (1992) señaló que la Lingüística basada en los corpus hacía posible nuevas aproximaciones a viejos problemas [...] poner en el terreno de las afirmaciones ideas que antes solo eran conjeturas o especulaciones (Torruella y Llisterri 1999:1).

La relación y el andamiaje del que hace mención Teresa Carbó (2007) entre una teoría y los datos que la manifiestan tiene que poder ser observada en el corpus. Si bien esta no es la realidad, sí es un modelo de ella, un modelo de lo que el investigador ha construido como su propio objeto de estudio.

1.3.4 El rechazo a la neutralidad

Del mismo modo en que hemos precisado que es imposible disociar la construcción del corpus de la constitución de la investigación, su diseño no se encuentra desligado o desvinculado de quien ha construido el objeto de estudio y el corpus mismo, por lo que es necesario explicitar los "valores y la toma de posición ideológica y política que acompañan la investigación, inconfesas e inconscientes muchas de las veces" (Carbó 2016: 593). Ahora que hemos enfatizado que la apuesta por la construcción de un cuerpo y de un objeto de estudio que quede capturado en él supone posturas epistemológicas con respecto a cómo acercarse, cómo observar y cómo construir el análisis mismo, podremos comprender por qué Carbó insiste:

> Estoy convencida de que el asunto de la evidencia empírica, de su conversión en datos válidos, plausibles y elocuentes, y de su organización en un corpus de análisis, conlleva temas complejos de teoría y andamiaje conceptual, mucho más que de metodología como repertorio preestablecido de procedimientos o maneras técnicas de hacer (2016: 14).

Cuando un investigador toma decisiones teóricas y de andamiaje conceptual no puede hacerlo con base en nada, ni en una neutralidad exigida y mal entendida alegando (y confundiendo) que es lo mismo la neutralidad que la objetividad. Lo que es peor, muchas veces la malentendida neutralidad se vuelve el pretexto para más bien encubrir y silenciar una postura (siempre hay una, incluso considerarse neutral es ya tomar una postura) en lugar de la sinceridad de declararla. Sobre la neutralidad, Coronado (2016) afirma:

> En el proceso de reflexión sobre mis motivos descubrí con gran alegría que en el fondo de este interés estaba en juego la comprensión de mi propia identidad social como mexicana, una identidad ambigua aparentemente distante de la de los pueblos indios contemporáneos [...] Podía elegir entre muchos eventos, producciones discursivas orales y escritas, manifestaciones visuales, comportamientos individuales o colectivos, recuerdos, entrevistas, prácticas repetitivas y eventos únicos. Mi avidez contradecía la sabiduría popular de mi cultura y en lugar de escuchar el dicho *el que mucho abarca poco aprieta* decidí seguir mi convicción, ahora fundada en las ciencias de la complejidad, de que es imposible entender las partes sin mirar el conjunto y simultáneamente de que solo mirando las partes y sus conexiones podemos tratar de aproximarnos al todo (36-39).

Ya desde el 2002, Teresa Carbó había precisado algunos de los puntos más importantes que develan la imposibilidad de la neutralidad, por ejemplo, el de la relación del tema con el investigador. Sobre el asunto de la objetividad, señala:

> esta intimidad, digo, no neutraliza nunca, ni debe llegar a hacerlo, la fuerza vital y movilizadora que una disposición básica de asombro (intriga, sorpresa), una curiosidad pertinaz y un deseo profundo de entender imprimen a la acción investigativa y a su creatividad metodológica y operacional. Sin embargo, tan provechosa cercanía cognoscitiva representa el final, o casi, de la historia de una investigación. En la vida real, el proceso comienza de una manera bastante menos precisa y más insidiosa. (Carbó 2002: 17)

Si toda investigación se construye desde una postura y teoría específica, esto supone que todo estudio supone una forma de mirar el fenómeno, una forma de construirlo y plantearlo, una que se encuentra determinada, también, justamente por el lugar desde donde esta acción se realiza. El corpus es, necesariamente, el resultado también de ese lugar en donde estamos posicionados para plantear tal investigación.[5]

[5] Existen asuntos muy interesantes con respecto a la neutralidad que no abordamos por la naturaleza de este trabajo, sin embargo, recomendamos ampliamente la revisión del exquisito trabajo "La falsa neutralidad de la neutralidad ideológica" del gran maestro Adolfo Sánchez Vázquez (1984).

1.4 Recuento general

Trabajar con corpus suele ser visto, desde una óptica lingüística como un problema de cantidad o, dicho de mejor modo, como un problema de representatividad. Pese a que es evidente que esta es una cuestión fundamental en la constitución de un corpus, supone una postura errónea limitarlo a eso.

Los fenómenos que se quieren estudiar, muchas veces, no solo son complejos, sino altamente frecuentes y ocurren en la realidad de modo poco ordenado y organizado. La metodología de análisis de corpus es una propuesta con dimensiones metodológicas que ha sido postulada, defendida y bien practicada (en estudios serios y científicos) con el objetivo de hacer observable y posible de estudiar, de manera sólida, rigurosa y científica, bajo objetivos concretos, un objeto de estudio que suele aparecer como inconmensurable.

Las características que debe tener un corpus variarán mucho dependiendo del enfoque o marco teórico, del objeto de estudio y, por supuesto, de los objetivos concretos de la investigación. Lo que puede aparecer como un corpus no pertinente para ciertos fines y objetivos, puede serlo para otros. De ahí que antes hayamos insistido en la relación que en una metodología de estudio de corpus se establece entre el fenómeno, el objeto de estudio, los objetivos y el mismo corpus.

En el caso de este tipo de investigaciones lo que podemos discutir, refutar y debatir es si la metodología de análisis de corpus es pertinente o adecuada a partir del fenómeno de estudio y los objetivos que el investigador ha definido y diseñado en una investigación concreta, así como la pertinencia, congruencia, consistencia y relevancia del corpus utilizado con respecto a los objetivos perseguidos, los resultados encontrados y las interpretaciones que se le dan a esos datos.

Siempre que sea necesario estudiar un fenómeno que se presenta en una cantidad amplia de manifestaciones, cuando los datos de donde podemos obtener la información que necesitamos se encuentran capturados en materialidades textuales que no han sido creadas para nuestra investigación o bien cuando necesitamos obtener un corpus con una población que nos permita encontrar datos en una muestra parcial, se hará necesario este tipo de posicionamiento como decisión consciente que puede obedecer a cuestiones de tiempo, de operatividad real, de posibilidad de acceso a los datos o de postura epistemológica. Esto que puede quedar muy claro, supone, cuando se hace con seriedad, un conjunto de retos importantísimos que serán el objeto de gran parte de este libro.

El primero de ellos es tener definidos con claridad y bien acotados tanto el objetivo general como los objetivos particulares, las preguntas de investigación y el objeto de estudio. Sin este conjunto de elementos no se puede pensar en el corpus y viceversa. Pese a lo que muchas veces se piensa, lo importante no es que, orillados por las prisas, redactemos esos

elementos de manera clara, sino que los tengamos definidos y sustentemos explícitamente por qué los encontramos plausibles, lógicos y coherentes.

Sólo cuando ya hay precisión en estos aspectos se puede contestar la pregunta que da título a este apartado y que no es menor. Cuando ya se tienen determinados esos elementos, hay que reflexionar ampliamente con respecto a estas preguntas: ¿De qué otro modo se podría abordar el mismo objeto de estudio (que no sea por medio de análisis de corpus)? ¿Qué aportarían esos otros modos al análisis que no se lograría si solo usamos análisis de corpus? ¿Qué no se puede hacer con esos otros análisis que sí nos permite el análisis de corpus? ¿Esto que se pretende investigar/estudiar/analizar solo puede hacerse desde esta forma?

¡Claro! No son preguntas que se contesten rápido; de hecho, en realidad lo que hacen es disparar un proceso más consciente y fino para poder entender lo que tiene que considerar al constituir o construir[6] el corpus. Y, por supuesto, eso debe formar parte de la explicación explícita de la metodología y de la defensa de nuestro trabajo.

Finalmente, hay que advertir que, cuando se afirma que "el corpus es lo de menos" o "en realidad el corpus no importa tanto", lo que se intenta decir con estas frases es que el investigador está interesado en un fenómeno en el que, sus hipótesis y supuestos, le hacen pensar que tal fenómeno (sea o no lingüístico) debe presentarse o comportarse casi de manera similar, con las mismas funciones y con el mismo grado de importancia sin que sea relevante en dónde sea observado, pero de ninguna manera se está afirmando que no sea importante el corpus en términos de diseño de investigación, simplemente se está planteando un trabajo desde otro posicionamiento metodológico. Como bien apuntaron Benveniste (1997) y Coseriu (1955) desde hace muchos años, las direccionalidades en el estudio de la lengua varían dependiendo de la unidad y el nivel del fenómeno que nos interesa.

Evidentemente, cuando queremos hacer un estudio de un fenómeno lingüístico muy preciso, supongamos, del marcador del discurso, entonces, estamos aún en un momento de falta de definición de los objetivos y alcances concretos de una investigación. Equivale a decir "quiero construir una casa", pero seguir ignorando una serie de criterios operativos que nos permitan conocer qué tipo de casa, con qué materiales, etc.

Todo lo que se ha dicho en este capítulo sirve para llegar a la conclusión de que no hay una receta ni un método infalible para construir un corpus, cuando se le comprende en su definición especializada, justamente los análisis de corpus suponen el reto (que el investigador ha de superar, no siempre sin tener varios errores y fallos) de construir un puente, un camino que puede que incluso no exista y que, por lo tanto, no supone la aplicación de

[6] En este trabajo se parte de que sí existen diferencias entre ambos procedimientos, como se abordará más adelante.

una serie de pasos ya conocidos. Es posible incluso que un mismo investigador tenga que construir un camino distinto para la construcción de corpus distintos de diversas investigaciones. ¿Por qué? Porque el fenómeno de estudio, la naturaleza de este es distinto y, por lo tanto, no podemos esperar que funcione el camino que ya ha recorrido y construido para otros casos.

Una de las maneras que ayuda para mantener el ánimo y superar estos retos, que acompaña en este acertijo que solo ha de poder resolver quien se ha propuesto tal prueba es acercarse y conocer las muchas maneras en que otros han resuelto sus propios retos en su esfuerzo por presentar el cuerpo más adecuado para capturar aquello que su investigación reclama.

Un número de la revista *Estudios de Lingüística Aplicada* (número 46, publicado en 2007) está dedicado por completo al tema del corpus. El lector encontrará en él una excelente introducción escrita por Teresa Carbó y Elin Emilsson. En "La elocuencia de los cuerpos", Teresa Carbó insiste en la importancia de la obra negra que supone el diseño del camino para llegar a un corpus. A lo largo de los artículos allí reunidos, el lector encontrará una discusión explícita y abierta sobre la manera en la que diversas investigadoras hicieron este camino. Se encuentra allí un cúmulo de ejemplos que muestran cómo se han enfrentado una variedad de retos para construir corpus multimodales, corpus de textos periodísticos, de discursos políticos, corpus para acercarse al cambio lingüístico en lenguas indígenas, entre otros. Sugerimos ampliamente al lector que recurra a este volumen.

Para aumentar el repertorio, en cada capítulo de este libro se ofrece una bibliografía final con trabajos en los que el lector podrá conocer acerca de los procedimientos que otros investigadores han reportado para la constitución de sus corpus. A falta de recetas no tenemos más que asirnos a las experiencias de otros y otras, no esperando encontrar en ellas nuestras respuestas, sino esperando encontrar una retroalimentación estimulante que nos impulse a seguir encontrando nuestro propio camino.

1.5 Reflexiones recomendadas

De acuerdo con la información reunida en el recuento de este capítulo y a partir de las discusiones y reflexiones en nuestra práctica con respecto al diseño de los corpus en distintas investigaciones, consideramos que hay algunas preguntas que disparan y ayudan en el tránsito hacia su diseño:

¿Cuál es el objetivo general o la pregunta central que aborda mi investigación?

¿Qué tipo de manifestaciones del uso del lenguaje me interesan y cómo se relacionan con el objeto de estudio de mi trabajo?

¿De qué otro modo se podría abordar el mismo objeto de estudio (que no sea por medio de análisis de corpus)?

¿Qué elementos aportarían esos otros modos al análisis que no se pueden obtener si lo hacemos con análisis de corpus?

¿Qué no nos permiten destacar esos otros análisis que sí nos permite el análisis de corpus?

¿Esto que quiero investigar/estudiar/analizar solo puede hacerse por medio de un análisis de corpus? ¿Por qué?

Es indispensable que las reflexiones y posibles respuestas a estas cuestiones se aborden en un espacio donde podamos discutirlas con otros colegas ye especialistas.

1.6 Para leer más acerca de la definición específica del corpus

Biber. D. 1993. Using register-diversified corpora for general language studies, *Computational Linguistics,* 19/2, 219-243.

Briz Gómez, A. y M. Albelda Marco. 2009. Estado actual de los corpus de lengua española hablada y escrita, *El español en el mundo*, Instituto Cervantes, (ed.)165-225. Madrid: Instituto Cervantes.

Carbó, T. 2001a. La constitución del corpus en análisis del discurso. *Escritos. Revista del Centro de Ciencias del Lenguaje* 23: 17-47.

—. 2001b. Tocar el lenguaje con la mano. Experiencias de método. *Revista Latinoamericana de Estudios del Discurso* 1(1): 43-67.

—. 2001c. El cuerpo herido o la constitución del corpus en análisis de discurso. *Escritos* 23, 17-47.

—. 2002. Investigador y objeto. Una extraña/da intimidad. *Iztapalapa* 53, 15-32.

—. 2004. Protocolos de investigación en análisis de discurso y consolidación del campo disciplinario. *Discurso, teoría y análisis* 26, 121-30.

Carbó, T. (ed.) 2007. Corpora, conceptos y métodos en análisis de discurso. *ela. Estudios de Lingüística Aplicada,* 46 monográfico. México: unam, Centro de Enseñanza de Lenguas Extranjeras.

— 2016. Introducción. La elocuencia de los cuerpos. *Estudios de Lingüística Aplicada, 0*(46). doi:https://doi.org/10.22201/enallt.01852647p.2007.46.576

Carbó, T. y E. Salgado. 2013. El itinerario de un corpus multimodal para escrutar el desempeño presidencial reciente en México (2006-2012), *Estudios del discurso en América Latina. Homenaje a Anamaría Harvey*, 527-550. Bogotá: aled.

Gutiérrez, S., L. Guzmán y S. Sefchovich. 1988. "Discurso y Sociedad" Capítulo IX, *Hacia una metodología de la reconstrucción*. México: Porrúa-unam.

Harris, Z. 1952. Discourse Analysis. *Language* 28 (1): 1-30.

Kennedy, G. D. 1998. *An Introduction to Corpus Linguistics*. Londres: Longman.

Kock, J. de. 2001. *Lingüística con corpus. Catorce aplicaciones sobre el español*. Salamanca: Universidad de Salamanca.

Kress, G. y T. Van Leeuwen. 2001. *Multimodal discourse. The modes and media of contemporary communication*. Londres y Nueva York: Bloomsbury Academic.

—. 2003 [1998]. Front Pages: (The Critical) Analysis of Newspaper Layout, *Approaches to media discourse*, A. Bell y P. Garret (eds.), 186-219. Reino Unido: Blackwell Publishers.

McEnery, T. y A. Wilson. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 2.ª edición de 2001.

Schifrin, D., D. Tanen, y H. Hamilton (eds.). 2001. *The Handbook of Discourse Analysis*. Malden: Blackwell.

Scollon, R. 2003 [2001]. Acción y texto: para una comprensión conjunta del lugar del texto en la (inter)acción social, el análisis mediato del discurso y el problema de la acción social, *Métodos de análisis crítico del discurso*. R. Wodak y M. Meyer (eds.), 205-266. Barcelona: Gedisa.

Shiro, M. 2012. El método tampoco viene del aire. *Revista Latinoamericana de Estudios del Discurso* 12(2): 3-6.

Sinclair, J. M. (ed.) 1987. *Looking up: an Account of the COBUILD Project in Lexical Computing.* Londres-Glasgow: Collins.

Sinclair, J. M. 1996. Preliminary Recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P.

Sinclair, J. M., Payne, J. y Pérez, Ch. (eds.). 1996. Corpus to Corpus: A Study of Translation Equivalence. *International Journal of Lexicography* 9 (3): Autumn.

