SELECCIÓN Y EVALUACIÓN DE FUENTES DE INFORMACIÓN PARA LA ELABORACIÓN DE UN CORPUS LINGÜÍSTICO

103698 – Trabajo de Fin de Grado

Grado en Traducción e Interpretación
Curso académico 2015 - 2016

Estudiante: Núria Valls Rodríguez
Tutora: Mª Pilar Cid Leal

Facultad de Traducción e Interpretación
Universitat Autònoma de Barcelona

DATOS DEL TFG

Título: Selección y evaluación de fuentes de la información para la elaboración de un corpus.

Autor: Núria Valls Rodríguez.

Tutor: Mª Pilar Cid Leal.

Centro: Facultad de Traducción e Interpretación.

Estudios: Grado en Traducción e Interpretación.

Curso académico: 2015-16.

PALABRAS CLAVE

Corpus, fuentes bibliográficas, gastronomía, cocina, alimentación.

Corpus, bibliographical sources, gastronomy, cuisine, food.

Resumen del TFG

El presente trabajo pretende analizar y seleccionar diversas fuentes bibliográficas que sirvan de ayuda para hacer un corpus lingüístico, concretamente en el ámbito de la cocina y la gastronomía. Para ello, se presenta una primera parte introductoria sobre la teoría de los corpus lingüísticos y sus elementos más destacados como por ejemplo su definición, su historia y sus particularidades más relevantes. En los siguientes capítulos, se desarrolla el análisis y selección de fuentes bibliográficas útiles para toda persona que quiera hacer un corpus lingüístico siguiendo los criterios de evaluación de fuentes de información basándonos en unos determinados parámetros con el fin de valorarlas y comentarlas y con el propósito de crear una selección bibliográfica favorable y práctica.

Summary

This project aims to analyse and select various bibliographical sources which should help to make a linguistic corpus, particularly in the field of cooking and gastronomy. To do this, an introductory part on the theory of linguistic corpus and its salient elements such as its definition, its history and its most important characteristics is presented. In the following chapters, analysis and selection of useful bibliographic sources for anyone who wants to make a linguistic corpus is developed following the evaluation criteria of information sources based on certain parameters in order to evaluate them and discuss them with the purpose of creating a favourable and practical bibliographic selection.

Aviso legal

Legal notice

AGRADECIMIENTOS

Me gustaría agradecer a varias personas la ayuda y el apoyo incondicional que me han dado a lo largo de la creación del presente Trabajo de Fin de Grado.

En primer lugar, agradezco a mis padres que hayan estado a mi lado durante todo el proceso y elaboración del TFG, por haber hecho que este reto personal también fuese suyo. Especialmente le agradezco a mi madre el tiempo dedicado a la lectura y a la crítica de mi trabajo, y a darme una visión objetiva para mejorarlo.

En segundo lugar, agradezco a mi tío, Ignacio, por brindarme toda la ayuda e información necesaria para poder hacer el trabajo más interesante. También agradecerle las ideas y cambios propuestos que, sin duda, han hecho que el proyecto tenga una forma diferente.

Finalmente, y no por eso menos importante, quiero agradecer a mi tutora, Mª Pilar Cid, su tiempo, cercanía y ayuda que han hecho que una pequeña idea pudiese convertirse en este proyecto.

A todos, muchas gracias.

ÍNDICE

1. INTRODUCCIÓN....1
2. LOS CORPUS LINGÜÍSTICOS....4
2.1. ¿QUÉ ES UN CORPUS LINGÜÍSTICO?....4
2.2. HISTORIA DE LOS CORPUS LINGÜÍSTICOS....8
2.3. CLASIFICACIÓN DE LOS CORPUS....10
2.3.1 Colecciones de textos....10
2.3.2 Niveles de corpus....11
2.3.3. Tipología de corpus....11
2.4. ASPECTOS PARA DISEÑAR UN CORPUS....17
2.4.1. Aspectos generales....17
2.4.2. Aspectos específicos de los corpus orales....19
2.5. CORPUS DE ESPAÑOL DE ESPAÑA Y AMÉRICA....19
2.6. LOS CORPUS, INTERNET Y LA TRADUCCIÓN....21
3. CONSTRUCCIÓN (HIPOTÉTICA) DE UN CORPUS....23
4. SELECCIÓN Y EVALUCIÓN DE FUENTES DE INFORMACIÓN....25
a) Obras sobre teoría y construcción de corpus lingüísticos....26
b) Recopilación de obras que compondrían el corpus....33
a. Recetarios....34
b. Libros de teoría de técnicas culinarias....37
c. Revistas gastronómicas....38
5. CONCLUSIONES....39
6. BIBLIOGRAFÍA....42

1. INTRODUCCIÓN

Desde hace tiempo me han llamado mucho la atención las diferencias lingüísticas que se observan entre los variados dialectos del español que se utiliza en los diversos países de habla hispana. Es fascinante que un mismo idioma pueda ser tan cambiante según el lugar del mundo en que se habla. En consecuencia empecé a plantearme la siguiente pregunta: ¿Cómo puedo saber cómo se dice o cómo se usa un término en diferentes sitios? Y, mientras como estudiante de traducción me introducía y aprendía nuevos idiomas, la curiosidad por mi propia lengua seguía estando presente. También fue durante el período de estudios cuando empecé a oír el concepto corpus lingüístico, tan lejano y extraño al principio, aunque finalmente me di cuenta que los corpus eran una herramienta muy útil que podía ser usada para aprender sobre mi lengua, así como también para la traducción profesional.

En el momento en que me planteé el tema del Treball de Fi de Grau (TFG) creí que sería una buena idea trabajar sobre los corpus en español de España y de América. A partir de aquí, y después de consultar a mi tutora, decidí que sería muy interesante hacer el trabajo sobre la búsqueda y la evaluación de fuentes de información para la construcción de un corpus lingüístico. Es decir, investigar cómo se empieza desde cero un corpus, buscar las fuentes más útiles para la construcción de un (hipotético) corpus y evaluarlas.

Este tema tiene un gran interés para los investigadores, porque antes de empezar cualquier proyecto, es muy importante saber encontrar las fuentes adecuadas y específicas para crear un corpus lingüístico. Es relevante decir que un proyecto como este puede ayudar tanto a estudiantes, profesores como a cualquier persona que se interese por hacer un corpus, ya que es muy difícil encontrar un marco teórico sin tener ninguna referencia o ayuda, pues hay mucha información y es difícil precisar cuál es la más adecuada. Por esta razón, comprendí que podría hacer una aportación bastante necesaria si elaboraba una recopilación de fuentes de la información sobre este tema. Partiendo de estas ideas, decidí, con ayuda de mi tutora, que este trabajo podría ser de gran interés para los estudiosos que trabajan en el campo de la lingüística.

La metodología utilizada para hacer el trabajo ha constado de dos partes muy diferenciadas. Para empezar, lo más importante era hacer un marco teórico, para así poder entender con más profundidad de qué trata un corpus lingüístico, ya que no solo es un conjunto de textos; la historia de estos, cuándo empezaron y cómo surgió la

necesidad de crear esta herramienta; los tipos de corpus que existen y para qué sirven cada uno de ellos; los corpus ya existentes de español de España y América; los aspectos que se tienen que tener en cuenta antes de empezar a hacer un corpus y, finalmente, como se usan los corpus en la traducción y en Internet.

La utilidad de hacer este marco teórico ha sido, aparte de entender qué son los corpus, poder analizar e investigar los corpus ya existentes, similares al que yo me imagino que sería útil diseñar en un futuro.

Para hacer el marco teórico he utilizado libros, documentos, ponencias, etc. de autores especializados en corpus lingüísticos. Mi trabajo se ha basado en encontrar la información más adecuada y acertada y recopilarla en un documento. De esta forma en el trabajo podemos encontrar la historia de los corpus, las clasificaciones de los diferentes tipos de corpus explicadas, los aspectos útiles para diseñar un corpus, una clasificación de los diferentes corpus existentes más utilizados e información sobre el uso de los corpus, la tecnología y la traducción.

El siguiente paso después de este marco teórico era definir las características que debería tener el corpus hipotético, y de esta forma poder buscar las fuentes de información adecuadas para la elaboración de este corpus especifico. Después de investigar mucho y viendo que no existían corpus relacionados con este tema, decidí que mi corpus hipotético fuese un corpus de gastronomía, cocina y alimentación del español de España y América. Pensé que sería muy interesante poder diseñar un corpus con las características que yo creyese más adecuadas sin dejarme influir por un corpus que ya estuviese construido.

En primer lugar, inicié la selección y evaluación de las fuentes de información necesarias para la creación de un corpus y específicamente del corpus hipotético presentado. Este apartado es el de mayor envergadura, ya que era difícil buscar fuentes de la información sobre la teoría de los corpus lingüísticos. Finalmente me decidí por ocho fuentes que son los que a mí me han aportado más información a lo largo del trabajo. Se trata de libros de teoría sobre corpus lingüísticos, ensayos sobre cómo diseñar un corpus y las diferentes características que pueden tener, directorios sobre corpus existentes y también el corpus PRESEA.

A continuación, elegí las fuentes más representativas que usaría para hacer mi corpus, es decir los textos que recopilaría para procesar en el corpus. Esta fue la tarea más complicada, porque quería encontrar documentos que se pudiesen procesar con

facilidad a formato utf-8, que es el formato que se usa en los corpus para poder formatearlos y hacer el etiquetaje.

Para decidir qué tipos de textos podrían formar parte de la selección bibliográfica final me tuve que fijar en las especificaciones que quería que tuviese mi corpus, de esta forma decidí dividirlo en tres partes: recetarios, libros de técnicas culinarias y revistas gastronómicas, todos estos de los diferentes países que habla española de América y España.

Una vez seleccionados los documentos que quería evaluar, hice unas fichas con las diferentes especificaciones que creía que debían cumplir los textos, esto me ayudó a ver cuáles de los textos que había escogido eran más adecuados para hacer un corpus lingüístico.

El trabajo está dividido en tres partes principales: el marco teórico, donde podemos encontrar una definición exhaustiva de los corpus, la historia, una clasificación de la tipología, los aspectos para diseñarlos, un listado de algunos corpus existentes y la relación entre los corpus y la traducción. La segunda parte, donde se explican las características principales del corpus (hipotético) y la tercera parte que trata de la selección y evaluación de las fuentes de la información. En esta podemos encontrar obras de teoría y construcción de corpus lingüísticos y una recopilación de las obras que compondrían el corpus.

Espero que mi aportación teórica pueda servir de ayuda a todos los profesionales que trabajan en este campo de la lingüística y la traducción.

2. LOS CORPUS LINGÜÍSTICOS

La palabra «corpus» puede tener diferentes significados. Proviene del latín y sus acepciones son, entre muchas otras: cuerpo humano o de un animal, cuerpo de un objeto, masa, etc. Sin embargo, «corpus» posee también otro significado más abstracto: conjunto, total, corporación. Es este segundo significado el que da el nombre al corpus lingüístico tal y como lo conocemos hoy: un conjunto de textos.

Según el *Diccionario de la Lengua Española* (2001) de la Real Academia:

> "Conjunto lo más extenso y ordenado posible de datos o textos científicos, literarios, etc., que pueden servir de base a una investigación".

