Lingüística de corpus

Diana Alejandra Hincapié Moreno
Julio Alexánder Bernal Chávez

Instituto Caro y Cuervo

Lingüística de corpus

Diana Alejandra Hincapié Moreno
Julio Alexánder Bernal Chávez

Instituto Caro y Cuervo


VERITAS • LIBERABIT • VOS
INSTITUTO CARO Y CUERVO

Julio Alexánder Bernal Chávez

Diana Alejandra Hincapié Moreno

Lingüística de corpus

ISBN 978-958-611-372-4 (e-Book) .

INSTITUTO CARO Y CUERVO
SEDE CASA DE CUERVO
Calle 10 4–69, Bogotá

IMPRENTA PATRIÓTICA
Sede Yerbabuena
Autopista Norte, km 9, 300 m

Contenido

Introducción

> Los avances en las ciencias del lenguaje y sus interdisciplinas deben beneficiarse del uso adecuado de las evidencias empíricas provenientes de diversas fuentes (protocolos de verbalización, textos originales, elicitación de datos, técnicas estadísticas, mecanismos introspectivos, etc.); aún más, mayor robustez se conseguirá si se emplea más de un medio de aproximación al fenómeno en indagación. La información concurrente recolectada así fortalece y provee resultados certeros que justifican el desarrollo acumulativo del conocimiento científico (Parodi, 2008, p. 94).

Hasta 1970, la lingüística de corpus (LC) la estudiaba únicamente un reducido número de investigadores y académicos, e incluso se utilizaba casi de manera exclusiva para el análisis de la lengua inglesa; pero con el paso del tiempo, los cambios de paradigmas lingüísticos y la incursión de la tecnología en el campo de las ciencias humanas, la LC se ha constituido hoy día en una metodología lingüística en auge y de gran valor, en virtud de las facilidades que brinda para recolectar, sistematizar, analizar y explotar muestras de lengua real o en uso.

Aunque existe bibliografía sobre la LC, la mayoría de esta se encuentra en inglés o se ha escrito con base en la experiencia de investigadores españoles, pero al ser una metodología joven es largo el camino teórico y práctico que queda por recorrer. Pensando en este camino, con el presente libro se busca delimitar un área poco estudiada hasta el momento en Latinoamérica, brindándoles a sus lectores herramientas que les permitan comprender la metodología, reflexionar sobre esta y aplicarla.

La escritura de este libro nace dentro del proyecto de investigación del Grupo de Lingüística de Corpus del Instituto Caro y Cuervo (ICC). A lo largo de su historia, el Instituto ha desarrollado investigaciones sobre lengua española que, por su magnitud e importancia, se deben preservar, divulgar y explotar; tal es el caso del *Atlas lingüístico y etnográfico de Colombia*[1], los estudios del habla culta[2] y *El español hablado en Bogotá*[3]. Querer cumplir estos objetivos nos llevó casualmente a la lingüística de corpus, ya que reconocemos en esta las posibilidades para preservar, digitalizar, almacenar, sistematizar, explotar y poner al servicio del público académico y general los materiales de las investigaciones.

Sin embargo, para poder hacer uso de esta metodología y preservar los

archivos resultantes de las investigaciones del Instituto era necesario conocerla a fondo, lo que implicaba una formación teórica antes de la práctica. Es así como en 2013, pensando en la apertura de este grupo de investigación y la preservación del material del ICC por medio de la creación de corpus lingüísticos, se comenzó con una indagación sistemática sobre la bibliografía existente en LC, las universidades, facultades y grupos de investigación que trabajan con esta metodología en el mundo, las publicaciones dedicadas al tema y los corpus existentes.

Una vez recopilada esta información y con una base de datos constantemente alimentada, hubo necesidad de explorar textos teóricos, para lo cual durante ocho meses de lectura se extrajeron citas y se hicieron comentarios de diferentes textos sobre la LC. De manera paralela, se desarrollaba en el grupo de investigación el mismo proceso de formación teórica y lectura en el campo de lingüística computacional, lo que facilitaba la discusión de conceptos y la aclaración de dudas.

A finales de 2013, ya culminado el proceso de lectura, se elaboró la macroestructura de un artículo que reflejaría el estado del arte de la LC, pero con una característica especial: una visión sobre la perspectiva de esta metodología en y sobre lengua española. El proceso de escritura comenzó en 2014, con revisión y retroalimentación constantes por parte de los autores, en las que los comentarios y anotaciones iban y venían capítulo por capítulo; se culminó con la escritura, ya no de un artículo, sino de un libro.

La obra contiene en forma general los siguientes temas: definición de la lingüística de corpus; definición, características y tipología de los corpus; historia de la LC; usos de los corpus lingüísticos; creación de corpus, y relación entre la LC y la lengua española.

La idea es que este libro permita a estudiantes, profesores e investigadores aproximarse de un modo sencillo y claro a la lingüística de corpus, con el propósito de comenzar a emplearla en las investigaciones lingüísticas en el país y en Latinoamérica, además de pensar y construir corpus representativos de las diferentes variedades del español e incluso de las lenguas aborígenes americanas.

Estudiar la LC y construir corpus es una tarea que no solamente atañe a lingüistas. Es una labor interdisciplinaria que permite construir conocimiento desde diversas perspectivas y que incluso involucra a entidades gubernamentales e industriales; gubernamentales, de manera que posicionen el país y la lengua española, dadas sus características culturales y demográficas, por medio de

recursos lingüísticos como los corpus y con recursos académicos como la producción científica, resultado de investigaciones basadas en LC, e industriales, en cuanto a la necesidad de crear herramientas informáticas, material didáctico y diccionarios, entre otros, basados en lingüística de corpus.

Para cerrar, cabe tener en mente el enunciado de Mar Cruz Piñol: “El trabajo con corpus repercute en las aplicaciones de la lingüística, en la metodología de la investigación y en los propios fundamentos teóricos del estudio del lenguaje” (2012, p. 28). Esperamos que tras la lectura del presente libro, todo lector pueda reconocer el impacto que esta metodología tiene, y aplicar sus principios a investigaciones venideras y construcciones de corpus futuros.

---

1. Flórez et al., 1982.

2. González y Otálora, 1986.

3. Montes et al., 1998.

Definición de la lingüística de corpus

¿Qué es la lingüística de corpus? Autores como Geoffrey Leech (1991) argumentan que la LC es una teoría lingüística con base en tecnologías, mientras que Tony McEnery (2001) deja de lado esta concepción teórica del lenguaje y opta por definirla como una metodología para el análisis de la lengua, definición aceptada en el mundo académico *stricto sensu*.

La bibliografía sobre LC es amplia, especialmente en lo que se refiere a las producciones en lengua inglesa; en el caso del español y en Latinoamérica, Chile es uno de los países que más trabajos han realizado en LC. Venegas (2010, p. 26) y Parodi (2010, p. 14), investigadores chilenos, coinciden en que la lingüística de corpus constituye un conjunto de principios metodológicos apoyados en técnicas estadísticas y computacionales para estudiar datos reales de la lengua.

En nuestro caso, partimos de que la lingüística de corpus es una metodología que se encarga de sistematizar y analizar conjuntos extensos de datos orales, escritos o visuales de una o varias lenguas, ordenados con criterios lingüísticos, literarios, culturales y sociales, con el propósito de dar cuenta de la lengua en uso, valiéndose de herramientas computacionales y estadísticas que facilitan el acceso, almacenamiento y análisis de los datos desde concepciones diversas.

La LC basa su aplicación en lo siguiente:

- La lengua en uso como insumo (corpus conformados por muestras reales de lengua oral o escrita).
- El análisis sistemático de la lengua (análisis que se ajusta a un conjunto de reglas estrictas de recolección, almacenamiento y anotación).
- La posibilidad de trabajar desde un enfoque cualitativo o cuantitativo en una investigación (por ejemplo, desde las observaciones e intuiciones de los investigadores y desde resultados cuantificables, como listas de palabras).

Dadas estas tres características, la LC toma gran fuerza cuando el funcionalismo lingüístico, como reacción al generativismo, le da importancia a la función comunicativa y social del lenguaje y no se centra —tal como lo hacía el generativismo— en un solo aspecto, como la sintaxis o la explicación de

estructuras y principios del lenguaje desde la perspectiva de adquisición individual[4]. Giovanni Parodi comenta al respecto:

> Algunos de estos aspectos resultaron descuidados desde los estrechos límites del estructuralismo saussureano y del generativismo chomskiano, debido —en parte— a que el uso de la lengua (*parole* o actuación, según corresponda) era considerado demasiado cambiante e impredecible y, por consiguiente, inadecuado como objeto de ciencia. Desde la LC, con el despuntar del medio siglo XX, son muchos los lingüistas que anhelan indagar el uso lingüístico, tal como es producido, comunicado y comprendido entre hablantes/escribientes y oyentes/lectores reales y en situaciones concretas y particulares (2008, p. 97).

Además, la LC comienza a darles gran valor no solo a la lengua escrita sino también a la lengua oral, puesto que su materia prima es la lengua en uso. Adicionalmente, la inclusión de técnicas estadísticas y de herramientas computacionales para el procesamiento y el análisis de la información hace de los datos evidencia científica mucho más objetiva, pues se pasa de la intuición del investigador como única partida al análisis y la explotación de datos cuantificables, lo que lleva a la posibilidad de unir técnicas cuantitativas y cualitativas; esto permite tener un acercamiento y hacer un análisis más amplio de los datos, ya que puede cubrir varios aspectos de la lengua, desde lo formal hasta lo social.

Algunas disciplinas pueden usar la lingüística de corpus, desde diversos enfoques y con aproximaciones cuantitativas y cualitativas; ejemplo de esto son trabajos como *Metaphor in Discourse*, de Elena Semino (2008); *A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press*, de Paul Baker (2008); *British Sign Language Corpus Project*, del Economic and Social Research Council (2008), y *Corpus Method and Diagnostic Questionnaire for Chronic Pain*, de Elena Semino (2013), en los que se demuestra cómo se puede llegar a conclusiones por medio del lenguaje en áreas como el análisis del discurso, la política, la economía e incluso la medicina.

Las bases de aplicación de esta metodología, que a la vez se constituyen en ventajas, son las siguientes:

- Prioridad a la lengua en uso escrita y oral.
- Aproximación a los datos de una manera cuantitativa y cualitativa.
- Herramienta apta para diferentes disciplinas.

A propósito del tema, Tony McEnery (2014) establece, en su curso virtual

Corpus Linguistics: Method, Analysis, Interpretation, que el trabajo con corpus permite tomar como base grandes cantidades de datos, lo que muestra las tendencias de la lengua en uso; revela fenómenos o casos que serían difíciles de encontrar a simple vista o por intuición, e igualmente facilita la investigación, puesto que las herramientas computacionales ahorran tiempo y son bastante precisas.

La lingüística de corpus no siempre fue una metodología de fácil implementación. En los años cincuenta y sesenta, por ejemplo, la recolección, sistematización, anotación y análisis de datos lingüísticos demandaban mucho tiempo y capital humano, por varias razones: los procesos debían llevarse a cabo manualmente, el papel podía dañarse con facilidad, era necesario tener amplios espacios para archivar los documentos y un orden estricto para no confundirlos, los investigadores debían contar una por una las palabras de los textos para saber de qué material disponían, además de que tenían que analizar cada dato para así determinar las características semánticas, sintácticas y morfológicas de cada término[5].

Pero con la llegada de la era tecnológica, los computadores y los programas informáticos se pusieron a disposición de la LC, de tal manera que la construcción y la explotación de corpus se convirtieron en procesos más rápidos, seguros y confiables. Por esto la LC se concibe en la actualidad como *lingüística de corpus computacional*, y aunque no se use en el nombre constantemente la palabra "computacional", se da por sentado que se habla de corpus digitales, no solo por el modo en que están almacenados y presentados, sino porque los computadores, los sistemas informáticos, los *softwares* y hasta la web se convirtieron en elementos básicos para las investigaciones basadas en corpus.

Es común encontrar los términos *lingüística de corpus computacional* y *lingüística computacional de corpus*, y es más común aún creer que se refieren a lo mismo, y si bien están altamente relacionados y ambas lingüísticas hacen uso la una de la otra, son términos diferentes. Por un lado, la *lingüística de corpus computacional* toma herramientas computacionales (*hardware* y *software*) para construir y explotar corpus, mientras que la *lingüística computacional de corpus* toma los corpus desarrollados por la *lingüística de corpus computacional* para así estudiar el lenguaje natural y crear modelos lógicos aplicados a varios programas informáticos, los cuales permiten que las máquinas puedan procesar lenguaje natural y formar parte de situaciones comunicativas, como los programas de reconocimiento de voz, de procesamiento de texto y la traducción automática. Es así como podemos lograr que teléfonos móviles ejecuten tareas

por medio del reconocimiento de voz o que nuestros computadores corrijan automáticamente los textos que escribimos.

Además de cumplir como herramienta para diversas disciplinas, existen también enfoques disponibles para el trabajo con corpus[6]. Tognini-Bonelli (2001) los denomina *corpus-based* (basado en corpus) y *corpus-driven* (guiado por corpus). Giovanni Parodi plantea la siguiente explicación:

> en el primer caso, el objetivo es el manejo de un método ("basado en corpus") que permita poner a prueba categorías o ejemplificar teorías y descripciones ya formuladas […] En el segundo caso, el lingüista busca ir más allá de los ejemplos para dar sustento a sus argumentos; así, desde el enfoque "guiado por el corpus" de la lingüística de corpus, la teoría no existe de manera independiente de la evidencia (2010, p. 47).

En otras palabras, en una aproximación *basada en corpus* el investigador conoce la teoría, tiene hipótesis y lo que busca es validarlas o rechazarlas mediante los datos del corpus, en tanto que en la segunda opción, en el enfoque *guiado por corpus*, es la observación de ciertos patrones o fenómenos encontrados en un corpus la que lleva a la formulación de una o varias hipótesis, lo que no significa que una investigación no pueda valerse de ambos enfoques.

En general, la LC es una herramienta que permite recopilar, almacenar y explotar grandes cantidades de textos con información lingüística natural; además, pone al investigador en el papel de observador y analista de datos, y le da la posibilidad de valerse de herramientas informáticas que arrojan información sobre patrones lingüísticos (colocaciones, frecuencias, concordancias, etc.[7]), el enriquecimiento de los textos con información extra (procesos de anotación[8]) y el análisis de múltiples parámetros al mismo tiempo. Aunque los corpus no representan la lengua en su totalidad ni explican los fenómenos lingüísticos (tarea de los investigadores), sí contienen datos objetivos que permiten la descripción de la lengua en uso, el análisis sistemático y la posibilidad de trabajar desde diversas disciplinas.

---

4. Véase el apartado "Historia de la lingüística de corpus" para profundizar en la relación de la LC y el funcionalismo lingüístico.

5. Si bien es cierto que en la actualidad aún se hacen análisis y anotaciones manuales en corpus pequeños y por decisión de los investigadores, la mayoría de los corpus —en especial los de grandes dimensiones— se valen de herramientas informáticas y lógico-matemáticas para llevar a cabo estos procesos.

6. Para información detallada sobre el concepto de *corpus*, véase el apartado Definición de corpus.

7. Veáse Glosario.

8. Ibid.

Definición de corpus

Según el *Diccionario de la lengua española* (Real Academia Española, 2001), un corpus corresponde a un "Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación". A partir de esta definición, diferentes recopilaciones de textos podrían denominarse corpus, cualquier conjunto de datos serviría como material lingüístico para una investigación, por lo que dicho material arrojaría resultados confiables; pero esto, en términos prácticos, no es correcto; en tal sentido, es necesario aclarar que existen tres tipos de colecciones textuales:

- El archivo (informatizado).
- La biblioteca de textos (electrónicos).
- El corpus.

El archivo informatizado tiene como objetivo principal la conservación de material. Esta primera colección hace referencia a uno o más conjuntos de textos en soporte digital, con características diversas, incluyendo fechas, estructuras y temas variados[9]. Por su parte, la biblioteca de textos electrónicos[10] corresponde a una o varias colecciones de textos digitales, almacenados en un formato estándar y organizados según áreas del conocimiento humano[11] para su fácil acceso; y por último, un corpus informatizado se refiere a un conjunto de textos en formato digital, al igual que los anteriores, pero recolectados, almacenados y sistematizados de acuerdo con criterios lingüísticos.

Lo que diferencia principalmente un corpus de otras colecciones de textos son los criterios de selección y sistematización, los cuales se ven reflejados en la información que acompaña los datos lingüísticos. Los criterios pueden ser externos e internos. Los externos corresponden a información paratextual, es decir, datos que hacen referencia al marco en el que el texto se produce como forma de comunicación, conocidos también como *metadatos*[12], entre los que están los nombres de los autores e información sobre la situación comunicativa, el nivel social de los participantes, el año de producción, etc. Estos datos facilitan las tareas de recuperación de la información.

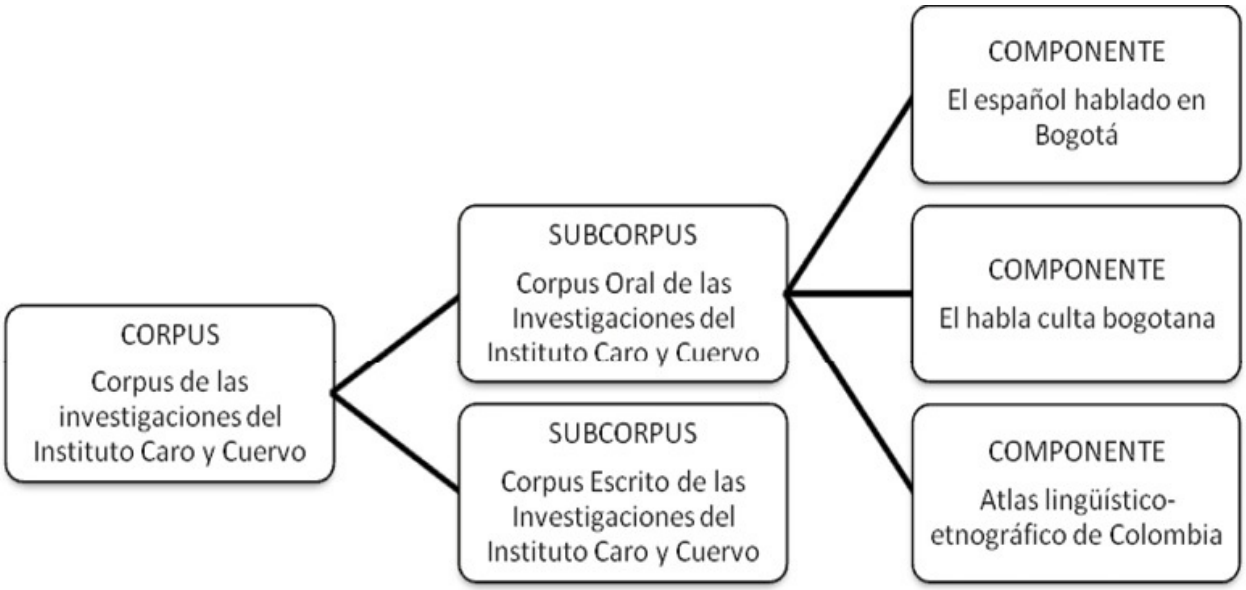

Figura 1. Corpus de las Investigaciones del Instituto Caro y Cuervo.