Torruella, J. y Llisterri, J. 1999. "Diseño de corpus textuales y orales" En Blecua, J. M., Clavería, G., Sánchez, G. y Torruella, J. (eds). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e informática, Departamento de Filología Española, Universidad Autónoma de Barcelona: Editorial Milenio.

Van Leeuwen, T. 2005. *Introducing Social Semiotics*. Londres: Routledge.

Verón, E. 1971. Ideología y comunicación de masas: la semantización de la violencia política, *Lenguaje y comunicación* social, VVAA, 91-133. Buenos Aires: Ediciones Nueva Visión.

—. 1987 [1981]. *Construir el acontecimiento. (Los medios de comunicación masiva y el accidente en la central nuclear de Three Mile Island*). Barcelona: Gedisa.

Wenger, E. 1998. *Communities of practice*. Cambridge: Cambridge University Press.

Williamson, R. 2007. El diseño de un corpus multimodal. *ELA. Estudios de Lingüística Aplicada* 46: 207-31.

2. Tipos de corpus

2.1 Introducción

Una vez que hemos problematizado los planteamientos metodológicos y epistemológicos del corpus, podemos reflexionar acerca de sus tipos y sus diversas clasificaciones. Conocerlos ayuda mucho para las reflexiones y el trabajo que supone el diseño de nuestro propio corpus de investigación. En primer lugar, porque tener un panorama general de las distintas clasificaciones y tipos de corpus ofrece al investigador el conocimiento de una gama muy amplia de posibilidades, lo que contribuye a que tenga más elementos para poder establecer criterios y diseñar el corpus que su investigación necesita. Además, como se ha advertido en el capítulo anterior, no se puede pensar en la selección del corpus como un proceso aislado de la construcción y el diseño de toda la investigación misma, por lo que es importante reflexionar acerca de los distintos tipos de corpus pensando siempre en que no todos son igual de pertinentes para distintos objetivos ni todos funcionan para todas las investigaciones. Tanto conocer como analizar los tipos de corpus supone, al mismo tiempo, tener mayor consciencia y solidez en la propuesta metodológica de la investigación.

En este capítulo se hablará sobre los tipos de corpus que algunos autores han propuesto y las potencialidades y limitaciones que cada uno de ellos tienen. Pondremos mucho énfasis en destacar que, a partir de los distintos tipos de clasificaciones, existen corpus que son más o menos pertinentes, idóneos o adecuados para una investigación, ya que un tipo de corpus siempre impone limitaciones, pero, al mismo tiempo, supone aciertos y cierta pertinencia en relación con la investigación que se ha postulado. En ese sentido, si uno parte de la pregunta ¿cuál de estos tipos de corpus supone las restricciones más adecuadas para nuestros objetivos y la pertinencia más idónea para nuestros objetivos y preguntas?, estamos partiendo de un punto más fructífero. Desde nuestro punto de vista, se ha puesto mucha atención en las clasificaciones, aunque la mayoría de los autores al hablar de ellas lo hacen a partir de un solo criterio y no se reflexiona ni discute el hecho de que existen distintos criterios y propuestas para tal clasificación. El segundo aspecto que aparece un poco descuidado en el estado de la cuestión es que al hablar sobre los tipos de corpus no se explicita cómo los distintos tipos de corpus suponen correlaciones (más o menos pertinentes) dependiendo de los objetivos de una investigación.

El hecho de que los corpus puedan ser utilizados en investigaciones de distintas disciplinas y para muy diversos objetivos hace que en realidad existan muchas y variadas clasificaciones de los tipos de corpus que, incluso, parten de taxonomías distintas. A eso se suma que, en distintas fuentes, se utilizan nombres o conceptos distintos, por lo que, hay que

decirlo, es difícil el acceso a la literatura con respecto a este tema, pues se corre el riesgo de acabar más confundido que antes de leer este tipo de trabajos. Para evitar que al acercarnos a la literatura especializada todo se nos presente como un conjunto desordenado de propuestas, hablaremos, en primer lugar, de los diferentes tipos de criterios que se han usado (en distintas propuestas) para establecer la tipología de corpus y, al final, se ofrecerá, junto con una síntesis que proponemos, una reflexión sobre algunos aspectos que nos ayuden a usar esta información con mayor claridad para nuestro objetivo.

Frente a este panorama, nos parece necesaria una sistematización breve y panorámica que ponga el acento en las diferencias entre los distintos criterios de categorización, lo que, además supone el beneficio de ofrecer al lector una guía para identificar los distintos conceptos o términos que encontrará en distintas tradiciones y enfoques. Evidentemente, el lector atento irá descubriendo que hay tipos de corpus que le resultan más atractivos o más fructíferos para su investigación que otros y, por lo tanto, hacia el cierre de este bloque, se podrá retomar la reflexión para entender cómo lo expuesto nos ayuda para nuestro proceso de diseño de un corpus.

2.2 Tipos de clasificaciones y corpus

Como ya advertimos, los tipos de corpus que se han postulado parten de supuestos o criterios clasificatorios distintos que no siempre son excluyentes, por lo que revisaremos los más comunes. Es importante advertir que a lo largo de todo este apartado se presenta una síntesis y estructuración de muchas fuentes, todas ellas reportadas al final de este capítulo.[7]

2.2.1 Por el tipo de soporte en el que se presenta el contenido o material del corpus

Existen clasificaciones de tipos de corpus que toman como criterio clasificatorio los tipos de soporte de los datos, evidentemente estas parten de las profundas diferencias entre reunir materiales orales, escritos y de otra naturaleza. Es muy distinto y supone retos diversos seleccionar un corpus de pinturas, fotografías, videos o volantes de propaganda política que uno de relatos orales. A eso se suma el hecho de que no es lo mismo analizar un corpus oral que uno escrito, y, por supuesto, es mucho más complejo y supone mayores retos trabajar con corpus multimodales o que recuperan y reúnen materiales diversos en donde puede haber lenguaje oral o escrito junto con otros elementos como imágenes, tipografías de letras, etc. [8] Con respecto a este criterio, se postulan los siguientes tipos de corpus:

[7] Hay que aclarar que no incorporamos la propuesta de corpus multilingües de Sinclaire, Payne y Pérez (1996) debido a que, de acuerdo con la definición que el autor da a este tipo de corpus nos parece que habla más bien de un acervo y no de un tipo de corpus. Lo mismo ha pasado con lo que el autor llama corpus oportunistas.

[8] Multimodalidad en el sentido en que lo postulan Kress y Van Leeuwen (2001): cuando el objeto de estudio no se comprende ni analiza solo en su dimensión lingüística verbal (como unidad del lenguaje), sino como parte

a) **Corpus orales:** Son aquellos en donde el conjunto de muestras que se seleccionan para el análisis está constituido por audios en los que lo que se privilegia es mantener y respetar el formato y los fenómenos de oralidad con todos los tipos de fenómenos que le caracterizan (tono, acento, pausas, silencio, velocidad, etc.). Aunque puede haber variaciones en la decisión, consciente, de qué elementos de todos los que están presentes en la comunicación oral (incluyendo los paralingüísticos) se necesitan capturar y cuáles no. Esto no quiere decir que un corpus oral no pueda ser transcrito, sino justamente que un corpus oral que se transcribe respetará e indicará con la mayor fidelidad posible los fenómenos orales por lo que, en caso de que se les resguarde en una plataforma de transcripción especializada y no solo de audio, será imprescindible establecer una serie de criterios de qué fenómenos fonético fonológicos y prosódicos queremos conservar, cómo se van a indicar en la transcripción e incluso con qué tipo de programas se van a archivar o procesar ciertos fenómenos. Además, supone siempre que la reunión de los materiales requiere del uso de programas o de aparatos que permiten grabar y conservar audio. Generalmente, este tipo de corpus se presentan en investigaciones en las que, evidentemente, interesan fenómenos que están presentes solo en los registros orales o bien fenómenos atados a niveles como el fonético fonológico o la conversación natural.

b) **Corpus escritos:** Son aquellos en donde el conjunto de la muestra que se selecciona proviene o es una manifestación de tradiciones escritas (que se producen y consumen en su forma escrita o como oralidad secundaria) y en los que, por tanto, se conserva esta característica al juntar las manifestaciones en este tipo de plataforma (generalmente programas de procesamiento de texto).[9] También puede ser que los corpus escritos sean resultado de una recuperación de datos en grabación oral, pero que, por el tipo de investigación que estamos realizando y debido al objeto de estudio, se requiera de una transcripción que solo refleje el lenguaje verbal escrito y los fenómenos que están en el uso del lenguaje verbal en otros niveles, por lo que no requerimos que el corpus respete y mantenga los fenómenos de oralidad. En estos casos los discursos o manifestaciones que constituyen el corpus tan solo se transcriben y no se indican ni marcan tantos elementos como en los corpus orales; sin embargo, el investigador debe reportar por qué decidió no dar cuenta en sus transcripciones de estos elementos y justificarlo claramente, así como informar los criterios de la transcripción, si es

de otra serie de expansiones semánticas como el formato, los colores, el tipo de letra, los gestos, etc. que también (y en algunos casos aún más que el uso del lenguaje verbal) construyen significación.

[9] La oralidad secundaria hace referencia a una manifestación en la que se ha utilizado la escritura para fingir cierto grado de oralidad (Ong 1997).

que se marcaron ciertos fenómenos en la transcripción y con qué claves o símbolos o si, por otro lado, tan solo se realizó una transcripción ortográfica o simplemente se tomaron las versiones ya en formato escrito de ciertas fuentes. Un ejemplo de este tipo de corpus es el realizado por el Instituto de Investigaciones Jurídicas con los discursos de toma de posesión presidencial.[10]

c) **Corpus multimodales:** Son aquellos en los que el conjunto de materiales que se reúne o agrupa mantiene una plataforma multimodal (es decir, no ponen el foco en reunir solo lenguaje verbal oral o escrito). Por ejemplo, un corpus puede estar conformado por tres videos, por cinco portadas de periódicos o por tres películas. En todos estos casos tenemos corpus multimodales en los que es importante (en tanto que será objeto del análisis) el formato mismo en el que viven estos fenómenos, pues los elementos multimodales (el tipo y tamaño de las letras, el lado hacia el que se carga una imagen, etc.) se vuelven significativos también para acercarnos al fenómeno que queremos analizar. Actualmente ya existen programas que permiten que se cargue, por ejemplo, un video y que este se analice en varios niveles indicando observaciones, notas o etiquetas en los segundos en los que es necesario (por ejemplo, ELAN, software de transcripción lingüística multimodal diseñado por el Instituto Max Planck o bien el software propuesto por los lingüistas sistémico funcionales).[11] Por supuesto, un corpus multimodal supone muchos más retos con respecto a la manera en la que se va a reunir el material y cómo será trabajado, y, generalmente, son requeridos en análisis con enfoques semióticos, comunicativos o de otras disciplinas. Un ejemplo de corpus multimodal es el usado en el trabajo de Antúnez Piedra y Mateo Ruiz (2016).[12]

Definir la clase de corpus que necesitamos a partir del tipo de soporte depende, fundamentalmente, de la naturaleza y características del objeto de estudio central de la investigación: si queremos analizar marcas prosódicas de cortesía, evidentemente necesitamos un corpus oral con criterios muy detallados en la indicación de las marcas prosódicas, pero si nos interesa un tipo de marcador del discurso, podemos necesitar un corpus oral o escrito, esto depende de en qué dimensión o variante queremos observar el uso de ese marcador y en qué periodo, pero es posible que tal acercamiento pueda prescindir de los datos paralingüísticos de la oralidad, por ejemplo. Por supuesto, si lo que queremos es analizar el uso de los colores en las portadas de periódicos, necesitamos que el corpus se reúna y mantenga en soportes que permitan que no se pierdan datos acerca de eso mismo. Es decir, el tipo de corpus que

[10] Se puede consultar en https://archivos.juridicas.unam.mx/www/bjv/libros/6/2720/4.pdf
[11] Véase http://www.isfla.org/Systemics/index.html
[12] Disponible en https://dialnet.unirioja.es/servlet/articulo?codigo=7740499

podemos escoger (dependiendo del tipo de soporte) basa su pertinencia en una relación estrecha con el objeto o fenómeno de estudio y las dimensiones que se quieren analizar (mismas que determinan el tipo de soporte). ¿Cuál es el más pertinente? El tipo de corpus por soporte que permite conservar y mantener de la mejor manera las dimensiones o elementos en donde se encuentran los datos que necesitamos para la investigación y el conjunto de elementos, segmentos y niveles que componen o están en juego en el fenómeno a analizar será el más pertinente en tanto que ayuda a conservar con el mayor rigor y meticulosidad y hacer observable lo que más interesa del fenómeno de estudio.

2.2.2 Por el periodo de los datos que reúne

Otro criterio que se ha utilizado para establecer una clasificación de corpus es el de periodo o tiempo de los datos que contiene. De acuerdo con este criterio, un corpus puede ser de los siguientes subtipos:

a) **Corpus sincrónicos:** Son aquellos en los que se reúne un conjunto de textos o de datos para estudiar un fenómeno de manera sincrónica, es decir, en un momento específico o en un corte de tiempo determinado y corto, por lo que solo permite observar el fenómeno a la luz de ese momento y, por supuesto, los datos obtenidos solo revelan el comportamiento de un fenómeno o de ciertas correlaciones dentro del periodo seleccionado por lo que los resultados son válidos dentro de esa temporalidad o momento. Ejemplo de este tipo de corpus es el CREA (Corpus de Referencia del Español Actual).[13]

b) **Corpus diacrónicos:** Son aquellos en los que se reúne un conjunto de materiales o textos que pertenecen a momentos distintos, representativamente hablando, en periodos largos de tiempo (con cortes que se explican, especifican y justifican muy claramente entre los distintos periodos), pues el objetivo es obtener datos u observar un fenómeno y la manera en que han cambiado, evolucionado o se ha mantenido a lo largo del tiempo. El CORDE (Corpus Diacrónico del Español) es un buen ejemplo de este tipo.[14]

Evidentemente el objetivo de una investigación siempre debe establecer (lo ideal, de hecho, es que se diga explícitamente) si el fenómeno o el objeto de estudio interesan con una óptica sincrónica o diacrónica, pues el cumplimiento del objetivo depende justamente de que el corpus permita que este coincida con el planteamiento metodológico de la investigación. Cabe resaltar que el corpus es siempre una especie de muestra, de ejemplo de un fenómeno más amplio;

[13] Disponible en https://www.rae.es/banco-de-datos/crea
[14] https://www.rae.es/banco-de-datos/corde

sin embargo, este recorte o muestra debe corresponder con el tipo de enfoque sincrónico o diacrónico con el que se han planteado las preguntas de investigación, hipótesis y objetivos. Evidentemente si quiero observar la manera en la que ha surgido y se ha usado el marcador evidencial *dizque* un corpus que solo reúna los usos de esta forma en los tuits de cuentas de Twitter de medios de comunicación nacionales en el año 2021 no me va a permitir observar cuándo surgió y si ha habido cambios en su uso más allá de ese año y de ese tipo de manifestaciones. La pertinencia para justificar el tipo de corpus por periodo de tiempo se soporta en la correlación entre los objetivos y preguntas de investigación y la representación o cantidad de datos que necesito para cumplir con ellos.