Según el *Gran Diccionario de Uso del Español Actual* (2006), el corpus es un:

> "Conjunto de textos, procedentes del lenguaje oral o escrito o de ambos, recopilados de fuentes variadas y ordenados y clasificados según determinados criterios, de tal manera que, sobre ese conjunto, es posible realizar estudios e investigaciones lingüísticas o literarias".

2.1. ¿QUÉ ES UN CORPUS LINGÜÍSTICO?

Para entender qué son los corpus es importante entender la lingüística de corpus (de ahora en adelante LC). Esta es muy distinta de la mayoría de los otros temas que pueden estudiar en la lingüística, ya que no es estrictamente sobre el estudio de cualquier aspecto particular de la lengua, sino que más bien es un área que se centra en un conjunto de procedimientos o métodos, para el estudio de la lengua (aunque al menos una escuela importante de lingüistas no está de acuerdo con la caracterización de la LC como metodología, sino que piensan que es una teoría (De Kock, 2011). Los procedimientos en sí mismos aún están en desarrollo, y siguen siendo un conjunto poco definido. Así que podemos ver que la LC actualmente constituye un enfoque metodológico para el estudio de las lenguas y que presenta oportunidades revolucionarias para la descripción, el análisis, y la enseñanza de discursos de todo tipo. También brinda una base empírica para el desarrollo de materiales educativos y metodológicos de diversa índole así como para la construcción de gramáticas, diccionarios y otros, tanto discursos generales como especializados, orales y escritos. Con esto podemos decir que la LC constituye un conjunto o colección de principios metodológicos para estudiar cualquier dominio lingüístico y que se caracteriza por

brindar sustento a la investigación de la lengua en uso, a partir de corpus lingüísticos que se ayudan de la tecnología computacional y los programas informáticos.

Es importante destacar que el desarrollo de la LC también ha dado lugar, o al menos ha facilitado, la exploración de nuevas teorías de la lingüística.

Antes de explorar el impacto de corpus en la lingüística en general, sin embargo, volvamos a la observación de que la LC se centra en un grupo de métodos para el estudio de idiomas. Esta es una observación importante pero debe matizarse. La LC no es un conjunto monolítico tal como lo son la fonética, la sintaxis, la semántica, etc. sino que es un método de investigación que puede ser empleado en todas las ramas o áreas, podríamos decir que es un campo heterogéneo.

A pesar de esto hay otras teorías sobre la LC. Por ejemplo, Taubert (2015) propugna una postura empirista radical en la que los corpus sólo deben ser objeto de análisis en sí mismos, desligados de sus productores, no permitiendo así el uso de categorías provenientes de otras esferas del conocimiento.

> "Para la lingüística de corpus, el significado de un texto o de un segmento textual es independiente de las intenciones de sus hablantes (su autor). La dislocación del hablante/autor de un texto distingue el lenguaje escrito (grabado) del lenguaje oral. En el lenguaje oral, el hablante está usualmente presente y si existe un fallo de comunicación, preguntamos: "¿Qué quieres decir?" y no: "¿Qué significa esto?"" (Taubert, 2005: 5)

Otros científicos como Leech (1992) afirman que la LC no es un campo ni un área de estudio, sino que es un terreno determinado por el foco especial en los corpus con base en metodologías radicalmente diferentes, producto de la incorporación de los avances tecnológicos. Sinclair (1991) y Simpson y Swales (2001) argumentan que la LC es una técnica o una tecnología, cuyo fundamento es el corpus mismo.

Una vez vistas las diferentes teorías sobre la LC y hemos explicado de que se trata, podemos definir lo que hoy en día se entiende por corpus lingüístico, aunque no es una tarea simple. En palabras de Johansson (1991) es, «a body of texts put together in a principled way, often for the purposes of linguistic research»[1].

[1] Traducción propia: Un conjunto de textos unidos conforme a una serie de principios, a menudo con el propósito de investigación lingüística.

Esta definición se puede enriquecer con la propuesta por Crystal (1991: 32):

> "Una colección de datos lingüísticos, ya sea de textos escritos o de transcripciones de habla grabada, los que pueden ser utilizados como punto de partida para descripciones lingüísticas o como un medio de verificaciones de hipótesis acerca de una lengua."

Además de ser un conjunto de textos hay que tener en cuenta la manejabilidad mediante procedimientos electrónicos.

> "Un corpus lingüístico es un conjunto de datos lingüísticos (pertenecientes al uso oral o escrito de la lengua, o ambos) sistematizados según determinados criterios, suficientemente extensos en amplitud y profundidad de manera que sean representativos del total del uso lingüístico o de algunos de sus ámbitos, y dispuestos de tal modo que puedan ser procesados mediante ordenador con el fin de obtener resultados varios y útiles para a descripción y análisis". (Sánchez *et al*, 1995: 8-9)

Además de estas definiciones podemos incluir la de Marcos Marín (1994: 80), en la que se indican las características a las que toda colección de textos debería someterse para ser considerada un "corpus" o, en la terminología del autor, un "archivo digital" (debemos tener en cuenta la antigüedad de la cita y cita y los cambios a todos los niveles que se han producido desde 1994):

1. Debe estar en soporte electrónico. El soporte puede ser un disco láser, discos en un ordenador, disquetes (puede reducirse mucho espacio compactando), cintas, discos removibles, y lo que nos traiga el futuro.
2. Los usuarios han de poder recuperar esa información electrónicamente, bien por conexión directa al ordenador que la almacena, bien por conexión a través de modem telefónico, por red de correo electrónico, o por servicio de distribución de cintas o discos magnéticos o láser.
3. Deben existir reglas que establezcan con claridad el modo de acceder a la información almacenada, emplearla, respetar el copyright y los derechos de autor y evitar la concurrencia ilícita de intereses en el mercado externo, es decir, el mal uso del material almacenado.
4. Los textos han de estar codificados de forma estándar, lo que supone la codificación y uso que reduzcan al máximo la ambigüedad.
5. Los textos codificados han de estar clasificados según una tipología textual también perfectamente delimitada y estandarizada, para facilitar su recuperación.

Esta definición se puede contrastar con la del *Expert Advisory Group on Language Engineering Standards* -EAGLES- (1996a), donde se propone algunas recomendaciones para que un corpus pueda considerarse como tal:

1. El corpus debe ser lo más extenso posible de acuerdo con las tecnologías disponibles en cada época.
2. Debe incluir ejemplos de amplia gama de materiales en función de ser lo más representativo posible.
3. Debe existir una clasificación intermedia en los géneros entre el corpus en total y las muestras individuales
4. Las muestras deben de ser tamaños similares.
5. El corpus, como un todo, debe tener una procedencia clara.

Del mismo modo, Bieber, Rappen, Clark y Walter (2001: 50) consideran que las ventajas que ofrece un estudio basado en un corpus son:

1. Adecuada representación del discurso en su forma de ocurrencia natural en muestras amplias y representativas a partir de textos originales.
2. Procesamiento lingüístico (semi)automático de los textos mediante el uso de computadores. Ello permite análisis más amplios y profundos de los textos mediante conjuntos de rasgos lingüísticos caracterizados.
3. Mayor confiabilidad y certeza en los análisis cuantitativos de los rasgos lingüísticos en grandes muestras de textos.
4. Posibilidad de resultados acumulativos y replicables. Posteriores investigaciones pueden utilizar los mismos corpus u otros pueden ser analizados con las mismas herramientas computacionales.

Como podemos observar hay varias características relevantes comunes a la hora de construir un corpus:

1. Extensión.
2. Formato.
3. Representatividad.
4. Diversificación.
5. Marcado o etiquetado.
6. Procedencia.
7. Tamaño de las muestras.
8. Clasificación y adscripciones de tipos disciplinar, temático, etc.

Con todo esto es interesante apuntar lo que dice Leech (2002): un corpus puede ofrecer una información detallada acerca de una lengua particular, pero es imposible recolectar un corpus que abarque toda una lengua.

Finalmente, podríamos definir razonablemente los corpus como la forma de hacer frente a un conjunto de textos de lectura mecánica que se considera una base adecuada sobre la que estudiar un conjunto específico de preguntas de investigación. El conjunto de textos o corpus tratado suele ser de un tamaño que desafía el análisis manual y visual por sí solo dentro de cualquier periodo de tiempo razonable. Es la gran escala de los datos utilizados que explica el uso del texto legible por una máquina.

Un apunte interesante que deberíamos añadir es que, a menos que utilicemos un ordenador para leer, buscar y manipular los datos, trabajar con grandes conjuntos de datos no sería factible debido al tiempo que necesitaría un analista humano, o un equipo de analistas, para buscar a través del texto. Sin duda, es muy difícil, por no decir casi imposible, buscar en un gran corpus manualmente, de forma que se pueda garantizar que no haya ningún error. Asimismo, podemos decir que este trabajo sería muy difícil de hacer para un humano porque los corpus incluyen herramientas que permiten a los usuarios buscar a través de ellos rápidamente y de forma fiable, cosa que no se podría hacer sin ordenadores.

2.2. HISTORIA DE LOS CORPUS LINGÜÍSTICOS

A principios del siglo XX ya existían proyectos dignos de atención que se basaban en un trabajo manual lento e increíblemente minucioso con datos lingüísticos. Por ejemplo podemos hablar de las líneas de concordancia manuales (el análisis de todas las apariciones de una palabra en su contexto) que eran la herramienta tradicional de trabajo en la elaboración de ediciones críticas de libros; entre las que se puede destacar el *Corpus Theomisticus* (concordancia de la obra de Santo Tomás de Aquino) que se inició a finales de los años 40 por el Padre Busa, pionero de la informática lingüística.

En la segunda mitad del siglo XX es cuando la historia de los corpus empieza a ganar importancia y empiezan tomar la forma que conocemos actualmente. En este momento es cuando se desarrollaron corpus textuales de la lengua inglesa: el primero fue SEU (*Survey of English Usage*, 1959), dirigido por Randolph Quirk, un corpus inglés británico oral transcrito elaborado aún sin ordenadores (en fichas de cartón) pero con la

intención de ser informatizado. El SEU fue la base para otros corpus que fueron una referencia para los estudios lingüísticos angloamericanos.

W. Nelson Francis y Henry Kučera dieron el impulso de los corpus con la creación del *Brown University Standard Corpus of Present-Day American English* (también llamado *Brown Corpus*) en 1964. Este corpus fue una selección cuidadosamente compilada del inglés americano, con un total de aproximadamente un millón de palabras extraídas de una amplia variedad de fuentes. Kučera y Francis compilaron una obra muy rica y variada, que combina elementos de la lingüística, la enseñanza de idiomas, la psicología, la estadística y la sociología. La mayor aportación que hicieron fue que este era procesable mediante ordenadores. Un corpus muy similar al anterior, tanto en el género de las obras usadas como en la extensión, era el *Lancaster-Oslo/Bergen Corpus* (1978), este del Reino Unido. Dos años después se hizo el *London-Lund Corpus of Spoken English*, la versión electrónica del SEU. Después de la aparición de estos tres corpus nos situamos en la era que Javier Pérez Guerra (1998 y 1999) denomina *lingüística de corpus d.c.* (después de los computadores), así que estos tres proyectos mencionados se podrían considerar los «corpus de primera generación».