Los datos internos, por su parte, se refieren a elementos lingüísticos (morfemas, fonemas, lexemas o cualquier unidad o categoría lingüística), fenómenos lingüísticos en cualquier nivel de la lengua, tales como yeísmo, seseo, dequeísmo, apócope, metonimia, onomatopeya, etc., o patrones lingüísticos[13] que corresponden a estructuras lingüísticas y paralingüísticas que utilizamos para organizar el discurso, inmersas dentro de los textos. Aunque existen autores que consideran que los corpus solo deben ser objeto de análisis en sí mismos, está claro que una de las ventajas de aproximación con corpus es que permiten la inclusión de diversas disciplinas[14], tales como la sociolingüística, la pragmática, la fonética, el análisis del discurso y la semántica, lo que hace posible enriquecer los corpus mediante el uso de categorías provenientes de varias áreas del conocimiento.

A la vez, los corpus están divididos en *subcorpus* y *componentes*. Los subcorpus son las divisiones que se efectúan dentro del corpus en general; por ejemplo, un corpus denominado Corpus de las Investigaciones del Instituto Caro y Cuervo (figura 1) podría contar tanto con el Subcorpus Oral de las Investigaciones del Instituto Caro y Cuervo, como con el Subcorpus Escrito de las Investigaciones del Instituto Caro y Cuervo. Además, los corpus —y, por ende, los subcorpus— están formados por componentes, los cuales hacen referencia a colecciones de muestras de la lengua o de textos que comparten un criterio lingüístico; por citar un caso, una variedad como *El español hablado en*

Bogotá podría ser un componente del Corpus del Instituto Caro y Cuervo.

Las definiciones que se encuentran de corpus son diversas. En este caso haremos un recorrido por algunas, para así determinar las características más relevantes de los corpus y articular nuestra propia definición. Por ejemplo, Francis, Kučera & Mackie definen corpus como "[...] a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis" (1982, p. 7). Estos autores exigen que la colección de textos sea representativa. En palabras de Biber, la representatividad[15] "refers to the extent to which a sample includes the full range of variability in a population" (1993, p. 243), dejando claro que un corpus no pretende dar una visión total de una o varias lenguas, sino que busca ofrecer una muestra de ellas, o de una variedad determinada, que permita investigaciones o estudios basados en datos objetivos. Al ser los corpus representaciones y muestras reales de una lengua, pueden validar, ejemplificar o dar pie a diferentes teorías o hipótesis.

Una segunda definición dada por Sinclair hace especial énfasis en que los textos que conforman los corpus se deben producir en situaciones reales, es decir, deben ser textos naturales[16]: "[...] a corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language" (1991, p. 171). La tercera definición hace referencia a que la recolección, organización y sistematización de los datos están dadas por criterios específicos; así lo deja ver Mercado en su definición: "Colección de textos, reunidos según unos criterios precisos, eventualmente estructurados y enriquecidos con información adicional, en vista de una explotación teórica o práctica" (2008, p. 7).

Encontramos una última definición de corpus: "[...] recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con la finalidad de poder ser tratados mediante procesos informáticos y destinados a reflejar el comportamiento de una o más lenguas" (Torruela & Llisterri, 1999a, p. 7). Llama la atención que parte de la anterior definición corresponde al carácter computacional del corpus, ya que debido al tamaño de los corpus actuales se requiere que su almacenamiento sea en medios digitales y que el tratamiento y el análisis de la información se hagan mediante procesos informáticos. Por ende, la cuarta característica corresponde a la naturaleza computacional.

En términos generales, las características que permiten la definición de corpus[17] son las siguientes:

- Muestra representativa de la lengua.
- Textos producidos en situaciones reales.
- Criterios explícitos de organización.
- Naturaleza computacional.

Colecciones de textos como el *International Corpus of English*[18], *British National Corpus*[19], crea[20], Corde[21], *Corpus del español*[22] o el *British Sign Language Corpus*[23] corroboran en la práctica las características anteriormente enunciadas y sustentan que un corpus es un conjunto extenso de datos escritos, orales o visuales tomados de textos naturales y representativos de una o varias lenguas, ordenados con criterios lingüísticos, literarios, culturales y sociales, los cuales dan cuenta de la lengua en uso; almacenados, sistematizados y analizados con la ayuda de herramientas computacionales.

---

9. Para más información sobre el concepto de *Archivo*, revisar C. Martín (2009). *Temas de biblioteconomía: concepto y función de archivo. Clases de archivos. El sistema archivístico español*, y para *Archivo digital,* revisar: C. Lacombe (2011). *Archivos digitales*.

10. Véase artículo A. Sorli y A. Merlo (2000). *Bibliotecas digitales (I): colecciones de libros de acceso público*.

11. Las áreas del conocimiento humano corresponden a la división del conocimiento en materias. El sistema de clasificación del conocimiento más usado en las bibliotecas es el denominado Clasificación Decimal Universal (CDU), propuesto por Melvil Dewey. Las áreas que propone son obras generales, filosofía y psicología, religión, ciencias sociales, ciencias puras, ciencias aplicadas, arte, lengua y literatura, y geografía e historia.

12. Término desarrollado en los apartados “Características de un corpus” y “Diseño y elaboración de corpus”. Véase el Glosario.

13. Para más información sobre *patrones lingüísticos,* véase V. Soler (2007). *Patrones lingüísticos para la búsqueda de información conceptual en el corpus textual especializado de la cerámica TXTCerama*.

14. Véase el apartado “Usos de los corpus”.

15. Para más información sobre *representatividad*, véase el apartado “Características de un corpus”.

16. Los textos naturales hacen referencia a textos producidos en situaciones comunicativas reales, es decir, conversaciones, emisiones de radio, artículos científicos, novelas, etc. Para información más detallada, véase el apartado “Características de un corpus”.

17. Para obtener información más detallada sobre las características de los corpus, véase el apartado “Características de un corpus”.

18. http://ice-corpora.net/ice/.

19. http://www.natcorp.ox.ac.uk/.

20. http://corpus.rae.es/creanet.html.

21. http://corpus.rae.es/cordenet.html.
22. http://www.corpusdelespanol.org/.
23. http://www.bslcorpusproject.org/.

Características de un corpus

Las características que definen un corpus y lo diferencian de cualquier otra colección de textos son cualidades que actúan entre sí de manera complementaria y que permiten conocer las posibilidades que trae la LC para la investigación lingüística. Así las cosas, se puede decir que[24]:

1. *Un corpus es una muestra de lengua*[25]. Los corpus son porciones de lenguas o de variedades lingüísticas capaces de representar sus tendencias o características. Un corpus no puede mostrar la totalidad de una lengua, puesto que es imposible recolectar todas las producciones realizadas en un idioma, pero sí es posible almacenar textos que evidencien el comportamiento de una lengua y que se constituyan como referencia.
2. *Las muestras de un corpus son reales*[26]. Lo que un corpus busca es ser una fuente confiable, con datos que permitan el estudio de la lengua natural. Por esto, los textos que componen un corpus, ya sean orales, escritos o visuales, se deben producir en situaciones comunicativas naturales y con un propósito comunicativo auténtico, aunque para la creación de algunos corpus se registran muestras de lengua de personas con características específicas de edad, sexo y profesión, entre otras, o se llevan a cabo actos comunicativos delimitados a partir de un tema o de un contexto determinado, e incluso en otros casos se hacen pruebas monitoreadas, en las que el investigador pide la lectura de enunciados, palabras o sonidos, y los graba mientras detecta y analiza los fenómenos producidos.
3. *Los corpus relacionan la teoría y los datos*[27]. Si bien los corpus son conjuntos de textos sin conceptos, explicaciones o definiciones, sí se construyen con criterios específicos y teniendo claro de dónde se toman los textos y por qué se ha elegido esta procedencia; por ejemplo, un corpus de referencia del español, aunque compuesto por diferentes géneros, se encuentra estructurado de acuerdo con criterios específicos textuales, diatópicos y sincrónicos, entre otros, lo que hace que se

convierta en un modelo de la realidad de la lengua. Este modelo se sustenta en procesos estadísticos[28] que permiten que los datos dibujen y corroboren la estructura y el funcionamiento que en la teoría se tiene.

4. *Brindan información adicional*[29]. Una de las características de los corpus es que no solamente cuentan con los textos que los conforman, sino que además poseen información adicional que enriquece los datos. Diríamos que hay *información de datos externos e internos* y *anotación*. La información de datos externos e internos corresponde a los *metadatos*, información que identifica la procedencia y las características de los textos, y a su vez permite hacer búsquedas específicas dentro de un corpus; van desde el número de hablantes, el año de producción, la tipología textual, hasta la duración, entre otros. Por su parte, los datos internos pueden corresponder a información sobre el aspecto físico del documento, como la estructura.
5. La anotación corresponde a la inclusión de datos que buscan enriquecer el corpus con información lingüística adicional; es así como cada elemento de un corpus puede tener una etiqueta en la que se expliquen sus características fonéticas, morfológicas, léxicas, etc. La anotación no es una característica primordial de un corpus, puesto que existen corpus *no anotados* o *planos*, pero esta información adicional permite hacer búsquedas más específicas dentro de los corpus.
6. *Facilitan la extracción de datos homogéneos y cuantificables*[30]. Gran parte de la cualidad cuantitativa de los corpus está dada por el componente lógico-matemático utilizado en los procedimientos para el análisis de la información. Los corpus son una muestra de la lengua real, e incluso el número de apariciones de fenómenos lingüísticos se constituye en información relevante que se puede generalizar para la lengua o la variedad. Alguna información distribucional o estadística que se puede extraer con los procedimientos lógico-matemáticos son las *frecuencias de ocurrencias*, referidas a la frecuencia de aparición de morfemas, palabras, expresiones o patrones gramaticales, entre otros, y de *coocurrencias*, referidas a la frecuencia de aparición de estos elementos dentro de un contexto específico; por ejemplo, la locución *a pesar* puede aparecer de manera frecuente acompañada por la preposición *de*, lo que da como resultado la expresión *a pesar de*.

7. *Tienen varias posibilidades de composición*[31]. Los corpus pueden estar compuestos por materiales orales, textuales o multimodales; estos últimos son aquellos textos que recogen modalidades variadas de comunicación, como el lenguaje de señas, las grabaciones en video de situaciones comunicativas, expresiones faciales, etc. Un corpus puede contener uno, dos o más tipos de textos; esto es, hay corpus que son orales o corpus que están compuestos por textos orales y escritos o por videos y textos. Además de la composición de un corpus de acuerdo con el medio de producción, también es variada su composición según la extensión de las muestras. Partiendo de las necesidades y objetivos de la creación del corpus, los textos que este contiene pueden ser muestras completas, como un libro entero o fragmentos, esta elección se realiza cuidando también los parámetros de equilibrio[32].
8. *Su tamaño puede variar*[33]. No existe un número exacto de palabras o textos que determine el tamaño perfecto de un corpus. El tamaño está dado por los objetivos del corpus, las necesidades de cada investigación y los recursos electrónicos de los que se disponga para el almacenamiento del corpus. Si bien es cierto que una cantidad mayor de datos permite potencialmente abarcar una porción mayor de la lengua, lo que en verdad importa es que el tamaño esté pensado con base en muestras diversificadas y balanceadas, pues un corpus que no sea representativo sirve de muy poco[34].
9. *Son representativos y diversos*[35]. Se dice que un corpus es representativo, puesto que por más grande que sea no puede contener toda una lengua o variedad, pero sí puede representarla. La representatividad se refiere a la capacidad que tiene un corpus para comportarse como un modelo de la lengua, mostrando sus partes y sus tendencias, constituyéndose así como una referencia.
10. Hablar de representatividad puede tender a la subjetividad, pues dependiendo de la experiencia lingüística de cada persona puede ser o no representativo; por esto hay que estar muy atentos a los objetivos que tiene la construcción del corpus y a la variedad o lengua que se busca representar. Para abandonar un poco la visión subjetiva de esta característica, también se puede echar mano de datos estadísticos; gracias a que la representatividad está muy ligada al equilibrio[36] es posible decidir el porcentaje de los textos que componen el corpus, de acuerdo

con la realidad. Por ejemplo, si se quiere crear un corpus del español oral de Bogotá, deberían recogerse muestras de todas las zonas de la ciudad proporcionales a la población de cada zona.

11. Para que un corpus sea representativo tiene que ser, a su vez, diverso; es decir, el corpus debe contener registros o categorías textuales variadas, clasificaciones internas, ya sean temáticas, de género, disciplina o cualquier otra categoría; de esta manera, se asegura que se abarque un amplio segmento de la lengua, se alcance un mayor grado de representatividad y, por ende, resultados más confiables y generalizables. Además, en muchos estudios basados en lingüística de corpus la comparación es importante, ya que permite encontrar patrones, rasgos comunes o rasgos distintivos y estas comparaciones son posibles gracias a la diversidad de registros dentro de un mismo corpus.
12. *Deben tender al equilibrio*[37]. Con equilibrio nos referimos a recoger muestras proporcionales en tres aspectos: representatividad, variedad y tamaño. Representatividad en cuanto a que las muestras deben ser reflejo de las variedades que se encuentran en la lengua real, por ejemplo si estamos construyendo un corpus oral de referencia del español de Colombia, no sería consecuente con esta característica que el corpus contuviera un 60 % de muestras de conferencias académicas y un 40 % de muestras de conversaciones espontáneas, primero porque en la realidad de la lengua no se producen más conferencias que conversaciones espontáneas, y segundo porque también afectaríamos el segundo aspecto, el de variedad. Para que un corpus sea equilibrado en las variedades, es clave que exista una porción similar de textos en cada registro o género que conforma el corpus, esto es, que el porcentaje de prensa sea parecido al porcentaje de literatura y, a su vez, al de textos académicos. Para seguir en la línea del equilibrio, también es conveniente que las muestras sean de igual o similar tamaño, lo que significa que la mayoría de los textos contenidos en el corpus deben tener una longitud parecida o un número semejante de palabras.
13. Cuando un corpus es equilibrado, es posible explotarlo desde muchos más enfoques y para diferentes trabajos; además, facilita la comparación entre registros o géneros. En ciertos casos, el tema del equilibrio queda en la teoría, ya que no es fácil construir un corpus de tales características. Cuando esto ocurre, es importante conocer en detalle la composición del

corpus, para así extraer datos cuantitativos de un modo correcto y no presentar conclusiones erróneas.

14. *Su formato es digital*[38]. Aunque en la historia de la lingüística de corpus han existido corpus físicos y se han creado algunos por medio de procesos manuales, en la actualidad los corpus se conciben de manera digital. La digitalización de los corpus permite que aumente su tamaño, puesto que la capacidad de almacenamiento es más elevada y los procesos de sistematización y análisis más sencillos pueden ser manipulados por un mayor número de personas sin que el corpus sufra daño, y el análisis estadístico y lingüístico se puede llevar a cabo mediante la ayuda de herramientas computacionales.
15. *Los corpus han de ser de fácil acceso*[39]. La digitalización ayuda también a que los corpus estén disponibles para un grupo amplio de personas. Existen algunos como el *Corpus del español*, de Mark Davies[40], que son de libre uso y se encuentran en la web. Otros, por ejemplo, aunque cuentan con una versión *online*, requieren un registro previo, pero igual pueden utilizarse; en algunos casos hay que pagar para acceder a ellos y otros simplemente son privados.
16. Pero con fácil acceso no solo nos referimos al carácter público o privado de los corpus, sino a la facilidad de acceder a los datos por medio de diferentes programas, ya que no de mucho serviría el almacenamiento de millones de palabras cuando al acceder a ellas no se les puede aplicar ninguna forma de análisis. Por ejemplo, los programas de concordancias[41] permiten obtener listas de frecuencias de palabras o expresiones con diferentes criterios, como aparición por lema, por contexto, etc., lo cual facilita también el acceso a la información.

    En términos generales, un corpus debe constituirse como una muestra de lengua real con diferentes posibilidades de composición que relaciona la teoría y los datos, brinda información adicional a la explícita en los textos, facilita la extracción de datos homogéneos y cuantificables, no se rige por un tamaño estándar establecido, es representativo y diverso, tiende al equilibrio, es digital y de fácil acceso. Estas características hacen de los corpus fuentes de datos aptas para investigaciones lingüísticas.

24. La elección de estas características están sustentadas en los parámetros de Eagles (1996) y de Parodi (2008), quienes también proponen unas características específicas para que un corpus se considere como tal. Además de esto, cada característica cuenta con el apoyo de autores que en su momento han hablado sobre dichas cualidades.

25. Autores como McEnery & Wilson (2012) y Parodi (2008) hablan sobre esta característica.

26. Autores como Parodi (2008), McEnery & Wilson (2012) y Venegas (2010) se refieren a esta característica.

27. Autores como Venegas (2010), Torruela & Llisterri (1999) y Gries (2009) hablan sobre esta característica.

28. Véase la característica número cinco de este mismo apartado: “Facilitan la extracción de datos homogéneos y cuantificables”.

29. Autores como Parodi (2008), Rafel & Soler (2003), Gries (2009)this article se refieren a tal característica.

30. Autores como Gries (2009), McEnery & Wilson (2012), Torruela & Llisterri (1999) y Rojo (2008) hablan sobre esta característica.

31. Autores como Parodi (2010) hablan sobre esta característica.

32. Véase la característica 9 de este mismo apartado: “Deben tender al equilibrio”.

33. Autores como Parodi (2010), Rojo (2008) y Leech (1991) hablan sobre esta característica.

34. Véanse las características 8 y 9 de este mismo apartado: “Son representativos y diversos” y “Deben tender al equilibrio”.

35. Autores como McEnery, Xiao & Tono (2006), Hrušková (2008), Gries (2009), Mercado (2008), McEnery & Wilson (2012), Procházková (2006), Rafel & Soler (2003) y Parodi (2008 y 2010) hablan sobre esta característica.

36. Véase la característica 9 de este mismo apartado: “Deben tender al equilibrio”.

37. Autores como Gries (2009) y Baquero (2010) hablan sobre esta característica.

38. Autores como McEnery & Wilson (2012), Parodi (2010), Rojo (2008), Venegas (2010) y Rafel & Soler (2003) hablan sobre esta característica.

39. Autores como McEnery & Wilson (2012) y Leech (1991) hablan sobre esta característica.

40. http://www.corpusdelespanol.org/.

41. Los *programas de concordancias* corresponden a herramientas computacionales de análisis textual, que generan listas de ocurrencias de palabras que generalmente van juntas. Algunos programas son AntConc, WConcord y MicroConcord.

Tipología de los corpus

La creación de un corpus responde a diferentes objetivos o finalidades, como por ejemplo obtener información sobre una lengua en general, un periodo de tiempo específico, una variedad lingüística, cambios en la lengua, un género literario o un tema, entre otros. Estos objetivos o finalidades determinan los criterios de construcción, y por ende se constituyen en los principales parámetros para establecer tipologías de corpus.

Hablamos de tipologías y no de tipología porque en realidad no existe una sola clasificación establecida. Llisterri y Torruela (1999) presentan una tipología según el porcentaje y la distribución de los tipos de texto, según la especificidad de los textos, según la cantidad de texto que se recoge de cada documento, según la codificación y la anotación y según la documentación que acompaña los textos. Procházková (2006) habla de corpus orales, corpus multimodales, corpus de textos, corpus sincrónicos, diacrónicos, monolingües, multilingües, corpus históricos, de referencia, monitores y dialectales. Por su parte, Milka Vilayandre (2006) establece la tipología de corpus a partir de siete parámetros principales: la modalidad de la lengua, el número de lenguas a que pertenecen los textos, el tamaño o cantidad de textos que conforman el corpus, el carácter abierto o cerrado del corpus, la variedad lingüística o el grado de especialización de los textos, el período temporal que abarcan los textos y el tratamiento aplicado al corpus. A su vez, Mercado (2008) propone una tipología de los corpus según porcentaje de distribución de los diversos tipos de textos que los componen, especificidad de los textos, cantidad de textos que recogen, tipo de codificación y anotaciones añadidas al texto, y contenido.