2.2.3 Por el tipo de unidades que se reúnen o estudian

No en todas las investigaciones se usan corpus que estén conformados por textos o unidades comunicativas completas (lo que en la tradición estadounidense se conoce como *corpora*), hay casos en los que el corpus tan solo es un listado de las muestras que fueron localizadas en un corpus más amplio y que cumplen con ciertos criterios. De acuerdo a este enfoque, podemos encontrar los siguientes tipos de corpus:

a) **Corpus de unidades textuales o comunicativas completas (también llamados *corpora* o textuales):** Son aquellos en los que cada una de las unidades que compone el corpus representa un texto (oral o escrito), dispositivo o unidad comunicativa completa. Un ejemplo es el CORPHU (Corpus del Habla de Puebla) en donde se reúnen documentos (completos) que reflejan diversos registros de los hablantes de Puebla.[15]

b) **Corpus de muestras:** Son aquellos en los que las unidades que conforman el corpus están compuestas de muestras, segmentos que, bajo ciertos criterios, han sido seleccionados y que son los que se analizarán a profundidad. Por ejemplo, todos los verbos emocionales que se localizaron en el CORDE[16] al aplicar criterios específicos de búsqueda. Para quienes constituyen este tipo de corpus es fundamental establecer los criterios operativos que usaron para encontrar y seleccionar el listado de muestras, así como los criterios con los que se formó el corpus del que sacaron o tomaron las muestras. Igual de importante es que se reporte qué parte del contexto previo y posterior se recuperó, en caso de que se haga, y por qué. Un ejemplo de un trabajo con este tipo de corpus es el realizado por Galindo Flores (2023) y como se puede observar en su trabajo, es imprescindible que en este tipo de corpus se explique bajo

[15] http://www.corpus.unam.mx/geco/portal/index/corphu
[16] https://www.rae.es/banco-de-datos/corde

qué criterios se tomaron las muestras o datos y por qué se ha tomado ese número de datos y no otro.

c) **Corpus léxicos:** Son aquellos que están constituidos por unidades menores organizadas de acuerdo a ciertos criterios y pautas vinculadas con el objeto de interés. Un ejemplo muy interesante es el Corpus Neológico de Morfolex Estudio de la morfología y el léxico del español.[17]

Generalmente, quienes hacen análisis del discurso necesitan corpus de unidades textuales completas, pues un discurso es una unidad supraoracional de la esfera de la comunicación; frente a esto, quienes analizan fenómenos muy específicos de un nivel de la lengua o de algún nivel de análisis de otra disciplina pueden solo necesitar muestras. La pertinencia en este caso dependerá del nivel o dimensión que necesitamos para hacer la investigación.

2.2.4 Por el grado de representatividad con que se puede observar el fenómeno de estudio en el material que se reúne

De acuerdo con la manera en la que el corpus con el que vamos a trabajar o que vamos a seleccionar refleja la magnitud y la totalidad del fenómeno que interesa, se suele subclasificar en:

a) **Corpus exhaustivos:** Son aquellos en los que las unidades que lo conforman permiten mostrar o reunir de manera exhaustiva los datos o manifestaciones del fenómeno de estudio. Suelen ser corpus muy amplios y nutridos con distintas fuentes y criterios. Un ejemplo es el CORPES XXI (Corpus del Español del Siglo XXI), cuya versión actual cuenta con 357 000 documentos, que suman algo más de 381 millones de formas ortográficas, procedentes de textos escritos y de transcripciones orales.[18]

b) **Corpus representativos:** Son aquellos en los que las unidades que lo conforman son solo ejemplos del fenómeno de estudio. La selección de la muestra de esos ejemplos obedece a razones metodológicas que se vinculan con la profundización del análisis y los alcances temporales de la investigación planteada. Sin embargo, es fundamental que el investigador deje claros los criterios que sustentan que los ejemplos seleccionados sean representativos del fenómeno que se quiere estudiar o bien que al menos permitan un primer acercamiento de donde se obtendrán resultados que son poco generalizables, pero no por ello poco importantes. El trabajo de Ángeles Chargoy (2019) es un ejemplo en el que la autora ha argumentado por qué solo utilizar tres

[17] https://sites.google.com/site/morfolex/corpus-neol%C3%B3gico
[18] https://www.rae.es/banco-de-datos/corpes-xxi

periódicos de circulación nacional como fuente de su corpus debido a la representatividad que estos suponían.

c) **Corpus equilibrados:** Son corpus en los que existe un trabajo consciente e intencional para que las variedades de las esferas de la comunicación o registros que están incluidos tengan proporciones parecidas a las de cada clasificación. El equilibrio bien puede presentarse con alguna otra variable. Evidentemente no siempre es posible lograr balance, pero la pertinencia de los corpus equilibrados dependerá de los objetivos de la investigación. Hay muchos ejemplos de investigaciones en donde se utiliza el mismo número de unidades que representen al sexo masculino que al femenino en un afán de equilibrio. En el Corpus Oral de Referencia de la Lengua Española Contemporánea (CORLEC)[19] se ha hecho un esfuerzo para que haya cierto equilibrio entre los tipos textuales que han quedado representados ahí.

d) **Corpus piramidales:** En estos la representatividad o proporcionalidad de alguna de las variables va de más a menos a partir de criterios defendibles y justificables. Un ejemplo es el corpus paralelo bilingüe francés español del Centro Virtual Cervantes pues reconoce en sus criterios cierta piramidalidad en los niveles.[20]

e) **Corpus de monitoreo:** Son corpus en los que una cantidad de textos reunidos bajo criterios claros está en constante renovación con el objetivo de que refleje de manera actualizada ciertos fenómenos. La base de datos que se ha construido en ADESSE (Base de datos de Verbos, Alternancia de Diátesis y Esquemas Sintáctico Semánticos del Español)[21] se nutre de un corpus de monitoreo Arthus,[22] a partir de él es que se han extraído los datos que pueden consultarse en esta base.

f) **Corpus comparables:** Reúnen una serie de unidades en las que se comparten características comunicacionales, de esfera de la comunicación o género discursivo, pero en distintas lenguas. Un buen ejemplo de este tipo de corpus puede verse en ACCURAT corpus of comparable sentences.[23]

g) **Corpus especializados:** Aquellos en los que se reúnen materiales con base en observar de manera especializada un fenómeno en concreto. La propuesta de Sánchez Martín (2018) para estudiar la lengua de la geometría es un buen ejemplo de este tipo de corpus.[24]

[19] www.lllf.uam.es/ESP/Info Corlec.html
[20] https://cvc.cervantes.es/lengua/biblioteca_fraseologica/n5_durante/gonzalez_rey_07.htm
[21] http://adesse.uvigo.es/index.php/ADESSE/Inicio
[22] http://adesse.uvigo.es/data/corpus.php
[23] https://www.lt-innovate.org/lt-observe/resources/accurat-corpus-comparable-sentences
[24] Disponible en https://ojsspdc.ulpgc.es/ojs/index.php/PhilCan/article/view/884/902

En estos casos la pertinencia de estas posibilidades de tipos de corpus depende completamente del tipo de representatividad y la cantidad que exige la investigación que hemos planteado y, por ende, del grado de generalización que queremos alcanzar.

2.2.5 Por el grado de análisis o trabajo de etiquetado que contiene o no

Otro criterio que permite clasificar los corpus depende de qué tanto trabajo de etiquetamiento y marcaje incluye el corpus. El etiquetado consiste en marcar cierta información (de distintos niveles) de los datos contenidos en un corpus; por ejemplo, indicar la categoría gramatical de todas las palabras que lo componen o el tipo de oraciones por la actitud del hablante de todas y cada una de las oraciones que hay en él.

a) **Corpus no etiquetado.** Muchos de los corpus que se constituyen para investigaciones, si bien se ofrecen al lector al final del trabajo o en algún soporte digital complementario no incluyen marcas de información y categorías de ninguna índole, sino simplemente el conjunto de los materiales reunidos. Un ejemplo de este tipo de corpus es el que se ofrece en el Anexo 1 del trabajo de Cruz Bueno (2016).

b) **Corpus etiquetado.** En algunos casos los corpus sí se marcan (en muchas ocasiones se marca más de un tipo de dato), y precisamente el etiquetado permite hacer búsquedas más refinadas o bien extraer cierto tipo de información con más exactitud. Para conocer un ejemplo de este tipo de corpus (pese a que hay muchos ejemplos) recomendamos consultar el trabajo de Castillo Rodríguez, Díaz Lage y Rubio Martínez (2020) ya que en él se podrán consultar también definiciones más precisas del etiquetado.

La pertinencia de diseñar un corpus etiquetado o no dependerá completamente de los objetivos que tenga el corpus en sí mismo y las funciones para las que ha sido pensado.

2.2.6 De acuerdo a si reportan o no su procedencia y criterios

Existen corpus que reportan las fuentes de donde se han obtenido (e incluso los pasos que se siguieron) los materiales que se han reunido y otros que no lo hacen. Del mismo modo existen investigaciones de análisis de corpus en los que se argumentan tanto los criterios como la pertinencia de cada uno de ellos con respecto a los objetivos de la investigación (e incluso los retos y problemas que la elección de los criterios supuso), pero también hay trabajos que no lo hacen. Como el lector ya podrá suponer, consideramos que es inadecuado presentar investigaciones de análisis de corpus en las que no se reportan las fuentes de donde se han obtenido los materiales del corpus (o los instrumentos para obtener los materiales) y en las que no se argumentan las decisiones y criterios para el diseño del corpus.

2.2.7 Por la función que cumple en la investigación

Los corpus también se pueden clasificar a partir de la función que juegan en la investigación. Aunque lo más común es que el corpus constituya el cuerpo de datos que se analizará, también es posible que desempeñe otras funciones, y, de acuerdo con esto, podemos hallar los siguientes tipos (Sinclair, 1996)

a) **Corpus directo:** es el nombre que recibe el conjunto de materiales que será analizado en su totalidad para la investigación que se está realizando. En varios de los ejemplos que hemos citado ya los corpus son directos, como en Gómez Gordillo (2021) y Cruz Bueno (2016).

b) **Corpus control/ contraste:** en algunas ocasiones, debido a la naturaleza de la investigación, se requiere además un corpus control, es decir, un conjunto de materiales que se constituye o se obtiene sin que se cumpla uno de los criterios que tenemos en el corpus directo. Esto se hace con el objetivo de tener más material que permita observar si las variables de nuestras preguntas o hipótesis efectivamente tienen un impacto. En estos casos se estudiará y analizará tanto el corpus directo como el corpus control, poniendo énfasis en las manifestaciones correlacionadas con la presencia o ausencia de una o más variables. Tal es el caso del corpus control de Modesto Torres (2022).

c) **Corpus testigo:** se suele utilizar este término cuando un corpus se constituye con el fin de recopilar una muestra de un fenómeno que se atestigua justo con el corpus y del que tan solo se abstraen los fenómenos concretos que interesan, sin que sea necesario analizar todo el corpus sino solo el o los fenómenos que nos importan. El corpus en este caso tiene las funciones de mostrar que el fenómeno existe al menos en un periodo específico y de obtener de ahí muestras más pequeñas que son las que sí se analizarán. La tesis de Galindo Flores (2023) justamente toma como corpus testigo el CORPES, de donde obtuvo la lista de muestras que analizaría a profundidad.

d) **Corpus para usar en una intervención de obtención:** en algunos casos, las investigaciones que se realizarán requieren que el investigador constituya un corpus para poder implementar un instrumento de obtención. Por ejemplo, si queremos hacer un estudio acerca de la manera en la que hablantes no nativos del español comprenden las ironías en los titulares periodísticos, necesitamos, primero, constituir un corpus adecuado que reúna un número suficiente y apropiado para exponer a personas que cumplan con el requisito de no ser hablantes nativos del español a esos titulares y luego implementar ciertas técnicas para que expresen y reunamos la información obtenida de cómo comprenden o no esas ironías. En este caso, el corpus no sirve para ser el material de análisis, sino para acompañar la aplicación de un instrumento que nos

permita obtener los datos que necesitamos y que sí son los datos que vamos a analizar. Ejemplo de este tipo de corpus es el que utilizó Pérez Alvarado (2022) para analizar la comprensión de metáforas en hablantes del español como segunda lengua.

e) **Corpus paralelos:** el corpus paralelo es un recurso lingüístico consistente en textos de dos lenguas (en algún formato electrónico adecuado) que están alineados a cierto nivel de granularidad; generalmente a nivel de párrafo, aunque también a nivel de sección, página o incluso a veces de palabra. Un ejemplo es OPUS Open Parallel Corpora. [25]

2.2.8 Por el procedimiento que nos lleva a él

Finalmente consideramos que en ninguna fuente se ha hecho explícito con la debida profundidad que los corpus pueden ser diferentes por el método que se ha utilizado para su construcción. No es lo mismo seleccionar con criterios claros y pertinentes, un grupo de manifestaciones o unidades comunicativas que tener que implementar algunos instrumentos para que, con nuestra propia intervención, se produzca u obtenga un corpus. De hecho, para nosotros es fundamental y de ahí que se haya advertido desde el inicio con una nota distinguir el término con el que hacemos referencia al proceso general de diseño metodológico de corpus de una manera laxa, mientras que se puede utilizar el término constituir únicamente en los casos en los que justo se hace eso, se constituye un corpus, porque existen otras formas (que no son la constitución) tales como la replicación o la recolección de corpus.

Por ello, nos parece importante agregar que, de acuerdo con este criterio, los corpus pueden ser de los siguientes subtipos:

a) **Corpus constituido:** son aquellos que se forman seleccionando una serie de unidades que ya han sido producidas en sus propias y muy particulares esferas y condiciones de comunicación y que el investigador reúne *a posteriori* para una investigación.[26] Es decir, cuando lo que hacemos es establecer una serie de criterios que justifica que hayamos tomado esos materiales y no otros, en realidad lo que estamos haciendo es constituir un corpus. En este tipo de corpus cada uno de los criterios (sobre los que hablaremos más adelante) debe explicarse y argumentarse con la claridad y el detalle suficientes como para que, si alguien más, partiendo del mismo universo o conjunto de materiales existentes, aplica esos criterios, llegue a la misma selección. Esto da solidez a la constitución del corpus. Un ejemplo de un corpus constituido es el de Gómez Gordillo (2012).

[25] https://opus.nlpl.eu/

[26] Usamos el concepto de *esferas* en el sentido en que lo usa Bajtín (1995) cuando habla de esferas de la comunicación.

b) **Corpus recolectado u obtenido:** en este tipo, el investigador, como parte de la metodología de la investigación, define y crea uno o más instrumentos con los que recabará en un grupo o población bien delimitado y con criterios claros, una serie de materiales o datos. Precisamente el resultado o los datos obtenidos con la aplicación de tales instrumentos constituye su corpus. A diferencia del anterior, en este caso el investigador crea instrumentos para que las unidades que constituirán su corpus sean generadas dentro del mismo proceso de investigación. En este tipo de corpus es fundamental que el investigador reporte los instrumentos creados y su pertinencia para obtener el corpus deseado, así como que describa con detalle la manera en la que tales instrumentos se usaron para obtener cada una de las unidades del corpus. Justamente el trabajo, ya mencionado de Cruz Bueno (2016) es ejemplo de un corpus recolectado.

c) **Corpus retomados:** un corpus retomado es aquel que se constituyó o se obtuvo en una investigación previa y que el investigador retoma para su propio estudio. En estos casos se debe reportar cada uno de los procedimientos, criterios y pasos metodológicos con los que en la investigación previa se constituyó el corpus. Además, es importante aclarar que cuando se retoma otro corpus, se puede retomar de manera completa (es decir, usando la totalidad del corpus de la investigación previa) o bien de manera parcial (solo se toman algunas de las unidades que conformaban el corpus retomado). Por supuesto la decisión de retomar el corpus de manera completa o parcial debe ser una decisión metodológica y bien argumentada de manera explícita y reportando siempre las fuentes.

d) **Corpus replicados:** son aquellos que se constituyen o se obtienen replicando la metodología o los criterios que se reportaron en la constitución o recolección de una investigación previa, pero a diferencia del corpus retomado, dará por resultado, siguiendo los mismos pasos y criterios, un conjunto de materiales diferentes. Por ejemplo, se pueden retomar los criterios de constitución de un corpus de un estudio previo que se hizo en otro país para reunir un corpus de entrevistas hechas a presidentes, o bien se pueden replicar los instrumentos y criterios usados en una investigación en otro estado para obtener narraciones de incidentes de conductores de vehículos con ciclistas, pero en un país o ciudad o con una población distinta. Estos corpus permiten profundizar y comparar contrastar los resultados de estudios previos.

Es de suma importancia que quede claro que estas clasificaciones no son excluyentes, sino combinables, es decir, a partir de los objetivos de investigación y del diseño de construcción del objeto de estudio podemos determinar que necesitamos un corpus oral, exhaustivo, sincrónico, textual, constituido y especializado. ¿De qué depende eso? De que cada una

de estas características sea coherente y pertinente con la investigación misma; dicho de otra manera, que suponga que el corpus permita mirar en los grados, niveles y dimensiones oportunos el fenómeno que interesa. Como se ha mencionado, cada tipo de corpus supone, para tener rigor, cumplir con la explicación detallada de ciertos elementos metodológicos, esto hará que el corpus de la investigación pueda ser replicable y que sea mucho más sólido.

