CAPÍTULO 4

CORPUS LINGÜÍSTICOS

4.1 GENERALIDADES

Un corpus lingüístico es un conjunto de textos almacenados en formato electrónico y agrupados con el fin de estudiar una lengua o una determinada variedad lingüística. Su objetivo es constituirse en elementos de referencia para el estudio de una frase concreta o un cierto aspecto de una lengua.

4.1.1 TIPOLOGÍA

Básicamente se pueden distinguir dos tipos de corpus: *corpus orales y corpus textuales*. [Cervantes]

A) CORPUS TEXTUALES

Existen dos criterios para la clasificación de este tipo de corpus: según los objetivos para los que han sido creados y según su contenido.

- Corpus de la lengua general con fines generales

Su objetivo principal es constituir una fuente de información textual del español para fines diversos. Dentro de este grupo tenemos:

- *"Corpus de Referencia del Español Actual" (CREA)*. Desarrollado por el Instituto de Lexicografía de la Real Academia de la Lengua Española, contendrá textos literarios, periodísticos, científicos y técnicos, así transcripciones de grabaciones de la lengua oral y de medios de comunicación correspondientes a los últimos veinticinco años (1975-1999).

- "*Corpus Diacrónico del Español*" *(CORDE)*. Desarrollado también por el Instituto de Lexicografía de la Real Academia de la Lengua Española, recogerá textos de la lengua española desde sus orígenes hasta 1975.

- "*Archivo de textos hispánicos de la Universidad de Santiago*" *(ARTHUS)*. Incluirá textos literarios, periodísticos y transcripciones de la lengua oral de diferentes períodos de la historia de la lengua española.

- "*CUMBRE*". Es un conjunto de datos lingüísticos representativos del uso del español contemporáneo recogidos por la editorial SGEL SA y supervisados por A. Sánchez (Universidad de Murcia).

- *"Corpus de español de la República de Argentina"* y *"Corpus Chileno de Referencia"*. Abarcan una gran variedad de tipos de textos del español escrito en Argentina y Chile, respectivamente.

En la Tabla 4.1 aparecen de forma resumida los corpus de la lengua general con fines generales junto con el tamaño y el estado actual en el que se encuentran.

Proyecto	Tamaño actual	Fases intermedias	Tamaño final	Estado
ARTHUS	3 millones			En desarrollo
CORDE			70 millones (1997)	En desarrollo
Corpus Argentina			2 millones	Finalizado
Corpus Chile			2 millones	Finalizado
CREA		100 millones(1997)	200 millones(2000)	En desarrollo
CUMBRE	8 millones		30 millones	Finalizado

Tabla 4.1. Corpus de la lengua general con fines generales[1]

- Corpus de la lengua general con fines específicos

Pretenden dar respuesta a un propósito concreto, como el estudio de determinados aspectos de la gramática o el léxico de la lengua, la extracción de datos estadísticos, el estudio del comportamiento lingüístico de una determinada población de hablantes, análisis comparativos de diversas variedades lingüísticas, o el desarrollo y evaluación de sistemas de procesamiento del lenguaje. En este grupo pueden incluirse:

- "*TANGORA*". Recopilado por IBM España. Su finalidad ha sido la extracción de datos estadísticos para el modelo de lenguaje utilizado en el proyecto TANGORA.

- "*SISCOOR*" (Sistema de consultas coordinadas). Desarrollado en la Universidad Politécnica de Valencia, contiene textos científicos y técnicos.

- "*CorVerifSDGEE*" (Corpus de verificación del sistema de diccionarios y gramáticas electrónicos del español). Es un corpus textual directamente relacionado con el Sistema de Diccionarios y Gramáticas Electrónicos del Español. Se ha desarrollado en la Universidad Autónoma de Barcelona y está en constante ampliación.

[1] Todos los tamaños que aparecen en las Tablas 4.1, 4.2 y 4.3 se refieren a millones de palabras.

- "*CRATER*". Proyecto europeo consistente en textos de la IUT (International Telecommunications Union) en español, francés e inglés. Este corpus está disponible en la Universidad Autónoma de Madrid.

- La Universidad Pompeu Fabra está elaborando una colección especializada de textos técnicos multilingües con anotación estructural.

- Actualmente están en proyecto dos corpus para el estudio de diferentes aspectos gramaticales de la lengua: "*Gramática Española-Corpus de contraste*" (Universidad de Salamanca) y "*Valencias verbales del español*" (Universidad de Murcia).