Si bien todas las tipologías anteriormente enunciadas son válidas y logran representar los tipos de corpus existentes, en el presente texto incluimos las tipologías enunciadas en una propuesta propia, en la que se articulan varias características y se profundiza en diversos rasgos. A continuación se presenta la propuesta de tipología de corpus según siete criterios de clasificación: medio de producción de los textos, número de lenguas, especificidad de los textos, distribución de los textos, tamaño de las muestras recogidas, información extra de los textos y documentación que los acompaña.

Tipología de los corpus

Criterio de clasificación	Tipología	Subnivel
Medio de producción de los textos	Corpus escrito	
	Corpus oral	Corpus para la descripción fonética de la lengua
		Corpus para el desarrollo de tecnologías del habla
		Corpus oral
	Corpus multimodal	
Número de lenguas	Corpus monolingües	
	Corpus bilingües	Corpus bilingüe comparable
		Corpus bilingüe paralelo
		Corpus bilingüe alineado
	Corpus multilingües	Corpus multilingüe comparable
		Corpus multilingüe paralelo
		Corpus multilingüe alineado
Especificidad de los textos	Corpus general	
	Corpus especializado	
	Corpus genérico	
	Corpus canónico	
	Corpus cronológico	Corpus diacrónico o histórico
		Corpus sincrónico
Distribución de los textos	Corpus grande	
	Corpus equilibrado	
	Corpus piramidal	
	Corpus cerrado	
	Corpus abierto o monitor	
Tamaño de las muestras recogidas	Corpus textual	
	Corpus de referencia	
	Corpus léxico	
Información extra de los textos	Corpus simple	
	Corpus anotado	
Documentación que acompaña los textos	Corpus no documentado	
	Corpus documentado	

Medio de producción de los textos

Según el medio de producción de los textos que componen un corpus se puede decir que existen corpus escritos como el *Corpus diacrónico del español (Corde)*[42], corpus orales como el *Corpus oral de referencia del español contemporáneo*[43] y corpus multimodales como el *British Academic Spoken English* (*base*)[44].

Corpus escrito

Los corpus escritos —también llamados textuales[45]— están constituidos por textos o muestras de lengua escrita. Es uno de los tipos de corpus más comunes, puesto que su recolección es más sencilla en comparación con los corpus orales o multimodales, debido a que muchos textos ya están digitalizados, y de no ser así solo se requiere un proceso de escaneo por ocr[46]. Sus fuentes pueden provenir de libros, revistas, prensa, artículos, textos de internet, entre muchos otros.

Corpus oral

Un corpus de este tipo está formado por muestras de lengua oral, que corresponden a señales de voz, transcripciones y, en algunos casos, a ambas. Podemos dividir los corpus orales en corpus para la descripción fonética de la lengua; uno de los ejemplos es The Chain Corpus[47], corpus para el desarrollo de tecnologías del habla como The Carnegie Mellon Communicator Corpus[48], y corpus orales como cola o Corpus oral del lenguaje adolescente[49].

Corpus para la descripción fonética de la lengua

Estos corpus se constituyen a partir de grabaciones y transcripciones fonéticas realizadas en condiciones acústicas óptimas, y la mayoría de las veces con una preparación previa respecto al contenido de las muestras. En este tipo de corpus las grabaciones pueden ser inventarios de los sistemas fonético-fonológicos de la lengua, frases aisladas, textos leídos, habla espontánea y grabaciones de medios de comunicación.

Corpus para el desarrollo de tecnologías del habla

El objetivo de estos corpus es ayudar en el desarrollo de aplicaciones en el ámbito de las tecnologías del habla. Se construyen de acuerdo con la aplicación que se está creando, se componen por la señal sonora y algunas veces por transcripciones que permiten la elaboración de modelos estadísticos del lenguaje. Las muestras pueden provenir de sonidos aislados, inventarios de unidades

fonéticas, grabaciones específicas con números generalmente utilizadas en programas de reconocimiento de voz, habla espontánea, diálogos que ayudan a desarrollar servicios automáticos por teléfono, frases diseñadas con la aparición de ciertos sonidos y *logatomes* o palabras sin sentido, pero fonológicamente bien formadas.

Corpus oral

Este tipo de corpus oral hace referencia al que se organiza por lo regular con propósitos netamente lingüísticos. Se construye a partir de grabaciones de muestra oral o sus transcripciones, en un primer momento ortográficas. El objetivo de estos corpus es reflejar una lengua o variedad a partir de los usos de la lengua hablada, ya sea discursos, conferencias, conversaciones, habla espontánea, etc. El *Corpus oral y sonoro del español rural* (*Coser*)[50] puede ser un ejemplo claro de este tipo de corpus.

Corpus multimodal

El material que forma parte de estos corpus combina dos o más medios de producción, es decir, pueden estar constituidos por texto, sonido, imagen o video. De esta manera, los datos pueden contener información prosódica, kinésica, contextual, etc. Sus fuentes son usualmente documentales, lenguaje de señas y videoconferencias, entre otras.

Número de lenguas

Un corpus puede contener muestras de una o más lenguas, dependiendo del objetivo que tenga. Según el número de lenguas, encontramos corpus monolingües, corpus bilingües y corpus multilingües.

Corpus monolingüe

El objetivo de este corpus es dar cuenta de una lengua o una variedad lingüística. Por tal motivo, los datos o textos que lo conforman corresponden a una sola lengua.

Corpus bilingüe

Los corpus bilingües recogen muestras de dos lenguas que no necesariamente comparten criterios de selección o son traducciones. Dependiendo de estas dos situaciones, se puede hablar también de corpus bilingües comparables y corpus bilingües paralelos.

Corpus bilingüe comparable

El objetivo de un corpus de estas características es comparar el comportamiento de dos lenguas en situaciones comunicativas similares. Por esto recoge textos parecidos y con criterios de selección compartidos.

Corpus bilingüe paralelo

En el caso de estos corpus, los textos ya no solamente comparten criterios de selección, sino que corresponden a traducciones en las dos lenguas. Estos corpus son muy utilizados en el campo de la traducción.

Corpus bilingüe alineado

En un corpus bilingüe alineado encontramos, al igual que en un corpus paralelo, los textos traducidos, pero su presentación se hace de manera que los textos, párrafos y frases de una lengua aparezcan paralelos a los textos traducidos, lo que facilita el análisis y la comparación. Son de especial utilidad en contextos bilingües.

Corpus multilingüe

Los corpus multilingües contienen información de tres o más lenguas, información que no responde necesariamente a los mismos criterios de selección o a la traducción de todos los textos en las diferentes lenguas. A su vez, los corpus multilingües se dividen en corpus comparables, corpus paralelos y corpus alineados.

Corpus multilingüe comparable

Contiene información similar de tres o más lenguas, que responden a criterios de selección parecidos pero que no son traducciones.

Corpus multilingüe paralelo

Esta colección corresponde a textos con los mismos criterios de selección y traducidos en tres o más lenguas.

Corpus multilingüe alineado

Funciona de la misma manera que un corpus bilingüe alineado, solo que los mismos textos o traducciones se encuentran en tres o más lenguas. Son muy útiles en contextos multilingües, como la Unión Europea.

Especificidad de los textos

De acuerdo con la especificidad de los textos que componen un corpus se

puede decir que existen corpus generales, corpus especializados, corpus genéricos, corpus canónicos y corpus cronológicos.

Corpus general

Un corpus general recoge muestras diversas y equilibradas, para así poder representar una lengua o variedad en su totalidad y en las situaciones comunicativas más frecuentes.

Corpus especializado

Su objetivo es representar un tipo particular de lengua o un sublenguaje, como el lenguaje médico, el de niños de 4 a 10 años o el lenguaje científico.

Corpus genérico

El objetivo de este tipo de corpus es aportar datos para la descripción y comparación de un género textual específico frente a otros, motivo por el cual recopila textos pertenecientes a un solo género: poemas, ensayos, novelas, etc.

Corpus canónico

Un corpus canónico recoge todos los textos producidos por un mismo autor, sin importar el género o registro; de este modo, es el autor quien determina los textos que configuran el corpus.

Corpus cronológico

Esta clase de corpus determina su principal parámetro de conformación a partir de características temporales, con el objetivo de estudiar la lengua o una variedad dentro de un periodo específico. Entre los corpus cronológicos encontramos los corpus diacrónicos o históricos y los corpus sincrónicos.

Corpus diacrónico o histórico

Los corpus diacrónicos sirven como fuente para la observación y descripción de los cambios de una lengua o variedad a través de periodos largos y sucesivos. Por esto recogen textos que abarquen siglos, por ejemplo datos del español desde el siglo XV hasta el siglo XIX.

Corpus sincrónico

El corpus sincrónico permite el estudio de una lengua o variedad en un punto particular del tiempo, por ejemplo el *Corpus del español mexicano contemporáneo*[51], que abarca el periodo de 1921 a 1974. Por lo general, sirve para comparar variedades o lenguas y su recolección es mucho más fácil que en la construcción de un corpus histórico, puesto que se limita a una sola etapa.

Distribución de los textos

El número, el porcentaje y en general la distribución de las muestras dentro de un corpus determinan si se habla de un corpus grande, un corpus equilibrado, uno piramidal, un corpus cerrado o un corpus monitor.

Corpus grande

Se habla de corpus grandes en comparación con otros, pues no existe una cifra determinada que indique si es o no grande. El fenómeno de corpus con un número elevado de elementos se da gracias a las facilidades computacionales de almacenamiento, organización y análisis de información. Es posible que por el tamaño este tipo de corpus deje un poco de lado los parámetros de equilibrio y representatividad.

Corpus equilibrado

Este tipo de corpus recoge el mismo número o una porción similar de muestras para representar las diferentes variedades, géneros, registros, fuentes, etc.

Corpus piramidal

Un corpus piramidal se divide en distintos niveles: un primer nivel reúne pocas variedades temáticas, pero muchos textos; un segundo nivel abre un poco el abanico de las variedades temáticas, pero reduce el número de textos, y así sucesivamente.

Corpus cerrado

Un corpus cerrado tiene un tamaño definido antes de su recopilación, un tamaño ya sea en número de palabras o de textos, y al alcanzar esta cifra se da por terminado. El tamaño lo definen, de acuerdo con su criterio, quienes lo construyen.

Corpus abierto o monitor

Este es un corpus dinámico, que si bien puede tener un número fijo de elementos, como en el caso del corpus cerrado, se actualiza periódicamente, de manera que mantiene la misma cantidad de información pero ingresando datos más actuales y excluyendo datos antiguos cada cierto tiempo. En materia de representatividad, el ideal es que los datos que se ingresen tengan características similares a los datos que se desechan, aunque por la naturaleza viva de la lengua muchas veces esta premisa no se da; por ejemplo, si en la década de los ochenta la prensa física tenía una presencia lingüística muy fuerte, es posible que al

actualizar el corpus con datos del año 2000, las entradas de un blog remplacen los datos de la prensa.

Tamaño de las muestras recogidas

Aparte de la cantidad de muestras recogidas y su distribución, otro factor determinante en la tipología de los corpus es el tamaño de dichas muestras. Según este criterio, se pueden definir tres clases de corpus: textual, de referencia y léxico.

Corpus textual

Las muestras de estos corpus son los textos completos, esto es, recogen novelas, artículos, conversaciones o cualquier producción comunicativa en su totalidad.

Corpus de referencia

A diferencia del corpus textual, las muestras que conforman estos corpus corresponden a fragmentos de textos. El tamaño del fragmento no está estandarizado, sino que responde a la apreciación de quienes construyen el corpus; sin embargo, al construir un corpus de referencia se deben tener en cuenta aspectos de equilibrio y representatividad, ya que el objetivo de un corpus de referencia es proporcionar información de una lengua o una variedad de la manera más completa posible. Para que el corpus sea equilibrado y representativo se determinan el número de palabras por fragmento, el número de muestras tomadas de la misma fuente, género, registro, se seleccionan fragmentos de partes variadas del texto y se busca que la distribución sea similar, de modo que logre representar la variedad.

Corpus léxico

Al igual que un corpus de referencia, las muestras del corpus son fragmentos, pero el interés de quienes lo construyen está en el léxico, por lo cual los fragmentos tienden a ser más pequeños pero con una longitud invariable.

Información extra de los textos

Una de las características de los corpus es que brindan información adicional a la que el texto por sí solo nos puede proporcionar. Parte de esta información se

da mediante el proceso de anotación[52], y depende del objetivo, las facilidades y necesidades que tengan los investigadores a la hora de construir el corpus. Que un corpus cuente o no con información extra es un criterio tipológico que da como resultado corpus simples y corpus codificados o anotados.

Corpus simple

Los corpus simples corresponden a aquellos que no tienen ninguna información lingüística adicional, simplemente se encuentran los textos ordenados y en un formato neutro llamado *plain text* (texto simple), que permite la lectura de computadores y humanos, puesto que es solo texto sin formato, es decir, sin negrita, cursiva, fuentes o códigos adicionales.

Corpus codificado o anotado

Un corpus codificado o anotado es aquel en el que cada uno de los textos cuenta con etiquetas que contienen información adicional, ya sea sobre elementos estructurales como enunciación del título, cambio de párrafo, cambio de capítulo, lo que indica que es un corpus codificado, o con información lingüística, caso en el cual estaríamos hablando de un corpus anotado. La anotación puede realizarse contemplando diferente información, como categoría gramatical, estructura sintáctica, lema, turnos de habla y fenómenos fonéticos. La anotación y sus categorías pueden variar, dependiendo del tipo de corpus que se construye y del interés que se tiene sobre este.

Documentación que acompaña los textos

En esta clasificación se dispone de dos categorías: corpus no documentados y corpus documentados, como el *World Atlas of Language Structures*[53].

Corpus no documentado

Los textos que conforman estos corpus no cuentan con archivos relacionados como imágenes, descripciones del corpus o de sus componentes, que acompañen o amplíen de alguna manera los datos que el corpus contiene. Lo que no quiere decir que no puedan ser corpus anotados.

Corpus documentado

A diferencia del anterior, un corpus documentado vincula archivos adicionales dtd (*Document Type Definition*) para describir los componentes del texto o para entrelazar información de los datos que permita conocer más

profundamente los materiales del corpus. Estos documentos suelen describir rasgos específicos de tipologías textuales, de fenómenos contenidos en los corpus o sencillamente bibliografía relacionada.

Como nota final, vale la pena aclarar que un corpus no responde a un único criterio (medio de producción de los textos, número de lenguas, especificidad de los textos, distribución, tamaño de las muestras, información extra, documentación que acompaña los textos); sino que responde a una característica por criterio, es decir, un corpus puede ser oral, monolingüe, anotado, etc. De esta manera los objetivos de creación se sustentan los unos en los otros y el corpus resultante termina abarcando y definiendo más la variedad o lengua que representa.

---

42. http://corpus.rae.es/cordenet.html.

43. http://www.lllf.uam.es/ESP/Info%20Corlec.html.

44. http://www2.warwick.ac.uk/fac/soc/al/research/collect/base/.

45. En este caso preferimos el término *escrito* y no *textual*, ya que puede generar confusiones con los corpus que más adelante denominamos textuales y que hacen referencia a aquellos que toman textos completos para construir la colección.

46. OCR (*Optical Character Recognition*) se refiere a un proceso de digitalización de textos y conversión de estos en caracteres que pueden ser procesados por un computador.

47. http://chains.ucd.ie/corpus.php.

48. http://repository.cmu.edu/cgi/viewcontent.cgi?article=2394&context=compsci.

49. http://www.colam.org/om_prosj-espannol.html.

50. http://www.lllf.uam.es:8888/coser/.

51. http://www.corpus.unam.mx:8080/cemc/.

52. Véase el apartado “Características de un corpus” para *anotación*.

53. http://wals.info/

Historia de la lingüística de corpus

La historia de la lingüística de corpus ha sido escrita teniendo en cuenta cambios en paradigmas lingüísticos y desarrollos en el área de la tecnología computacional. Al tiempo que la LC responde a estos cambios y se expande como metodología en el área de las humanidades, los corpus construidos reciben también el impacto de estas transformaciones, lo que se ve reflejado en términos de tamaño, composición y explotación.

Los primeros trabajos basados en aproximaciones de lingüística de corpus datan del siglo xix. En 1857 se inició la construcción del *Oxford English Dictionary* por parte de la Philological Society of London, trabajo que en 1878 retomaría la Oxford University Press. La creación de este diccionario se basó en la toma de citas como ejemplos lexicográficos y en la selección de datos textuales para la elaboración del diccionario, todo de manera manual. En 1897, J. Kading, lingüista alemán, trabajó fuertemente en la constitución de un corpus de cerca de once millones de palabras procedentes de la lengua alemana, con el fin de analizar la distribución de las letras y sus secuencias.

A comienzos del siglo xx, la necesidad de estudiar lenguas no documentadas —como las amerindias— hace que lingüistas se acerquen al trabajo con datos reales y recurran a los hablantes nativos para así obtener muestras, acercamiento que permitió describir y generar hipótesis sobre tales lenguas. Algunos trabajos de esta época son *Handbook of Native American Indian Languages* de Franz Boas (1911), *Language* de Leonard Bloomfield (1933) y *The Structure of English* de Charles Fries (1952).

En la década de los cincuenta aparece The Survey of English Usage[54], el primer centro de investigación dedicado al trabajo con corpus, en el que Randolph Quirk comienza la creación de un corpus del inglés británico oral con sus correspondientes transcripciones, conocido como el *Survey of English Usage* (*SEU*) o *Corpus de Quirk,* un conjunto de un millón de palabras grabado en cintas de carrete, transcrito manualmente y organizado en tarjetas de papel. Por otra parte, John Rupert Firth comenzaba a introducir el término *colocación* en el ámbito de la lingüística de corpus*,* con el que se refería a la ocurrencia sistemática de dos o más palabras dentro de un contexto, concepto que aún hoy se tiene en cuenta para la explotación y análisis de corpus.

Sin embargo, en 1950 no solo se estaban construyendo los primeros corpus sino que un nuevo paradigma emergía de la voz de Noam Chomsky en el área del lenguaje: el generativismo. Este movimiento lingüístico opacó el nacimiento de los estudios basados en corpus y disminuyó el impacto de tales investigaciones. Parodi (2008, p. 99) dice al respecto:

> […] diversos investigadores coinciden en apuntar que la lingüística generativa constituyó una influencia decisiva y hegemónica en el devenir científico de las ciencias del lenguaje, diluyendo o debilitando el desarrollo de posturas que abordaban el estudio del lenguaje desde ópticas diversas; en particular, desde opciones que no coincidían en una definición idealizada del lenguaje ni de metodologías de índole hipotético deductivo (Francis, 1979; Conrad & Biber, 2001; Chafe, 1992; Sinclair, 1991; Leech, 1991; Kennedy, 1998; McEnery & Wilson, 1996; Moreno, 1998)

El generativismo considera el lenguaje como una facultad innata en el ser humano y por tanto lo estudia desde una perspectiva mentalista, concentrándose en la *competencia*[55] y no en la *actuación*[56] del hablante.