Es mucho más adecuado plantear la pregunta ¿qué tipo de corpus reclama mi investigación a partir de sus propias pautas, objetivos y preguntas?, que plantear esta cuestión como ¿qué tipo de corpus uso? Porque en esta segunda interrogante se están obviando las reflexiones y retos que supone la pertinencia entre la investigación y el corpus que requiere.

2.3 Reflexiones recomendadas

Al igual que al cierre del capítulo anterior, consideramos que es útil, para disparar las reflexiones y la toma de decisiones con respecto a corpus, plantearse cuestiones como las siguientes:

¿Necesito que el material conserve rasgos de oralidad o multimodalidad o basta con que esté escrito para tener acceso a los datos y fenómenos que me interesan?

¿Analizaré las unidades completas como manifestaciones discursivas o solo uno o unos fenómenos muy precisos que aparecen en el corpus?

¿Requiero tener todas las manifestaciones del periodo sincrónico o diacrónico que vas a estudiar o, por el contrario, creo que es suficiente trabajar únicamente con algunas?

¿Quiero analizar el fenómeno en un momento específico o a través de un periodo de tiempo?

¿El fenómeno de interés es observable y queda bien representado en el tipo de corpus que he pensado?

¿Voy a decidir qué materiales que ya existen se integran en el corpus o creo que debo crear un instrumento para obtenerlos pues no hay materiales *ad hoc* para observar el fenómeno? ¿Existen corpus anteriores o ya realizados que pudieran ser útiles o existen procedimientos ya utilizados previamente que puedes replicar para obtener el corpus que necesitas?

¿El resultado de los materiales que recopilé es lo que se analizará directamente y permite cumplir todos los objetivos o requiero reunir otros materiales con características similares, pero con alguna variación para cumplir el objetivo?

¿El corpus cumple alguna otra función en la investigación que no sea la de funcionar como el material de análisis?

2.4 Para leer más acerca de los tipos de corpus

Aarts, J. y W. Meijs (eds.). 1986. *Corpus Linguistics II*. Ámsterdam: Rodopi B.V.

Aarts, J., de Haan, P. y Oostdijk, N. (eds.) 1993. English Language Corpora: Design, Analysis and Exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Neijmen 1992. Ámsterdam: Rodopi B.V.

Aijmer, K. y B. Altenberg (eds.). 1991. *English Corpus Linguistics*. Londres: Longman.

Alvar Ezquerra, M. y Villena Ponsoda, J. A. 1994. Estudios para un Corpus del Español. Anejo *Analecta Malacitana*. *Revista de la Sección de Filología de la Facultad de Filosofía y Letras* 7. Universidad de Málaga: Grafur.

Armstrong, S. (ed.) 1994. *Using Large Corpora*. Cambridge: mit Press.

Atkins, B., Clear, J. y Ostler, N. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1): 1-16.

—. 1992. Corpus design criteria. *Literary and Linguistic Computing*. Journal of the Association for Literary and Linguistic Computing 7(1): 1-16.

Baker, M. 1993. Corpus Linguistics and Translation Studies — Implications and Applications, *Text and Technology*: In honour of John Sinclair, M. Baker, G. Francis y E. Tognini-Bonelli (eds.), 233-252. Ámsterdam: John Benjamins Publishing Company.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. Londres y Nueva York: Continuum.

Barlow, M. 1996. Corpora for Theory and Practice. *International Journal of Corpus Linguistics* 1 (1): 1-38.

Barnbrook, G. 1993. *The Automatic Analysis of Dictionaries: Parsing Cobuild Dictionaries,* M. Baker, G. Francis, y E. Tognini-Bonelli (eds.): 313-331.

Baud, R. *et al*. 1998. Extracting Linguistic Knowledge from an International Classification, *Division of Biomedical Informatics*, Nashville: Vanderbilt University.

Baugh, S., Harley, A. y Jellis, S. 1996. The Role of Corpora in Compiling the Cambridge Dictionary of English. *International Journal of Corpus Linguistics*, 1 (1): 39-60.

Beckmann, F. y G. Heyer (eds.). 1993. *Theorie und Praxis des Lexikons*. Berlin: Walter de Gruyter.

Benjamins, V., D. Fensel y A. Gómez. 1999. Knowledge Management Through Ontologies. Documento disponible en http://www.aifb.uni-karlsruhe.de/WBS/broker/inhalt-wp.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243-257.

Borja Albi, A. 2007. Corpora for Translators in Spain. The cdj-gitrad Corpus and the gentt Project, *Incorporating Corpora: The Linguist and the Translator*, G. Anderman y M. Rogers (eds.), 243-265. Clevendon: De Gruyter.

Briz Gómez, A. y M. Albelda Marco. 2009. Estado actual de los corpus de lengua española hablada y escrita, *El español en el mundo*, Instituto Cervantes, (ed.)165-225. Madrid: Instituto Cervantes.

Cabré, M. T. 2007. Constituir un corpus de textos de especialidad: condiciones y posibilidades, *Les corpus en linguistique et en traductologie*, M. Ballard y C. Pineira-Tresmontant (eds.), 89-106. Arras: Artois Presses Université.

Couthard, M. (ed.) 1994. *Advances in Written Text Analysis*. Londres: Routledge.

Johansson, S. 1998. On the role of corpora in cross-linguistic research, *Corpora and cross-linguistic research: Theory, method, and case studies*, S. Johansson y S. Oksefjell (eds.), 3-24. Ámsterdam: Rodopi B.V.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Londres y Nueva York: Longman.

Llisterri, J. y J. Llisterri. 1999. Diseño de corpus textuales y orales, *Filología e informática. Nuevas tecnologías en los estudios filológicos*, J. M. Blecua *et al*. (eds.), 45-77. Barcelona: Editorial Milenio.

McEnery, T. y A. Wilson. 1996. *Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics*. Edimburgo: Edinburgh University Press.

—. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Oostdijk, N y P. de Haan (eds.). 1994. *Corpus-based Research into Language. In Honour of Jan Aarts* No. 12. Ámsterdam: Rodopi B.V.

Sánchez, A. 1995. *cumbre: Corpus Lingüístico del Español Contemporáneo. Fundamentos, Metodología y Aplicaciones*. Madrid: sgel.

Sinclair, J.M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

—.1992. Trust the Text, *Advances in Systemic Linguistics*, pp. 5-19., M. Davies y L. Ravelli (eds.), 5-19. Londres: Pinter.

Teubert, W. 1996. Comparable or Parallel Corpora? *International Journal of Lexicography* 9 (3): 238-265.

Tognini-Bonelli, E. 1996b. *Corpus Theory and Practice*. Birmingham: twc.

3. Universo, acervo y corpus

3.1 Introducción

Los objetivos de nuestra investigación, así como el conocimiento de las características generales y peculiares de distintos tipos de corpus son elementos fundamentales que ayudan a delinear algunos elementos que, por supuesto, hacen que se tenga más certeza y claridad que cuando uno solo sabe que necesita un corpus. Sin embargo, la reflexión y consideraciones necesarias no terminarán de estar completas si no tenemos un conocimiento preciso, crítico y reflexivo acerca de la manera en la que nuestro fenómeno de interés se manifiesta, como un hecho en el mundo, en un **universo** y cómo algunas de las muestras de esas expresiones pueden quedar plasmadas en un **acervo** y/o en un **corpus**. Es decir, la comprensión de la importancia en el diseño del corpus está incompleta si no entendemos que el corpus, siempre, sin importar sus características, establece una relación con el conjunto total de elementos, unidades o fenómenos que nos interesa analizar y que, para poder diseñar[27] de manera seria y sólida una investigación y el corpus de ella, tal relación debe estar clara. Aún más, el camino del universo al corpus, cuando se incorpora como una reflexión argumentada, dota de gran solidez al trabajo. Por ello mismo, ahora revisaremos estos tres conceptos fundamentales para todo aquel que quiere diseñar un corpus, sin importar el tipo de características (de las revisadas en el capítulo anterior) que considere las más pertinentes. En la bibliografía acerca de corpus, si bien estos conceptos aparecen definidos, nos parece que se ha dado muy poca importancia y espacio a entender (salvo en Carbó 2001a, 2001b, 2001c, 2002, 2004, 2007 y 2016) la importancia de estos conceptos como elementos que orientan el proceso de construcción del corpus de la investigación, es decir, pese a que se ha enfatizado su valía teórica, se ha obviado su estrecha relación con su aplicación en el trabajo de construcción del corpus. De ahí que, en este apartado, insistamos en esta dimensión.

3.2 Definiciones básicas

Un universo es el conjunto total de algo, en este caso de las manifestaciones o de los fenómenos que nos interesan como objeto de estudio. Un universo está conformado por un conjunto (abierto o cerrado) de unidades que comparten al menos un rasgo o característica que los hacen ser parte de un conjunto total con algo en común. Por ejemplo, existe un universo de

[27] Hasta antes del capítulo anterior habíamos usado diseñar o construir el corpus en términos generales, ahora ya sabemos que no es lo mismo constituir un corpus que obtener un corpus, etc., por lo que a partir de ahora usaremos el término construir para referirnos de manera general a los distintos procesos en que se llega a un corpus.

discursos de toma de posesión de los presidentes mexicanos, si bien no es un universo infinito (pues hasta el momento en que escribo estas líneas México ha tenido 68 presidentes, de los cuales todos rindieron un discurso de toma de protesta, pero solo los últimos 15 lo han hecho bajo el formato discursivo de toma de protesta presidencial que conocemos actualmente), sí es un universo abierto, pues cada seis años ingresa un nuevo elemento, sin embargo, aunque está abierto, no crece con celeridad.

Ahora, supongamos que nuestra investigación se centra en el uso de emoticonos en las conversaciones de WhatsApp, el universo de conversaciones en este formato y con emoticones, no solo es inmenso, sino que está abierto y día tras día crece exponencialmente. De ahí que, de entrada, sea más sencillo plantear el estudio del uso de emoticonos en las conversaciones en WhatsApp entre padres e hijos adolescentes o entre parejas sentimentales en donde los participantes tienen un alto grado de estudios (así estamos partiendo de un universo, también inmenso y en constante crecimiento, pero más acotado). Existen otros casos en los que, por ejemplo, existen universos de fenómenos que ya están cerrados: si me interesan los discursos de Samir Flores contra la termoeléctrica en Morelos, el universo de manifestaciones está cerrado, debido a que (por desgracia) Samir fue ejecutado extrajudicialmente y ya no habrá más manifestaciones que se sumen a ese universo.

Los universos pueden no estar condicionados solo por el mismo tipo de género o subgénero, situación comunicativa, orador o tema, sino, a veces, también por un elemento común mucho más amplio, por ejemplo, existe un universo de todos los documentos oficiales con respecto a la Reforma Energética (unos serán discursos parlamentarios, otros son propuestas de leyes, otros son dictámenes, etc.).

El universo de un fenómeno de estudio deberá determinarse invariablemente a partir de los objetivos de la investigación, pero los universos pueden estar constituidos como tales a partir de diferentes criterios: temáticas, tipo de manifestaciones discursivas, tipo de soporte que se utiliza, género o subgénero, creador u orador, por el tipo de ceremonia o evento que se realiza por medio de ellos, por el tipo de fenómeno lingüístico, etc. Además, los universos pueden estar acotados a un periodo de tiempo específico dependiendo del enfoque sincrónico o diacrónico de la investigación (véase Capítulo 2).

Tener claridad en el universo del fenómeno o manifestación lingüística que nos interesa es importante, pero además en todo momento hay que verificar que mi comprensión del universo suponga, al menos, la presencia de un rasgo central de mi objeto de estudio en la totalidad de manifestaciones. En ese sentido el investigador puede establecer si el fenómeno que quiere comprender, analizar y/o describir forma parte de un universo abierto o cerrado (dependiendo de si se seguirán sumando elementos a él o no) y si es un universo exponencial o controlado (dependiendo de la rapidez con la que crece). Esto, aunque pare-

ce obvio, supone también tener conciencia de que los universos cerrados (que no el análisis de ellos como fenómenos) resultan más manejables. Incluso puede ocurrir que ya hayan sido recabados o reunidos en otros archivos o investigaciones, mientras que los universos abiertos y en construcción exponencial son más complejos de manejar y es menos frecuente que se encuentren sistematizados.

Todas las unidades del universo de las manifestaciones de una investigación son susceptibles, tentativamente, de formar parte de la muestra o corpus del estudio, en tanto que constituyen una manifestación concreta y particular que permite observarlo.

Un acervo, por otro lado (que suele confundirse una y otra vez con un corpus), constituye una colección de ciertos (algunos, muchos) elementos de un universo, pero tal colección no ha sido trabajada o reunida con criterios específicos para constituir tal cual un corpus ni para servir a una investigación: más bien es el resultado de un interés que nos lleva a ir recolectando unidades sin poner mucho cuidado o mucha atención en los criterios seguidos y para qué fines lo estamos haciendo (véase Carbó 2001c). Muchos investigadores y estudiantes, una vez que sentimos interés por un fenómeno, empezamos a guardar materiales o unidades que nos vamos encontrando, sin embargo, véase cómo ir reuniendo lo que yo me encuentro no es un criterio metodológico congruente con los lineamientos y pautas de una investigación específica.

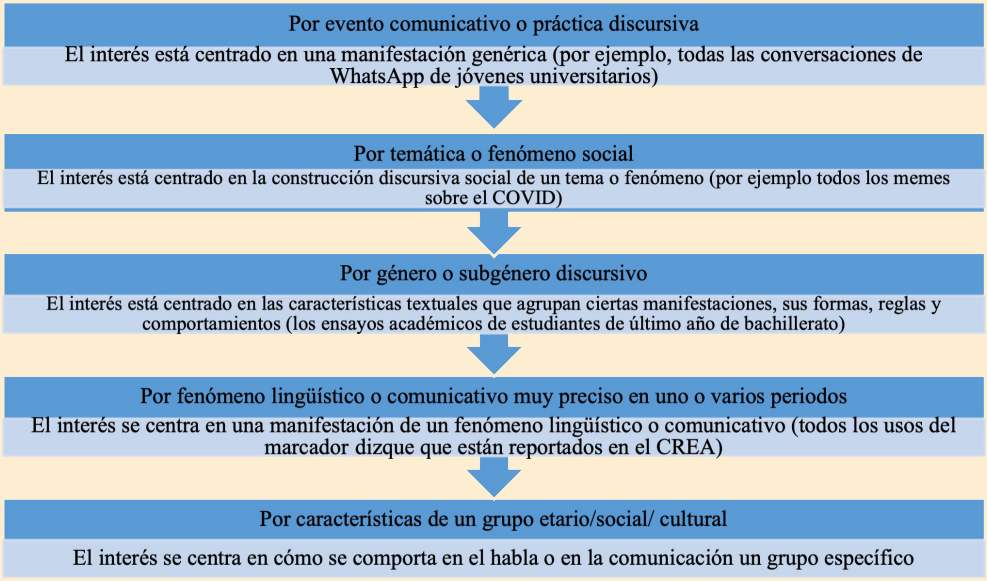
Es posible que en algunas investigaciones resulte útil que, en un primer momento, cuando aún no hay mucha certeza en el diseño de los objetivos y las preguntas de investigación, se recopilen o reúnan sin criterios muy claros, sino más bien azaroso, los fenómenos que comparten ciertas condiciones (por ejemplo, todos los titulares de las notas sobre feminicidios que aparecen en los periódicos que leo, en las notas que comparten mis contactos en redes sociales o que encuentro en los puestos de periódicos cuando salgo a la calle). Lo que es importante es tener claridad en el hecho de que lo que se está reuniendo es un acervo y no un corpus; de alguna manera, esto supone un avance, ya que el acervo puede convertirse en una fuente para el corpus o bien en una especie de corpus de contraste.

Para Carbó (2002), el acervo es un concepto más amplio que el del corpus:

> En cuanto al acervo, por extenso que se llegue a recopilar el territorio que acabará siendo de la incumbencia de nuestro escrutinio sistemático es siempre solo un ángulo, no más, del asunto (un costado, una vista, un perfil), acotado por el grado de apertura que permite el compás de una cierta mirada, sobre el continuo indiferenciado del tejido significante que confronta al estudioso con su "aparecer así ahora aquí", en una cierta morfología y no otra, en un lugar/momento determinado y particular y solo en él. (20).