La segunda generación de corpus se puede situar a mediados de la década de los 80, cuando el avance tecnológico introduce un nuevo elemento: el reconocedor óptico de caracteres (OCR). Este permite digitalizar los textos y agilizar el proceso de captación de muestras. Por consiguiente, el aumento del volumen de datos recopilados es considerable. Además, en esta etapa se consigue un logro muy importe: la expansión de los corpus como herramientas de estudio y análisis. De esta época se puede destacar el *Longman-Lancaster English Language Corpus*, que es el resultado del trabajo en colaboración entre *Longman Publishers* y la Universidad de Lancaster. Consta de unos 30 millones de palabras del inglés escrito, procedentes de textos publicados.

A principios de los años 80 del siglo XX John Sinclair dirige el proyecto COBUILD en la Universidad de Birmingham, con la colaboración de la editorial Collins. La finalidad del proyecto consistía en recopilar un corpus de textos almacenados en ordenador (7 millones de palabras), para la elaboración de un diccionario y el estudio de la lengua. El principal resultado consistió en la producción del diccionario de la lengua inglesa *Collins COBUILD Dictionary*, al cual le sucedieron numerosas publicaciones dedicadas a la enseñanza, el aprendizaje y el estudio de la lengua inglesa. Posteriormente la editorial decidió aumentar el tamaño del corpus hasta conseguir los 200 millones de palabras, convirtiéndose en *The Bank of English Corpus*. Así aparece la

tercera generación de corpus, que son proyectos de gran envergadura que dan lugar a los «Mega-corpus». Estos se caracterizan por las ingentes cantidades de datos compilados que el avance de la tecnología permite almacenar, procesar y analizar de forma cada vez más rápida y exhaustiva. La mayoría de estos proyectos están respaldados por importantes editoriales y son destinados igualmente a fines comerciales.

Por lo que refiere a los corpus en español, la situación es completamente diferente:

> "Existe una gran diferencia entre el número de iniciativas dedicadas a la recopilación de córpora en lengua inglesa y en lengua española, así como las dimensiones de éstos. La diferencia es tan grande que desafía cualquier comparación o paralelismo que se pretenda establecer y muestra con claridad que, en lo que se refiere a la lengua española, es necesario que se promuevan más (y más variadas) iniciativas para la creación y distribución de recursos lingüísticos". (M. Chantal Pérez, 2002).

Existen algunos corpus disponibles como recogía el informe sobre recursos lingüísticos para el español preparado por el Instituto Cervantes (1996), actualizado dos años después, la mayoría de los proyectos están aún en fase de desarrollo. Sin embargo, del estudio realizado por Llisterri y Garrido (1998) se desprende que la participación española en proyectos de ingeniería lingüística e industrias de la lengua es cada vez mayor.

2.3. CLASIFICACIÓN DE LOS CORPUS

Una vez hemos establecido el concepto de corpus, pasamos a comentar algunos de los principales tipos. Este apartado vamos a dividirlo en tres subapartados: los tipos de colecciones de textos, los niveles de corpus y la tipología de corpus.

2.3.1 Colecciones de textos

Como hemos podido ver, en la lingüística los corpus se refieren de forma general al ámbito de recopilación de textos y hay que distinguir al menos tres tipos diferentes de recopilaciones, según el grado de especificación en los criterios de selección:

- Archivo/colección (*Archives*): un repertorio de textos en formato informático en el que los textos no están relacionados ni coordinados de forma alguna.
- Biblioteca de Textos Electrónicos (*Electric text library*): una colección de textos en formato informático que poseen un formato estandarizado y siguen ciertas

convenciones en cuanto al contenido, pero sin rigurosas limitaciones de selección.

- Corpus Informatizado (Computer corpus): una recopilación de textos seleccionados según criterios lingüísticos, codificados de modo estándar y homogéneo, con la finalidad de ser tratados mediante procesos informáticos y destinados a reflejar el comportamiento de una o más lenguas.

Como podemos ver los dos primeros tipos de recopilaciones no implican una selección u ordenación siguiendo los criterios lingüísticos, pero los corpus sí. Estos criterios pueden ser internos, que hacen referencia a patrones lingüísticos presentes en los textos; o externos, que hacen referencia a datos de los autores, la función comunicativa, el nivel social de los participantes, etc. (Sinclair 1996: 5).

2.3.2 Niveles de corpus

Podemos encontrar diferentes niveles de corpus dependiendo de la selección de textos destinada a constituirlo (Torruella y Llisterri, 1999):

- Corpus: es un conjunto homogéneo de muestras de la lengua de cualquier tipo que se toman como modelo de un estado de la lengua predeterminado. Estos deben permitir mejorar el conocimiento de las estructuras lingüísticas de la lengua que representan una vez analizados.
- Subcorpus: suele ser una selección estática de textos, derivada de un corpus más general y complejo.
- Componentes: es una colección de muestras de un corpus o de un subcorpus, las cuales responden a un criterio lingüístico especifico muy concreto, son muy homogéneos comparados con los corpus y los subcorpus.

2.3.3. Tipología de corpus

Podemos encontrar diferentes clasificaciones dependiendo de los autores y los estudios, aunque algunas tienen muchos puntos en común pero la nomenclatura es diferente. Algunas de estas clasificaciones tienen más tipos de corpus en sus clasificaciones. Aquí vamos a presentar dos tipos entre las muchas que existen.

Autores como Torruella y Llisterri (1999) o J. Sinclair (1996) han propuesto estas clasificaciones en función de una serie de criterios, aunque en la práctica no siempre está clara ni se hace explícita la tipología de un corpus.

1. Según el canal de producción.
2. Según el grado de representatividad.
3. Según la especificad de los textos.
4. Según el porcentaje y la distribución de los textos.
5. Según los idiomas incluidos.
6. Según el nivel de análisis.

Esta es la clasificación que da el *Expert Advisory Group on Language Engineering Standards* -EAGLES- (1996b):

1. Corpus de referencia (Reference corpus).
2. Corpus monitor (Monitor corpus).
3. Corpus oral (Spoken corpus).
4. Corpus de fragmentos textuales (Sample corpus).
5. Corpus especiales, especializados y corpus diseñados con fines especiales.
6. Corpus bilingüe (o multilingüe).

Hay que tener en cuenta que con frecuencia estos criterios vienen determinados por la finalidad u objetivo que se persigue con el corpus: el estudio de la obra de un autor o de la producción literaria de una época determinada; la descripción de una lengua en general (por ejemplo, el español contemporáneo) o de una variedad, sublenguaje o aspecto lingüístico concreto (por ejemplo, textos técnicos), la obtención de un determinado producto comercial (por ejemplo, un diccionario).

A continuación hacemos una enumeración general los tipos de corpus basándonos en clasificaciones de diferentes autores[2]:

1. Según el canal de producción o la modalidad. Podemos distinguir entre corpus textuales (también llamados escritos), corpus orales y corpus mixtos.
    - Los corpus textuales están formados únicamente por muestras procedentes de la modalidad escrita de la lengua.
    - Los corpus orales únicamente recogen muestras de lengua hablada, que pueden ser transcripciones de grabaciones, se hace lo posible para mantener en el texto toda la información oral (silencios, risas, dudas, etc.) y en algunos casos la transcripción va acompañada de la

[2] Sobre la tipología de corpus puede verse Rafael y Soler (2001); Sinclair (1996); Torruella y Llisterri (1999); Vargas (2006); y Villayandre (2006).

grabación original (En el punto 2.3.3.1. hablamos sobre la tipología especifica de los corpus orales).

- Los corpus mixtos combinan ambas modalidades, aunque siempre favoreciendo la lengua escrita, ya que su obtención es menos costosa que la de la lengua oral que, además, siempre requiere un proceso posterior de transcripción de las grabaciones. Algunos corpus que pertenecen a este tipo tienen el 90% de sus textos escritos y el 10% orales.

2. Según el grado de representatividad, los corpus pueden ser textuales, léxicos y de referencia.
   - Los corpus textuales son aquellos que incluyen textos enteros, sin fragmentar.
   - Los corpus léxicos, también llamados *simple corpus*, incluyen fragmentos muy pequeños de cada documento.
   - Los corpus de referencia son aquellos formados por fragmentos de textos, habituales en los corpus que quieren proporcionar una información lo más completa posible sobre una lengua y tienen que incluir textos de diferentes géneros, temáticas, etc.

3. Según la especificidad de los textos, los corpus pueden ser generales, genéricos, canónicos, sincrónicos, diacrónicos o históricos, cronológicos o periódicos y especializados.
   - Los corpus generales pretenden reflejar la lengua o variedad lingüística de la forma más equilibrada posible; cuantos más tipos de textos, modalidades (textos orales, textos escritos), géneros y materias, mejor.
   - Los corpus genéricos que incluyen textos correspondientes a un único género, como podría ser novela, teatro o poesía, para caracterizar ese género frente a otros).
   - Los corpus canónicos que recogen la obra completa de un autor.
   - Los corpus sincrónicos que se centran en una etapa lingüística determinada, generalmente actual.
   - Los corpus diacrónicos o históricos que recogen muestras de textos de diferentes etapas temporales, para poder observar la evolución de la lengua.

- Los corpus cronológicos o periódicos que recogen muestras de textos de una época determinada, normalmente histórica, para poder estudiar ese período lingüístico.
- Los corpus especializados se ocupan de un ámbito especializado del uso de la lengua, normalmente para la elaboración de materiales relacionados con las lenguas de especialidad.

4. Según el porcentaje y la distribución de los textos o los límites establecidos, los corpus se clasifican en: corpus cerrados, corpus abiertos o monitor, corpus equilibrados y corpus piramidales.
   - Los corpus cerrados constan de un número finito de palabras, que se establece de forma previa a la recopilación del corpus. Una vez alcanzado ese número o límite, el corpus se da por finalizado. Este tipo de corpus son útiles cuando interesa estudiar fenómenos estáticos o estados de lengua.
   - Los corpus abiertos o corpus monitor, son corpus dinámicos, que se mantienen en constante crecimiento, normalmente mediante la introducción periódica de nuevas cantidades de textos según unas proporciones previamente definidas. Son un material excelente para los estudios diacrónicos, para observar tendencias de uso, cambios de significado, frecuencias de distribución, etc.
   - Los corpus equilibrados son los corpus que recogen la misma proporción de los diferentes tipos de textos.
   - Los corpus piramidales son los corpus que distribuyen las muestras en estratos piramidales, de modo que, si en un primer nivel hay muchas variedades y pocos textos, conforme decrece una proporción aumenta la otra.

5. Según los idiomas incluidos o el número de lenguas, los corpus se clasifican fundamentalmente en monolingües, bilingües o multilingües, corpus comparados y corpus paralelos.
   - Los corpus monolingües están compuestos por textos en una sola lengua. Se recopilan con el objetivo de dar cuenta de una lengua o variedad lingüística en general (o de un subconjunto de la misma).

- Los corpus bilingües o multilingües están formados por textos en dos (bilingües) o más lenguas (multilingües) sin que, en principio, sean traducciones unos de otros.
- Los corpus comparables están formados por una selección de textos en más de una lengua o variedades de la misma, que comparten aspectos parecidos en cuanto a su contenido y a criterios de selección. Se utilizan sobre todo para comparar variedades de la lengua en estudios contrastivos.
- Los corpus paralelos contienen textos en más de una lengua pero, a diferencia de los anteriores, se trata de los mismos textos y sus traducciones o equivalentes en una o más lenguas. Son especialmente útiles en los estudios de traducción y en entornos bilingües o multilingües.