Uno de los principios del generativismo es la creatividad lingüística, la cual consiste en la creación de enunciados infinitos con un número de elementos finitos; en palabras de Chomsky:

> L'aspect créateur de l'utilisation du langage reflète les possibilités infinies de la pensé et de l'imagination. Le langage offre des moyens finis mais des possibilités d'expression infinies, qui ne subissent d'autres règles que celles de la formation du concept et de la phrase, règles qui sont en partie spécifiques et idiosyncratiques, mais en partie aussi universelles, et telles que l'humanité tout entière en soit dotée (1966, p. 56).

Con base en esta idea, Chomsky niega toda credibilidad de resultados basados en corpus, argumentando que no existe ningún repertorio finito de datos que pueda dar cuenta de un objeto infinito como la lengua y que, por tanto, un corpus no contendrá todas las construcciones lingüísticas posibles[57].

Esta visión sobre el lenguaje hace que los lingüistas generativistas no se interesen en observar y estudiar la lengua en uso, desestimando el estudio de la lengua a través de corpus y además considerándolos parciales, finitos y no representativos. Chafe afirma al respecto: "One consequence of the modular view is that its adherents are not particularly interested in observing the everyday use of language, since they believe that whatever is most interesting about language exist independently of its use" (1992, p. 81).

Parodi hace referencia a las palabras de Sinclair (1991), con las que enuncia los efectos del enfoque generativista:

> Sedienta por falta de información adecuada, la lingüística languideció —de hecho— se volvió totalmente introvertida. Se hizo una moda mirar hacia adentro de la mente más que hacia la sociedad. La intuición se volvió la clave y se enfatizó la similitud de la estructura del lenguaje y varios modelos formales. El rol comunicativo del lenguaje fue escasamente mencionado (2008, p. 100).

Pero aunque el generativismo opacara los esfuerzos de la lingüística de corpus a causa de sus principios teóricos[58], en los años sesenta, con la llegada de los computadores, se continúa con la creación de algunos corpus, como el SEU y el *Brown Corpus*[59], los cuales gracias a la tecnología de la época se almacenaban en tarjetas perforadas, muchas de las cuales pudieron leerse en nuevos computadores y permitieron el trabajo con grandes cantidades de datos: un millón de palabras, que para la época era un acervo de gran amplitud.

Los avances de la LC se vieron reflejados en el SEU, no en cuanto a avances en el campo tecnológico sino en cuanto al proceso de anotación, debido a que Crystal y Quirk se dedicaron a anotar prosódica y paralingüísticamente[60] este corpus. En 1964 se dio inicio al *Brown Corpus,* colección textual informatizada creada por Henry Kučera y W. Nelson Francis, compuesta por un millón de palabras representativas del inglés americano, trabajos que continuarían su desarrollo en los siguientes quince años.

La década de los setenta fue una época decisiva para la lingüística de corpus, pues por una parte el funcionalismo lingüístico hizo que se le prestara atención al uso del lenguaje, y por otro lado, los avances informáticos permitieron el procesamiento de grandes volúmenes de datos; en suma, la lingüística de corpus volvió a nacer. La lingüística funcional nace como una crítica frente al generativismo, argumentando que este paradigma es idealista y que no ofrece herramientas para comprender la realidad de la lengua. Parodi señala en referencia a esto:

> El giro racionalista cognitivo que se impone desde el generativismo tiende a opacar de cierto modo el empirismo imperante y, en algunos casos, teñido de influencia conductista. Las bases contextualistas (o también externalistas), enmarcadas en paradigmas socioculturales del lenguaje, proveían un andamiaje para la lingüística de corpus tradicional, la que comienza a enfrentar una oposición desde el nuevo escenario interdisciplinario. Ahora bien, si bien es cierto que el generativismo aportó de manera crucial en materias nucleares acerca de la naturaleza del lenguaje humano, no es menos cierto que —entre otras— la visión idealizada del lenguaje (a saber, el estudio de la competencia lingüística) mantuvo un objeto de estudio casi único y se vieron difuminadas algunas investigaciones focalizadas en el estudio del lenguaje en uso (de la *performance*) y de la investigación de la variabilidad lingüística. Ello produjo una cierta discontinuidad o pérdida de impacto de ciertas líneas de investigaciones en lingüística (2008, p. 100).

El funcionalismo propugna el estudio de la lengua en uso: cómo se produce, cómo se comunica, cómo se entiende, quiénes son los participantes y cómo se desarrolla el acto comunicativo. McEnery y Hardie hablan sobre el funcionalismo lingüístico en relación con la corriente generativista:

> Functionalism, in a nutshell, is the rejection of this precept: functionalists investigate language form, but explain it with reference to the functions to which language is put. Language

> is not seen as an abstract, isolated system, but one that is *used* to communicate meaning, and which is shaped by the ways it is used, by the context in which it occurs and by the structure of human cognition. "Functionalism" in this broad sense covers a set of approaches to the theory of language sharing these features, including functional linguistics, cognitive linguistics and language typology. The emphasis on language in use makes functionalism compatible with corpus linguistics in a way that formalist linguistics is not (2011, p. 168).

La nueva concepción sobre los estudios de la lengua hace que los lingüistas, antropólogos, sociólogos e incluso psicólogos, ahora preocupados por los fenómenos de la comunicación, basen sus estudios en la creación y explotación de corpus, puesto que así pueden obtener pruebas empíricas y reales de hipótesis planteadas, o estudiar desde un conjunto de datos real los fenómenos que les interesan.

Sumados a la lingüística funcional, los avances informáticos de *software* y *hardware* fortalecieron el desarrollo y el uso de la LC. La tecnología informática brindó la posibilidad de construir y almacenar corpus de millones y billones de palabras, de realizar operaciones computacionales sobre grandes cantidades de datos y, por tanto, analizar los datos por medio de herramientas como etiquetadores morfosintácticos y programas semiautomáticos. Desde este momento, los corpus se concibieron como digitales.

El SEU, creado en la década de los cincuenta, se vio beneficiado por la era tecnológica. J. Svartik tomó los datos que se encontraban en el SEU y los digitalizó, creando así el *London-Lund Corpus of Spoken English*, que dio origen en 1985 a una de las gramáticas más relevantes del inglés: *A Comprehensive Grammar of the English Language*, escrita por Randolph Quirk, Sidney Greenbaum, Geoffrey Leech y Jan Svartvik, más adelante remplazada por la *Cambridge Grammar of the English Language*. El *Brown Corpus* también se valió de las herramientas computacionales para así publicar en 1979 su versión anotada, gracias a un programa de etiquetado de *part-of-speech*[61], creado por Green y Rubin. Tras la aparición del *Brown Corpus*, basado en el inglés americano, se creó el *Lancaster-Oslo-Bergen Corpus*[62] en 1978, el cual tiene las mismas características concentradas en el inglés británico y cuya versión etiquetada aparece en 1986.

A partir de 1980 ya no se habla de corpus sino de *megacorpus*, pues debido a las condiciones tecnológicas era mucho más fácil almacenar millones de datos, razón por la cual los corpus pasaron de tener un millón de datos a 450 millones; además, contenían los textos completos y no solo fracciones de estos, en diferentes registros, variedades e incluso de fuentes escritas y orales. Entre los llamados *megacorpus* encontramos *Bank of English* o *Cobuild Corpus*,

Cambridge English Corpus, *Longman/Lancaster English Corpus* y *British National Corpus*[63]. A partir de estas colecciones se crearon gramáticas y diccionarios.

Ya para los años noventa, aunque se continuó con la creación de *megacorpus*, no solo en inglés sino en diferentes lenguas, aparece una nueva modalidad de corpus: los corpus especializados[64]. Estos corpus, de tamaño más pequeño, contienen datos enfocados en algún tema, alguna variedad, o simplemente se centran en grupos específicos de hablantes; otras tipologías de corpus que comenzaron a tomar fuerza fueron los corpus diacrónicos[65], los cuales se encargan de estudiar una época o tiempo determinado, y los corpus monitores[66], que son actualizados constantemente. Además, los corpus pasaron de ser materiales construidos y explotados por grupos de investigación de varias universidades, como la Universidad de Lancaster, la Universidad de Birmingham y la Autónoma de Madrid, a convertirse en un material comercial que permite la creación y la explotación de diferentes tecnologías computacionales, como traductores automáticos y programas de reconocimiento de voz; algunos de estos corpus han sido el *Carnegie Mellon Communicator Corpus* o el Ceudex.

Durante los últimos años la lingüística de corpus se ha establecido como una metodología utilizada por varias lenguas, ya no solamente por el inglés. Lenguas como el español, el francés, el portugués, el mandarín, el polaco, el coreano, el checo o el húngaro cuentan con corpus de diversas características: generales, diacrónicos, sincrónicos, para fines específicos, monitores y documentados. El recorrido que le queda a la LC aún es largo, pues si bien ya ha permeado disciplinas diferentes de la lingüística, y ha conciliado las visiones generativistas y funcionalistas, existen lenguas como las lenguas indígenas latinoamericanas, que todavía no cuentan con corpus que faciliten su preservación y descripción, dadas sus características de oralidad y la poca difusión de la LC en el contexto académico latino.

La lingüística de corpus cuenta actualmente con asociaciones dedicadas al trabajo basado en corpus, como la Asociación Española de Lingüística de Corpus[67] (Aelinco), Asociación Lingüística Sistémico-Funcional de América Latina[68] (Alsfal), American Association for Corpus Linguistics[69] (aacl), International Quantitative Linguistics Association[70] (iqla), International Archive of Modern and Medieval English[71] (Icame) y Asociación Española de Lingüística Aplicada[72] (Aesla).

Existen también centros de investigación universitarios, como el Centre for

Corpus Liguistics[73] de la Universidad de Portsmouth, el Centre for Corpus Research de la Universidad de Birmingham[74], el Centre for English Corpus Linguistics[75] de la Universidad Católica de Louvaina, el University Centre for Computer Corpus Research on Language[76] de la Universidad de Lancaster, entre muchos otros, que también se dedican al trabajo mediante corpus.

En la actualidad, las herramientas tecnológicas no solo sirven para el almacenamiento y explotación de datos, sino que se constituyen en corpus; en otras palabras, la posibilidad de la web como un gran corpus es una de las opciones que poco a poco empiezan a llamar la atención de lingüistas. Aunque se habla de desventajas, como la dependencia de buscadores comerciales sin propósitos lingüísticos, el continuo cambio de resultados obtenidos, las dificultades que se pueden tener respecto a los derechos de autor o el carácter privado de algunos documentos, es una modalidad que brinda acceso a gran diversidad de textos, a bajos costos y de fácil acceso, razones por las cuales es un camino que se está comenzando a explorar y recorrer.

El avance de la LC se ha visto reflejado en los corpus, pues en su almacenamiento y diseño se pasó del trabajo manual al computacional; de corpus generales se crean ahora corpus especializados, diacrónicos, históricos y monitores, entre otros; de corpus simples a corpus anotados[77] y de corpus anotados manualmente a una anotación semiautomática o automática, por supuesto con una revisión de los investigadores. Con el paso de los años, la lingüística de corpus se ha convertido en una herramienta para diferentes disciplinas interesadas en el lenguaje, así como en una metodología que permite crear y probar hipótesis, describir la lengua y construir sistemas de procesamiento de lenguaje natural.

---

54. http://www.ucl.ac.uk/english-usage/index.htm.

55. La *competencia lingüística* hace referencia al conocimiento de la lengua adquirido por un hablante.

56. La *actuación lingüística* es el uso que un hablante da a la lengua. Está dada por la competencia y por factores sociales y culturales.

57. En la actualidad, reconocemos y sabemos que un corpus no puede contener todos los elementos y construcciones de una lengua, pero que cumple con la característica de la *representatividad* (véase el capítulo “Características de un corpus”).

58. McEnery y Hardie (2011, p. 168) enuncian claramente tres principios de la teoría generativista que dificultaban el trabajo conjunto con la lingüística de corpus: “The distinction between competence and performance, the rejection of corpus data reliance on introspection, and the view of language as an autonomous cognitive system”.

59. http://icame.uib.no/brown/bcm.html.

60. La *anotación prosódica* y *la anotación paralingüística* corresponden a la adición de datos o etiquetas a los elementos del corpus que hacen referencia a categorías en estas dos dimensiones. En la anotación prosódica podemos encontrar etiquetas respecto a la melodía, el acento, las pausas y el ritmo, entre otras; y en la anotación paralingüística es posible trabajar categorías relacionadas también con la entonación y la pronunciación y aspectos como la risa, el llanto, el suspiro, etc., que cuales reflejan emociones del entrevistado.
61. *Part-of-speech* (*Etiquetado gramatical* en español) corresponde a la asignación de una etiqueta a cada uno de los datos del corpus, la cual indica la categoría gramatical a la que corresponde el elemento según el contexto.
62. http://www.helsinki.fi/varieng/CoRD/corpora/LOB/ información y manual sobre el *Lancaster-Oslo-Bergen Corpus*.
63. http://www.natcorp.ox.ac.uk/.
64. Véase el apartado “Tipología de los corpus”.
65. Ibid.
66. Ibid.
67. http://www.um.es/aelinco/.
68. http://www.pucsp.br/isfc/alsfal/espanol/Inicio.html.
69. http://aacl.sdsu.edu/.
70. http://www.iqla.org/index.html.
71. http://icame.uib.no/.
72. http://www.aesla.org.es/es.
73. http://www.port.ac.uk/corpus-linguistics/.
74. http://www.birmingham.ac.uk/research/activity/corpus/index.aspx/.
75. http://www.uclouvain.be/en-cecl.html.
76. http://ucrel.lancs.ac.uk/.
77. Para información sobre las características de los corpus enunciados, véase el apartado “Tipología de los corpus”.

Usos de los corpus

En términos generales, los datos contenidos en un corpus los puede utilizar cualquier interesado en el estudio del lenguaje, para describir y analizar la lengua y establecer o corroborar hipótesis desde diferentes teorías o aproximaciones.

En este orden de ideas, las principales ventajas que la lingüística de corpus ofrece[78] a sus usuarios son las siguientes:

1. Prioridad a la lengua en uso, escrita y oral.
2. Aproximación a los datos de una manera cuantitativa y cualitativa.
3. Uso como herramienta apta para diferentes disciplinas.

Al respecto, Geoffrey Leech propone: “In Corpus Linguistics, the only requirement is that such questions should be capable of being answered by observing what is attested in corpus data” (2011, p. 161).

A continuación, se presentan las posibilidades generales que un corpus ofrece a cualquier usuario o investigador, tales como opciones de búsqueda, colocaciones y concordancias, entre otras, y los usos específicos que se les da a la lingüística de corpus y a los corpus en los estudios realizados por distintas disciplinas, tales como la lexicografía, la dialectología, la lingüística histórica, etc.

Usos generales y posibilidades que ofrecen los corpus

Las posibilidades de uso de un corpus dependen en gran medida de dos factores:

1. Las características propias del corpus, tal como la *anotación*[79].
2. Las herramientas computacionales y la interfaz.

Las características del corpus determinan las clases de búsquedas y resultados que se pueden hacer y obtener. Por ejemplo, si un corpus está anotado

morfológicamente se podrán hacer búsquedas que arrojen listas de sustantivos, verbos, artículos, etc. Otros factores determinantes en las búsquedas son las herramientas computacionales empleadas y la interfaz, ya que estas facilitan la interacción del usuario con los datos, establecen las búsquedas que se pueden hacer dentro del corpus y determinan la manera gráfica en la que se presentan los resultados.

Mediante las búsquedas dentro del corpus podemos consultar desde letras hasta frases, listas de datos, frecuencias de aparición, concordancias, colocaciones y palabras claves; de ahí en adelante, el empleo que se les dé a los datos depende de las necesidades y objetivos de los usuarios o investigadores.

Búsquedas

En términos generales, los corpus funcionan con base en la posibilidad de efectuar búsquedas de diferentes categorías. Dentro de un corpus podemos buscar letras, palabras, partes de palabras o frases, fonemas, elementos gramaticales (verbos, artículos, sustantivos, adverbios, adjetivos, etc.), sintácticos (sintagma nominal, sintagma verbal, etc.) y cualquier tipo de sondeo más específico, dependiendo de las etiquetas[80] que contengan el corpus y la anotación. El sistema de búsqueda en el que se basan la creación, construcción y explotación de corpus, dado por herramientas computacionales, permite hacer las búsquedas requeridas en diversos momentos y recuperar información que de otro modo se habría perdido.

Listas de datos

Los resultados de un corpus se muestran generalmente en forma de listas, que pueden ser de palabras, lemas, categorías gramaticales, etc.[81], o de frecuencias de aparición, colocaciones, concordancias o palabras claves[82]. La ventaja de este sistema es que los elementos pueden aparecer por orden alfabético, por orden de frecuencia[83] o incluso combinando las dos opciones; esto brinda la posibilidad de comparar listas de elementos dentro de un corpus o listas de diferentes corpus, lo que en muchos casos puede arrojar información sobre estructuras lingüísticas que son más comunes en algunos registros que en otros, pues todo depende del tipo de corpus que se compare.

Frecuencias de aparición

Los índices de frecuencias se constituyen en el elemento con mayor tradición en los estudios basados en lingüística de corpus debido a la cantidad de empleos que se les pueden dar, como la creación de glosarios, de diccionarios, de material

didáctico para la enseñanza de lenguas, la creación de hipótesis en el área de análisis del discurso, etc.

La frecuencia de aparición conlleva un proceso de revisión automática del contenido de un corpus, por medio de la cual se determina el número de veces que un elemento, ya sea palabra, categoría gramatical, lema, combinaciones de letras, frases o combinación de elementos en una cadena lingüística, aparece dentro de un corpus. Las frecuencias nos dan una idea clara respecto a la importancia y el uso de una palabra en una lengua o dentro de los textos y géneros de un mismo corpus.

Rafel y Soler dan dos ejemplos claros sobre el funcionamiento de las frecuencias en los corpus:

> En un corpus con los textos clasificados temáticamente, dos palabras presentan frecuencias similares. Sin embargo, una de ellas concentra casi todas sus apariciones en un tipo temático (por ejemplo, en matemáticas, o bien por derecho, o bien por psicología, etc.), mientras que la otra se presenta repartida más o menos equitativamente entre la totalidad de los grupos. De forma inmediata diríamos que la segunda palabra tiene un carácter más general en el vocabulario que la primera; esta, en cambio, podría tratarse con bastante probabilidad de una palabra específica de una determinada materia […] la frecuencia es un dato absoluto, cuyo valor depende fuertemente de la extensión del corpus a que se refiere. Supongamos que un determinado elemento léxico *a* aparece 50 veces en un corpus de 50.000.000 de palabras, mientras que otro elemento léxico *b* aparece también 50 veces en un pequeño corpus de 5.000 palabras, aunque *a* y *b* tengan las mismas frecuencias en términos absolutos, su importancia relativa en cada uno de los dos corpus es bastante diferente: *a* aparece una vez cada millón de palabras, mientras que *b* aparece una vez cada cien palabras (2003, p. 63).