Es muy importante entender que, cuando se intenta establecer con precisión tanto el universo de manifestaciones de nuestra investigación como si existen acervos de él (lo que casi siempre se hace por medio de investigación documental), para el caso de los segundos debemos tener mucha claridad en que los criterios compartidos que constituyen el universo y acervo de nuestra investigación pueden fundamentarse en distintos criterios, como se muestra en el Esquema 2.

Esquema 2. Criterios que se pueden seguir para la reunión de acervos a partir de universos



Desde el momento en que queremos establecer el universo y el acervo (si es que existe o si es que queremos ir creando uno) de nuestra investigación, debemos tener la certeza de que estamos utilizando o pensando en las líneas más pertinentes en concordancia con el diseño metodológico para definirlos, de lo contrario, se pueden generar contradicciones

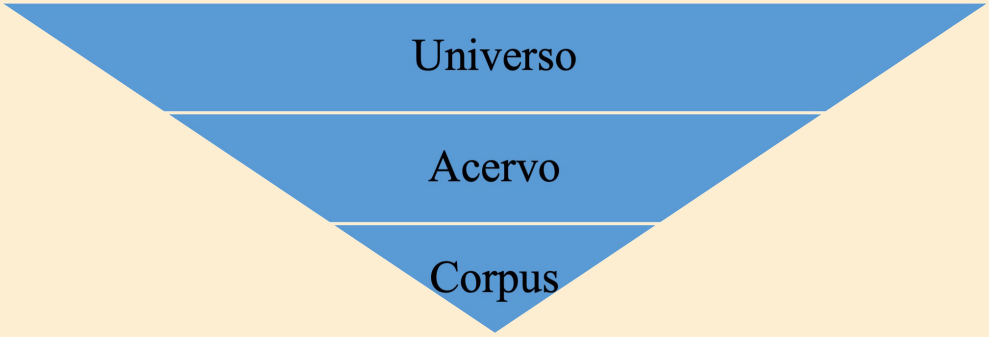
Como ya hemos visto (Capítulo 1) un corpus supone una selección guiada a partir de los planteamientos de una investigación en particular. De este modo, la representatividad y la pertinencia de esa selección depende absolutamente de la investigación que se está realizando, por lo que la posibilidad de diseñar un corpus más pertinente depende necesariamente de entender su diferencia con el universo y el acervo.

3.3 Del universo al corpus

La metodología de selección para el diseño de nuestro corpus, implica, necesariamente, que se reflexione seriamente y con mucha profundidad en estos conceptos y que tenga clara la manera en la que, recorriendo un camino que inicia en un universo de manifestaciones, aplicó una serie de criterios claros que le permitieron seleccionar solo algunas unidades de ese universo a partir de criterios que hacen explícita la pertinencia y adecuación, no solo para observar el fenómeno que interesa, sino para poder cumplir con los objetivos de la investigación, poder responder las preguntas planteadas o poder confirmar o rechazar las hipótesis. Aquí es imprescindible entonces que cada una de las características que tiene el corpus (véase Capítulo 2) coincida no solo con las necesidades de la investigación, sino que además se sustente en un argumento claro que explique cómo han ido quedando elementos del universo fuera hasta llegar al corpus.

La claridad en este recorrido debe estar reportada con el mayor detalle posible en el apartado metodológico de nuestra investigación y supone que se logre comunicar cómo se ha diseñado un camino con forma de embudo que nos lleva, al final, a tener los materiales más adecuados para nuestra investigación, tal y como se muestra en el Esquema 3.

Esquema 3. Camino del universo al corpus.



La conciencia de la manera en la que tomamos cada una de las decisiones de recorte en este camino no solo nos permitirá defender y argumentar muy claramente la pertinencia y adecuación del material, sino que, además, con el conocimiento que se adquiere en este proceso el investigador o estudiante es capaz de describir la ruta seguida de manera tal que los criterios de constitución de un corpus sean replicables; es decir que si cualquier otra persona partiera de un universo igual, al ir aplicando los criterios claramente definidos, debe llegar al mismo corpus y, también, que si a partir de un universo diferente, un investigador

quisiera replicar la investigación con otro universo y/o corpus para hacer un análisis comparativo-contrastivo, fuera posible replicar exactamente el mismo procedimiento.

Ahora bien, no todas las personas constituyen un corpus, algunos lo recolectan o crean, es decir, se involucran en la producción del corpus, debido a que no existe un acervo o conjunto específico al alcance que cumpla con lo necesario de acuerdo con los objetivos. Optar por este tipo de procedimiento permite controlar más algunas variables, pero también en estos casos (en los que no se parte de un universo del que hay que tomar solo algunas unidades de manera justificada) el procedimiento consciente y congruente con las necesidades y postulados de la investigación es tan importante como cuando se constituye un corpus. De hecho, en la obtención de un corpus uno parte de no tener nada a tener un corpus pertinente.

Finalmente, en los casos en los que se va a retomar o replicar un corpus, el estudiante o investigador debe revisar si la investigación que va a replicar reportó con detalle este recorrido o si el corpus que va a retomar lo hace, de lo contrario se estará cayendo en algunos errores metodológicos en la investigación, pues si no revisamos cómo se ha hecho este camino puede que no nos percatemos de que el corpus que queremos retomar o replicar supone, en esa vía, decisiones que hacen que no sea pertinente o que no queramos reproducir para nuestra investigación.

3.4 Facetas en el camino para llegar al corpus

Los caminos en este recorrido suponen muchas variaciones y particularidades porque cada investigación es muy diferente y por lo tanto necesita corpus diferentes y supone retos y dificultades muy variables. Sin embargo, hay algunos elementos que pueden ayudar a pensar en cómo comenzar a emprender o incluso crear ese camino.

3.4.1 Para constituir un corpus

Muchas veces las investigaciones que, por sus planteamientos específicos, requieren de la constitución de un corpus suelen haber surgido debido a una inquietud o profundo interés por un fenómeno determinado que se observó empíricamente. A partir de este interés el investigador puede comenzar con una lluvia de preguntas que le ayuden a reflexionar acerca de la naturaleza del corpus que requiere su propia investigación. Ahora bien, si estamos en una etapa muy incipiente en la que aún no hemos hecho la construcción o el diseño metodológico de la investigación y solo tenemos esa intuición o inquietud por un fenómeno observado, podemos plantearnos preguntas como: ¿En qué tipo de materiales o soportes he observado con más insistencia eso que llama mi atención? ¿Qué tipo de manifestaciones del uso del lenguaje son las que me interesan porque creo que en ellas aparecerá eso que me interesa?

Por otro lado, cuando ya se tiene al menos una versión primera del diseño metodológico se cuenta con muchos más elementos para comenzar a imaginarnos el camino que hemos de recorrer puesto que ya tengo elementos que de alguna manera son pistas indiscutibles del tipo de corpus que reclama mi investigación. Ayuda comenzar a preguntarse ¿Ya hay acervos y archivos que reúnan materiales que me interesan puesto que forman parte del universo de las manifestaciones que son mi objeto de estudio? ¿Existen tesis, artículos especializados o libros que informen haber trabajado con el mismo objeto de estudio y que reporten ya corpus o acervos? ¿Existen investigaciones en las que se han postulado propuestas metodológicas de constitución de corpus para el mismo fenómeno que me interesa (aunque lo analicen o trabajen desde ópticas o aristas diferentes)? ¿Cuál sería el universo de mi objeto de estudio? ¿Qué características debe tener el corpus de mi investigación a partir de sus objetivos, preguntas, metodologías, marco teórico y postura?

Evidentemente, con esto, es posible que haya más claridad acerca de ciertos elementos del corpus que necesita mi investigación, pero el camino de depuración del universo hasta llegar al corpus más pertinente no se podrá completar hasta que determinemos criterios operativos para la selección (cosa que veremos con detalle en el siguiente capítulo).

3.4.2 Para recolectar o crear un corpus

Por otro lado, quien va a recolectar o crear su propio corpus, debe comenzar a plantearse preguntas que ayuden a pensar en términos concretos cómo es que lo va a obtener y qué tipo de datos quiere que queden plasmados en ese proceso: ¿con qué colaboradores debo trabajar para lograr que los materiales capturen la presencia del objeto o fenómeno que quiero estudiar? ¿Qué materiales e instrumentos necesito para poder obtener esos datos? ¿En qué lugar específico y con qué pasos particulares puedo aplicar esos instrumentos? El investigador no debe olvidar que, si este es su caso, debe explicar, también paso a paso, cómo piensa implementar la recolección, así como exponer claramente, tanto el instrumento como el procedimiento utilizado. Por ello, puede ser fructífero plantearse preguntas como: ¿cómo me imagino paso a paso el procedimiento en cada ocasión en que utilice el instrumento para obtener los datos? ¿Cómo resguardaré, documentaré o respaldaré esos datos que va a generar el colaborador? ¿Cómo daré las instrucciones a los colaboradores?

3.4.3 Para quienes van a replicar un corpus o una metodología de corpus

Finalmente, en caso de que el investigador lo que quiera es repetir una metodología para obtener un corpus o usar un instrumento que ya ha sido utilizado en investigaciones previas, debe comenzar a preguntarse si efectivamente replicar esa y no otra manera es lo más pertinente para su investigación y si no le impone límites que no estaban considerados en

el diseño de su investigación o incongruencias con los datos que va a obtener. ¿Si replico los mismos pasos o aplico el mismo instrumento siguiendo el mismo procedimiento obtendré datos que me permitan capturar el objeto de estudio con las dimensiones que más me interesan? ¿Qué posibilidades existen de que, al aplicar exactamente los mismos pasos o instrumentos, por el caso concreto en el que yo me encuentro, no obtenga muestras tan pertinentes como las de la investigación que quiero replicar? ¿Qué límites supone la implementación de la réplica de esa metodología o instrumento que me preocupan o me generan ciertas inquietudes? ¿Qué ventajas y congruencias con la investigación supone la réplica de esta metodología?

3.5 Reflexiones recomendadas

De la misma manera que exponer nuestras reflexiones en espacios en donde haya más colegas nos permite profundizar y mejorar nuestras decisiones con respecto al corpus, es muy útil argumentar en este tipo de espacios formativos cuál es el universo, acervo y corpus de nuestras investigaciones, así como argumentar la congruencia que estos suponen con el estudio que estamos realizando. A continuación, ofrecemos algunas preguntas que pueden alimentar estos ejercicios de reflexión colectiva:

¿Cuál es el objetivo general o la pregunta central que aborda mi investigación?
¿Qué tipo de manifestaciones del uso del lenguaje necesito observar?
¿Qué tipo de variables quieres controlar en esas manifestaciones?
¿Qué criterios de homogeneidad y heterogeneidad requiere tu investigación?
¿Hay acervos o archivos con las manifestaciones que necesitas?
¿Qué tipo de universo es el que reúne las manifestaciones que me interesan?
¿Cuáles considero que es pertinente que formen parte del corpus? ¿Por qué?

En los casos en los que se ha decidido recolectar el corpus:

¿En qué lugar y con qué participantes es posible recolectar el corpus? ¿Por qué?

Enumera paso a paso cómo piensas recolectar el corpus e indica la relación entre los procedimientos y las variables que quieres controlar.

¿Cuántas muestras quieres obtener? ¿Por qué ese número y no otro?

¿Cuál sería el universo, el acervo y el corpus para el caso específico de la investigación que estoy desarrollando?

En los casos en los que se ha optado por replicar o retomar un corpus

¿Por qué es lo más oportuno y pertinente retomar ese corpus?

¿Qué no podríamos lograr si no retomáramos ese corpus?

¿Por qué es pertinente con esta investigación la replicación del corpus?

¿Qué no podríamos lograr si no replicamos el corpus?

¿Qué se aporta a la investigación al replicar el corpus y a la disciplina o línea de investigación a la que se adscribe nuestro estudio?

¿El corpus que voy a retomar presenta una explicación clara del universo, acervo y criterios de corpus?

¿La decisión de replicar una metodología de construcción de corpus supone cambios en el universo, acervo y corpus con respecto al que se replica? ¿Cuáles son esos cambios?

3.6 Ejemplos de trabajos en donde se reporta el tránsito del universo al corpus

Consideramos igual de útil que el lector pueda consultar algunos ejemplos de trabajos en los que se ha hecho énfasis explícito en la transición del universo al corpus, por lo que para que esta sección quede mejor ilustrada recomendamos (además de los artículos ya referidos del número especial de la revista *ELA* sobre la constitución de corpus), la consulta de las siguientes fuentes:

Ángeles Chargoy, J. I. 2019. Imágenes de candidatos políticos en diferentes periódicos mexicanos: hacia su caracterización. Tesis de maestría, Universidad Nacional Autónoma de México.

Carbó, T. 2001c. El cuerpo herido o la constitución del corpus en análisis de discurso. *Escritos* 23: 17-47.

— 2002. Investigador y objeto. Una extraña/da intimidad. *Iztapalapa* 53: 15-32.

Carbó, T. (ed.) 2007. Corpora, conceptos y métodos en análisis de discurso. *ELA. Estudios de Lingüística Aplicada,* 46 monográfico. México: UNAM, Centro de Enseñanza de Lenguas Extranjeras.

Carbó, T. y E. Salgado. 2013. El itinerario de un corpus multimodal para escrutar el desempeño presidencial reciente en México (2006-2012), *Estudios del discurso en América Latina. Homenaje a Anamaría Harvey*, 527-550. Bogotá: ALED.

Gómez Gordillo, L. A. 2021. El discurso presidencial mexicano ante el Congreso Estadounidense (1947-2010): agentividad en la construcción de la relación México-Estados Unidos. Tesis de maestría, Universidad Nacional Autónoma de México.

Grave Aragón, A. D. 2021. Análisis sociopragmático de la conversación escrita basado en el interaccionismo. El caso de los malentendidos en WhatsApp. Tesis de maestría, Universidad Nacional Autónoma de México.

Mendoza Cruz, C. E. [en prensa]. La codificación de género a través de los recursos lingüísticos en el stand up mexicano. Un estudio de caso. Tesis de maestría, Universidad Nacional Autónoma de México.

3.7 Para leer más sobre universo, acervo y corpus

Carbó, T. 2001a. La constitución del corpus en análisis del discurso. *Escritos. Revista del Centro de Ciencias del Lenguaje* 23: 17-47.

—. 2001b. Tocar el lenguaje con la mano. Experiencias de método. *Revista Latinoamericana de Estudios del Discurso* 1(1): 43-67.

—. 2001c. El cuerpo herido o la constitución del corpus en análisis de discurso. *Escritos* 23: 17-47.

—. 2002. Investigador y objeto. Una extraña/da intimidad. *Iztapalapa* 53: 15-32.

—. 2004. Protocolos de investigación en análisis de discurso y consolidación del campo disciplinario. *Discurso, teoría y análisis* 26: 121-30.

—. (ed.) 2007. Corpora, conceptos y métodos en análisis de discurso. *ela. Estudios de Lingüística Aplicada,* 46 monográfico. México: unam, Centro de Enseñanza de Lenguas Extranjeras.

—. Introducción. La elocuencia de los cuerpos. *Estudios de Lingüística Aplicada, 0*(46). doi:https://doi.org/10.22201/enallt.01852647p.2007.46.576

Carbó, T. y E. Salgado. 2013. El itinerario de un corpus multimodal para escrutar el desempeño presidencial reciente en México (2006-2012), *Estudios del discurso en América Latina. Homenaje a Anamaría Harvey*, 527-550. Bogotá: aled.

Sinclair, J. M. 1996. Preliminary Recommendations on Corpus Typology. eagles Document eag-tcwg-ctyp/p.

Sinclair, J. M., Payne, J. y Pérez, Ch. (eds.). 1996. Corpus to Corpus: A Study of Translation Equivalence. *International Journal of Lexicography* 9 (3): Autumn.

4. Los criterios del corpus

4.1 Introducción

Ahora que tenemos una idea general de que el diseño del corpus supone recorrer un camino que parte de un universo de manifestaciones y que supone una serie de decisiones dependientes de la investigación misma para sacar elementos de ese universo hasta llegar a un corpus, nos podemos dar cuenta que una vez localizado el universo, el investigador debe reflexionar con respecto a qué del universo debe formar parte del corpus. Preguntas como ¿necesito todas las muestras del universo para cumplir los objetivos? ¿Puedo trabajar (por tiempos y alcances) todas? ¿Qué tipo de representatividad necesita mi investigación por sus planteamientos? Son ejemplos del tipo de trabajo que habrá que resolver. La reflexión de preguntas como estas muestra la importante decisión que uno está construyendo al ir del universo al corpus. La ejecución de la construcción de ese camino metodológico debe hacerse con el mismo grado de rigurosidad que hemos tenido hasta ahora en cada uno de los aspectos señalados: bajo la implementación de criterios operativos claros.