  Si además, para facilitar su explotación, los textos de un corpus paralelo están dispuestos unos al lado de otro por párrafos o frases, de tal forma que sea más fácil extraer las equivalencias de traducción (aquellos elementos que son traducciones mutuas), entonces se habla de corpus alineados.

6. Según el nivel de análisis podemos encontrar corpus simples o brutos, corpus lematizados, corpus analizados morfológicamente, corpus con información sintáctica superficial, corpus analizados sintácticamente y corpus etiquetados semánticamente.
   - Los corpus simples o brutos, llamados *raw corpus* en inglés, no han sido lematizados.
   - Los corpus lematizados son aquellos en los que cada forma se ha relacionado con su lema
   - Los corpus analizados morfológicamente son aquellos que para cada palabra se indica su categoría gramatical y, en ocasiones, más detalles morfológicos.
   - Los corpus con información sintáctica superficial llevan en cada palabra o sintagma una anotación con la función sintáctica, indicando solo los constituyentes principales, como SN, SV, etc.
   - Los corpus analizados sintácticamente, llamados *treebanks*, presentan en cada oración un análisis sintáctico exhaustivo.
   - Los corpus etiquetados semánticamente son aquellos en los que se especifica qué relaciones semánticas mantienen las palabras, como por ejemplo hiponimia e hiperonimia, o marcos semánticos.

Por supuesto, son posibles más tipologías, pero nos hemos limitado a mencionar aquellas que son más habituales y que están más claramente delimitadas.

2.3.3.1. Tipología especifica de los corpus orales

En el apartado anterior hemos definido los diferentes tipos de corpus y las características generales que presenta cada uno. Sin embargo, los corpus orales tienen una tipología específica. Podemos considerar tres tipos según Torruella y Llisterri (1999):

1. Los corpus para la descripción fonética de la lengua:

   Consisten tradicionalmente en materiales grabados en condiciones acústicas óptimas que permitan su posterior análisis experimental en el laboratorio. En estos casos solemos encontrar desde combinaciones de segmentos hasta fragmentos de habla espontánea, pasando por frases aisladas o por textos leídos. Lo que caracteriza este tipo de corpus es un cuidadoso diseño del contenido, basado en el inventario de elementos segmentales y suprasegmentales de la lengua y un tamaño relativamente reducido, debido a que no suelen realizarse grabaciones con un número elevado de hablantes (Albayzín, Base de datos para el reconocimiento del habla en español).

2. Los corpus para el desarrollo de sistemas en el ámbito de las tecnologías del habla:

   Estos son usados para el desarrollo y la validación de los sistemas de síntesis, reconocimiento y diálogo que han surgido en el campo de las tecnologías del habla. Este tipo de corpus tienen varios usos, algunos de ellos son para la conversión de texto a habla, para los sistemas de reconocimiento del habla o para los diálogos utilizados para desarrollar sistemas de interacción entre personas y máquinas (EUROM1 - Multilingual Speech Corpus).

3. Las transcripciones ortográficas de la lengua hablada:

   En estos corpus se trabaja con transcripciones ortográficas procedentes de entrevistas realizadas especialmente para el corpus, aunque el punto de partida sea una grabación, el corpus se trata con los mismos procedimientos que un corpus textual una vez transcrito (Corpus del Proyecto para el estudio sociolingüístico del español de España y de América).

2.4. ASPECTOS PARA DISEÑAR UN CORPUS

Una vez hemos visto los distintos tipos de corpus y las aplicaciones que tiene cada uno, debemos hablar sobre los principales aspectos que deben tenerse en cuenta en el diseño de un corpus. Primero vamos a centrarnos en las cuestiones generales que tienen que tener todos los corpus y después introduciremos las cuestiones específicas para los corpus orales (Torruella y Llisterri 1999).

2.4.1. Aspectos generales

1. Finalidad.

   Este es el primer aspecto que debemos tener en cuenta cuando vamos a diseñar un corpus, decidir cuál será su uso. El punto va a condicionar los siguientes pasos del diseño, ya que nos basaremos en esto para tomar decisiones sobre el mismo.

2. Límites del corpus.

   Aquí se van a establecer los límites geográficos, temporales y/o lingüísticos que va a tener el corpus. Se deben marcar la fecha de inicio y de fin, las lenguas que va a incluir y/o el área geográfica. Esto es muy importante porque puede variar mucho de un corpus a otro dependiendo de los límites que vamos a marcar.

3. Tipo de corpus.

   Para definir el tipo de corpus se tendrán que tener en cuenta unos parámetros: el porcentaje y distribución de los textos que lo componen, la especificidad de los textos, la cantidad de texto que se tome de cada documento para formar las muestras, la codificación y las anotaciones, y la documentación que le acompañe.

4. Proporción de los diferentes grupos temáticos del corpus.

   Aunque es muy difícil desde un principio definir este punto, creemos que es de gran importancia especificar los diversos tipos temáticos y las proporciones de cada uno.

5. Muestra y población.

   Se deber escoger a quien va a representar este corpus, ya que es casi imposible representar a toda una población y sus características.

6. Número y longitud de los textos de muestras

La selección se puede hacer de diferentes formas y, una vez estén establecidas las partes, hay que decidir la longitud de las muestras.

7. Captura de textos y etiquetado.

   Se tiene que tener en cuenta que la captura de textos puede tomar mucho tiempo, sobre todo si se trata de textos impresos en papel, ya que se tienen que escanear y usar un programa de reconocimiento automático de caracteres (OCR).

   Los textos deben ser etiquetados para que se pueda facilitar la posterior explotación del corpus.

8. Procesamiento del corpus.

   El corpus en sí no es suficiente para facilitar datos exhaustivos del comportamiento del lenguaje, y para poder aprovecharlo al máximo, se debe disponer de herramientas adecuadas, como la frecuencia de aparición de palabras, los índices y concordancias, la lematización, las colocaciones, etc.

9. Crecimiento del corpus y "Feedback".

   Es importante que, al utilizar el corpus, analicemos los resultados y detectemos sus puntos débiles para poder reajustarlo constantemente.

10. *Hardware y Software*.

    Al diseñar el corpus es importante tener en cuenta la infraestructura informática que se va a necesitar para poder desarrollarlo y explotarlo.

11. Aspectos legales.

    Esta cuestión se tiene que considerar ya que los derechos de autor (copyright) pueden hacer que no podamos difundir ni explotar ciertos documentos, debemos hacer ciertas consideraciones antes de usar un texto e informarnos sobre la normativa actual en cada país.

12. Presupuesto y etapas.

    Al final se deben establecer las diferentes etapas en que se va a realizar el proyecto, por lo tanto se debe realizar un presupuesto tanto de personal humano como de programas, ordenadores y la adquisición de derechos de autor.

2.4.2. Aspectos específicos de los corpus orales

Además parte de los aspectos generales descritos en el apartado anterior, los corpus orales tienen un diseño y fases que los otros no necesitan.

1. Adquisición de datos

   La adquisición de datos requiere la realización de grabaciones o la obtención de estas a través de la radio y la televisión o de archivos sonoros disponibles. Se tiene que tener en cuenta que dependiendo del tipo de corpus que queramos las grabaciones pueden tener una calidad menor o necesitar unas características específicas

2. Selección de locutores

   Esto puede variar en función del objetivo del corpus y se utilizarán criterios como el sexo, la edad, la procedencia, el nivel sociocultural, etc.

3. Procesamiento del corpus

   Una vez hayamos obtenido los materiales se suele hacer la transcripción ortográfica que a veces se acompaña de una transcripción fonética o fonológica.

Como hemos visto el diseño de un corpus es una tarea larga y complicada, pero atendiendo a todos estos aspectos se puede conseguir un corpus que colme nuestras expectativas.

2.5. CORPUS DE ESPAÑOL DE ESPAÑA Y AMÉRICA

A continuación vamos a hacer un listado de algunos corpus de la lengua española que incluyen variedades de España e América conjuntamente. Los corpus escogidos para esta lista son los más representativos, así como también los más conocidos. Dividimos entre corpus orales y corpus escritos.

- Corpus orales:
  - *Subcorpus Oral del Corpus de Referencia del Español Actual* (CREA), es un banco de datos de la lengua española desarrollado por la Real Academia. El corpus oral constituye aproximadamente un 10 % del CREA. En la versión 3.2, junio de 2008, es posible acceder a casi 9 millones de formas procedentes de trascripciones de la lengua hablada, con más de 1600 documentos. Se trata de textos procedentes de grabaciones de radio o de televisión transcritos y codificados.

- *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico*, son muestras paralelas de doce ciudades de habla española: México, Caracas, Santiago de Chile, Santafé de Bogotá, Buenos Aires, Lima, San Juan de Puerto Rico, La Paz, San José de Costa Rica, Madrid, Sevilla y Las Palmas de Gran Canaria.
- *Proyecto para el Estudio Sociolingüístico del Español de España y de América* (PRESEEA), se trata de un corpus con conversaciones grabadas entre el investigador y los informantes. Hay una selección de los informantes a partir de criterios sociolingüísticos para conseguir una muestra estratificada.

- Corpus escritos:
  - *Corpus de Referencia del Español Actual* (CREA). Cuenta, en su última versión (junio de 2008), con algo más de ciento sesenta millones de formas. Se compone de una amplia variedad de textos escritos, producidos en todos los países de habla hispana desde 1975 hasta 2004. Los textos escritos, seleccionados tanto de libros como de periódicos y revistas, abarcan más de cien materias distintas. Las formas proceden el 50% de España y el otro 50% de América.
  - *Corpus del Español del Siglo XXI* (CORPES XXI), actualmente cuenta con cerca de 100 millones de formas, se elabora a partir de textos orales y escritos. El material proviene tanto de medios impresos —libros y prensa— como de contenidos publicados en Internet o emitidos en canales de información audiovisual. El 30 % de las formas proceden de España y el 70 % restante de América.
  - *Corpus diacrónico del español* (CORDE), es un corpus textual de todas las épocas y lugares en que se habló español, desde los inicios del idioma hasta el año 1974. Cuenta en la actualidad con 250 millones de registros correspondientes a textos escritos de muy diferente género.
  - *Cumbre, Corpus lingüístico del español contemporáneo*, contiene 20 millones de palabras del español oral y escrito de España e América. Está compuesto de textos extraídos de libros diversos, de la prensa y de la radio y la televisión.

2.6. LOS CORPUS, INTERNET Y LA TRADUCCIÓN

Hablaremos ahora sobre la utilización del corpus en la práctica profesional de los traductores, que es uno de los tres usos descritos por Jiménez-Crespo (2009). Este investigador considera que el corpus puede ser útil a los traductores debido a que proporciona datos cuantitativos y cualitativos objetivos sobre los que fundamentar el proceso de toma de decisiones, ya sea durante la fase previa de documentación, durante el proceso de traducción o en su posterior evaluación.

Los traductores suelen acudir a Internet para solucionar problemas de traducción y obedecen a una doble tendencia al hacer esto: *Web for Corpus* (WfC) y *Web as Corpus* (WaC) (Fletcher 2007: 25). La primera propuesta significaría que el traductor utiliza la web para recopilar corpus, es decir, que se produciría una descarga material de contenido digital que posteriormente sería procesado mediante distintas herramientas. En cambio, en la segunda propuesta, *Web as Corpus*, el traductor realizaría una consulta más o menos inmediata directamente en la web.