Concordancias

Las concordancias[84] se obtienen por medio de herramientas informáticas que dan la posibilidad de arrojar resultados a manera de líneas en las que una palabra determinada aparece acompañada por elementos de sus contextos lingüísticos. En otras palabras, las concordancias son todas las apariciones de una misma palabra acompañada de los elementos anteriores o posteriores. El número de elementos que aparecen junto a estas palabras, ya sea anterior, posterior o ambos, está determinado por las herramientas computacionales con las que cuenta el corpus. A renglón seguido se presenta un ejemplo de las concordancias de la palabra *cualquiera*:

Concordancias de la palabra cualquiera

una pintura	cualquiera	que no sea costosa
cada vez que	cualquiera	se acerca a la mesa
pásame	cualquiera	de las cartas

	cualquiera	no puede ganar
que	cualquiera	quiera ir a

Los resultados de las concordancias pueden aparecer en orden alfabético, según el orden de las palabras anteriores o posteriores, o en algún orden definido por el usuario, claro está, si el programa computacional está diseñado para ello. En términos más generales, las concordancias nos muestran secuencias específicas de elementos como letras o palabras de diversa longitud.

Colocaciones

Las colocaciones guardan relación con las concordancias y las frecuencias. Una colocación se ve influenciada por estas dos categorías, ya que corresponde a la frecuencia de aparición de una palabra en compañía de otra. McEnery y Hardie definen colocación como "A co-ocurrence relationship between two words. Words are said to *collocate* with one another if one is more likely to occur in the presence of the other than elsewhere" (2011, p. 240).

Por ejemplo, *contar* y *cuento* pueden ser una colocación en un determinado corpus debido al número de veces que pueden aparecer juntas en frases como *cuéntame un cuento*, *él cuenta cuentos* o *la madre les cuenta cuentos a sus hijos*. Una definición más exhaustiva, contenida en *A Glossary of Corpus Linguistics*, de Baker y Hardie, reza así:

> Described by Firth (1957: 14) as 'actual words in habitual company', collocation is the phenomenon surrounding the fact that certain words are more likely to occur in combination with other words in certain contexts. A collocate is therefore a word which occurs within the neighbourhood of another word (2006, p. 36).

Para poder hablar de colocaciones la relación debe darse entre dos o más elementos, la distancia máxima entre ellos no puede superar las cinco palabras y la frecuencia debe ser alta[85].

Sobre el tema de las colocaciones, Tony McEnery (2014) habla en su curso *Corpus Linguistics: Method, Theory and Practice - Future Learn* sobre tres tipos de fenómenos adicionales: la preferencia semántica, la coligación[86] (*colligation*) y la prosodia del discurso. La preferencia semántica es la relación entre un campo semántico y un grupo de palabras semánticamente relacionadas; por ejemplo: falda, camisa, saco, pantalón corresponden a prendas de vestir. La coligación, por su parte, señala la ocurrencia entre una palabra y una categoría gramatical; es el caso de *ella* + verbo. Dentro de un corpus es posible que la aparición de la palabra *ella* vaya seguida de un verbo: *ella come mucho*, *ella sufre de impaciencia*, *veremos si ella quiere salir*, etc.

Y la prosodia del discurso o prosodia semántica, que corresponde a la manera en que las palabras en un corpus pueden relacionarse con una asociación positiva o negativa del hablante debido a las colocaciones, en palabras de McEnery y Hardie es esto: "Semantic prosody is the tendency exhibited by some words or idioms to occur consistenly with either positive or negative meanings" (2011, p. 250). En este caso, podríamos decir que dados los ejemplos encontrados en cierto corpus la palabra *muerte* se relaciona con *violenta*, *súbita*, *dolorosa*, lo que tiene una asociación negativa, aunque hipotéticamente también podría formar colocaciones con los términos *feliz* y *tranquila*.

Palabra clave

Este término corresponde a las palabras que aparecen en un corpus con un grado de frecuencia más alto del esperado y que, al ser comparadas con otro corpus del mismo tamaño o más grande, siguen siendo distintivas y relevantes, dada la frecuencia de aparición. Según el glosario del libro *Corpus Linguistics: Method, Theory and Practice*, de McEnery y Hardie, una palabra clave es esta: "A word that is more frequent in a text or corpus under study than it is in some (larger) reference corpus, where the difference in frequency is statistically significant" (2011, p. 245). Por ejemplo, en un corpus del español de Bogotá aparece con una alta frecuencia la palabra *jurgo*, que significa *un montón*; al realizar la búsqueda de esta palabra en un corpus del español de Colombia, encontramos que su frecuencia es baja y que su aparición se concentra en los textos representativos de Bogotá; en este caso, puede decirse que es una palabra representativa del español de Bogotá y que corresponde a una palabra clave. Desde allí se pueden llevar a cabo análisis y estudios más específicos.

El uso de los corpus según la disciplina

La lingüística de corpus y el uso de corpus pueden combinarse con casi cualquier disciplina o área interesada en el lenguaje, debido a la evidencia que puede brindar a las investigaciones; incluso la construcción de un corpus y su mantenimiento se convierten en una actividad interdisciplinar, en la que se necesitan lingüistas, ingenieros de sistemas y matemáticos, entre otros. Rafel y Soler explican el porqué de su versatilidad:

> El objetivo de la lingüística de corpus es la prospección y el procesamiento de corpus para la descripción, a partir de datos objetivos, de las estructuras y de las categorías (sintácticas, léxicas, morfológicas, etc.) de la lengua. Un corpus sirve, así, como elemento de contraste de hipótesis del lingüista, y al mismo tiempo, como un elemento que puede conducir determinadas

> investigaciones lingüísticas, por la inmediatez de los tipos de evidencia que proporciona (2003, p. 70).

La lingüística de corpus ha trabajado de la mano con la lingüística histórica, la lexicografía, la adquisición del lenguaje, la enseñanza de lenguas y la sociolingüística, entre otras áreas y disciplinas. Es importante aclarar que muchas más disciplinas pueden hacer uso de los corpus y que esto depende de la creatividad de los investigadores, las necesidades de la investigación y el alcance que los corpus puedan tener en la investigación misma. A renglón seguido se describen algunos usos:

Semántica

La semántica utiliza los corpus para describir, descubrir, despejar dudas y probar hipótesis respecto a la utilización de palabras o frases, y al sentido que tienen en diferentes contextos. Los estudios realizados desde la semántica tienen gran impacto sobre los estudios lexicográficos y de análisis del discurso[87].

Morfología y sintaxis

Con la ayuda de los corpus es posible describir, verificar y descubrir estructuras morfológicas y construcciones sintácticas dentro de una lengua; esto ayuda a la descripción general de una lengua específica, un género o un registro, y también puede servir de apoyo en áreas como la enseñanza de lenguas, en la que se hace necesario que los estudiantes aprendan elementos y estructuras en uso. El estudio de la morfología y la sintaxis desde una perspectiva histórica basada en corpus también permite observar, analizar y describir los cambios de las lenguas en estos dos niveles[88].

Dialectología y sociolingüística

Estas disciplinas utilizan corpus con el propósito de describir fenómenos sobre variaciones geográficas y grupos sociales, comparar dialectos o sociolectos en los diferentes niveles de la lengua[89], revelar características de grupos sociales particulares, identificar patrones pertenecientes a una zona geográfica o a un grupo social y comparar el habla según el género (femenino o masculino)[90].

Gramática

Aunque los estudios de gramática basados en corpus engloban las ramas de la lingüística anteriormente enunciadas, la unión de estos estudios con un acercamiento sustentado en el empleo de corpus permite la elaboración de gramáticas que describen la lengua en uso[91].

Lingüística histórica

El primer beneficio que recibe la lingüística histórica de la lingüística de corpus es la digitalización de libros antiguos y manuscritos, ya que con el estudio de este material se pueden hacer descripciones diacrónicas, observar cambios lingüísticos y crear hipótesis de cambios futuros, determinar fechas de aparición y desaparición de elementos en cada lengua, y por supuesto recolectar material para la construcción de diccionarios etimológicos[92].

Estilometría y literatura

La estilometría y la literatura utilizan numerosos datos, en los que se reúnen obras de autores importantes en ciertas épocas de la historia, se analizan textos para extraer frecuencias, concordancias y ejemplos de uso de palabras o construcciones lingüísticas, se establecen autorías a partir de análisis textuales y estilísticos, se describen obras reflejadas más adelante en ediciones críticas, y en términos generales, se estudian estilos literarios, autores, géneros y periodos históricos en la literatura. En conjunto con herramientas computacionales pueden crearse programas para detección de plagio y detección de autoría[93].

Análisis del discurso

El análisis del discurso es una de las disciplinas más beneficiadas por la lingüística de corpus, ya que esta le facilita el almacenamiento de grandes cantidades de datos, a los cuales se puede acceder una y otra vez; además, las herramientas informáticas permiten la detección de patrones lingüísticos, lo que en el análisis del discurso determina muchas de las hipótesis y conclusiones. Una de las principales ventajas de esta disciplina es que puede utilizarse en diferentes campos, por lo que en la actualidad se realizan estudios sociales, políticos, etc., con la ayuda de corpus de periódicos, noticias y documentos políticos, entre otros. Estos estudios tienen gran impacto en la sociedad en general y en los medios de comunicación en particular[94].

Psicolingüística y lingüística clínica

La psicolingüística se encarga de estudiar la comprensión, producción y adquisición del lenguaje. Existen corpus diseñados para el estudio de la adquisición del lenguaje, en los que los informantes son niños; los corpus también pueden funcionar como una fuente para determinar las frecuencias de uso de los elementos de la lengua, y desde allí diseñar pruebas para experimentos de procesamiento[95]. Por otro lado, el estudio de patologías puede trabajarse también desde datos recolectados en corpus, ya que estos permiten la

descripción de patologías del lenguaje, el reconocimiento de patrones en cada una de ellas y la reflexión respecto al trabajo que se puede llevar a cabo para detectar y manejar estos fenómenos[96].

Lingüística forense

Ramas como la lingüística forense se valen de técnicas estadísticas y de herramientas informáticas para la recolección de pruebas en los procesos de peritaje. Los corpus dan la posibilidad de comparar pruebas con datos, lo que hace que se detecten patrones morfológicos, sintácticos y semánticos cuando las pruebas son escritas, y fonéticos cuando la fuente probatoria es oral[97].

Traducción

Los estudios en traducción implementan las tecnologías del lenguaje para facilitar, agilizar y validar sus trabajos; por este motivo, los corpus[98] se convierten en una herramienta para comprobar la calidad de las traducciones, encontrar equivalencias entre lenguas y conformar bases de datos que permiten la automatización de estos procesos, puesto que aquellos contienen ejemplos reales de uso[99].

Lexicografía

Guillermo Rojo dice en su artículo denominado "Sobre la creación de diccionarios basados en corpus":

> El objetivo de un proyecto lexicográfico basado en corpus es, con toda claridad, recoger las palabras que figuran en un corpus representativo de la lengua o variedad lingüística sobre la que se trabaja y reflejar los significados realmente presentes en los textos, incorporando las marcas de uso correspondientes en cada caso (2009).

Este comentario deja ver la utilidad principal de los corpus en lexicografía: la creación de diccionarios. Los corpus contextualizan las palabras en uso, pueden determinar las entradas de un diccionario por medio de la frecuencia de aparición de los elementos y brindan un acceso instantáneo a datos actualizados. Algunos diccionarios creados a partir del trabajo con corpus son el *Diccionario del castellano del siglo xv en la corona de Aragón*[100] y el *Gran diccionario de uso del español actual*. También se pueden hacer listas de frecuencias por géneros o registros y diccionarios bilingües con ejemplos reales.

Fonética y fonología

La fonética y la fonología utilizan corpus orales para describir las lenguas segmental y suprasegmentalmente, estudiar fenómenos articulatorios y acústicos, clasificar acentos, comparar sistemas fonéticos, obtener datos para caracterizar

hablantes, trabajar acerca de la interferencia fonética en el aprendizaje de lenguas, etc. Además, junto con la lingüística computacional y los corpus, logran construir modelos de lenguaje natural aplicados a tecnologías del habla[101], como aplicaciones para la conversión de texto a habla.

Lingüística computacional

La lingüística computacional toma los corpus como insumo para la creación de herramientas computacionales que permiten la búsqueda, la recuperación, el análisis y la explotación de datos contenidos en textos electrónicos. Estas herramientas son las que hacen posible etiquetar, anotar, buscar frecuencias, colocaciones y concordancias en un corpus.

Igualmente, los corpus son necesarios en la creación de modelos de lenguaje que faciliten el reconocimiento de voz y la conversión de voz a texto o de texto a voz. Desde esta perspectiva, los corpus y la lingüística computacional entran en la dinámica de la industria de la lengua:

> Existen grandes oportunidades de mercado en ámbitos como la educación o el ocio, con la integración de tecnología lingüística en juegos, en divulgación del patrimonio cultural, en paquetes de entretenimiento educativo, en bibliotecas, entornos de simulación y programas de capacitación. Los servicios de información móvil, el *software* de aprendizaje de idiomas, los entornos de *e-learning*, las herramientas de autoevaluación y el *software* de detección de plagio son solo algunas de las áreas de aplicación en las que la tecnología lingüística puede desempeñar un papel importante (Melero, Badia & Moreno, n.d.-a, p. 7).

Enseñanza de idiomas

Los corpus pueden usarse fuera y dentro del aula, como un elemento de investigación o como una herramienta didáctica en clase. A partir de un corpus se pueden hacer diccionarios para aprendices, construir material didáctico como libros o ejercicios para la clase, crear exámenes; adicionalmente, los estudiantes pueden utilizarlos para acercarse a la lengua, descubrir y describir patrones, y corroborar construcciones que ocurren en la lengua.

Otra posibilidad es la creación o explotación de *corpus de aprendices*, los cuales reúnen muestras de textos o interacciones producidas por estudiantes de la lengua. A partir de ellos se puede crear el material anteriormente mencionado, estudiar la interlengua y analizar los errores que cometen los aprendices[102].

Tras este recorrido por los usos de la LC, es posible afirmar que toda aquella disciplina interesada en el lenguaje y con necesidad de datos reales de la lengua puede utilizar los corpus y la lingüística de corpus como herramientas en sus investigaciones.

---

78. Véase el apartado “Definición de la lingüística de corpus”.

79. La *anotación* es el proceso mediante el cual se explicitan categorías lingüísticas por medio de etiquetas que se añaden a los datos.

80. Una *etiqueta* corresponde a una secuencia de caracteres de algún tipo de lenguaje de marcado (xml, html, sgml), la cual contiene información acerca de los elementos del corpus, de un documento o del corpus en general.

81. Al igual que en *Búsquedas*, las listas de un corpus dependen de las etiquetas de este y de las herramientas informáticas utilizadas.

82. Todos estos términos (*frecuencias de aparición, colocaciones, concordancia* y *palabras claves*) se desarrollan en este mismo capítulo.

83. Número de veces que el elemento aparece dentro del corpus.

84. Las *concordancias* se denominan también en inglés *Key Word in Context* (KWIC).

85. No existe una cantidad exacta para determinar que una frecuencia es alta, ya que este valor depende del tamaño del corpus, aunque autores como Tony McEnery, en su curso *Corpus Linguistics: Method, Theory and Practice - Future Learn* (2014), afirman que el valor mínimo de frecuencia para determinar si es una colocación es de 10.

86. Dentro de la bibliografía revisada, en ningún texto se utiliza el término en español.

87. Uno de los trabajos basados en lingüística de corpus y semántica se denomina *Introducción al análisis de estructuras lingüísticas en corpus. Aproximación semántica* (Alcántara, 2007).

88. Un ejemplo de estudios morfológicos y sintácticos es Futuro perifrástico y Futuro morfológico en el Corpus sociolingüístico de la Ciudad de México (Lastra, 2008).

89. Fonético-fonológico, morfo-sintáctico, léxico-semántico y pragmático.

90. Para ampliar sobre la relación entre lingüística de corpus y psicolingüística, véase *Corpus Linguistics: Method, Theory and Practice - Future Learn,* de McEnery y Hardie (2011b, pp. 94-121).

91. Para más información sobre la relación entre lingüística de corpus y gramática, véase *Corpus Linguistics Investigating Language Structure and Use*, de Biber, Conrad y Reppen (1998, pp. 55-83).

92. Johannes Kabatek (2012) da una explicación completa sobre la relación entre la lingüística histórica y la lingüística de corpus en su texto llamado *¿Es posible una lingüística histórica basada en un corpus representativo?*

93. Dentro del reconocimiento de la estilometría, la lingüística forense y la lingüística de corpus se encuentran trabajos como el de López, Méndez, Sierra y Solórzano (2013), *Exploración de medidas estilométricas para atribución de autoría*.

94. Un ejemplo del trabajo conjunto entre lingüística de corpus y análisis de discurso lo presentan Palacios y Sierra (2011, pp. 386-398) y se denomina *Corpus para el análisis del discurso del concepto ad hoc-cracia*.

95. Para más información sobre la relación entre lingüística de corpus y psicolingüística véase *Corpus Linguistics: Method, Theory and Practice* de McEnery y Hardie (2011, pp. 192-224).

96. *Corpus lingüístico de definiciones de categorías semánticas de sujetos ancianos sanos y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina* (Peraita y Grasso, 2010) presenta un trabajo entre la lingüística clínica, la semántica y la lingüística de corpus.

97. Para ampliar sobre la relación entre lingüística de corpus y lingüística forense, véase *La lingüística forense y el uso de los corpus lingüísticos*, de Cicres (2011, pp. 517-524).

98. En traducción se utilizan especialmente corpus bilingües y multilingües, comparados y paralelos.

99. José Cortez (2010) expone las ventajas que ofrecen los corpus al área de la traducción en su trabajo *El corpus* ad hoc *como herramienta de traducción*.

100. http://ghcl.ub.edu/diccaxv/home/index/myLanguage:es.

101. Véase *Lingüística computacional* en este mismo apartado.

102. Uno de los fundamentos teórico-prácticos en español que muestran las posibilidades de la lingüística de corpus y la enseñanza de idiomas es *Lingüística de corpus y enseñanza del español como 2/L*, de Mar Cruz Piñol (2012).

La construcción de un corpus

Hoy en día, pueden llevarse a cabo diversas investigaciones sobre el lenguaje con corpus ya existentes; sin embargo, cuando lo que se quiere es preservar, almacenar y sistematizar un material en particular o se requieren datos con características específicas[103], se hace necesaria la construcción de nuevos corpus.

Así como los corpus pueden utilizarse para responder a problemas desde diferentes disciplinas, para su creación y mantenimiento se requiere también un conocimiento interdisciplinar, hace falta conocimiento lingüístico, informático, matemático y, dependiendo del tipo de corpus que se quiera construir, se necesita además conocimiento histórico, sociolingüístico, etc.

Dentro de la lingüística de corpus no existe un protocolo que determine paso a paso cómo crear un corpus. Kennedy (1998), por ejemplo, propone cinco momentos: diseño de corpus, planeación del sistema de almacenamiento, obtención de permisos, captura de textos y marcado; Atkins, Clear y Ostler (1992) plantean también cinco estadios: planeación, adquisición de permisos, captura de datos, manipulación de textos y desarrollo de corpus.

A continuación presentamos una propuesta con cinco momentos principales, que puede adaptarse a diversas necesidades investigativas:

1. El diseño de corpus.
2. La obtención de permisos y captura de datos.
3. La planeación y preparación del sistema de almacenamiento.
4. El procesamiento del corpus.
5. Las opciones de uso.

Diseño y elaboración de corpus

El diseño de corpus cuenta con tres pasos específicos: definición de objetivos, definición de la composición del corpus y los criterios de recolección, y por último, la elección de la tipología. Esta etapa de diseño, junto con la de

procesamiento, determina las posibilidades de utilización de un corpus.

Paso 1. Definir los objetivos[104]

Para empezar, es necesario aclarar cuál es la finalidad del corpus. Partiendo de aquí, los objetivos tanto del corpus como del proyecto determinan las características de la colección textual y establecen el tipo de uso y búsquedas que se pueden efectuar.