No debe olvidarse que cada uno de los aspectos que se abordan en esta sección debe realizarse, necesariamente, teniendo presente que el objetivo es diseñar un corpus pertinente, esto es, en el que sea posible realizar un análisis o estudio que se corresponda y tenga congruencia con los objetivos de la investigación para aumentar su solidez y validez. Así pues, cada que se toma una decisión para comenzar a sacar elementos del universo para poder llegar a un corpus se debe mantener cuidado en revisar que cada decisión tenga congruencia, pertinencia y coherencia argumentable con los objetivos, las preguntas de investigación, la hipótesis a corroborar o refutar, el tipo de corpus que se necesita y el universo planteado.

Esta fase del diseño del corpus se realiza por medio de la construcción de criterios operativos que al aplicarse nos permiten llegar al conjunto de unidades más propicio, adecuado, pertinente y suficiente. En este capítulo atenderemos las cuestiones referentes a este proceso.

4.2 Variables y criterios operativos

Una **variable** es una construcción teórica segmentada, casi siempre, por atributos como sexo, grupo etario, nivel de formación escolar, etc. Implica una o varias características que comparten (o no) todas las unidades mínimas de análisis y que revelan las dimensiones con las que se ha de observar, comparar y analizar el corpus. No es el objetivo de este libro profundizar acerca de las unidades y variables de análisis en una investigación (eso amerita un esfuerzo incluso más amplio), pero lo que es cierto es que los objetivos, las preguntas, así como las hipótesis de una investigación suponen necesariamente que existen **variables**

que en nuestro estudio o análisis utilizaremos para describir u observar con más detalle el fenómeno, comparar o contrastar sus distintas manifestaciones y analizarlas.

Uno de los riesgos que se corre en el camino del universo al corpus supone que, como resultado de un descuido, perdamos datos que sí eran importantes para el estudio; es decir, variables exigidas, de origen, por la investigación. Si en el camino que vamos a recorrer se pierden de vista las variables que estaban postuladas en el diseño de la investigación (habrá algunas que deben tener todas las unidades que voy a analizar, hay otras que cambian y permiten dividir en grupos el material) puede ocurrir que los criterios utilizados para avanzar el trayecto, en lugar de ayudar, ocasionen la pérdida de información fundamental, entonces también es importante listar esas variables y tenerlas a la vista y, cada que se defina un criterio, se revise que este no implique la pérdida de los datos que necesito para mantener las variables.

Por otro lado, un **criterio operativo** es una norma, regla o definición que permite aplicar un procedimiento de selección, localización o forma de proceder con la especificidad suficiente como para que cualquiera que lo reciba pueda ejecutarlo. Por ejemplo, si decimos que solo se dejaron aquellas unidades que cumplían con el criterio de ser oraciones debemos especificar qué criterio operativo usamos para definir qué no se quedaba, qué salía y qué se mantenía en el camino del universo al corpus. Por cada decisión tomada con respecto a las unidades o materiales que se quedan en el corpus debe haber un criterio operativo que debe estar bien descrito y definido.

Tanto las variables como los criterios operativos que explicitan el camino del universo al corpus dependen y deben estar claramente relacionadas con las preguntas y objetivos de investigación y, al mismo tiempo, con un criterio que explique las decisiones de conformación del corpus.

4.3 Los criterios que permiten llegar al corpus: Homogeneidad/heterogeneidad; internos y externos

Sin importar qué naturaleza tenga la investigación, lo que es cierto es que el diseño de un corpus juega con dos variables fundamentales que han de reflejarse en el tipo de criterios que vamos a argumentar: la homogeneidad y el contraste. Esto quiere decir que en un corpus deben existir elementos que todas las unidades de análisis comparten y que justo nos permiten hacer un análisis a partir de comparaciones (que explica que todas provengan de un mismo universo), pero también es necesario tener elementos de contraste que permitan buscar diferencias. Lo que debe ser igual y lo que debe ser distinto en el corpus se determinará a la luz de los objetivos, preguntas de investigación e hipótesis, pero debe estar explicado con tanto detalle que permita entender qué de todo lo que constituye el universo se va a ir quedando y qué no hasta llegar al corpus. Cada explicación debe indicar con claridad la pertinencia que

ese criterio guarda con el planteamiento metodológico de la investigación misma. Es cierto que en los casos en los que los corpus se retoman o replican, nosotros no hemos diseñado los criterios, pero sí debemos reportar los criterios postulados en la investigación de donde hemos retomado o decidido replicar el corpus. Además, en los casos en los que se consulta un corpus exhaustivo para obtener un corpus léxico, debemos reportar los criterios de búsqueda que aplicamos al corpus exhaustivo para obtener las unidades del corpus de análisis y, como ya se ha dicho, en todos los casos la reflexión explícita de la pertinencia de estos criterios debe incluirse.

Criterios de homogeneidad: conjunto de reglas, normas o decisiones que el investigador atiende para seleccionar materiales que tengan elementos comunes. La homogeneidad se puede basar en diferentes dimensiones de los materiales o unidades que estaban presentes en el universo y que por lo tanto están presentes en las que constituirán el corpus, por ejemplo: fuente, destinatario, contextos de transmisión, consumo y respuesta. Siempre se parte, primero, de una homogeneidad y la homogeneidad discursiva tiene que ver con las condiciones de producción, con las temáticas, con los oradores, con las características contextuales. Sin embargo, puede haber muchos otros criterios de homogeneidad. Lo importante es entender que al menos uno de los parámetros de selección de corpus debe sustentarse en la homogeneidad. Esto se debe a que los materiales que vamos a analizar deben compartir al menos una característica para que el estudio pueda realizarse (un mismo formato, un mismo evento, un mismo tipo de discurso, un mismo tipo de verbo, etc.).

Existe otra forma de manejar la variable de homogeneidad que es más compleja: el tema. En estos casos el tema dará homogeneidad, mientras que la forma dará contraste. La pertinencia de utilizar una variable de este tipo dependerá, como en otros casos, de los objetivos. Si se opta por esta forma debe haber una completa justificación de por qué se hace y de la correlación con los objetivos.

Criterios de contrastes o heterogeneidad: conjunto de reglas, normas o decisiones que el investigador atiende para seleccionar materiales que, ya teniendo criterios de homogeneidad, permitan al mismo tiempo que en el corpus queden elementos diferentes. Este tipo de criterios pueden ser determinados por tiempo o momento, por orador, por tipo de material, por fuente, etc. Es decir, los criterios de heterogeneidad pueden ser establecidos con tantos principios como los de homogeneidad, pero en este caso lo que hacen es que se reúnan materiales que, además de compartir algún o algunos rasgos, difieren en otros

Sin embargo, los criterios que explican y argumentan nuestro corpus no solo pueden clasificarse a partir de si permiten la incorporación de elementos de comparación y contraste, ya que también existe una tipología de criterios de acuerdo a los requisitos que las unidades que conforman el corpus deben cumplir.

Criterios internos: conjunto de reglas, normas o decisiones que el investigador juzga que deben estar presentes en cada una de las unidades que formará parte del corpus para que esta pueda incorporarse.

Criterios externos: conjunto de reglas, normas o decisiones que el investigador juzga necesarias pero que no están en el material que se ha seleccionado (de ahí que se les conozca como externos), sino que son definidos a partir de otras nociones, conceptos o teorías externas que son importantes para observar el fenómeno del que las unidades del corpus son muestras particulares.

4.4 Los criterios en distintos tipos de corpus

Aquellos que recolecten su propio corpus han de tener presente que, cuando den cuenta metodológica de los instrumentos y procedimientos que emplearon, deberán explicar los criterios de homogeneidad y heterogeneidad que guiaron el diseño de esos instrumentos y procedimientos para que los materiales obtenidos contuvieran no solo los datos requeridos por la investigación, sino también datos adicionales que los hagan comparables (homogeneidad) y contrastables (heterogeneidad).

Es más, muchas veces estos elementos de heterogeneidad y homogeneidad dependen no solo de los instrumentos sino también o (en otros casos) fundamentalmente de la selección de los colaboradores. A lo mejor se les va a pedir que todos hagan una narración o cuenten algo con el mismo tema (homogeneidad); no obstante, los colaboradores tienen diferencias significativas (edades, estratos sociales, género, etc.), y es necesario explicar por qué esas diferencias son coherentes y pertinentes para la investigación.

Debido a ello, en el apartado metodológico de la tesis o investigación se debe destinar una parte en la que explícitamente se destaquen elementos como:

a) A partir de los objetivos, preguntas de investigación e hipótesis se explique la ausencia de acervos o corpus ya constituidos y la necesidad de obtener el corpus como parte del trabajo de investigación por medio del diseño de instrumentos y la intervención del investigador.
b) Se explique claramente la conexión entre el instrumento que se preparó para ello, argumentando cada una de sus partes y justificando su pertinencia para obtener el corpus adecuado haciendo hincapié en el tipo de corpus que requería la investigación y la manera en la que cada una de las partes del instrumento se pensó y diseñó para obtener el material más idóneo para la investigación con datos que permitirán homogeneidad y contraste al mismo tiempo que arrojara lo necesario para observar el fenómeno de estudio o sus manifestaciones.

c) Se explique y detalle la fase de pilotaje del instrumento, así como las correcciones que se le hicieron como resultado del mismo. Generalmente, cuando se van a aplicar instrumentos para obtener un corpus con un grupo de colaboradores que han sido seleccionados con mucho cuidado, es recomendable que cuando se tenga la propuesta del instrumento y del procedimiento con el que se va a aplicar, se haga un pilotaje que permita ensayar y probar si su aplicación funciona, si no supone problemas imprevistos para los colaboradores o confusiones que impiden que realice lo que se le pide. Además, en el pilotaje se puede observar si efectivamente el instrumento arroja y permite al investigador recopilar los materiales con los datos y los elementos que él suponía que iba a obtener al diseñarlos.
d) Se presente el (los) instrumento (s) final(es). Muchas veces después del pilotaje se incorporan algunos cambios (menores o sustanciales) a partir de lo que se pudo observar y detectar en el pilotaje, por lo que es importante decir cuáles son las razones que, observadas en el pilotaje, motivaron tales cambios y por qué hacían más adecuado tanto el instrumento como el procedimiento para la investigación.
e) Una reflexión final en la que se explique cómo el instrumento resultó el más adecuado para llegar a un corpus que la misma investigación exigía.

Mientras que, por otro lado, quienes van a constituir un corpus, deben dar cuenta metodológicamente de la manera en la que llegaron a esa selección, por lo que deberán de explicar en algún apartado de la tesis o investigación los siguientes elementos:

a) A partir de los objetivos, preguntas de investigación e hipótesis tendrán que hablar de manera general del universo de manifestaciones de esa investigación.
b) Dar cuenta y descripción detallada de si existen acervos, colecciones y corpus previos que se relacionen con el fenómeno de estudio y hacer un análisis crítico de estos vinculándolo con la necesidad de constituir un corpus propio haciendo hincapié en el tipo de corpus que se necesita.
c) Presentar y argumentar los criterios de homogeneidad y heterogeneidad (internos y externos) que se eligieron, así como por qué son pertinentes para llegar al corpus.
d) Explicar cómo, al aplicar estos criterios, quedan fuera elementos del universo hasta llegar a aquellos que constituyen nuestro corpus.
e) Valorar críticamente la pertinencia y congruencia del corpus obtenido con estos criterios.

Un asunto muy importante es que, cuando se redacte el borrador metodológico, se ponga mucho cuidado en mencionar explícitamente que cada uno de los criterios que se hayan

seleccionado sean criterios operativos y no criterios generales, además de que se expliquen y definan lo más clara y concretamente posible. Un criterio general no aporta la información precisa necesaria para discriminar qué se queda y qué se va, bien porque son ambiguos o bien porque son demasiado generales. Por ejemplo, decir algo como "se quedarán todas las oraciones que presenten una metáfora" es un criterio general, porque no se está definiendo qué es una ironía o qué se entiende por ironía; un criterio operativo debe mencionar claramente los elementos que deben cumplirse para que un elemento sea seleccionado; por ejemplo, es distinto decir "se quedarán todas las oraciones que presenten una ironía, esto es, aquellas construcciones con un verbo principal en los que exista una afirmación con respecto a la cual el escritor u orador marca una distancia y que deben entenderse justamente en el sentido opuesto a lo dicho explícitamente".

Esto mismo se ha de aplicar para todo criterio en la selección de las unidades del corpus; no basta con decir "se seleccionaron todas las noticias que hablaban de las elecciones" pues hace falta especificar claramente qué criterios debe cumplir un texto para que sea seleccionado y esté en el corpus y otro no.

4.5 ¿Es mi corpus comparable?: la normalización del corpus

En algunos casos ocurre que, efectivamente, el investigador tiene que reconocer que, por las razones que sea, las unidades que constituyen su corpus son distintas en aspectos que le preocupan, porque de ser así, ello impide saber si los resultados de las comparaciones y contrastes son datos significativos o no. Por ejemplo, si tengo tres entrevistas a presidentes, pero una de ellas es mucho más larga que las otras, el hecho de que aparezcan más adversativas o concesivas en una de ellas puede no ser significativo, ¿por qué?, pues porque esa diferencia puede obedecer simplemente a la extensión, porque si las extensiones de los discursos fueran similares a lo mejor el número de fenómenos también lo sería. En efecto, este cuestionamiento es prudente y está bien fundamentado.

En estadística existe un proceso que se conoce como **normalización de frecuencias** y se refiere a que se elabora un procedimiento que permite, por un lado, dejar fuera los comportamientos de los datos demasiado atípicos o fuera del rango (para evitar que estos falseen los resultados) y, por otro lado, realizar un ajuste bastante sofisticado que precisamente lo que hace es permitir que se puedan comparar datos entre sí cuando sabemos que es posible que haya elementos que hagan parecer que son diferentes o no comparables y que no pudimos controlar en la selección del corpus.

La normalización de frecuencias es un mecanismo que bien se puede aplicar a la constitución de los corpus con el objetivo de evitar que el hecho de que algún elemento que no pudimos controlar quede más representado o haga parecer que tiene más significatividad

de la que realmente tiene. Como el objetivo de este libro no es ofrecer conocimientos ni básicos ni avanzados de estadística, tan solo diremos que es necesario que, una vez que el estudiante o investigador tenga su corpus, se pregunte seriamente si, debido al tipo de análisis que va a aplicar, es necesario normalizar las frecuencias o no, y que en caso de que note que puede haber algún elemento que falsee la significatividad de los resultados del análisis, o bien lo reporte en la investigación y diga que eso no es significativo debido a que los elementos no son comparables (pues a lo mejor solo interesa destacar algo cualitativamente), o bien recurra al procedimiento de normalización de frecuencias que ya ha sido utilizado para la constitución de corpus como muestran los trabajos de Molina Salinas (2020).

4.6 La subjetividad y neutralidad en el corpus

Otra de las preocupaciones constantes de quienes trabajamos con corpus tiene que ver con una falta de comprensión de la manera en la que la constitución del corpus se relaciona con la subjetividad o la neutralidad del investigador. Anteriormente (Capítulo 1) se enfatizó que, en los enfoques que trabajan con corpus se rechaza la falsa neutralidad. De hecho, habíamos mencionado que sobre la falsa neutralidad de las ciencias sociales ya Adolfo Sánchez Vázquez (1984) ha dicho todo lo que se puede decir: la neutralidad es también una postura que quiere fingir que no hay postura, pero siempre hay una. Como bien lo explicaron Sigal y Verón (1982):

> Comencemos por la cuestión de la cientificidad. Si el tratamiento al que hemos sometido nuestro "objeto" se pretende científico (o, en todo caso, responde a nuestra concepción de la cientificidad), las razones que nos llevaron a elegir dicho objeto son, sin ninguna paradoja, perfectamente subjetivas: este trabajo tiene su origen, su único origen, en la necesidad de comprender, aunque solo fuese de manera imperfecta, parcial y provisoria, lo que ocurrió en la Argentina en 1973-74. Confrontados a este interrogante nos vimos obligados, es verdad, a remontar el curso de la historia hasta 1943. Hemos dicho comprender: en ningún momento este trabajo ha sido imaginado por sus autores como un pretexto para "expresar" sus puntos de vista a propósito del peronismo. Lo cierto es que una buena parte de la literatura sobre los fenómenos políticos nos parece de naturaleza "expresiva": con mayor o menor felicidad y talento, el autor se complace en manifestar sus opiniones y saldar cuentas (2).