Según Jiménez-Crespo (2009: 2), el uso de *Web as Corpus* como recurso documental entraña el riesgo de justificar decisiones en traducciones erróneas o fijar el uso de anglicismos.

El concepto de *Web as Corpus* es muy similar en muchos aspectos a la idea del corpus monitor. Toma como punto de partida una gran colección de datos que está en constante crecimiento, y lo utiliza para el estudio de la lengua. Además de utilizar los motores de búsqueda estándar tales como Google para explorar la web como corpus, los investigadores también han desarrollado las interfaces diseñadas específicamente para apoyar este uso de la web. A pesar de esto podemos ver que la *Web as Corpus* tiene algunos problemas específicos. En contraste con la mayoría de los corpus, la web es una mezcla de textos elaborados y editados, y lo que podría denominarse caritativamente material «preparado casualmente». El contenido de la web además no se divide por género, por lo tanto el material que regresa de una búsqueda en Internet tiende a ser una masa indiferenciada, que puede requerir una gran cantidad de procesamiento para ordenar en grupos significativos de textos. Además, no hay duda de que los muchos textos en la web contienen errores de todo tipo y que hay mucho ruido en todos los niveles de la lengua en la web, lo que representa un problema importante que los usuarios de la web como corpus deben abordar. No obstante, la web, sin duda, proporciona un volumen sustancial de datos que puede ser seleccionado y preparado para producir un corpus adecuado para una amplia variedad de propósitos. Por esta

razón la industria de la traducción está prestando atención a la lingüística de corpus y a internet para mejorar la productividad y poder ofrecer memorias de traducción o alcanzar la plena producción automática. Esto dice Jaap van der Meer (2011a), fundador de TAUS, en un artículo, "The future is corpus linguistics".

Cuando hablamos de la traducción y los corpus, no podemos dejar de lado la traducción automática estadística, designada a veces por las expresiones Stat MT o SMT (del inglés *Statistical Machine Translation*), es un paradigma de traducción automática donde se generan traducciones basadas en modelos estadísticos y de teoría de la información cuyos parámetros se obtienen del análisis de corpus de textos bilingües.

En cuanto a la traducción y a los corpus también caben destacar los corpus *ad hoc*. Corpas (2004) lo define como «un corpus virtual que se compila puntualmente para la realización de un determinado encargo de traducción en cualquier dirección (directa, inversa o indirecta). En su diseño no prima tanto la cantidad como la calidad: por regla general, el corpus ad hoc no incluye un número demasiado elevado de textos, pero sí textos muy adecuados, equiparables al texto original en cuanto a la temática, el género y la variedad textual».

Por su parte, Zanettin (2002) considera que un corpus *ad hoc* es una colección de documentos de Internet, creada como respuesta específica a una necesidad de traducción. Además, afirma que a este corpus se puede añadir más material si es necesario y que no está destinado a formar parte de un corpus permanente.

Hacer un corpus *ad hoc* puede ayudarnos a verificar o rechazar decisiones tomadas basándonos en otras herramientas, como diccionarios. A veces éstos nos proponen varias soluciones terminológicas, algunas poco adecuadas, otras con un contexto insuficiente para determinar cuál es el término que necesitamos. En caso de duda, el corpus puede ser una base de datos útil para ayudarnos a decidir cuál es la solución más adecuada. Además, también podemos valernos de este tipo de corpus para buscar el equivalente a un término en otra lengua, basándonos en nuestras intuiciones, o simplemente buscando todas las palabras que contengan su misma raíz, un prefijo o un sufijo, etc.

Por otra parte, se trata de una herramienta muy valiosa para estudiar las convenciones textuales, que son tan importantes para conseguir que el texto meta sea aceptable en la lengua y en la cultura receptoras.

3. CONSTRUCCIÓN (HIPOTÉTICA) DE UN CORPUS

A continuación se detallan las características de un hipotético corpus. Lo denominamos hipotético porque la construcción de este no forma parte del presente TFG. Como decíamos en la introducción, el objetivo fundamental del trabajo es evaluar las fuentes de información que se usarían en el caso de construir el corpus. No obstante, para poder buscar y analizar dichas fuentes, antes hay que imaginar un corpus concreto.

El corpus de gastronomía, cocina y alimentación del español de España y América es el corpus hipotético con el que se trabaja a partir de este punto. Un corpus como el que se describe no existe en este momento y por esta razón creo que aportaría mucha información sobre las diferencia que hay entre los diferentes de países de habla española, además se podría analizar lingüísticamente una temática diferente a la que ya hay hechas.

La finalidad de este corpus es que sirva de ayuda para la traducción de textos relacionados con el área de la gastronomía, cocina y alimentación. Muchas veces los ingredientes y las técnicas que se utilizan en la cocina tienen nombres muy diferentes según las diferentes comunidades culturales.

El corpus no va a tener límites temporales, ya que las recetas tradicionales de cada país a veces se encuentran en libros más antiguos, pero siempre tendremos en cuenta que el lenguaje no sea demasiado diferente al actual. En cuanto a los límites geográficos, se centraría en España y en los países de habla hispana del continente Americano. Recogería textos que representen estos países en un tanto por ciento igual, más o menos el 4% en cada país.

La distribución de los diferentes tipos de textos que compondrían el corpus sería la siguiente:

Recetarios – 80%.

Revistas gastronómicas – 10%.

Libros de teoría de técnicas culinarias – 10%.

La especificidad de los textos serían referidos a cocina tradicional, procurando que no haya interferencias con la alta cocina (*Haute Cuisine*).Intentaría que la cantidad

de texto que se tomase de cada documento fuese lo más amplia posible, para que así el corpus fuese más representativo. El corpus tendría etiquetas declarativas y analíticas, para poder analizar los diferentes elementos estructurales, los aspectos diferenciales entre países, etc.

El corpus debería contar con frecuencia de aparición de palabras, índices y concordancias, lematización, análisis morfológico y detección de unidades recurrentes (*collocations*).

4. SELECCIÓN Y EVALUCIÓN DE FUENTES DE INFORMACIÓN

En este apartado presentamos las fuentes de información seleccionadas y evaluadas para la creación de un corpus de gastronomía, cocina y alimentación del español de España y América. Las fuentes han sido seleccionadas se clasifican en dos grupos: obras sobre teoría y construcción de corpus lingüísticos, y recopilación de obras que compondrían el corpus. Este último apartado estaría dividido en las tres partes que hemos especificado en el apartado anterior: recetarios (80%), libros de teoría técnicas culinarias (10%) y revistas gastronómicas (10%). Para llevar a cabo la tarea de selección se han seguido diversos parámetros y criterios que se enumeran y desarrollan a continuación:

- **Referencia bibliográfica completa.**

- **Descripción general de la obra.**

- **Utilidad:** cada fuente de información seleccionada cumple con las necesidades de documentación que puede tener una persona que quiera construir un corpus.

- **Fecha de publicación:** No solo se ha procurado que una obra no sea demasiado antigua en relación a su fecha de publicación sino que también se ha intentado tener en cuenta que haya sido actualizada en ediciones posteriores o revisada. Si la obra es antigua, se ha revisado la información para comprobar que aún esté actualizada.

- **Autoría:** Para la recopilación de obras que compondrían el corpus se ha tenido en cuenta la autoría de la misma, ya que es importante saber que los libros están escritos por personas expertas en gastronomía, que además conocen o son de los lugares específicos de la cocina de la que hablan.

Una vez establecidos estos parámetros, se ha creado una ficha de cada fuente de información en la que se presenta y analiza cada una de sus cualidades en forma de lista. En total, se han elaborado quince fichas de análisis, que están ordenadas alfabéticamente.

a) Obras sobre teoría y construcción de corpus lingüísticos

A continuación hay una selección de fuentes que son muy útiles para saber qué son los corpus y cómo se usan. Se pueden encontrar libros de teoría sobre corpus, que nos ayudan a entender qué son y los diferentes usos que pueden tener. Además hay documentos que proporcionan los criterios básicos para poder diseñarlos y construirlos de forma sencilla aunque no por ello menos rigurosa. También se ha incluido un directorio de corpus textuales del español, dónde podemos ver las diferentes características que tienen estos. Finalmente se puede encontrar un corpus, este nos sirve de ejemplo para ver cómo es un corpus y también cómo poder usarlo.

Estas obras se han escogido teniendo en cuenta la información que proporcionan. Se ha decidido que estas son las fuentes que dan más información y que esta está presentada de forma que es sencilla de entender pero sin perder la precisión. Además se ha creído que con estas fuentes se puede aprender todo lo necesario sobre los corpus y su construcción.

Referencia	Beeby A., Rodríguez, P., y Sánchez-Gijón, P. (2009). *Corpus use and translating: corpus use for learning to translate and learning corpus use to translate.* Amsterdam: John Benjamins Publishing Co. ISBN 978 90 272 2426 2.
Descripción	Este libro contiene siete artículos de diferentes autores sobre los corpus lingüísticos. Aunque todos tratan sobre un mismo tema, se refieren a diferentes aspectos, como por ejemplo: cómo usar los corpus y los software de recuperación en las clases de traducción, el uso de los corpus para la traducción de la prosodia semántica, los corpus virtuales como recursos de documentación para la traducción de documentos de seguros de viaje, etc.
Utilidad	Recomiendo uno de los artículos "Developing documentation skills to build do-it-yourself corpora in the specialised translation course" que es muy interesante para aprender qué habilidades hay que trabajar para poder construir un corpus lingüístico, concretamente para usarlos en la traducción especializada.
Fecha de publicación	Fue publicado en 2009, por lo tanto la información sigue siendo bastante actual.

Referencia	Listerri, J. (2015). *Corpus textuales en español*. Liceu.uab.es. Universitat Autònoma de Barcelona. <http://liceu.uab.cat/~joaquim/language_resources/lang_res/Corp_text_esp.html> [Consultado: 10 diciembre 2015].
Descripción	Esta página web contiene un listado de diferentes corpus lingüísticos textuales en español. De cada uno de ellos hay una pequeña descripción y una explicación de los contenidos que tienen.
Utilidad	Ayuda a encontrar corpus ya hechos para poder consultarlos, y de esta forma poder ver cómo son y cómo se usan. Así, si se quiere hacer un corpus o estamos interesados en ellos, podemos ver unos cuantos ejemplos que nos pueden facilitar la labor. Además nos puede ayudar a ver qué corpus hay hechos y cuáles creemos que faltan por elaborar.
Fecha de publicación	Es de 2015, pero en la misma página web se avisa que no se actualiza con regularidad.