Paso 2. Definir la composición y los criterios de recolección

En un segundo paso se definen la composición y los criterios de recolección, con lo que se hace preciso pensar en la representatividad[105], el tamaño que se quiere o el material del que se dispone, la variedad de la lengua que el corpus representará y la cronología a la cual pertenecen los textos[106]. Mercado (2008, p. 19) plantea siete criterios que hay que decidir al momento de definir la composición del corpus:

1. Tipo: oral o escrito.
2. Tipos de registros: literatura, prensa, etc.
3. Parámetros demográficos: edad, sexo, grupo, etc.
4. Época.
5. Medios de comunicación: libros, periódicos, correos electrónicos, etc.
6. Niveles lingüísticos: coloquial, formal, lengua infantil, publicitaria, etc.
7. Tipos de textos: novelas, poemas, reportajes, columnas, encuestas, etc.

Después de definir la composición del corpus siguiendo los anteriores criterios, se hace necesario determinar las pautas de recolección, para lo que se requiere:

1. Precisar de dónde se tomarán los textos,
2. Concretar el número de muestras,
3. Definir las secciones que se utilizarán de cada texto
4. Determinar la longitud de las muestras

La definición de las secciones y la longitud de las muestras que se utilizarán para la construcción del corpus dependen de los objetivos previos y de las

facilidades que existan para obtener las muestras. Torruela y Llisterri proponen tres maneras de definir las secciones:

> a) Al azar; b) dividiendo los textos en tres partes de extensión parecida y extrayendo de cada una de ellas las muestras en número y proporciones aproximadamente iguales; c) determinando la estructura externa de los textos y decidiendo qué niveles estructurales se usarán para el muestreo (un número determinado de palabras o de frases de cada capítulo, un número determinado de cada apartado, un número determinado de cada párrafo, etc.) (1999, p. 20).

En el caso de la longitud de las muestras, pueden tomarse los textos completos o fragmentos; se debe evitar caer en la extracción de los inicios o finales del texto, a no ser que ese sea el objetivo del corpus, ya que esto puede afectar la característica de representatividad. La longitud de los fragmentos puede hacerse escogiendo un número determinado de palabras o de oraciones con sentido, lo que se puede lograr si se toman fracciones delimitadas por puntos en el caso de material escrito o pausas en el caso de muestras orales.

Paso 3. Definir la tipología

Según los objetivos del corpus, se establece o corrobora la tipología del corpus tras la definición de todas las variables de composición. Dependiendo de las elecciones hechas en el *paso 2* puede hablarse de corpus escrito, oral, multimodal, monolingüe, bilingüe, multilingüe, general, especializado, genérico, canónico, cronológico, grande, equilibrado, piramidal, cerrado, abierto, textual, de referencia o léxico[107].

Obtención de permisos y captura de datos

Para poder usar los textos o grabaciones que se incluirán en los corpus, es necesario tener en cuenta los derechos de autor. Muchas veces para reproducir los textos, esto es, digitalizarlos, se requiere la autorización de los autores; todos los textos de un corpus deben estar bajo la protección de derechos de autor y, además, deben tener los permisos para el uso que se requiere. Es recomendable buscar asesoría legal en este momento de la construcción del corpus para prevenir futuros inconvenientes, y considerar que las leyes respecto a los derechos de autor varían según el país y el material que se maneje.

Para la captura de datos se pueden requerir bastante tiempo y dinero, dependiendo de la cantidad de datos que deban recogerse y de las fuentes de dónde se obtengan. Para crear un corpus resulta indispensable que todo el material esté digitalizado, ya sea oral o escrito. En el caso del material escrito, existen tres opciones de captura:

1. Mediante *Optical Character Recognition* (OCR), proceso que consiste en escanear los textos físicos mediante un sistema de reconocimiento de caracteres para digitalizar los textos.
2. Transcripción manual.
3. Datos ya digitales.

Si se opta por el uso de OCR es recomendable realizar un control del material obtenido; a su vez, la transcripción manual es muy utilizada en corpus orales, ya que las cintas requieren en su mayoría una transcripción ortográfica que en muchas ocasiones no puede hacerse mediante programas automáticos de reconocimiento de voz[108]; en cualquiera de los dos procesos se necesitan una o dos revisiones manuales por parte de los investigadores tras el proceso de digitalización o transcripción, pues así se pueden corregir los errores que la automatización pueda tener. En lo referente al uso de datos ya digitales, los costos y el tiempo se reducen; además, muchos de estos datos se pueden tomar de internet.

Planeación y preparación del sistema de almacenamiento

En esta fase se debe pensar acerca del tamaño total del corpus para así obtener el espacio de almacenamiento[109]; no hay que olvidar que sin espacio de almacenamiento no hay corpus. Tras la obtención del espacio, los investigadores deben asegurarse de almacenar de manera sistemática y ordenada los datos; se recomienda guardar cada texto como un archivo diferente y llevar una secuencia clara, lógica y sistemática[110], para que el corpus pueda empezar a ordenarse con base en datos externos, como el nombre o el número del archivo, y facilitar así la ubicación de la información.

Procesamiento del corpus

La interfaz

Después de contar con el espacio y el sistema de almacenamiento, es necesario pensar en la interfaz, que es el conjunto de programas que permiten extraer información del corpus y facilitan la interacción del usuario con los

datos. Rafel y Soler dicen al respecto: “La óptima utilización de un corpus está en relación directa con las capacidades de la interfaz con que se accede al mismo para la ejecución de procesos de selección y para la presentación de los diferentes tipos de resultados posibles” (2003, p. 67).

Hasta este momento, con el material organizado según criterios lingüísticos, almacenado, y una interfaz que permita abrir, descargar o reproducir los archivos, se podría decir que se cuenta con un corpus simple[111].

Codificación

La codificación es el proceso de conversión de caracteres del lenguaje natural a un lenguaje que se pueda procesar por medio de máquinas o sistemas que se valen de programas computacionales; en el momento en que se prepara un corpus para su procesamiento, hay que decidir el formato del texto y la codificación, en función de los programas que se pretenden utilizar. Es aconsejable buscar sistemas de codificación con alto número de caracteres[112], pues así no se requiere cambiar de codificación años después de la construcción del corpus; esto, sumado al uso de estándares[113], permite la reutilización del corpus.

En la codificación, debe elegirse un lenguaje de marcas o etiquetas que permita representar información adicional a la que contiene el texto[114] y conviene ceñirse a un estándar de codificación. A continuación se presentan algunas opciones estándares de codificación y lenguajes de marcado:

Estándares de codificación y lenguajes de marcado

Estándares de codificación	**Lenguajes de marcado**	**Localización**
TEI (Text Encoding Iniciative[1])	XML	http://www.tei-c.org/index.xml.
ELRA (European Language Resources Association[2])	HTML	http://www.elra.info/.
LDC (Linguistic Data Consortium[3])	SGML	https://www.ldc.upenn.edu/.
CES (Corpus Encoding Standard[4])		http://www.tei-c.org/Activities/Projects/co02.xml.
Eagles (Expert Advisory Group on Language Engineering Standards[5])		http://www.ilc.cnr.it/EAGLES/browse.html.

Marcado de metadatos

Con la elección del lenguaje de marcado y el estándar, se da comienzo al proceso de etiquetado o marcado de metadatos. Este proceso consiste en insertar

etiquetas para enriquecer los textos, las cuales deben contener información estructural de los textos, como origen, autor, año, tipo de texto, participantes, duración y calidad de la grabación, entre otras. Dichas categorías dependen de los intereses del grupo y del estándar que se elija, ya que estos manejan etiquetas determinadas, aunque no significa que no puedan crearse nuevas categorías.

La inserción de estas etiquetas y un mayor número de datos externos registrados facilitan las búsquedas cruzadas y con una cobertura de más variables, ya que, por ejemplo, pueden llevarse a cabo búsquedas de datos con dos, tres o más características al mismo tiempo, lo cual hace que los datos sean a su vez más precisos.

Anotación lingüística

Aun cuando un corpus simple permite el acercamiento a los datos de manera confiable, existen investigaciones que requieren análisis más complejos y exactos, lo cual se puede lograr mediante la anotación lingüística de los datos contenidos en el corpus. La anotación lingüística corresponde al proceso de etiquetado de las palabras pertenecientes a los textos, con el fin de incluir información lingüística adicional, ya sea sobre su carácter semántico, fonético, morfológico, pragmático, etc. En términos generales, cada palabra de un corpus anotado tiene una o varias etiquetas que indican sus características.

La anotación debe estar separada del texto como tal, es decir, que al borrar las etiquetas el texto debe permanecer intacto. Procházková señala algunos principios de la anotación que es conveniente seguir:

- La evaluación de las anotaciones debe ser posible sin el texto original.
- Las normas de anotación deben ser accesibles.
- Los anotadores y las circunstancias de la anotación deben ser conocidos.
- Los usuarios deben saber que las anotaciones pueden contener errores (2006, p. 11).

Los corpus anotados requieren un proceso específico de etiquetado. Julia Baquero define claramente este proceso:

> Los corpus anotados o etiquetados requieren una transformación del texto original de forma que se pueda acceder a él y extraer la mayor cantidad de información posible. Para ello, los corpus son sometidos a un procesamiento que incluye, entre otras, la posibilidad de dividirlo en la unidad más pequeña —el *token*— sobre la cual se aplica una etiqueta de carácter lingüístico mediante un programa denominado etiquetador. Este asigna automáticamente a cada unidad, por ejemplo, su categoría, su correspondiente lema, características morfológicas, información sintáctica, etc., a partir de un archivo de diccionario que el programa utiliza para asignar la

etiqueta adecuada a cada expresión (2010, p. 35).

El *token* al cual se refiere Baquero corresponde, en lenguaje computacional, a cada una de las cadenas de caracteres dividida por espacios; en otras palabras, un *token* es igual a una palabra. El proceso por el que los datos son divididos en *tokens* se llama *tokenización,* y tal como Julia Baquero señala, facilita los diferentes tipos de procesamiento como las *frecuencias de aparición*, *las colocaciones* y *las concordancias*[115], ya que separa cada uno de los elementos del corpus.

Tras el proceso descrito anteriormente, se puede comenzar con la denominada anotación lingüística, la cual representa un tipo de análisis particular y un corpus. Un corpus puede contar con uno o más tipos de anotaciones:

- Lematización. En este caso, cada palabra va acompañada por su lema.
- Anotación morfológica o *part-of-speech* (pos). Las palabras tienen una etiqueta que corresponde a información morfológica.
- Anotación sintáctica o *parsing*. Cada palabra tiene información sintáctica.
- Anotación fonética.
- Anotación fonológica.
- Anotación prosódica.
- Anotación pragmática.
- Anotación discursiva.

La anotación se puede llevar a cabo de manera automática[116] o de manera manual; sin importar el método que se utilice, siempre debe existir una fase de revisión del material anotado. A continuación se presentan algunas herramientas computacionales que permiten la anotación o el procesamiento de corpus:

Software para procesamiento y anotación de corpus[117]Crecimiento y monitoreo

Software	**Localización**
Aconcorde[6]	http://www.andy-roberts.net/coding/aconcorde.
A.nnotate[7]	http://a.nnotate.com/.
Antconc[8]	http://www.antlab.sci.waseda.ac.jp/software.html.
Anvil[9]	http://www.anvil-software.org/.
Concapp[10]	http://concapp.software.informer.com/.
Corpus search[11]	http://corpussearch.sourceforge.net/.
Corpus wizard[12]	http://www2d.biglobe.ne.jp/~htakashi/software/cw2e.htm.
Elan[13]	http://tla.mpi.nl/tools/tla-tools/elan/.
Exmeralda[14]	http://www.exmaralda.org/.
Freeling[15]	http://nlp.lsi.upc.edu/freeling/.
ParaConc[16]	http://www.paraconc.com/.
Praat[17]	http://www.fon.hum.uva.nl/praat/.
Simple concordance program[18]	http://www.textworld.eu/scp/.
Svm-tool[19]	http://www.lsi.upc.edu/~nlp/SVMTool/#.
Textstat[20]	http://neon.niederlandistik.fu-berlin.de/textstat/.
Transcriber[21]	http://trans.sourceforge.net/en/presentation.php.
Tree tagger[22]	http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
VISL tools[23]	http://beta.visl.sdu.dk/visl2/.
Wraetlic tools[24]	http://alfonseca.org/eng/research/wraetlic.html.
Xaira[25]	http://xaira.sourceforge.net/.

La creación de un corpus no termina con la anotación de los elementos; un corpus requiere un continuo monitoreo, sobre todo cuando se encuentra en una fase piloto, puesto que en la mayoría de los casos y con los comentarios provenientes de los usuarios deben reajustarse fragmentos del corpus, mejorarse la interfaz y alimentar el corpus con nuevos datos.

Algunos comentarios sobre corpus orales

La idea de un corpus oral es que pueda contener la máxima diversidad de situaciones comunicativas posibles, a no ser que se quiera construir un corpus especializado, en el cual se recogerían muestras de situaciones comunicativas específicas, tales como exposiciones, discusiones y locuciones radiales[118]. Para

capturar muestras orales, muchas veces se utilizan entornos acústicamente controlados, como laboratorios de fonética o cabinas insonorizadas para evitar la influencia de los ruidos del ambiente. Cuando se cuenta con estas opciones, debe grabarse intentando obtener la mejor calidad y pasando las muestras por *software* especializados que mejoren la calidad del sonido, sin alterar las muestras.

Una de las fases más importantes, no obligatoria, en la construcción de un corpus oral es la transcripción[119], ya sea ortográfica, fonética, prosódica, etc. Para esto, se suelen usar alfabetos basados en el Alfabeto Fonético Internacional (AFI) que se puedan procesar informáticamente, tales como el Speech Assessment Methods Phonetic Alphabet (Sampa)[120].

Lo que se pretende con las transcripciones es evidenciar turnos de habla, variaciones de pronunciación, pausas, identidad del hablante, superposición entre locutores, fenómenos segmentales y fenómenos suprasegmentales. Estas transcripciones deben sincronizarse con la grabación.

Las fases que se presentaron en este capítulo pueden verse alteradas y modificadas por los objetivos, recursos y tipo de material que se recolecte; igualmente, esta es una propuesta que se puede modificar, según las necesidades que se presenten en la construcción de cada corpus.

---

103. Las características específicas hacen referencia a número de muestras que se quieren, tipos de registro y cualidades de los hablantes, entre otras; estas características están dadas por los objetivos y el tipo de corpus que se quiere construir.

104. Uno de los objetivos de cualquier corpus debería ser la posibilidad de que sus recursos lingüísticos sean siempre reutilizables.

105. Véase el apartado “Características de un corpus”.

106. Para más información sobre estos parámetros, Camino Rea (2010) hace un recorrido por estos cuatro aspectos en su texto *Getting on with Corpus Compilation: from Theory to Practice*.

107. Para ampliar sobre los tipos de corpus y sus características, véase el apartado “Tipología de los corpus”.

108. Los programas de reconocimiento de voz son entrenados bajo el léxico de un corpus, pero para que uno de estos programas funcione de manera correcta para la transcripción de un corpus debe contener los datos que se encuentran en el corpus oral que va a transcribir. Por tal motivo, es muy difícil que este proceso se dé automáticamente, ya que se debe contar con un *software* que contenga las características específicas de los archivos orales que se van a trabajar.

109. Ya que los corpus son digitales se requiere contabilizar el tamaño del material para de esta manera adquirir el espacio informático de almacenamiento donde estará contenido el corpus, puede ser espacio en la nube o dispositivos de almacenamiento como discos duros.

110. La manera en la que se ordenan los datos es una decisión de quien crea el corpus, lo que se recomienda es que exista una secuencia lógica en la forma como se nombran los archivos para de este modo sistematizarlos, por ejemplo nombrar cada archivo con un número, el tipo de material y la procedencia:

1orcol (or = oral, col = Colombia). La denominación de los archivos depende de las características de estos.

111. Para más información sobre un corpus simple, véase el apartado “Tipología de los corpus”.

112. Los sistemas de codificación (ASCII, ASCII Extendido, Unicode) cuentan con un número de caracteres computacionales, los cuales representan los caracteres de las diferentes lenguas del mundo; a mayor número de caracteres contenidos por el sistema, más fácil la representación del lenguaje por medios computacionales.

113. Los estándares son modelos claros de criterios para la codificación, el etiquetado y la anotación de un corpus.

114. La información adicional hace referencia a los procesos de anotación y adición de metadatos, explicados en el capítulo denominado “Características de un corpus”.

115. Para más información sobre estas categorías, véase el apartado “Usos de los corpus”.

116. Depende del acceso que se tenga a herramientas computacionales, y los resultados de la precisión de estas.

117. Para más información sobre herramientas computacionales, ingresar a: http://linguistech.ca/Online+Tools+-+home, http://www.uow.edu.au/~dlee/software.htm o http://linguistlist.org/sp/SearchWRListing-action.cfm?subclassid=7223&SearchType=LF&WRTypeID=2.

118. Estas situaciones comunicativas se definen a partir de los objetivos del corpus.

119. La *transcripción* es un proceso en el cual la lengua hablada se representa con caracteres escritos, en el caso de la *transcripción fonética* se busca representar los sonidos del habla, y cuando se habla de una *transcripción prosódica*, se representan los fenómenos suprasegmentales como el acento, el ritmo y la entonación mediante caracteres gráficos.

120. http://www.phon.ucl.ac.uk/home/sampa/.

La lingüística de corpus y la lengua española

La lengua española es una importante herramienta de comunicación internacional. Según cifras del último informe del Instituto Cervantes, *El español: una lengua viva* (2014), 548 millones de personas son hablantes de español, ya sea como lengua materna, segunda lengua, extranjera con dominio nativo o limitado, o son estudiantes; se cree además que en tres generaciones el 10 % de la población mundial hablará español. Así mismo, es la tercera lengua más utilizada en la red, produce el 10 % del PIB mundial, y según la base de datos del ISSN el 5 % del total de las revistas son en español.

Aun así, el impacto del español en el mundo científico no responde a la magnitud de la lengua; a propósito del tema, Melero, Badia y Moreno señalan: "A pesar del peso demográfico del español, de su posición como lengua de comunicación internacional y de la demanda actual del español como segunda lengua, su competitividad como lengua científica es seriamente cuestionada por el inglés" (n.d., p. 14).

Esta situación se ve claramente reflejada en la relación existente entre la lengua española y la lingüística de corpus. Rojo (2008), en su texto *Lingüística de corpus y lingüística del español*, enuncia que la LC en el español se ha desarrollado de manera atrasada en comparación con otras lenguas —como el inglés—, pero gracias al esfuerzo de diferentes equipos de investigación en el mundo hispánico, hoy en día se hace uso de la LC en los estudios de lengua española, lo que no quiere decir que no falte bastante camino por recorrer.

Uno de los elementos que ayudan al posicionamiento de una lengua en el mundo es la cantidad y la calidad de sus recursos lingüísticos[121]. Rafel y Soler se refieren así a este material: "El desarrollo de grandes corpus de referencia se ha convertido en uno de los primeros objetivos que deben cumplir las lenguas de un peso cultural y demográfico más destacado" (2003, p. 59). De esta manera, se hace explícita la necesidad de más corpus del español; aunque en forma intrínseca los estudios en lengua española se han valido de corpus lingüísticos, no es una metodología ampliamente conocida, desarrollada en publicaciones y mucho menos utilizada, en especial por países diferentes de España.