Como podemos ver, el asunto de la cientificidad con la que se constituye el corpus no puede ser abordado sin reconocer que hay cierto grado de subjetividad en toda decisión. En realidad, la honestidad y solidez reside en querer ocultar esto bajo la falsa idea de la neutralidad ideológica, pero esto de ninguna manera significa (necesariamente) que se esté construi-

yendo o constituyendo un corpus a modo en el que el investigador solo va a confirmar sus propios prejuicios.

4.7 Para leer más sobre los criterios y la normalización del corpus

Carbó, T. 2001c. El cuerpo herido o la constitución del corpus en análisis de discurso. *Escritos* 23: 17-47.

—. 2002. Investigador y objeto. Una extraña/da intimidad. *Iztapalapa* 53: 15-32.

Pérez Barajas, A. E. y A. Hernández Díaz. 2020. [coords]. *Propuestas metodológicas para el trabajo y la investigación lingüística. Aplicaciones teóricas y descriptivas*. México: Universidad de Colima.

Sánchez Vázquez, A. 1984. La ideología de la neutralidad ideológica, *Ensayos marxistas sobre filosofía e ideología*, 138-164. México: Océano.

Sigal, S. y E. Verón. 1982. Perón: discurso político e ideología, *Argentina, hoy,* Rouquié, A. (comp.) Buenos Aires: Siglo XXI.

—. 1988. *Perón o Muerte. Los fundamentos discursivos del fenómeno peronista*. Buenos Aires: Hyspamérica.

Verón, E. 1987. La palabra adversativa, *El discurso político. Lenguajes y acontecimientos*, VVAA, Buenos Aires: Hachette Livre.

5. La presentación argumentada del corpus

5.1 Introducción

Finalmente, una vez que el investigador o estudiante ha logrado diseñar su corpus, tendrá todavía el largo camino para construir una metodología de análisis para ese corpus que le permita contestar sus preguntas, cumplir sus objetivos y/o rechazar o confirmar sus hipótesis; sin embargo, esto ya no es materia de este libro. Antes de terminar, destinaremos un último y breve apartado para hablar de algunos elementos con respecto a la presentación y argumentación del método de diseño de corpus de una investigación, lo que incluye algunos requisitos que, aunque rayan ya en los límites de este trabajo, no pueden ni deben ser obviados por el estudiante o investigador.

Como bien advierte Gabriela Coronado (2016) "al parecer preferimos el secreto y la complicidad" (37), pues no siempre se da cuenta detallada de cómo se ha logrado llegar a la selección de ese y no otro corpus, sin embargo, que en algunas ocasiones se prefiera el silencio no quiere decir que nadie haya dicho nada con respecto a la importancia de reportar cómo es que se ha llegado a postular que es esa la selección más pertinente y no otra.

5.2 Descripción y argumentación del proceso y los criterios metodológicos y su pertinencia

Como se vio en el capítulo anterior, tener claridad en el camino metodológico que uno recorre desde un universo (o de cero) hasta el corpus es un trabajo arduo que requiere de una fuerte reflexión y mucho trabajo de parte del investigador. Incluso, en muchos casos requiere de mucha creatividad puesto que casi nunca existe ya un camino o modo construido para hacerlo de acuerdo con la investigación particular que uno ha planteado. No obstante, igual de importante es reportar ese camino metodológico. Ahora, un trabajo de investigación (tesis, artículo, libro) es un producto, por lo que debe ser estructurado con otra lógica, por ejemplo, aunque es probable que desde el inicio de una investigación de análisis de corpus se defina el corpus, esto no tiene por qué ir al inicio del trabajo, pero sí es importante que se destine un apartado o capítulo en el que se reporte este elemento; esto dará mayor calidad al trabajo y permitirá que otros investigadores repliquen lo que uno ha hecho e incluso amplíen el conocimiento de lo observado.

Así pues, una vez que se decide en qué apartado o capítulo se abordará el asunto metodológico del diseño del corpus, el investigador o estudiante debe revisar que en él se hayan abordado todos los aspectos indicados en la Tabla 1.

Tabla 1: Algunos elementos que podemos incluir en el reporte metodológico de la construcción del corpus

Corpus constituido	1. Presentación general del corpus: señalar cuántos textos componen el corpus, señalar fechas y explicar el contexto sociopolítico en el que ocurren y descripción de las características del tipo de corpus. 2. Presentación y explicación de los criterios operativos con los que se constituyó el corpus y defensa de su pertinencia, así como de las fuentes de donde se obtuvieron sus unidades. 3. Balance de la pertinencia, congruencia y los límites que ese corpus supone con respecto a la investigación.
Corpus obtenido o recolectado	1. Presentación general del estado insistiendo en la ausencia de corpus, archivos y acervos del material que se necesita reunir en él. 2. Descripción de las características del corpus que la investigación requiere dependiendo de sus objetivos. 3. Descripción y presentación del instrumento diseñado para obtener o recolectar el corpus y explicación de su pertinencia para obtener el tipo de corpus que se requiere (incluyendo las fases de pilotaje). 4. Narración detallada de cómo se aplicó el instrumento para obtener los materiales que constituyen el corpus. 5. Presentación general del corpus obtenido (cuántos textos o unidades lo componen y en qué material o soporte se resguardaron).
Corpus retomado	1. Presentación general del corpus: indicar cuántos textos lo conforman, señalar fechas y el proyecto o investigación de donde fue tomado. 2. Presentación y explicación de los criterios operativos con los que se constituyó el corpus que estamos retomando. 3. Defensa de su pertinencia para la investigación (sustentar por qué se retoma). 4. Explicación clara y precisa (con los criterios operativos) de aquello que se recupera de ese corpus en caso de que no se retome todo el corpus y argumentación de por qué es pertinente tomar solo una selección.

Corpus replicado	1. Presentación general del estado insistiendo en la ausencia de corpus, archivos y acervos del material que se necesita reunir en el corpus. 2. Descripción y presentación del instrumento y la metodología de aplicación de este tal y como se plantearon en la metodología que vamos a replicar. 3. Narración detallada de cómo se implementó la replicación tanto del instrumento como de su aplicación para el caso concreto de nuestra investigación. 4. Presentación general del corpus (cuántos textos o unidades componen el corpus y en qué material o soporte se resguardaron).

Una vez que se haya dado cuenta de estos elementos tendremos una investigación de análisis o estudio de corpus mucho más sólida y confiable.

Ahora bien, aunque esto ya no compete a los objetivos de este libro, es necesario decir que, dentro de los requerimientos metodológicos que supone el trabajo en un análisis de corpus, se incluye también la necesidad de que una vez que se tenga el corpus de análisis, en algún apartado de la tesis o el producto de investigación, se realice la caracterización del corpus; esto implica una descripción detallada (pero general) del comportamiento, forma y características de las unidades recabadas.[28]

De la misma manera es necesario que en el apartado o capítulo metodológico se especifique claramente cuáles son las formas en que se segmentará y numerará cada una de las unidades que componen el corpus (sí es que se van a segmentar o dividir), así como el nivel y la unidad de análisis con los que se trabajará.

5.3 El acceso al corpus

Finalmente, no es un asunto menor decir que en una investigación es necesario que el o los investigador(es) ofrezca(n) al lector el corpus de análisis, pues un análisis debe ser siempre cuestionable y comprobable. Si no ofrecemos el corpus completo, estamos pidiendo a los lectores que nos crean. Es cierto que puede haber investigaciones cuyos corpus son demasiado amplios, sin embargo, actualmente existe la posibilidad de que el corpus se adjunte en un CD o se ponga a disposición de los lectores vía electrónica.

[28] En caso de que lo que se quiera hacer sea un análisis del discurso, es necesario que se haga una caracterización discursiva.

El investigador o estudiante no debe perder de vista que reunir esos materiales, organizarlos y prepararlos (transcribirlos, segmentarlos, etc.) es parte del trabajo que ha realizado en la investigación por lo que es una parte de las aportaciones que se hacen. Por ello es importante que el investigador se pregunte si vale la pena organizar y poner su corpus en algún espacio en donde se pueda consultar ampliamente, pues es esta una manera de que se amplíe el conocimiento y trabajo en nuestras disciplinas (claro, siempre que se rinda el debido crédito a quien construyó el corpus que se retoma para una nueva investigación).

Independientemente de esto, en la investigación se debe anexar (por escrito o digitalmente) el corpus íntegro, y, si así se considera, también se deberían agregar, en archivos separados, los materiales con el corpus segmentado para su análisis y trabajados en tablas, matrices, etc. Esto siempre da mayor soporte y solidez a nuestros trabajos.

Con esto hemos logrado decir al menos lo básico que ha de plantear y reflexionar cualquiera que desee elaborar una investigación con análisis de corpus. No hay manera de dominar este ejercicio si no es haciéndolo una y otra vez, realizando los ejercicios propuestos y levantándose ante cada una de las caídas que todos los que trabajamos con corpus sufrimos antes de poder llegar a tener ese cuerpo que reclamamos como el más adecuado y pertinente para nuestra investigación. Una vez que lo tengamos, habrá que recargar las fuerzas, hemos vencido uno de los retos, pero nos queda por delante la construcción de una metodología, soportada teóricamente, que le haga justicia a ese corpus que hemos diseñado con tanto cuidado y esmero.

5.4 Para leer ejemplos de reportes de metodología en la constitución de corpus

5.4.1 Sobre constitución de corpus en investigación de Análisis del discurso

Gómez Gordillo, L. A. 2021. *El discurso presidencial mexicano ante el Congreso Estadounidense (1947-2010): agentividad en la construcción de la relación México-Estados Unidos*. Tesis de maestría, Universidad Nacional Autónoma de México.

5.4.2 Sobre la constitución y retomado de corpus para análisis lingüísticos

Ángeles Chargoy, J. I. 2019. *Imágenes de candidatos políticos en diferentes periódicos mexicanos: hacia su caracterización*. Tesis de maestría, Universidad Nacional Autónoma de México.

Grave Aragón, A. D. [en prensa]. *Análisis sociopragmático de la conversación escrita basado en el interaccionismo. El caso de los malentendidos en WhatsApp*. Tesis de maestría, Universidad Nacional Autónoma de México.

5.4.3 Sobre la constitución de corpus para obtención de datos en enfoques experimentales

Gil Carrillo, I. O. 2019. *Variedad escalar, acceso léxico y contexto lingüístico: una aproximación experimental a la derivación de implicaturas escalares*. Tesis Doctoral, Universidad Nacional Autónoma de México.

Rincón Hernández, I. M. 2017. *Procesamiento de los verbos frasales con una marca aspectual en aprendientes adultos de inglés como segunda lengua*. Tesis de maestría, Universidad Nacional Autónoma de México.

Villaseñor García, G. K. 2012. *El efecto del silencio en la interpretación de respuestas a peticiones: un estudio de pragmática experimental*. Tesis de maestría, Universidad Nacional Autónoma de México.

5.4.4 Sobre el trabajo con corpus multimodales

Mendoza Cruz, C. E. [en prensa]. *La codificación de género a través de los recursos lingüísticos en el stand up mexicano.* Un estudio de caso. Tesis de maestría, Universidad Nacional Autónoma de México.

Bibliografía

Aarts, J. y W. Meijs (eds.). 1986. *Corpus Linguistics II*. Ámsterdam: Rodopi B.V.

Aarts, J., de Haan, P. y Oostdijk, N. (eds.). 1993. English Language Corpora: Design, Analysis and Exploitation. Papers from the Thirteenth International Conference on English Language Research on Computerized Corpora, Neijmen 1992. Ámsterdam: Rodopi.

Aijmer, K. y B. Altenberg (eds.). 1991. *English Corpus Linguistics*. Londres: Longman.

Alvar Ezquerra, M. y Villena Ponsoda, J. A.1994. Estudios para un Corpus del Español. Anejo *Analecta Malacitana*. *Revista de la Sección de Filología de la Facultad de Filosofía y Letras* 7, Universidad de Málaga: Grafur.

Ángeles Chargoy, J. I. 2019. *Imágenes de candidatos políticos en diferentes periódicos mexicanos: hacia su caracterización*. Tesis de Maestría, Universidad Nacional Autónoma de México.

Antúnez Piedra A. y Mateo Ruiz, M. 2016. Mecanismos de divulgación en un corpus multimodal de noticias de contenido económico, *e-AESLA Revista Digital*, No. 2, pp. 117-127.

Armstrong, S. (ed.) 1994. *Using Large Corpora*. Cambridge: mit Press.

Atkins, B., Clear, J. y Ostler, N. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7 (1): 1-16.

—. 1992. Corpus design criteria. *Literary and Linguistic Computing*. Journal of the Association for Literary and Linguis-tic Computing 7(1): 1-16.

Bajtín, M. 1995. *Estética de la creación verbal*. México: Siglo XXI.

—. 2005. *Problemas de la poética de Dostoievski.* México: fce.

Baker, M. 1993. Corpus Linguistics and Translation Studies — Implications and Applications, *Text and Technology*: In honour of John Sinclair, M. Baker, G. Francis y E. Tognini-Bonelli (eds.), 233-252. Ámsterdam: John Benjamins Publishing Company.

Baker, P. 2006. *Using Corpora in Discourse Analysis*. Londres y Nueva York: Continuum.

Barlow, M. 1996. Corpora for Theory and Practice. *International Journal of Corpus Linguistics* 1 (1): 1-38.

Barnbrook, G. 1993. *The Automatic Analysis of Dictionaries: Parsing Cobuild Dictionaries,* M. Baker, G. Francis, y E. Tognini-Bonelli (eds.): 313-331.

Baud, R. *et al*. 1998. Extracting Linguistic Knowledge from an International Classification, *Division of Biomedical Informatics*, Nashville: Vanderbilt University. Documento disponible en la red.

Baugh, S., Harley, A. y Jellis, S. 1996. The Role of Corpora in Compiling the Cambridge Dictionary of English. *International Journal of Corpus Linguistics*, 1 (1): 39-60.

Beaugrande, R.A de y Dresssler, W.U. 1997. *Introducción a la Lingüística del texto*. Barcelona: Ariel

Beckmann, F. y G. Heyer (eds.). 1993. *Theorie und Praxis des Lexikons*. Berlin: Walter de Gruyter.

Benjamins, V., Fensel, D. y Gómez, A. 1999. Knowledge Management Through Ontologies. Documento disponible en http://www.aifb.uni-karlsruhe.de/WBS/broker/inhalt-wp.

Benveniste, E. 1997. *Problemas de lingüística general*. Vol. 1. México: Siglo XXI Editores

Bernal Lengómez, J. 1985. *En torno a la lingüística textual*. Madrid: Centro Virtual Cervantes.

Bernárdez, E. 1982. *Introducción a la lingüística del texto*. Madrid: Espasa-Calpe.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243-257.

—. 1993a. «Using register-diversified corpora for general language studies». En *Computational Linguistics,* 19/2, pp. 219-243.

Blásquez Martínez, L. e Ignacio López Moreno. 2016. *Guía para la investigación cualitativa: etnografía, estudio de caso e historia de vida*. R. Güereca Torres (coord.). México: uam.

Borja Albi, A. 2007. Corpora for Translators in Spain. The cdj-gitrad Corpus and the gentt Project, *Incorporating Corpora: The Linguist and the Translator*, G. Anderman y M. Rogers (eds.), 243-265. Clevendon: De Gruyter.

Briz Gómez, A. y Albelda Marco, M. 2009. Estado actual de los corpus de lengua española hablada y escrita, *El español en el mundo*, Instituto Cervantes, (ed.)165-225. Madrid: Instituto Cervantes.

Cabré, M.T. 2007. Constituir un corpus de textos de especialidad: condiciones y posibilidades, *Les corpus en lin-guistique et en traductologie*, M. Ballard y C. Pineira-Tresmontant (eds.), 89-106. Arras: Artois Presses Université.