Referencia	Mcenery, T. y Hardie, A. (2012). *Corpus Linguistics: method, theory and practice.* Cambridge; New York: Cambridge University Press. (Cambridge textbooks in linguistics). ISBN 9780521547369; 9780521838511.
Descripción	Este manual describe los métodos básicos de la lingüística de corpus. Explica cómo se desarrolló esta disciplina y estudia los principales enfoques para el uso de los datos de corpus. Incluye una amplia gama de ejemplos para mostrar cómo estos datos han llevado a la innovación metodológica y teórica en la lingüística en general. Se dan explicaciones claras y detalladas de los aspectos clave de la metodología y la teoría de la lingüística de corpus contemporáneos. El libro tiene una narrativa estructurada y coherente que vincula el desarrollo histórico de este campo a temas de actualidad en la lingüística. Al final de cada capítulo hay actividades prácticas y preguntas para debate. Además hay un extenso glosario que proporciona un fácil acceso a las definiciones de todos los términos técnicos utilizados en el texto.
Utilidad	Aprender en profundidad sobre lingüística de corpus y sus principales usos. Una parte interesante que ayuda mucho a la comprensión del tema son los ejercicios al final de cada capítulo. Esto hace que los estudiantes puedan poner a prueba los conocimientos que han adquirido sobre lo que han leído.
Fecha de publicación	El libro fue publicado en 2012, por lo tanto está actualizado.

Referencia Parodi, G. (2010). *Lingüística* de corpus: *de la teoría a la empiria*. Madrid: Iberoamericana; Frankfurt am Main: Vervuert. (Lingüística Iberoamericana; 40). ISBN 9788484895015.

Descripción En este ensayo se discuten algunas de las nuevas conceptualizaciones que buscan definir la Lingüística de corpus. Se empieza hablando sobre la definición de la lingüística de corpus, seguidamente se ejemplifican algunos procedimientos y herramientas típicas de las investigaciones en lingüística. En el siguiente apartado se presenta un recurso computacional que encarna los principios de la LC, el sitio web *El Grial* que es una herramienta de etiquetaje morfosintáctico, base de almacenamiento de corpus e interfaz de consulta de corpus electrónicos. Los siguientes capítulos están dedicados a una investigación sobre el desarrollo e implementación de un análisis multidimensional a partir de un corpus especializado; y también nos dan una descripción del Corpus PUCV-2006 del Español Académico y profesional. Finalmente nos dan una selección de sitios web con corpus disponibles en línea y habilitados con herramientas computacionales.

Utilidad Esta obra es muy útil para aprender sobre la teoría de la lingüística de corpus, ya que está explicada usando citas de muchos autores clásicos en el ámbito de la lingüística y aporta la traducción al español de todas estas. También ofrece la investigación sobre una herramienta de etiquetaje morfosintáctico que puede ser de mucha ayuda para conocer el funcionamiento de otras herramientas y así poder usarlas para la elaboración de un corpus.

Fecha de publicación El libro fue publicado en 2010 por lo que es una obra relativamente actual.

Referencia	Pérez Hernández, M. Chantal (2002). *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento.* Tesis. Málaga: Universidad de Málaga. <http://elies.rediris.es/elies18/index.html> [Consulta: 15 de noviembre 2015].
Descripción	Hay información sobre como explotar los corpus textuales informatizados. Contiene una parte teórica sobre corpus: la definición, la tipología de corpus, etc. También hay una recopilación de proyectos de corpus en Europa y una explicación sobre las herramientas que se usan para los corpus. También podemos encontrar otros apartados sobre terminología, terminografía, el uso de un gestor de base de datos, el uso de corpus para extraer información, etc.
Utilidad	Nos puede ayudar mucho para entender el funcionamiento de los corpus. La primera parte del trabajo es la más útil ya que encontramos información básica de teoría de corpus y también sobre los tipos de corpus que existen, de esta forma podemos empezar a pensar qué características debería tener el nuestro corpus (hipotético).
Fecha de publicación	Fue publicado en 2002. No es muy actual, ya que puede ser que hayan habido algunos avances tecnológicos, pero la parte teórica sigue correcta.

Referencia	PRESEEA (2014). *Corpus del Proyecto para el estudio sociolingüístico del español de España y de América.* Alcalá de Henares: Universidad de Alcalá. <http://preseea.linguas.net> [Consulta 10 de Mayo de 2016]
Descripción	PRESEEA es un proyecto para la creación de un corpus de lengua española hablada representativo del mundo hispánico en su variedad geográfica y social. PRESEEA agrupa a cerca de 40 equipos de investigación sociolingüística comprometidos con una metodología común para reunir un banco de materiales coherente que posibilite su aplicación con fines educativos y tecnológicos. En la página web, a parte del corpus lingüístico, podemos encontrar muchos documentos sobre la metodología que usan para crear el corpus, recursos para hacer las transcripciones y los cuestionarios que usan para las entrevistas. Además cuenta con explicaciones sobre los equipos que están involucrados con el proyecto.
Utilidad	Ayuda a ver cómo funciona un corpus ya hecho. Este en concreto es muy interesante porque se puede ver como se ha hecho todo el proceso de la recopilación y construcción del corpus. Los documentos sobre la metodología usada dan mucha información sobre qué pasos se tienen que seguir para poder conseguir llegar al objetivo de la creación de un corpus.
Fecha de publicación	Se empezó a construir en 2014 y aún no se ha terminado, por lo tanto la información que aparece en la página web es actual.

Referencia	Torruellla, J. y Llisterri, J. (1999). "Diseño de corpus textuales y orales". En Blecua, J.M., Clavería, G., Sánchez, C. y Torruella, J. (eds.) *Filología e informática: nuevas tecnologías en los estudios filológicos*. Bellaterra: Seminario de Filología e Informática, Departamento de Filología Española, Universitat Autònoma de Barcelona. ISBN 8489790418. P 45-47. http://liceu.uab.cat/~joaquim/publicacions/Torruella_Llisterri_99.pdf [Consulta 10 de febrero de 2016].
Descripción	Presentan las pautas para obtener un corpus suficientemente organizado de la realidad que se quiera reflejar, para que pueda ser explotado con ciertas garantías de éxito. Nos da información general sobre qué son los corpus y nos delimita los distintos tipos de corpus y sus aplicaciones. Finalmente se explican los principales aspectos que se deberían tener en cuenta para diseñar un corpus, así como también se ofrecen algunas indicaciones sobre las principales etapas propias del proceso de elaboración de un corpus.
Utilidad	Este capítulo ayuda a conocer las pautas generales que hay que seguir para diseñar un corpus textual u oral. Además proporciona información sobre los corpus lingüísticos de forma muy concisa pero fácil de entender y también nos habla sobre la tipología de corpus que existen. Puede ser útil en el momento de empezar a trabajar con corpus para hacernos una idea general de las características que debemos tener en cuenta para elaborarlo y las decisiones que se deben tomar antes de empezar a hacerlo.
Fecha de publicación	El artículo fue publicado en el año 1999. Aunque sea una obra antigua, nos aporta mucha información de ayuda básica sobre el diseño de corpus, pero es cierto que debemos ser críticos y percatarnos de si ha surgido algún cambio a partir de lo presentado en el artículo.

Referencia	Villayandre Llamazares, M. *Lingüística Computacional II. Curso monográfico sobre Lingüística de corpus*. Universidad de León. P. 34-44. <http://fhyc.unileon.es/Milka/LCII/Corpus5.pdf> [Consulta: 15 de abril 2016].
Descripción	Es el apartado cinco de este libro y contiene varios criterios que se deben tener en cuenta para diseñar un corpus. Se divide entre criterios internos, como el tema y el estilo, y criterios externos, como la cronología, el origen, el estado, etc.
Utilidad	Estos criterios son muy útiles para seguir una pauta cuando tenemos que decidir las características del corpus. Da información muy detallada sobre cada criterio con ejemplos.
Fecha de publicación	No se sabe la fecha de publicación, no podemos saber si es actual o no.

b) Recopilación de obras que compondrían el corpus

En este apartado hay una selección de algunos de los recetarios, páginas web, blogs, revistas gastronómicas y libros de teoría de técnicas culinarias que podrían servir como ejemplo de documentos que se podrían incluir en este corpus.

Las fuentes que se han usado han sido escogidas de forma que se puedan apreciar todos los tipos de documentos que podrían formar parte del corpus. De estos libros solo se usaría el texto en la etapa de procesamiento, es decir, el texto en el que se ha eliminado título, hoja de copyright, índices, tablas, pies de fotos, bibliografías, etc. Los documentos seleccionados están en un formato que se pueda transformar fácilmente a .txt (utf-8), como .doc, epub o PDF. De esta forma el procesamiento posterior que se haría sería mucho más fácil.

Para hacer el corpus se usarían fuentes de todos los países de habla española: España y los veinte países de América: Argentina, Belice, Bolivia, Chile, Colombia, Costa Rica, Cuba, República Dominicana, Ecuador, El Salvador, Guatemala, Honduras, México, Nicaragua, Panamá, Paraguay, Perú, Puerto Rico, Uruguay y Venezuela (Belice también se cuenta porque, aunque no es la lengua oficial, es la lengua más usada en el país. En Puerto Rico el español es lengua cooficial con el inglés y por lo tanto también se ha añadido a la lista de países). De todos estos países se recopilarían todos los tipos de fuentes que hemos mencionado anteriormente.

De los recetarios de cada país, no solo se usarían libros de cocina tradicional, sino que también se incluirían todo tipo de recetarios: cocina moderna, cocina vegetariana, cocina fusión, etc. De esta forma el corpus podría contar con una gran cantidad de textos para analizar.

También se ha tenido en cuenta que las fuentes que se han escogido contengan, a parte de las recetas, glosarios de ingredientes y utensilios, información sobre medidas, técnicas y consejos de cocina.

En estas fichas no se ha añadido la fecha de publicación porque no es un aspecto que se deba analizar para el tipo de corpus (hipotético) que se ha propuesto.

A continuación vamos a poner algunos ejemplos de fuentes que se podrían usar en el corpus concreto que hemos planteado en los objetivos.

a. Recetarios

Referencia Celina (2016). *Recetas Salvadoreñas*. http://www.recetassalvadorenas.com/ [Consulta: 5 de mayo de 2016].

Descripción Tiene recetas tradicionales de El Salvador, que están divididas en varias categorías: recetas caseras, mariscos, repostería, etc. Cada una tiene una pequeña explicación, indicando cuando se come, de donde es tradicional, información sobre los ingredientes o incluso historias sobre el origen de los productos.

Utilidad Este blog tiene muchas recetas, pero podemos ver que está escrito por una persona que no es profesional de la cocina sino que lo hace como afición, por lo tanto debemos considerar la posibilidad de que haya errores lingüísticos.

Autoría La autora es Celina, una chica que se dedica a recopilar recetas de su país por diversión. Este dato tiene gran importancia porque no ha estado revisado de la misma forma que un recetario convencional y debemos tener cuidado y ser muy críticos en el momento de procesar los textos de esta fuente. A pesar de esto, el blog nos da un buen ejemplo del lenguaje usado en el país.

Referencia García, V. y Martínez Argüelles L. (2016). *Cocina Vegana*. Madrid: Ediciones Oberon. ISBN 9788441537620

Descripción Son recetas vegetarianas y veganas y el libro está dividido en siete capítulos: patés y untables; cremas, purés y sopas; tapas y picoteo; comida rápida de preparar; cereales; guisos, potajes y estofados, y dulces y postres. Al final del libro hay un apartado con datos nutricionales de cada receta.

Utilidad Nos aporta una visión diferente a la cocina tradicional y el uso de ingredientes y procesos de cocina nuevos. Además en todas las recetas hay trucos sobre cocción, conservación y uso de los ingredientes que son más poco convencionales. Las autoras nos aportan una visión más creativa y tecnológica de la cocina.

Autoría Las autoras son investigadoras, graduadas en nutrición humana y dietética. Por lo tanto, vemos que las recetas que aparecen en el libro están muy bien elaboradas en su vertiente nutricional.