La creación y explotación de corpus requiere también recursos y herramientas computacionales, puesto que la mayoría de las herramientas que se encuentran en línea se han diseñado para el trabajo con material en lengua inglesa. Y si bien muchos de estos programas se pueden utilizar con material en español, es necesario diseñar herramientas que soporten los análisis propios de la lengua, de modo que los estudios en lengua española y los investigadores puedan apoyarse en la LC como una metodología.

Como se enunció anteriormente, la producción científica en español es muy reducida; algunos de los autores de publicaciones relacionadas con corpus y lengua española son Leonel Ruiz Miyares (Cuba), Julia Baquero (Colombia), Víctor M. Castel, Ana María Miret, Rodolfo Bonino y Lina Grasso (Argentina), Giovanni Parodi, René Venegas y Manuel Contreras (Chile), Mariela Grassi, Marisa Malcouri y Javier Couto (Uruguay), Luis Lara, Pedro Martín Butragueño y Yolanda Lastra (México), Guillermo Rojo, M. Paz Battaner, M. Antonia Martí, Irene Castellón Masalles, Joaquim Rafel, Joan Soler, Joaquim Llisterri, Joan Torruella, Manuel Alcántara Pla, Mario Barcala, Antonio Briz, Marta Albelda, Teresa Cabré, Carmen Bach, M. Luisa Carrio, Miguel Ángel Candel-Mora, Mar Cruz Piñol, Manuel Ezquerra, Juan Villena, Francisco Marcos, Francisco Navarro, Chantal Pérez, Pamela Benítez, Antonio Ortiz, Herminia Peraita, Pilar Sánchez-Gijón y María Rosa Vila y Milka Villayandre (España). Esta lista permite ver que aunque sí existe producción académica sobre el español, queda claro que la mayor parte de esta producción proviene de España.

La relación de la LC y el español comienza claramente en 1964 con el "Proyecto de estudio coordinado de la norma lingüística culta de las principales ciudades de España e Iberoamérica"[122], con el cual se buscaba construir un gran corpus[123] oral representativo del español culto de varias ciudades españolas e iberoamericanas. Aun cuando esta iniciativa no se pensó desde la lingüística de corpus, se enmarcó en esos parámetros, por lo que permitió el estudio de diversos fenómenos lingüísticos y el contraste entre las variedades del español, obviamente sin los apoyos tecnológicos de la actualidad.

En 1969, Paul Garvin y la Universidad Mayor de San Marcos de Perú publican uno de los primeros textos en español dedicados a los fundamentos y las herramientas informáticas necesarios para los trabajos en LC, denominado *Breve introducción a la computación lingüística*, considerado pionero en su área.

Años después, aparecen en el panorama de los estudios lingüísticos del español varios proyectos que desembocarían en la construcción de algunos corpus. En 1991, el "Proyecto para el estudio sociolingüístico del español de

España y de América"[124] (Preseea) da inicio a sus actividades de recolección y construcción de un corpus del español hablado representativo en su variedad geográfica y social de diferentes ciudades hispanohablantes como Alcalá de Henares, Buenos Aires, Culiacán, Lérida, Mérida, Montevideo, San Juan de Puerto Rico, Valencia, Granada, Lima, Oviedo, Santiago de Chile, Valparaíso, La Habana, Miami, Pereira, Medellín, Bogotá, Monterrey, Quito y Zaragoza, entre otras. Cada subcorpus, representativo de una ciudad específica, se encuentra en un estado diferente, algunos están en proceso de recolección, otros en fase de transcripción y unos restantes ya analizados y con material publicado.[125] En 2014, Preseea tiene en su página web[126] un corpus con información de Alcalá de Henares, Caracas, La Habana, Lima, Madrid, Medellín, Monterrey, Montevideo y Valencia, catalogada según el sexo del informante (hombre-mujer), la edad y el nivel de estudios.

Hacia finales de los años noventa, la Real Academia Española (rae) pone a disposición del público en general, de manera virtual y gratuita, dos nuevos corpus: el *Corpus de referencia del español actual*[127] (CREA) y el *Corpus diacrónico del español*[128] (Corde). El CREA cuenta con más de 160 millones de palabras extraídas de textos orales y escritos entre 1975 y 2004, provenientes en un 50 % de fuentes españolas y el otro 50 % de fuentes americanas; esto refleja la falta de equilibrio en la representatividad del corpus, ya que para ser equilibrado y representativo debería tener muestras de cada país hispanohablante, según su porcentaje de producción lingüística.

De igual manera, el crea es considerado un corpus de gran importancia para el español por ser el primero de su tipo y por su tamaño. A su vez, el Corde cuenta con 250 millones de palabras tomadas de textos escritos de diferentes géneros, que datan de todas las épocas y lugares donde se ha hablado español, desde su consolidación como lengua hasta el año 1975. Este corpus ha servido como material para la construcción del *Nuevo diccionario histórico del español*[129] (actualmente, en proceso de elaboración).

En 2001, Mark Davies crea un corpus, gratuito y de libre acceso, denominado *Corpus del español*, con más de cien millones de palabras procedentes de registros escritos de los siglos XIII al XX y registros hablados de este último siglo. La interfaz[130] permite que el usuario realice búsquedas de palabras, frases, lemas, categorías gramaticales, colocaciones y frecuencias[131].

Desde 1990, la relación entre la lingüística de corpus y el español se ha estrechado; esto se puede ver en el número de asociaciones y eventos

relacionados de alguna manera con la LC y el español, la constitución de grupos de investigación, la utilización de esta metodología por diferentes universidades y la creación de diversos corpus.

Asociaciones que desarrollan eventos o propuestas desde la LC

Asociación	**Link**
American Association for Corpus Linguistics (AACL)	http://aacl.sdsu.edu/
Asociación Española de Lingüística de Corpus (Aelinco)	http://www.um.es/aelinco/
Asociación Española de Lingüística Aplicada (Aesla)	http://www.aesla.org.es/es
Asociación de Lingüística y Filología de América Latina (Alfal)	http://www.mundoalfal.org/
Asociación Lingüística Sistémico-Funcional de América Latina (Alsfal)	http://www4.pucsp.br/isfc/alsfal/espanol/Inicio.html

Si bien existen más de cinco asociaciones dedicadas al trabajo con la lingüística, son las nombradas anteriormente las que de alguna manera desarrollan procesos o eventos relacionados con la LC. De las cinco, una se dedica especialmente al trabajo con esta metodología (Aelinco), dos de ellas son españolas (Aelinco y Aesla), dos son latinoamericanas (Alfal y Alsfal), y una realiza sus actividades desde Estados Unidos, enfocada principalmente en estudios sobre lenguas como el inglés y el español (AACL).

Eventos relacionados con la LC

Evento	**Institución organizadora**
Congreso Internacional de Lingüística de Corpus	Aelinco
AACL	American Association for Corpus Linguistics
Jornada de Corpus Lingüístics: Constitució, Etiquetatge i Explotación	Universidad Pompeu Fabra
Escuela Internacional de Verano de Lingüística de Corpus	Universidad Pompeu Fabra
Jornada de Divulgación de la Lingüística de Corpus/Corpus Linguistics: An Introductory Seminar and Workshop	Universidad de Salamanca

En 2014, el único evento activo fue el Congreso Internacional de Lingüística de Corpus, en su sexta edición; por su parte, el AACL se lleva a cabo cada dos años, por lo cual el último evento se celebró en la ciudad de San Diego (California), en 2013. Las Jornadas de Corpus Lingüístics ofrecidas por la Universidad Pompeu Fabra no han tenido continuidad desde finales de los

noventa, y la Escuela Internacional de Verano (2010), al igual que la Jornada de Divulgación, ha tenido una sola presentación, efectuada en 2007. Estos datos demuestran la falta de trabajo conjunto en el ámbito de la lengua española con respecto a la LC, pues mientras que en inglés se cuenta con eventos anuales como Corpus Linguistics Conference[132], Summer School in Corpus Linguistics[133], Workshop on Annotation[134] y el International Workshop on Treebanks and Linguistic Theories[135], en español solo existe un evento anual.

Grupos de investigación y universidades que trabajan con la LC

Dependencia	**Universidad**	**País**
Centro de Lingüística Teórica	Universidad Autónoma de Barcelona	España
Centre de Llenguatge i Computació (CliC)	Universidad de Barcelona	España
Corpus Multilingüe de Economía y Negocios (Comenego)	Universidad de Alicante	España
El Institut Universitari de Lingüística Aplicada (IULA)	Universidad Pompeu Fabra	España
Grupo de Análisis de las Lenguas de Especialidad (GALE)	Universidad Politécnica de Valencia	España
Grupo de Estructuras de Datos y Lingüística Computacional	Universidad de Las Palmas de Gran Canaria	España
Grupo de Fonética	Universidad Autónoma de Barcelona	España
Grupo para el Estudio de la Historia Lingüística Iberoamericana	Universidad de Valladolid	España
ILSE Grupo de Investigación	Universidad de Almería	España
Instituto Interuniversitario de Lenguas Modernas Aplicadas de la Comunidad Valenciana (Iulma)		España
Laboratorio de Lingüística Informática	Universidad Autónoma de Madrid	España
Research Group for Multidimensional Corpus-based Studies in English (Muste)	Universidade da Coruña	España
Grupo de Investigación Procesos de Gramaticalización en la Historia del Español (Programes)	Universidad Complutense de Madrid	España
Valencia Español Coloquial (Val. Es.Co)	Universidad de Valencia	España
Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía (Lacell)	Universidad de Murcia	España
Lingüística Aplicada Computacional, Enseñanza de Lenguas y Lexicografía (Lacell)	Universidad Nacional Autónoma de México	México
Grupo de Ingeniería Lingüística (GIL)	Universidad Nacional Autónoma de México	México
Grupo de Ingeniería Lingüística (GIL)	Universidad Nacional Autónoma de México	México

Laboratorio de Estudios Fónicos	El Colegio de México	México
Escuela Lingüística de Valparaíso	Universidad de Valparaíso	Chile
Grupo de Lingüística Hispánica	Universidad de los Andes (Mérida, Venezuela)	Venezuela
Instituto de Investigaciones Lingüísticas	Universidad de Costa Rica	Costa Rica
Grupo de Investigación en Lingüística de Corpus	Instituto Caro y Cuervo	Colombia
Grupo de Investigación en Traducción y Nuevas Tecnologías	Universidad de Antioquia	Colombia

En cuanto a los grupos de investigación y la divulgación de la LC en el ámbito académico del español, se encuentran en la actualidad (2014) veinticuatro equipos de trabajo que se dedican a la creación o explotación de corpus, o al trabajo desde la lingüística computacional basado en corpus[136]. De los veinticuatro grupos, más de la mitad se encuentran en territorio español; solamente nueve son latinoamericanos, con una fuerte presencia mexicana.

Corpus nacionales

	Corpus	**País**	**Localización**
1	*Corpus del español mexicano contemporáneo*[1] (CEMC)	México	http://www.corpus.unam.mx:8080/cemc/.
2	*Corpus histórico del español de México*[2] (CHEM)	México	http://saussure.ii.unam.mx/chem/.
3	*Corpus lingüístico de referencia de la lengua española en Chile*[3]	Chile	http://www.lllf.uam.es/ESP/Chile.html.
4	*Corpus lingüístico de referencia de la lengua española en Argentina*[4]	Argentina	http://www.lllf.uam.es/ESP/Argentina.html.

En español existen cuatro corpus representativos de la variedad hablada de cada país. México cuenta con el *Corpus del español mexicano contemporáneo* y el *Corpus histórico del español de México,* los dos con acceso gratuito en línea, mientras que Chile tiene un *Corpus de referencia del español de Chile* y Argentina un *Corpus de referencia del español de Argentina,* los dos donados al *Corpus del español*, de Mark Davies. Se encuentran también corpus nacionales como *American National Corpus*[137] (Estados Unidos), *British National Corpus*[138] (Inglaterra), *Thai National Corpus*[139] (Tailandia), *Hungarian National Corpus*[140] (Hungría), *Cˇ eský národní korpus*[141] (República Checa) y *Hellenic National Corpus*[142] (Grecia), entre otros.

Corpus del español

	Corpus		**Corpus**
1	ABC	42	Corpus oral del lenguaje adolescente (COLA)
2	Adquisición, desarrollo y representación de categorías semánticas en niños de edad escolar	43	Corpus oral peninsular
3	Albayzín	44	Corpus oral y sonoro del español rural (Coser
4	Alfal	45	Corpus para el estudio del español hablado en Santiago de Compostela - CSC
5	Almecor	46	Corpus sociolingüístico de Mérida, Venezuela (CSMV)
6	Análisis de la conversación de la Universidad de Alcalá de Henares (Acuah)	47	Corpus textual del español periodístico
7	Análisis del discurso oral	48	Cráter
8	Análisis del discurso público actual (ADPA)	49	Cumbre
9	Briscoe	50	DIES - RTP
10	Caracas 77	51	Diferencias individuales en la adquisición del lenguaje
11	Caracas 87	52	DIMEx100
12	CATE	53	Disponibilidad léxica de los adolescentes
13	Cedel2	54	El corpus virtual de la red
14	Ceudex	55	El Grial
15	Corpus 92	56	El Mundo 1994-1995
16	Corpus anotado con relaciones discursivas - RST Spanish Treebank	57	Espal
17	Corpus de contextos definitorios (Corcode)	58	FAE-Esp Can
18	Corpus de conversación coloquial - Valencia español coloquial	59	Frecuencia de elementos léxicos en manuales de preescolar
19	Corpus de documentos coloniales (Mérida, Venezuela)	60	Gaudí
20	Corpus de documentos españoles anteriores a 1700	61	Hamburg Corpus of Argentinean Spanish (HaCASpa)
21	Corpus de encuestas de Asunción de Paraguay (CEAP)	62	Hopinion
22	Corpus de las sexualidades en México (CSMX)	63	LAN
23	Corpus de referencia del español actual (CREA)	64	Legebidium
24	Corpus de verificación del sistema de diccionarios y gramáticas electrónicos del español (CorVerifSDGEE)	65	Lejes

25	Corpus de vocabulario del niño de 6 a 14 años	66	Léxico informatizado del español (Lexesp)
26	Corpus del español	67	Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico (MC-NLCH)
27	Corpus del español actual (CEA)	68	Multext
28	Corpus del español mexicano contemporáneo (CEMC)	69	Número
29	Corpus del habla de Almería	70	PA 85/86 - Corpus de dígitos
30	Corpus del Nuevo Diccionario Histórico del Español (NDHE)	71	PA 85/86 - Corpus de letras
31	Corpus diacrónico del español (Corde)	72	Proyecto para el Estudio Sociolingüístico del Español de España y de América (Preseea)
32	Corpus digital del español colonial mexicano (Corecom)	73	Spanish FrameNet (SFN)
33	Corpus histórico del español de México (CHEM)	74	Spatis
34	Corpus informatizado: Textos del español de Uruguay (Corin)	75	Tangora
35	Corpus lingüístico de definiciones de categorías semánticas de personas mayores sanas y con la enfermedad de Alzheimer	76	Telémaco
36	Corpus lingüístico de referencia de la lengua española en Argentina	77	TIC-0448/89
37	Corpus lingüístico de referencia de la lengua española en Chile	78	UAM-Treebank
38	Corpus lingüístico en ingeniería (CLI)	79	Variedades urbanas andaluzas (VUA)
39	Corpus oral de la variedad juvenil universitaria del español hablado en Alicante (COVJ)	80	Vestel
40	Corpus oral de referencia del español contemporáneo (Corlec)	81	Vox-Bibliograf
41	Corpus oral de referencia del español contemporáneo		

Además de los ya conocidos Corde, CREA y *Corpus del español*, existen diversos corpus en español hechos con diferentes fines; por ejemplo, algunos corpus realizados por México, Venezuela y Paraguay se donaron al CREA, como el CEAP, *Caracas 77* y *Caracas 87*. Los corpus que se presentan en la tabla anterior (tabla 8) no corresponden a todos los existentes en lengua española, pero sí muestran un espectro del lugar donde se encuentra la LC en relación con la lengua. La mayoría de los corpus son representativos de la variedad ibérica, lo que denota poco desarrollo de recursos lingüísticos representativos de otras variedades del español. Cabe señalar que la mayor parte de ellos se puede consultar a través de internet.

Si bien la relación entre la LC y el español es cada vez más fuerte, todavía

son diversos los campos en los que se puede explorar. El uso de esta metodología brinda la posibilidad de conocer mejor las características de las variedades de la lengua española; además, esta relación ofrece oportunidades como las que Berber (2011) plantea: oportunidad de innovación, interdisciplinariedad, creación de comunidad investigativa en lengua materna y exploración del contexto local.

Al español le hacen falta corpus representativos de cada país, diccionarios de frecuencias, gramáticas basadas en usos reales, corpus de aprendices y *software* especializados en la lengua; adicionalmente, el área científico-académica del español debe superar las barreras metodológicas y tecnológicas para ser más competitivos en el campo de la investigación lingüística, divulgar efectivamente publicaciones en español y hacer que la LC forme parte de los currículos de los pregrados y posgrados relacionados con la lingüística. De esta manera, la relación entre la LC y el español se consolidará.

---

121. Con recursos lingüísticos se hace referencia a la literatura, los diccionarios y los corpus, entre otros.

122. Desde el 2003 este proyecto se conoce con el nombre de "Proyecto de la norma culta hispánica Juan M. Lope Blanch".

123. El término *corpus* no se utilizó durante el desarrollo del proyecto, pero el resultado fue la constitución de un corpus no digital.

124. http://preseea.linguas.net/.

125. Esta información se encuentra actualizada a 2014.

126. http://preseea.linguas.net/Corpus.aspx.

127. http://corpus.rae.es/creanet.html.

128. http://corpus.rae.es/cordenet.html.

129. http://web.frl.es/DH/org/login/Inicio.view.

130. La *interfaz* corresponde al programa info7rmático que permite la interacción del usuario con el corpus.

131. Para más información sobre *colocaciones* y *frecuencias*, véase el apartado "Características de un corpus" o el Glosario.

132. http://ucrel.lancs.ac.uk/.

133. http://ucrel.lancs.ac.uk/summerschool/corpusling.php.

134. http://www.ling.uni-potsdam.de/acl-lab/law2014/.

135. http://tlt13.sfs.uni-tuebingen.de/.

136. Es posible que existan más grupos que trabajen con la LC, pero estos veinticuatro son los que más trabajos han desarrollado a partir de la metodología.

137. http://www.americannationalcorpus.org/OANC/index.html.

138. http://www.natcorp.ox.ac.uk/.

139. http://www.arts.chula.ac.th/~ling/TNC/.
140. http://corpus.nytud.hu/mnsz/index_eng.html.
141. https://www.korpus.cz/.
142. http://hnc.ilsp.gr/en/.

Consideraciones finales

> Así las cosas, el desarrollo de la LC continúa en un marco extraordinariamente interesante y en ebullición. Las implicancias que la perspectiva teórica que (ya sea *profunda* o *superficial*) pueda traer consigo (Hunston & Thompson, 2006) anuncian —en alguna medida— que estamos en medio de un proceso de cambios y ajustes, y avanzando hacia una mirada cada vez más compleja y enriquecida de los objetos de estudio. Miradas que ciertamente potencian las indagaciones empíricas del lenguaje y de las lenguas particulares, desde múltiples puntos de mira y haciendo confluir aproximaciones antes impensadas (Parodi, 2008, p. 118).