Carbó, T. 2001a. La constitución del corpus en análisis del discurso. *Escritos. Revista del Centro de Ciencias del Lenguaje* 23: 17-47, disponible en http://cmas.siu.buap.mx/portal_pprd/work/sites/escritos/resources/LocalContent/31/1/carbo.pdf

—. 2001b. Tocar el lenguaje con la mano. Experiencias de método. *Revista Latinoamericana de Estudios del Discurso* 1(1): 43-67.

—. 2001c. El cuerpo herido o la constitución del corpus en análisis de discurso. *Escritos* 23: 17-47.

—. 2002. Investigador y objeto. Una extraña/da intimidad. *Iztapalapa* 53: 15-32.

—. 2004. Protocolos de investigación en análisis de discurso y consolidación del campo disciplinario. *Discurso, teoría y análisis* 26: 121-30.

—. (ed.) 2007. Corpora, conceptos y métodos en análisis de discurso. *ELA. Estudios de Lingüística Aplicada,* 46 monográfico. México: UNAM, Centro de Enseñanza de Lenguas Extranjeras.

—. 2016. Introducción. La elocuencia de los cuerpos. *Estudios de Lingüística Aplicada, 0*(46), pp. 13-30.

Carbó, T. y E. Salgado. 2013. El itinerario de un corpus multimodal para escrutar el desempeño presidencial reciente en México (2006-2012), *Estudios del discurso en América Latina. Homenaje a Anamaría Harvey*, 527-550. Bogotá: aled.

Castillo Rodríguez, C., Díaz Laje, J.M. y Rubio Martínez, B. 2020. Compilación y análisis de un corpus de estudiantes etiquetado: un estudio basado en corpus de usos de adjetivos. *Círculo de Lingüística Aplicada a la Comunicación*, *81*, 115-136. https://doi.org/10.5209/clac.67932

Charaudeau, P. y D. Maingueneau. 2005. *Diccionario de análisis del discurso*. Buenos Aires: Amorrortu Editores.

Coronado, G. 2016. El *corpus* del delito: la cultura como hipertexto, *ela*, núm. 46, pp. 33-61.

Coseriu, E. 1955. Determinación y Entorno: Dos problemas de una lingüistica del hablar, *Romanistisches Jahrbuch*, vol. 7, no. 1, 1955, pp. 29-54.

Cruz Bueno, E. 2016. *Sola contra el mundo, pero no indefensa: un estudio de caso de la construcción de los frames identitarios de madre, pareja y estudiante: trabajadora en tres historias de vida de mujeres madres solteras cabezas de hogar (msch) en el Distrito Federal*, tesis de maestría, Universidad Nacional Autónoma de México.

Curcó, C. 2021. *Semántica. Una introducción al significado lingüístico en español.* Routledge: New York.

Couthard, M. (ed.) 1994. *Advances in Written Text Analysis*. Londres: Routledge.

Fonte, I. 2008. Analizar un caso histórico en un *corpus* de discursos periodísticos: Cuba y Estados Unidos (1906-1921), *ela,* núm. 46, pp. 63- 82.

Foucault, M. 2005. *Vigilar y castigar. Nacimiento de la prisión*, trad. Aurelio Garzón del Camino, Bs. As., Siglo XXI.

Franco Trujillo, E. D. y Molina Salinas, C. 2020. "Una metodología para la elaboración de un corpus con fines lexicológicos: el caso del proyecto galea" en *Propuestas metodológicas para el trabajo y la investigación lingüística. Aplicaciones teóricas y descriptivas*, Pérez Barajas, E. y Hernández Díaz, A. [coords.] Universidad de Colima: México, pp. 779-797.

Galindo Flores, A. C. 2023. *Construcciones con verbos de apoyo y nombres fisiológicos en español*, Tesis de maestría, Universidad Nacional Autónoma de México.

Gil Carrillo, I. O. 2019. *Variedad escalar, acceso léxico y contexto lingüístico: una aproximación experimental a la derivación de implicaturas escalares*. Tesis Doctoral, Universidad Nacional Autónoma de México.

Glaser, B. G. y Strauss, A. L., 1967, *The Discovery of Grounded Theory. Strategies for Qualitative Research*. Chicago, Aldine.

Gómez Gordillo, L. A. 2021. *El discurso presidencial mexicano ante el Congreso Estadounidense (1947-2010): agentividad en la construcción de la relación México-Estados Unidos*. Tesis de Maestría, Universidad Nacional Autónoma de México.

Grave Aragón, A. D. [en prensa]. *Análisis sociopragmático de la conversación escrita basado en el interaccionismo. El caso de los malentendidos en WhatsApp*. Tesis de Maestría, Universidad Nacional Autónoma de México.

Gregorio de Mac, M. I. d., y Rébola de Welti, M. C. 1992. *Coherencia y cohesión en el texto* (1.ª ed.). Buenos Aires: Plus Ultra.

Gutiérrez, S., L. Guzmán y S. Sefchovich. 1988. Discurso y Sociedad. *En Hacia una metodología de la reconstrucción*, capítulo IX. México: Porrúa-unam.

Harris, Z. 1952. Discourse Analysis. *Language* 28 (1): 1-30.

Huffschmid, A. 2016. De los cuerpos al *corpus*. Una experiencia de investigación en torno al discurso zapatista y sus ecos en el mundo, *ela*, núm. 46, pp. 83-114.

Johansson, S. 1998. On the role of corpora in cross-linguistic research, *Corpora and cross-linguistic research: Theory, meth-od, and case studies*, S. Johansson y S. Oksefjell (eds.), 3-24. Ámsterdam: Rodopi B.V.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. Londres y Nueva York: Longman.

Kress, G. y T. Van Leeuwen. 2001. *Multimodal discourse. The modes and media of contemporary communication*. Londres y Nueva York: Bloomsbury Academic.

—. 2003 [1998]. Front Pages: (The Critical) Analysis of Newspaper Layout, *Approaches to media discourse*, A. Bell y P. Garret (eds.), 186-219. Reino Unido: Blackwell Publishers.

Kock, J. de 2001. *Lingüística con corpus. Catorce aplicaciones sobre el español*. Salamanca: Universidad de Salamanca.

Llisterri, J. y J. Llisterri. 1999. Diseño de corpus textuales y orales, *Filología e informática. Nuevas tecnologías en los estudios filológicos*, J. M. Blecua *et al*. (eds.), 45-77. Barcelona: Editorial Milenio.

Martín Peris, E., *et al.* 2004. *Diccionario de términos clave de ele*. Madrid: sgel. Disponible en: http://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/indice.htm

McEnery, T. y A. Wilson. 1996. *Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics*. Edimburgo: Edinburgh University Press.

—. 1996. Corpus Linguistics. Edinburgh: Edinburgh University Press.

Mendoza Cruz, C. E. [en prensa]. *La codificación de género a través de los recursos lingüísticos en el stand up mexicano.* Un estudio de caso. Tesis de Maestría, Universidad Nacional Autónoma de México.

Meyer, M. 2003. Acción y texto: para una comprensión conjunta del lugar del texto en la (inter)acción social, el análisis mediato del discurso y el problema de la acción social, *Métodos de análisis crítico del discurso*, R. Wodak y M. Meyer (eds.), 205-266. Barcelona: Gedisa.

—. 2003b. Capítulo 2 entre la teoría, el método y la política: la ubicación de los enfoques relacionados con el acd, pp. 35-59 en Wodack, R. y Meyer M. Comps 2003. *Métodos de análisis crítico del discurso*. Barcelona: Ariel.

Modesto Torres, L. A. 2023. Polémica Pública en el Proceso Constituyente de la Ciudad de México. Tesis Doctoral. Universidad Autónoma Metropolitana.

Murgunova, E. 2013. La lingüística del texto, *De la lingüística científica a la lingüística textual*, H. Ocaña Dayar et al (eds.). La Habana: Editorial Pueblo y Educación.

O' Halloran, K. 2012. Análisis del discurso multimodal. *Revista Latinoamericana de Estudios del Discurso*, 12(1), 75-97.

Ong, W. 1997. *Oralidad y escritura. tecnologías de la palabra*, [traducción de Angélica Scherp], Fondo de Cultura Económica, México, 1987, 2da. imp. 1997

Oostdijk, N y P. de Haan (eds.). 1994. *Corpus-based Research into Language. In Honour of Jan Aarts*. núm. 12. Ámsterdam: Rodopi B.V.

Pérez Alvarado, P. 2022. *Malentendidos del lenguaje metafórico en aprendientes de español como segunda lengua,* tesis de maestría. Universidad Nacional Autónoma de México.

Pérez Barajas, A. E. y A. Hernández Díaz. 2020. [Coord]. *Propuestas metodológicas para el trabajo y la investigación lingüística. Aplicaciones teóricas y descriptivas*. Colima: Universidad de Colima.

Rébola de Welti, M.C y G. de Mac M. I. 1997. *Coherencia y Cohesión en el texto*. Buenos Aires: Plus Ultra.

Rincón Hernández, I. M. 2017. *Procesamiento de los verbos frasales con una marca aspectual en aprendientes adultos de inglés como segunda lengua*. Tesis de Maestría, Universidad Nacional Autónoma de México.

Salgado Andrade, E. 2007. Un corpus discursivo para entender el presidencialismo en México. *ela*, núm. 46, pp. 149-175.

Salgado López, M. 2016. *Redes léxico-semánticas y la construcción identitaria en los discursos de toma de posesión presidencial de Cárdenas a Peña Nieto*. Tesis de Maestría, Universidad Nacional Autónoma de México.

—. 2020. *Presidencialismo, tecnocracracia y modernidad: Construcción ecoica de aliados y enemigos en entrevistas a Díaz Ordaz, Ernesto Zedillo y Peña Nieto*. Tesis Doctoral, Universidad Nacional Autónoma de México.

Sánchez, A. 1995. *cumbre: Corpus Lingüístico del Español Contemporáneo. Fundamentos, Metodología y Aplicaciones*. Madrid: sgel.

Sánchez Martín, F. J. 2018. Corpus especializado para el estudio de la lengua de la geometría española del siglo xvii, *Philologica Canariensia,* núm. 24, pp. 113-130.

Sánchez Vázquez, A. 1984. La ideología de la neutralidad ideológica, *Ensayos marxistas sobre filosofía e ideología*, 138-164. México: Océano.

Schifrin, D., D. Tanen y H. Hamilton (eds.). 2001. *The Handbook of Discourse Analysis*. Malden: Blackwell.

Scollon, R. 2003 [2001]. Acción y texto: para una comprensión conjunta del lugar del texto en la (inter)acción social, el análisis mediato del discurso y el problema de la acción social, *Métodos de análisis crítico del discurso*, R. Wodak y M. Meyer (eds.), 205-266. Barcelona: Gedisa.

Shiro, M. 2012. El método tampoco viene del aire. *Revista Latinoamericana de Estudios del Discurso* 12(2): 3-6. en aled.

Sigal, S. y E. Verón. 1982. Perón: discurso político e ideología, *Argentina, hoy,* Rouquié, A. (comp.), Buenos Aires: Siglo XXI.

—. 1988. *Perón o Muerte. Los fundamentos discursivos del fenómeno peronista*. Buenos Aires: Hyspamérica.

Sinclair, J. M. (ed.) 1987. *Looking up: an Account of the cobuild Project in Lexical Computing.* Londres-Glasgow: Collins.

—. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

—. 1992. Trust the Text, *Advances in Systemic Linguistics*, pp. 5-19., M. Davies y L. Ravelli (eds.), 5-19. Londres: Pinter.

—. 1996. *Preliminary Recommendations on Corpus Typology. eagles Document* eag-tcwg-ctyp/p, disponible en http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Sinclair, J. M., Payne, M. J. y Pérez, Ch. (eds.). 1996. Corpus to Corpus: A Study of Translation Equivalence. *International Journal of Lexicography* 9 (3).

Strauss, A. L. (1987): Qua/ítative Ana/ysis for Social Scientists. Cambridge: Cambridge University Press.

Strauss, A. L., 1987, *Qualitative Analysis for Social Scientists.* Cambridge, Cambridge University Press.

Teubert, W. 1996. Comparable or Parallel Corpora? *International Journal of Lexicography* 9 (3): 238-265.

Tognini-Bonelli, E. 1996b. *Corpus Theory and Practice*. Birmingham: twc.

Torruella, J. y Llisterri, J. (1999). "Diseño de corpus textuales y orales" En Blecua, J. M., Clavería, G., Sánchez, G. y Torruella, J. (eds). *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e informática, Departamento de Filología Española, Universidad Autónoma de Barcelona, Editorial Milenio.

Van Dijk, T. 1978. *La ciencia del texto*. Barcelona: Paidós. Wikipedia, diversas páginas de internet

—. 1980. *Estructuras y funciones del discurso. Una introducción interdisciplinaria a la lingüística del texto y a los estudios del discurso*. México: Siglo XXI Editores.

—. 1999. *El análisis crítico del discurso*. Barcelona: Antropos

Van Leeuwen, T. 2005. *Introducing Social Semiotics*. Londres: Routledge.

Verón, E. 1971. Ideología y comunicación de masas: la semantización de la violencia política, *Lenguaje y comunicación* social, VVAA, 91-133. Buenos Aires: Ediciones Nueva Visión.

—. 1987a. La palabra adversativa, *El discurso político. Lenguajes y acontecimientos*, VVAA. Buenos Aires: Hachette Livre.

—. 1987b [1981]. *Construir el acontecimiento. (Los medios de comunicación masiva y el accidente en la central nuclear de Three Mile Island*). Barcelona: Gedisa.

Villaseñor García, G. K. 2012. *El efecto del silencio en la interpretación de respuestas a peticiones: un estudio de pragmática experimental*. Tesis de Maestría, Universidad Nacional Autónoma de México.

Wenger, E. 1998. *Communities of practice*. Cambridge: Cambridge University Press.

Williamson, R. 2007. El diseño de un corpus multimodal. *ela. Estudios de Lingüística Aplicada* 46: 207-31. http://mexico.cnn.com, www.jornada.unam.mx, www.presidencia.gob.mx

Wodack, R. y Meyer M. [comps]. 2003. *Métodos de análisis crítico del discurso*. Barcelona: Ariel.

La constitución del corpus,

fue editado en diciembre de 2023 por el Posgrado en Lingüística de la
Universidad Nacional Autónoma de México,
siendo coordinadora Carmen Curcó.
La composición tipográfica,
en tipos Garamond Premier Pro de 16:13,
11:14, 10.5:14 y 9:10.8 puntos
fue realizada por Diego García del Gállego.
La edición estuvo al cuidado de Carmen Curcó y Melanie Salgado López.
El diseño de portada es obra de Diego García del Gállego
y la ilustración de la portada es de Rosa María C. Dies.

LA CONSTITUCIÓN DEL CORPUS

Algunas reflexiones teórico prácticas para investigaciones con análisis de contenido o de corpus

El objetivo de este trabajo es ofrecer un material, introductorio y básico, acerca de las reflexiones teórico-prácticas que son fundamentales en el proceso de selección o diseño de un corpus (sin importar si este es de manifestaciones del lenguaje verbal escrito u oral o de otra naturaleza multimodal o semiótica) para una investigación. Del mismo modo, el lector podrá acercarse a una revisión panorámica y general tanto de los enfoques teóricos como de algunos ejemplos que se relacionan con la selección del corpus. En todos los casos se hace énfasis en las correspondencias que este proceso entabla con el diseño de la investigación misma.

En ese sentido, el libro pretende funcionar como un material de acercamiento que guíe a estudiantes, profesores e investigadores para identificar los retos fundamentales de diseño metodológico de las investigaciones que apuestan por aproximarse a un fenómeno mediante el análisis de contenido o de corpus (se sitúen o no dentro de la Lingüística), el tipo de congruencia y concordancia que hay que cuidar en estos enfoques y algunas definiciones, posturas teóricas y ejemplos que ayudan para la comprensión y reflexión de este trabajo. Por ello mismo, este material permite conocer el panorama necesario para el diseño de un corpus, así como para desarrollar una postura crítica, alimentada por el conocimiento de fuentes reconocidas, ante las tareas que el investigador o estudiante debe enfrentar. Con el objetivo de que tal cometido se cumpla, en el trabajo se encontrarán nociones, definiciones, ejemplos, reflexiones y un listado de fuentes en el que se puede profundizar en los aspectos abordados.

La Colección **Breviarios de Lingüística** publica textos breves sobre temas selectos de lingüística, así como estudios específicos recientes con una dimensión didáctica. Se propone difundir propuestas académicas generadas en nuestro programa, pero también recibe trabajos externos.