Referencia Gironella De'Angeli, A. y De'Angeli, J. (1988). *Gran libro de la cocina mexicana*. México, D.F.: Ediciones Larousse. ISBN 9789706071972.

Descripción Es un recetario de comida tradicional mexicana. Se divide en diferentes apartados, como comida sencilla, comida con especias, etc. Además cuenta con un apartado de glosario con las explicaciones de diferentes ingredientes y también tiene un apartado con los nombres de diferentes utensilios que se usan en la cocina mexicana.

Utilidad Contiene una gran cantidad de recetas y explicaciones. Además podemos añadir la información del glosario que nos puede ayudar para la creación del corpus.

Autoría Alicia Gironella De'Angeli, es una reconocida chef mexicana, conocedora de las raíces gastronómicas de su país y además una ávida investigadora de la cocina mexicana. Este hecho hace que el libro sea muy representativo de la gastronomía de México.

Referencia Ortega, Simone (2008). *1080 recetas de cocina*. Madrid: Alianza Editorial. ISBN: 9788420691855.

Descripción Es un recetario de cocina española. No solo incluye 1080 recetas típicas españolas sino que también tiene un calendario de productos alimenticios, menús semanales e información complementaria sobre las cantidades de comida usuales, el tiempo de cocción, consejos, trucos y términos usuales de la cocina.

Utilidad Aporta una gran cantidad de recetas. Además se pueden encontrar explicaciones sobre las técnicas culinarias que también sirven como información adicional que se puede añadir al corpus.

Autoría La autora de este libro es Simone Ortega, es de origen francés pero vivió toda su vida en España. Por lo tanto podemos ver que la autora tiene muchos conocimientos sobre la comida tradicional española y podemos fiarnos de que estas recetas son fieles a la tradición del país.

Referencia *La cocina Peruana.* (2008). Lima, Perú. Lexus editores S.A. ISBN: 9789972209499.

Descripción Este libro está compuesto por recetas de Perú divididas en diferentes apartados: entradas, sopas, platos principales, postres, licores y refrescos y salsas. Además podemos encontrar una explicación sobre los principales ingredientes que se usan en la gastronomía del país, consejos sobre cómo cocinar y manipular algunos ingredientes y un glosario con diferentes tipos de cocciones y nombres de platos.

Utilidad Hay una gran cantidad de recetas y explicaciones. Además podemos añadir información del glosario que nos puede ayudar para la creación del corpus.

Autoría En la obra no figura ningún autor individual. Está confeccionado por el Departamento de Creación Editorial de Lexus Editores. Esto no nos da tanta seguridad sobre que el contenido de la obra sea tan fiable como los otros libro que están hechos por cocineros e investigadores, aun así como es una editorial de Lima nos asegura que está hecha en el mismo país.

b. Libros de teoría de técnicas culinarias

Referencia Ferrando Valverde, F. (2012). *Técnicas culinarias.* Valencia: España Editorial Brief. ISBN 978-84-15204-34-3.

Descripción Está dedicado a la teoría de la cocina y a las técnicas culinarias. Podemos encontrar una gran variedad de vocabulario con explicaciones sobre menaje, cantidades y equivalencias, preelaboraciones y cocciones, presentación de platos, explicaciones sobre cada tipo de alimento, sus cocciones y formas de conservación. Por último encontramos un apartado con las técnicas básicas de preparación de distintos platos.
Esta obra nos presenta las bases de la cocina de forma sencilla y usando pocos tecnicismos.

Utilidad Enseña las técnicas culinarias más básicas, da mucha información interesante sobre cómo trabajar en una cocina y cuáles son los conocimientos fundamentales que debemos tener para empezar a cocinar.
Se puede usar para elaborar el corpus ya que esta información sobre gastronomía y alimentación no se puede encontrar en ningún recetario con tanto detalle como aquí.

Autoría La autora es Fina Ferrando Valverde. Comenzó sus estudios en la Escuela de Hostelería y Turismo Altaviana de Valencia y los completó en el Centro de Estudios e Investigación de Ciencias Domésticas (CEICID) en Madrid. Trabajó durante más de veinte años como docente y jefa del Departamento de Cocina y Servicios de la escuela Altaviana. Además ha seguido formándose y perfeccionando su técnica en dos de las escuelas culinarias más importantes de Europa: Le Cordon Bleu y la Escuela Ferrandi.
Como podemos ver la carrera profesional de la autora es muy dilatada, lo que asegura que el libro proporciona información contrastada.

c. Revistas gastronómicas

Referencia *Revista Chef Oropeza Día A Día*. (2012) Oropeza Comunicaciones Culinarias. Ciudad de México: México.

Descripción Contiene varios apartados donde se puede encontrar información muy variada sobre comida y alimentación, uso de diferentes ingredientes, descripción de técnicas y recetas.

Utilidad Da una visión diferente de los recetarios y nos aporta información adicional que podemos incluir en nuestro corpus.
Al tratarse de una revista hay mucha información en imágenes que no podrá ser usada en el corpus ya que no se puede procesar con facilidad.

Autoría El autor de la revista no es una persona sino que se trata de Oropeza Comunicaciones Culinarias, por lo tanto es un grupo de personas que se dedican a buscar la información adecuada para publicar.

5. CONCLUSIONES

Si bien estrictamente hablando cualquier colección de textos puede ser llamada un "corpus", en la lingüística moderna este término ha pasado a denominar enormes cantidades de textos escritos u orales compilados de forma sistemática y representativa.

El hecho de que hoy en día los corpus sean informatizados permite al usuario (lingüista, docente, estudiante o usuario general) recurrir a herramientas de búsqueda que le permitan obtener resultados a una velocidad y con un precisión muy grande; y al diseñador de corpus a enriquecerlo mediante la anotación de información extra de diversos tipos: sintáctica, semántica, pragmática, etc.

El *boom* que ha experimentado la LC en los últimos años se ha introducido en el mundo de las publicaciones comerciales hasta tal punto que algunas editoriales cuentan hoy día con colecciones o series específicamente basadas en la explotación de corpus para el diseño y redacción de diccionarios, gramáticas, libros de uso de la lengua y libros de texto, y todo esto con el objetivo de combinar un enfoque prescriptivo de la lengua (cómo debería usarse una lengua) con información empírica descriptiva (cómo se usa una lengua).

Es evidente que en la actualidad, los corpus se han convertido en una herramienta imprescindible para el trabajo de los lingüistas, traductores e intérpretes. Ahí reside su importancia, ya que facilitan mucho el trabajo del análisis lingüístico que realizan estos profesionales. También gracias a la utilización de estas herramientas, actualmente se están llevando a cabo diferentes proyectos de uso cotidiano que no podrían haber existido sin la intervención de los corpus, como los diccionarios, los traductores automáticos o los sistemas de reconocimiento de voz (en automoción, en telefonía, etc.).

A su vez la tecnología ha impulsado y mejorado la elaboración de los propios corpus gracias a los programas informáticos, que permiten ampliar las bases de datos de las que se nutren estas herramientas y hacen que la tarea de analizar los textos sea más asequible.

Otro aspecto relevante relacionado con los corpus es el hecho de que la globalización de las comunicaciones ha llevado a la ampliación del campo de aplicación de los corpus, ya que hay mucha más influencia e interacción entre las diferentes lenguas y sus variaciones. Es decir, antes la información a la que se accedía para

elaborar un corpus era bastante limitada; en cambio, actualmente se pueden analizar muchos aspectos diferentes en muchas lenguas diferentes porque podemos acceder con mucha más facilidad a los textos que nos permitirán elaborar el corpus.

Debido a la gran proliferación de análisis teóricos e interpretaciones sobre los corpus se hace necesario elaborar una recopilación de fuentes que ayude a comprender mejor todos los aspectos relacionados con los corpus.

Este ha sido el principal objetivo del Trabajo de fin de grado, crear una selección de fuentes de la información para la creación de un corpus lingüístico que fuera útil para las personas interesadas en este tema.

Para poder conseguir una selección adecuada de fuentes fue importante crear un marco teórico que me ha permitido aprender y profundizar sobre los corpus. Además este puede ayudar a cualquier persona interesada en este tema a acceder a unas ideas básicas para consultar y para entender esta herramienta.

Gracias a mi trabajo de síntesis, las personas que necesiten conocer qué es un corpus pueden acercarse a las diferentes teorías y adquirir unas ideas básicas sobre el tema que les pueden servir de puente para ampliar más tarde sus conocimientos con la selección de fuentes propuesta.

En este trabajo se ha hecho evidente que la variedad de tipologías y características de los corpus es tan enorme que la persona que va a hacer un corpus necesita tener una formación básica para poder seleccionar el tipo de corpus que quiere diseñar. Ahí es donde entra en funcionamiento la selección bibliográfica que he propuesto. En ella se pueden encontrar documentos de muchos tipos, como teoría de corpus, tipología de corpus existentes, características y pasos que se deben tener en cuenta para diseñar un corpus, etc. Estas fuentes van a dar toda la información necesaria para poder empezar a usar y a trabajar con los corpus lingüísticos.

Además la inclusión de ejemplos de muestras que se podrían usar en el corpus (hipotético) propuesto es de gran ayuda para poder apreciar qué características pueden tener estos textos.

Con este trabajo se ve de forma clara los pasos que hay que seguir para poder construir un corpus. Entender la teoría es el primer paso y después usar las fuentes seleccionadas para poder profundizar en el tema y encontrar cuáles son las características que se tienen que tener en cuenta para poder seguir adelante con el diseño del corpus.

Finalmente, me gustaría hacer una valoración personal. Creo que he cumplido con el objetivo que me propuse al iniciar este trabajo. He aprendido mucho sobre un tema que pensaba que sería difícil de entender y esto me ha aportado una gran satisfacción. Durante el proceso de elaboración del proyecto me he encontrado con algunas dificultades que han hecho que el trabajo fuese un reto importante que me ha estimulado a seguir trabajando en este tema. Fue en este momento complicado, cuando me di cuenta de que este trabajo sería una aportación buena y necesaria para todos aquellos que quieren introducirse y trabajar en el ámbito de los corpus.

6. BIBLIOGRAFÍA

- **Obras generales:**

Alcántara-Plá, P. (2016). *Lista de corpus*. <http://inicios.es/corpus/> [Consulta: 8 diciembre 2015].

Alonso Jiménez, E. (2012). *Linguee y las nuevas formas de traducir*. Universidad Pablo de Olavide. <https://rio.upo.es/xmlui/bitstream/handle/10433/851/2013_Alonso_Linguee_Skopos.pdf?sequence=1> [Consulta: 12 diciembre 2015].

Aroutiounova, C. (2010). "Traducir con corpus" [Blog] *El placer de traducir*. <http://www.elplacerdetraducir.com/?s=Ejemplo+de+c%C3%B3mo+elaborar+un+corpus+ad+hoc> [Consulta: 10 de noviembre de 2015].

Bieber, D., Reppen, R., Clark, V. y Walter, J. (2001). "Representing spoken language in university settings: The desing and constructions of the spoken component of the T2K-SWAL Corpus", en Simpson, R., Swales, J. (eds.): *Corpus Linguistics in North America*. Ann Arbor: Univerity Michigan Press. ISBN 9780472033584. P 48-57.