La lingüística de corpus se constituye en una metodología para la investigación y el análisis de datos de la lengua en uso. Su campo de aplicación se expande cuando se recurre a herramientas informáticas, ya que permiten el almacenamiento, la sistematización y la explotación de grandes cantidades de material lingüístico; dicha metodología toma cada vez más fuerza y son más los corpus que se crean día tras día. A propósito de esto, Rafel y Soler comentan:

> En la actualidad, la cantidad de corpus existentes y de proyectos de constitución de corpus crece cada día, hasta el punto de que se hace difícil dar una relación de los mismos. Hay direcciones web específicas que están actualizadas periódicamente, donde puede encontrarse información sobre diferentes corpus (2003b, p. 59).

Algunas de estas páginas web en las que se puede encontrar información sobre corpus y herramientas para su explotación son:

- http://www.meta-share.org/.
- http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/
- http://www.ling.ohio-state.edu/~dickinso/corpus.html.
- http://ucrel.lancs.ac.uk/#sec.
- https://www.ldc.upenn.edu/.
- http://www.uow.edu.au/~dlee/CBLLinks.htm.
- http://www.helsinki.fi/varieng/CoRD/corpora/.

Con este libro se busca, además de brindar bases teóricas sobre la LC y los corpus, despertar el interés de estudiantes, profesores, académicos e investigadores, para que exploren y utilicen la lingüística de corpus en sus clases, proyectos e investigaciones, de manera que se difunda esta metodología

en el panorama latinoamericano; por esto, nuestras consideraciones finales van encaminadas hacia lo que falta hacer en esta área:

- Los corpus deben ser novedosos, representativos, variados y reutilizables. Construir un sinnúmero de corpus similares, que incluyan o trabajen los mismos fenómenos, solo desemboca en la acumulación, no sistemática, de recursos lingüísticos.
- Las lenguas con más peso cultural y demográfico requieren un corpus de referencia. Por eso, el español necesita corpus de cada variedad lingüística.
- Dada la variedad lingüística latinoamericana, se hace necesario documentar las diferentes lenguas, lo que sugiere la creación de nuevos corpus.
- La LC y la explotación de corpus abren el espectro investigativo a diferentes áreas interesadas en el lenguaje, no solamente a la lingüística; estudios que responden a múltiples necesidades pueden resultar del análisis de datos lingüísticos de la lengua en uso.
- Aunque existen iniciativas como TEI (Text Encoding Iniciative)[143], se debe buscar la estandarización de parámetros de construcción de corpus; de este modo, el material lo pueden utilizar investigadores de diversos campos y disciplinas.
- La oferta de herramientas computacionales para la explotación de corpus es variada, pero no suficiente; por esto se requiere la creación de nuevas tecnologías, especialmente para el trabajo de la lengua española, con especial énfasis en los procesos de anotación.
- La explotación de la web como corpus requiere atención, sobre todo de estudiantes, profesores, académicos e investigadores hispanohablantes.
- Las publicaciones científicas y académicas en español, tanto en libros como en revistas, deben verse más permeadas por la LC, así como sucede en lenguas como el inglés.
- Los programas universitarios de pregrado, posgrado e investigación relacionados con el lenguaje deben incluir en sus currículos materias relacionadas con la lingüística de corpus y la lingüística computacional, especialmente en Latinoamérica. Parodi comenta al respecto:

- La superación de la barrera metodológica y tecnológica no puede esperar si queremos, efectivamente, producir investigación competitiva y de primer orden, acompañada de publicaciones indexadas de amplia difusión en nuestra lengua. La docencia de pregrado y de posgrado exige que así sea, para que —entre otros— la superación de la brecha digital deje de ser una utopía y el acceso al conocimiento especializado esté disponible democráticamente (2010, p. 166).
- Los departamentos universitarios de lingüística, ingenierías e informática pueden trabajar en proyectos conjuntos, de manera que se formen expertos en el área de la LC.
- La investigación en lingüística de corpus es una tarea de la academia, las entidades gubernamentales e incluso las industriales.
- La unión entre centros universitarios, editoriales y empresas de tecnología pueden contribuir a la creación y explotación de corpus y, por supuesto, a la creación de nuevas herramientas informáticas.

---

143. Véase el Glosario.

Glosario

Anotación
Adición de información lingüística (fonética, morfológica, semántica, etc.) a cada uno de los elementos de un corpus.

Anotación paralingüística
Adición de información a cada uno de los elementos de un corpus, sobre datos no lingüísticos que acompañan las situaciones comunicativas, como signos fisiológicos o emocionales, el volumen de la voz y el ritmo.

Anotación prosódica
Adición de información de elementos paralingüísticos propios de la oralidad, como la entonación, las pausas, el ritmo y los acentos, entre otros, a cada uno de los elementos de un corpus.

Archivo informatizado
Conjunto de textos en soporte digital, de características diversas en cuanto a fechas, estructuras y temas, que busca la conservación de material textual.

Biblioteca de textos electrónicos
Colecciones de textos digitales, almacenados en un formato estándar y organizados según áreas del conocimiento humano, con el fin de facilitar las búsquedas.

Coocurrencia
Apariciones frecuentes de diferentes elementos lingüísticos dentro de un mismo contexto.
Ejemplo: la palabra *dinero* en un corpus de revistas financieras tiene una elevada frecuencia de aparición, acompañada de las palabras *lavado* y *de*, formando la expresión *lavado de dinero*.

Codificación
Proceso de conversión del lenguaje natural a caracteres susceptibles de procesamiento por programas computacionales.

Coligación (*Colligation*)
Secuencia de palabras en la que un término léxico coocurre a menudo con una

categoría gramatical.

Colocación
Secuencia de términos léxicos que coocurren frecuentemente en una misma lengua.

Componente
Constituyente de un corpus que corresponde a colecciones de muestras de lengua que comparten un mismo criterio lingüístico, como la variedad, el registro y la procedencia.

Concordancia
Lista de todas las ocurrencias de una palabra o término específico dentro de un contexto o número determinado de elementos que la acompañan antes o después de su aparición.

Corpus
Conjunto de textos en formato digital, recolectados, almacenados y sistematizados de acuerdo con criterios lingüísticos como muestra representativa de una lengua o variedad.

Enfoque basado en corpus (*corpus-based*)
Forma de trabajo desde la lingüística de corpus en la que el investigador conoce la teoría, tiene hipótesis y busca validarlas o rechazarlas mediante los datos del corpus.

Enfoque guiado por corpus (*corpus driven*)
Forma de trabajo desde la lingüística de corpus en la que a partir de la observación de patrones o fenómenos encontrados en un corpus se llega a la formulación de hipótesis.

Estándar de codificación
Referencia que permite entender, manejar y guiar los procesos y códigos empleados por un *software*. Algunos estándares para manipulación de corpus son *Expert Advisory Group on Language Engineering Standards* (Eagles) y *Text Encoding Initiative* (TEI).

Etiqueta
Secuencia de caracteres de algún lenguaje de marcado (xml, hatml, sgml) que contiene información adicional acerca de los elementos del corpus, los documentos o un corpus en general.

Etiquetador
Programa computacional encargado de adicionar cualquier tipo de información extra al corpus y sus elementos.

Etiquetado gramatical o morfológico (*part-of-speech,* pos)
Proceso de anotación en el que se asigna una *etiqueta* a cada palabra del corpus donde se indica la categoría gramatical, según el contexto.

Frecuencia
Número de veces que un mismo elemento (morfema, palabra, expresión, patrón gramatical) aparece dentro de un corpus.

Funcionalismo
Enfoque a la teoría lingüística que busca explicar la lengua a partir de referencias de uso.

Interfaz gráfica
Programa computacional que permite y facilita la interacción del usuario con el corpus.

Lenguaje de Marcas de Hipertexto (*Hypertext Markup Language, html*)
Sistema de codificación utilizado para agregar etiquetas que indican al navegador cómo estructurar y mostrar contenido, especialmente en la web.

Lingüística computacional
Disciplina de la lingüística aplicada y la inteligencia artificial encargada del estudio, diseño y elaboración de modelos computacionales capaces de simular las habilidades lingüísticas del ser humano.

Metadato
Información estructurada que describe el contenido y las características de los datos, los textos y los corpus, y que a su vez permite hacer búsquedas dentro de la colección.

Palabra clave (*key word*)
Término que, por su alta frecuencia de aparición en comparación con otros corpus, se convierte en propio y representativo del corpus al que pertenece.

Programa de concordancias
Herramientas computacionales de análisis textual que generan listas de ocurrencias de palabras que generalmente van juntas.

Recursos lingüísticos
Material propio y representativo de una lengua o variedad, como la producción literaria, los diccionarios y los corpus.

Representatividad
Rasgo ideal de un corpus para comportarse como un modelo de la lengua mostrando sus partes y tendencias, y constituyéndose en una referencia.

Sistema de codificación
Lenguaje compuesto por caracteres computacionales capaz de representar los caracteres propios de las diferentes lenguas. Algunos sistemas de codificación son ASCII, ASCII Extendido y Unicode.

Subcorpus
División de un corpus en porciones más pequeñas con características comunes y que pueden funcionar de manera independiente.

Token
Unidad informática o componente léxico (palabra) compuesto por caracteres propios de algún lenguaje de programación, en los que se divide cada uno de los textos de un corpus.

Transcripción
Proceso manual o automático en el cual la lengua hablada se representa con caracteres escritos. Puede ser fonética cuando se representan los sonidos del habla y prosódica cuando se representan, mediante caracteres gráficos, fenómenos suprasegmentales de la lengua como la entonación y el acento.

Bibliografía

Alcántara, M. (2007). *Introducción al análisis de estructuras lingüísticas en corpus. Aproximación semántica*. Madrid: UAM Ediciones.

Atkins, S., Clear, J. & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing*, *7*(1), 1-16. doi:10.1093/llc/7.1.1.

Baker, P., Gabrielatos, C., KhosraviNik, M., Krzyzanowski, M., McEnery, T. & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3), 273-306. Doi:10.1177/0957926508088962.

Baker, P. & Hardie, A. (2006). *A Glossary of Corpus Linguistics*. Manchester: Edinburgh University Press.

Baquero, J. (2010). *Lingüística computacional aplicada*. Bogotá: Universidad Nacional de Colombia.

Berber, T. (2011). Corpus linguistics in South America. En *Perspectives on Corpus Linguistics* (pp. 29-45). Amsterdam/Philadelphia: John Benjamins Publishing.

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

Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

BNC Consortium. (2007). British National Corpus [Text]. Recuperado el 7 de marzo de 2014 de http://www.natcorp.ox.ac.uk.

Chafe, W. (1992). The Importance of Corpus Linguistics to Understanding the Nature of Language. En *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991* (pp. 79-97). Estocolmo: Walter de Gruyter.

Chomsky, N. (1966). *Linguistique cartésienne*. París: Seuil.

Cicres, J. (2011). La lingüística forense y el uso de los corpus lingüísticos. En *Actas del III Congreso internacional de lingüística de corpus. Tecnologías de la información y las comunicaciones: presente y futuro en el análisis de corpus*. Valencia: Universidad Politécnica de Valencia.

Corpus Linguistics: Method, Analysis, Interpretation - Future Learn (2014).

Course, Lancaster University. Recuperado de https://www.futurelearn.com/courses/corpus-linguistics/todo/241.
Cortez, Godínez, J. (2010). El corpus *ad hoc* como herramienta de traducción. En *Memorias del VI Foro de Estudios en Lenguas Internacionales*. Chetumal: Universidad de Quintana Roo.
Cruz, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: Arco Libros.
Davies, M. (s.f.). Corpus del español. Recuperado el 7 de marzo de 2014 de http://www.corpusdelespanol.org.
Economic and Social Research Council (2008). BSL Corpus Project. Recuperado el 28 de febrero de 2014 de http://www.bslcorpusproject.org.
Flórez, L., Montes, J., Mora, S., Rodríguez, M., Figueroa, J. & Lozano, M. (1982). *Atlas lingüístico-etnográfico de Colombia (ALEC)*. Bogotá: Instituto Caro y Cuervo.
Francis, N., Kučera, H. & Mackie, A. W. (1982). *Frequency analysis of English usage: lexicon and grammar*. Boston: Houghton Mifflin.
González, A. & Otálora, H. (1986). *El habla de la ciudad de Bogotá: materiales para su estudio*. Bogotá: Instituto Caro y Cuervo.
Gries, S. (2009). What is Corpus Linguistics? *Language and Linguistics Compass*, *3*(5), 1225-1241. doi:10.1111/j.1749-818X.2009.00149.x.
Grupo de Tecnología del Habla de la Universidad Politécnica de Madrid (s.f.). Corpus lingüísticos. Recuperado de http://lorien.die.upm.es/juancho/pfcs/AJP/cap4.pdf.
Hrušková, J. (2008). *Los corpus crea y Corde en el contexto de los corpus lingüísticos*.
ICE Teams (1990). International Corpus of English (ice). Recuperado el 7 de marzo de 2014 de http://ice-corpora.net/ice.
Instituto Cervantes (2014). *El español: una lengua viva*. Madrid: Instituto Cervantes.
Kabatek, J. (2012). ¿Es posible una lingüística histórica basada en un corpus representativo? Recuperado de https://www.academia.edu/2299020/_Es_posible_una_linguistica_historica_l
Kennedy, G. (1998). *An introduction to corpus linguistics*. Londres, Nueva York: Longman.
Lastra, Y. (2008). Futuro perifrástico y futuro morfológico en el corpus sociolingüístico de la Ciudad de México. Presentado en el XV Congreso Internacional de la Alfal. Montevideo.

Leech, G. (1991). The state of the art in corpus linguistics. Recuperado el 9 de agosto de 2013 de http://ccl.pku.edu.cn/doubtfire/CorpusLinguistics/Introduction/The%20state%
Leech, G. (2011). Principles and applications of Corpus Linguistics. En *Perspectives on Corpus Linguistics* (pp. 155-170). Amsterdam/Philadelphia: John Benjamins Publishing.
López, F., Méndez, C., Sierra, G. & Solórzano, J. (2013). Exploración de medidas estilométricas para atribución de autoría. Presentado en el III Seminario de Lingüística Forense. México, D.F.
Maher, J. & Groves, J. (2007). *Chomsky para todos*. Barcelona: Paidós.
McEnery, T. (2001). *Corpus Linguistics: An Introduction*. Manchester: Edinburgh University Press.
McEnery, T. & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice.* Cambridge, Nueva York: Cambridge University Press.
McEnery, T. & Wilson, A. (2012). ICT4LT Module 3,4 Corpus Linguistics. Recuperado el 9 de agosto de 2013 de http://www.ict4lt.org/en/en_mod3-4.htm.
McEnery, T., Xiao, R. & Tono, Y. (2006). *Corpus-based Language Studies: An advanced resource book*. Londres, Nueva York: Routledge.
Melero, M., Badia, T. & Moreno, A. (s.f.-b). *La lengua española en la era digital*. Barcelona: Springer.
Mercado, H. (2008). *Fundamentos de la lingüística de corpus*.
Montes, J., Mora, S., Espejo, M., Figueroa, J., Lozano, M., Ramírez, R. & Duarte, G. (1998). *El español hablado en Bogotá*. Bogotá: Instituto Caro y Cuervo.
Palacios, M. & Sierra, G. (2011). Corpus para el análisis del discurso del concepto ad hoc- cracia. En *Actas del III Congreso Internacional de Lingüística de Corpus. Tecnologías de la información y las comunicaciones: presente y futuro en el análisis de corpus*. Valencia: Universidad Politécnica de Valencia.
Parodi, G. (2005). Discurso especializado y lingüística de corpus: hacia el desarrollo de una competencia psicolingüística. *Boletín de Lingüística*, *23*, 61-88.
Parodi, G. (2007a). Lingüística de corpus: puntos de mira. En *Lingüística de corpus y discursos especializados: puntos de mira* (pp. 13-30). Valparaíso: Ediciones Universitarias de Valparaíso.
Parodi, G. (2007b). *Working with Spanish corpora*. Londres, Nueva York:

Continuum.
Parodi, G. (2008). Lingüística de corpus: una introducción al ámbito. *RLA. Revista de Lingüística Teórica y Aplicada*, *46*(1), 93-119. doi:10.4067/S0718-48832008000100006.
Parodi, G. (2010). *Lingüística de corpus: de la teoría a la empiria*. Madrid/Frankfurt: Iberoamericana.
Parodi, C. & Carrera, M. (2011). *Informe de las actividades del proyecto para la historia del español de América*. Madrid: Alfal.
Peraita, H. & Grasso, L. (2010). Corpus lingüístico de definiciones de categorías semánticas de sujetos ancianos sanos y con la enfermedad de Alzheimer. Una investigación transcultural hispano-argentina. Madrid, Buenos Aires: Fundación bbva.
Procházková, P. (2006). Fundamentos de la lingüística de corpus. "Concepción de los corpus y métodos de investigación con corpus". Recuperado de http://prochazkova.de/fundamentos_de_la_ling%C3%BC%C3%ADstica_de_
Rafel, J. & Soler, J. (2003). El procesamiento de corpus. La lingüística empírica. En *Las tecnologías del lenguaje* (p. 295). Barcelona: Editorial uoc.
Rea, C. (2010). Getting on with Corpus Compilation: from Theory to Practice. *ESP World*, *9*.
Real Academia Española (2001). *Diccionario de la lengua española* (22.a ed.). Madrid: Espasa.
Real Academia Española (s.f.-a). Corpus de referencia del español actual (crea). Recuperado el 7 de marzo de 2014 de http://corpus.rae.es/creanet.html.
Real Academia Española (s.f.-b). Corpus diacrónico del español (Corde). Recuperado el 7 de marzo de 2014 de http://corpus.rae.es/cordenet.html.
Rojo, G. (2008). Lingüística de corpus y lingüística del español. Presentado en el XV Congreso de la Alfal. Montevideo.
Rojo, G. (2009). Sobre la construcción de diccionarios basados en corpus. *Revista Tradumàtica*. Recuperado de http://webs2002.uab.es/tradumatica/revista/num7/articles/02/02art.htm.
Semino, E. (2008). *Metaphor in discourse*. Cambridge, UK; Nueva York: Cambridge University Press.
Semino, E. (2013). *Corpus methods and a questionnaire for the diagnosis of pain symptoms*. Presentado en Ucrel crs, Lancaster University.
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
Soler, V. (2007). Patrones lingüísticos para la búsqueda de información

conceptual en el corpus textual especializado de la cerámica TXTCerama (p. 14). Presentado en Jornades de Foment de la Investigació. Valencia. Recuperado de http://www.uji.es/bin/publ/edicions/jfi10/trad/14.pdf.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam/Philadelphia: John Benjamins Publishing.

Torruela, J. & Llisterri, J. (1999a). Diseño de corpus textuales y orales. En *Filología e informática: nuevas tecnologías en los estudios filológicos* (pp. 45-77). Barcelona: Milenio.

Venegas, R. (2010). *Lingüística de corpus: métodos y herramientas para el análisis del discurso escrito*. Recuperado de http://www.slideserve.com/ellie/ling-stica-de-corpus-m-todos-y-herramientas-para-el-an-lisis-del-discurso-escrito.

Viana, V., Zyngier, S. & Barnbrook, G. (2011). *Perspectives on corpus linguistics*. Amsterdam/Philadelphia: John Benjamins Publishing.

Villayandre, M. (2006). Lingüística de corpus. Recuperado el 9 de agosto de 2013 de http://fhyc.unileon.es/Milka/LCII/LC1.htm.