Corpas Pastor, G. (2004). "La traducción de textos médicos especializados a través de recursos electrónicos y corpus virtuales" en L. González & P. Hernúñez (Eds.), *Las palabras del traductor. Actas del II Congreso Internacional «El español, lengua de traducción»*, 20 y 21 de mayo, Toledo 2004. Bruselas: Comisión Europea/ESLETRA, P. 137-164.

Cruz Piñol, M. (2012). *Lingüística de corpus y enseñanza del español como 2/L*. Madrid: ARCO/LIBROS, S.L. ISBN 9788476358504.

Crystal, D. (1991). *A dictionary of linguistics and phonetics*. London: Blackwell. ISBN 9781405152969.

De Kock, Josse (ed.). (2001). *Lingüística con corpus: catorce aplicaciones sobre el español*. Serie gramática Española, 1. Apuntes Metodológicos, 7. Salamanca: Universidad de Salamanca. ISBN 9788478008988.

EAGLES. (1996a). *Synopsis and comparison of morphosyntactic phenomena encoded inlexicons and corpora. A common proposal and applications to European languages*. Pisa: ILC-CNR.

- (1996b). *Preliminary Recommendations on Corpus Typology*.) EAG-TCWG-CTYP/P.

Fletcher, W. H. (2007). "Concordancing the Web: Promise and problems, tools and techniques" en Hundt, M., Nesselhauf, N. y Biewer, C. (eds.). *Corpus linguistics and the web*. Amsterdam: Rodopi, ISBN 9789042021280. P. 25-46.

Jiménez-Crespo, M. A. (2009). “El uso de corpus textuales en localización” en *Revista Tradumàtica*, 7, P. 1-15.

Kennedy, G. (1998). *An introduction to corpus linguistics*. Londres: Longman. ISBN 9780582231542.

Kuebler, S. y Zinsmeister, H. (2015). *Anotated corpora. Corpus linguistics and linguistically annotated corpora*. Londres: Bloomsbury. ISBN 9781441164476.

Listerri, J. (2015). *Corpus de lengua oral en español*. Liceu.uab.es. Universitat Autònoma de Barcelona. <http://liceu.uab.es/~joaquim/language_resources/spoken_res/Corp_leng_oral_esp.html#> MC-NLCH [Consultado:10 diciembre 2015].

Listerri, J. (2015). *Corpus textuales en español*. Liceu.uab.es. Universitat Autònoma de Barcelona. <http://liceu.uab.cat/~joaquim/language_resources/lang_res/Corp_text_esp.html> [Consultado: 10 diciembre 2015].

Marcos Marín, F. (1994). *Informática y Humanidades.* Madrid: Gredos. ISBN 9788424916657.

Menéndez-Barzanallana Asensio, R. *Informatica aplicada a la traducción*. Murcia: Universidad de Murcia <http://www.um.es/docencia/barzana/TEI/Informatica-Aplicada-a-la-Traduccion-Introduccion.html> [Consulta: 25 noviembre 2015].

McEnery T. y Hardie A. (2012). *Corpus Linguistics: method, theory and practice.* Cambridge: Cambridge University Press. ISBN 9780521547369.

Parodi, G. (2010). *Lingüística* de corpus: *de la teoría a la empiria*. Madrid: Iberoamericana. ISBN 9788484895015.

Pérez Guerra, J. (1998). *Introducción a la lingüística de corpus. Un ejercicio con herramientas informáticas aplicadas al análisis textual.* Santiago de Compostela: Tórculo Edicións. ISBN 9788484080022.

Pérez Hernández, M. Chantal (2002). *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento*. Tesis. Málaga: Universidad de Málaga. <http://elies.rediris.es/elies18/index.html> [Consulta: 15 de noviembre 2015].

PRESEEA (2008). Marcas y etiquetas mínimas obligatorias [Versión 1.2]. *PRESEEA, Proyecto para el estudio sociolingüístico del español de España y de América.* <http://preseea.linguas.net/Portals/0/Metodologia/Marcas_etiquetas_minimas_obligatorias_1_2.pdf> [Consulta: 28 noviembre 2015].

Rea Rizzo, M. (2008). *El inglés de las telecomunicaciones: estudio léxico basado en un corpus específico.* Director: Sánchez Pérez, Aquilino. Tesis. Murcia: Universidad de Murcia. Departamento de Filología Inglesa. <http://hdl.handle.net/10803/10819> [Consulta: 4 de noviembre 2015].

Real Academia Española (2001). *Diccionario de la lengua española* (22.a ed.). Madrid, España. ISBN 9788467041897.

Real Academia Española, *Corpus diacrónico del español* (CORDE) Madrid: Real Academia Española. <http://www.rae.es> [Consulta: 28 noviembre 2015].

Real Academia Española. *Corpus de referencia del español actual* (CREA) Madrid: Real Academia Española. <http://www.rae.es/recursos/banco-de-datos/crea> [Consulta: 28 noviembre 2015].

Real Academia Española. *Corpus del español del siglo XXI* (CORPES) Madrid: Real Academia Española. <http://www.rae.es> [Consulta: 28 noviembre 2015].

Sánchez, A. (2006). *Gran diccionario de uso del español actual.* Madrid: SGEL. ISBN 9788497782241.

Sánchez, A. *et al* (1995). *Corpus lingüístico del español contemporáneo: CUMBRE*. Madrid: SGEL. ISBN 9788471435460.

Sánchez, A., Sarmiento, R., Cantos, P., y Simón, J. (1995). *Cumbre. Corpus lingüístico del español contemporáneo. Fundamentos y aplicaciones.* Madrid: SGEL. ISBN 9788471435460.

Samper, J. A., Hernández Cabrera, C. E., y Troya, M. (Eds.). (1998). *Macrocorpus de la norma lingüística culta de las principales ciudades del mundo hispánico* (MC-NLCH). Edición en CD-ROM. Las Palmas de Gran Canaria: Servicio de Publicaciones de la Universidad de las Palmas de Gran Canaria. ISBN 8489728887.

Taubert, W. (2005). "My versión of corpus linguistics" en *International Journal of Corpus Linguistics* 10. ISSN 1384-6655. P. 1-13.

Torruellla, J. y Llisterri, J. (1999). "Diseño de corpus textuales y orales" en Blecua, J.M., Clavería, G., Sánchez, C. y Torruella, J. (Eds.) *Filología e informática: nuevas tecnologías en los estudios filológicos*. Barcelona: Seminario de Filología e Informática, Departamento de Filología Española. Universitat Autònoma de Barcelona, Editorial Milenio < http://liceu.uab.cat/~joaquim/publicacions/Torruella_Llisterri_99.pdf> [Consulta: 8 diciembre 2015].

Van der Meer, J. (2011a). "The future is Corpus Linguistics" en *TAUS Translation Automatization User Society* <https://www.taus.net/think-tank/articles/event-articles/the-future-for-translators-looks-bright-but-they-will-have-to-reinvent-the-profession-first> [Consulta: 8 noviembre 2015].

Varela Vila, T. "Córpora ad hoc en la práctica traductora especializada: aplicación al ámbito de las enfermedades neuromusculares" en *Construcción eficiente de recursos lingüísticos multilingües (INCITE08PXIB302179PR)* Universidad de Murcia <http://www.um.es/lacell/aelinco/contenido/pdf/55.pdf> [Consulta: 26 noviembre 2015].

Villayandre Llamazares, M. (2008). *Lingüística con corpus (I)*. Universidad de León ISBN 9788476357859.

Villayandre Llamazares, M. *Lingüística Computacional II. Curso monográfico sobre Lingüística de corpus*. Universidad de León. P. 34-44.

Vivaldi Palatresi, J. (2009). "Catálogo de herramientas informáticas relacionadas con la creación, gestión y explotación de corpus textuales" en *Revista Tradumatica*, nº 7 <http://webs2002.uab.es/tradumatica/revista/num7/articles/10/10central.htm> [Consulta: 5 noviembre 2015].

Zanettin, F. (2002). "DIY Corpora: The WWW and the Translator" en B. Maia, J. Haller & M. Urlrych (Eds.), *Training the Language Services Provider for the New Millennium*. Oporto: Facultade de Letras, Universidade do Porto. <http://www.federicozanettin.net/DIYcorpora.htm >. [Consulta: 25 noviembre 2015].

- **Obras recomendadas en el apartado 4:**

Beeby A., Rodríguez, P., and Sánchez-Gijón, P (2009). *Corpus use and translating: corpus use for learning to translate and learning corpus use to translate.* John Benjamins Publishing Co.: Amsterdam. ISBN 978 90 272 2426 2.
Bornia, L. (2001). *La cocina dominicana*. (15ª ed.) Santo Domingo: Editora Taller. ISBN 978-0963554819.

Celina (2016). *Recetas Salvadoreñas*. http://www.recetassalvadorenas.com/ [Consulta: 5 de mayo de 2016].

Ferrando Valverde, F. (2012). *Técnicas culinarias*. Valencia: Brief. ISBN 978-84-15204-34-3.

García, V. y Martínez Argüelles L. (2016). *Cocina Vegana*. Madrid: Ediciones Oberon. ISBN 9788441537620

Gironella De'Angeli, A. & De'Angeli, J. (1988). *Gran libro de la cocina mexicana*. México, D.F.: Ediciones Larousse. ISBN 9789706071972.

Listerri, J. (2015). *Corpus textuales en español*. Liceu.uab.es. Universitat Autònoma de Barcelona.
<http://liceu.uab.cat/~joaquim/language_resources/lang_res/Corp_text_esp.html>
[Consultado: 10 diciembre 2015].

McEnery, T. y Hardie, A. (2012). *Corpus Linguistics: method, theory and practice*. Cambridge; New York: Cambridge University Press. (Cambridge textbooks in linguistics). ISBN 9780521547369; 9780521838511.

Ortega, Simone (2006). *1080 recetas de cocina.* Madrid: Alianza Editorial. ISBN: 9788420691855.

Parodi, G. (2010). *Lingüística* de corpus: *de la teoría a la empiria*. Madrid: Iberoamericana. ISBN 9788484895015.

Pérez Hernández, M. Chantal (2002). *Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento.* Tesis. Málaga: Universidad de Málaga. <http://elies.rediris.es/elies18/index.html> [Consulta: 15 de noviembre 2015].

Torruellla, J. y Llisterri, J. (1999). "Diseño de corpus textuales y orales". En Blecua, J.M., Clavería, G., Sánchez, C. y Torruella, J. (Eds.) *Filología e informática. Nuevas tecnologías en los estudios filológicos*. Bellaterra: Seminario de Filología e Informática, Departamento de Filología Española, Universitat Autònoma de Barcelona. ISBN 8489790418. P 45-47. http://liceu.uab.cat/~joaquim/publicacions/Torruella_Llisterri_99.pdf [Consulta 10 de febrero de 2016].

La cocina Peruana. (2008). Lima, Perú: LEXUS EDITORES S.A. ISBN: 9789972209499.

Revista Chef Oropeza Día A Día nº 31. (2012). Oropeza Comunicaciones Culinarias. Ciudad de México: México.

Villayandre Llamazares, M. *Lingüística Computacional II. Curso monográfico sobre Lingüística de corpus*. Universidad de León. P. 34-44. <http://fhyc.unileon.es/Milka/LCII/Corpus5.pdf> [Consulta: 15 de abril 2016].