- Con el mismo fin se ha desarrollado el "*AGLE* " (Archivo Gramatical de la Lengua Española), que contiene más de 100.000 citas recogidas por el gramático español Salvador Fernández Ramírez y editado por el Instituto Cervantes.

- La editorial Diccionarios SM está recopilando un conjunto de textos literarios, periodísticos, científicos y técnicos, así como transcripciones de medios de comunicación para uso en trabajos lexicográficos.

- Bibliograf SA está desarrollando un corpus con fines predominantemente lexicográficos que incluye una amplia variedad de textos.

Proyecto	Tamaño actual	Tamaño final	Estado
CorVerifSDGEE	3 millones	En ampliación	En desarrollo
Diccionarios SM	60.000		En desarrollo
LEXESP		2.5 millones	Finalizado
TANGORA		320 millones	Finalizado
VOX-BIBLIOGRAF	10.352.337		En desarrollo

Tabla 4.2. Corpus de la lengua general con fines específicos

- Corpus de un sublenguaje con fines específicos

  En esta categoría tenemos:

- *"LAN"*. Corpus de textos técnicos elaborado por la empresa Micro Focus SA, con el objetivo de validar herramientas de etiquetado de textos en español en el marco de los proyectos SPAL y DISUF, "*Diccionarios de la lengua española"* y MORFEO-SP, *"Morfología estándar operativa de la lengua española"*.

- *"LEGEBIDUN"* (Software para el cotejo y composición simultánea de textos bilingües). Con este corpus la Universidad de Deusto se propone reunir un gran número de textos administrativos y legales para formar un corpus bilingüe español-euskera.
- Corpus de transcripción de la lengua oral para el estudio de "*La norma lingüística culta de la lengua española hablada en Madrid*", creado por el Consejo Superior de Investigaciones Científicas.

- "*Corpus 92, Lengua escrita por aspirantes a estudios universitarios*", P.A.A.U. (Universidad Pompeu Fabra). Su objetivo es caracterizar el texto académico escrito por estudiantes que han cursado la enseñanza secundaria.

- "*Corpus contrastivo español/francés*" (Universidad de Sevilla), que permitirá el análisis contrastivo y de errores en traducciones del español y del francés.

- "*LEJES* ". Proyecto desarrollado conjuntamente en las universidades de Granada y Bonn y dirigido al estudio del léxico jurídico.

- En el campo del estudio histórico del vocabulario español se están desarrollando dos proyectos en la Universidad Autónoma de Barcelona: la "*Informatización del Diccionario Crítico Etimológico Castellano e Hispánico de J. Coromina y J.A. Pascual*" y el "*Archivo Informatizado de Textos Jurídicos Medievales (AITJUM)*".

- "*Corpus de vocabulario del niño de 6 a 14 años (Diccionario de frecuencias)*" de la Universidad de Granada. Tiene por finalidad determinar el vocabulario usual de los niños de estas edades.

- "*Representación de categorías semánticas en niños ciegos de nacimiento de edad escolar (EGB)*", llevado a cabo por la Universidad Nacional de Educación a Distancia.

Sublenguaje	Proyecto	Tamaño actual	Tamaño final	Estado
Textos académicos escritos por estudiantes	Corpus 92, Lengua escrita por aspirantes a estudios universitarios	2 millones		En desarrollo
Lenguaje infantil	Corpus de vocabulario del niño de 6-14 años		8.937	Finalizado
	Representación de categorías semánticas en niños ciegos		18.000 – 20.000	Finalizado
Textos periodísticos	Corpus textual del español periodístico			En desarrollo
Textos legales	AITJUM			En desarrollo
	LEJES		5.000	En desarrollo
Textos académicos	Corpus textual plurilingüe especializado	5 millones		En desarrollo
	CRATER		5.5 millones	Finalizado
	LEGEBIDUN			En desarrollo
	LAN		26.000	Finalizado
Libros de texto	Frecuencia de elmentos léxicos en manuales de preescolar		50.226	En desarrollo

Tabla 4.3. Corpus de un sublenguaje con fines específicos

- Corpus de un sublenguaje con fines generales

Dentro de este grupo cabe mencionar el "*Corpus Textual del Español Periodístico*" actualmente en desarrollo en la Universidad Autónoma de Barcelona.

B) **CORPUS ORALES**

Los corpus orales se pueden clasificar en dos grandes categorías:

1. Corpus para el estudio de la lengua oral

Tienen como objetivo principal caracterizar desde un punto de vista lingüístico la lengua hablada. Debemos distinguir:

• Corpus generales

Dentro de este grupo se dispone en la actualidad del "*Corpus de Referencia del Español Contemporáneo*" (Universidad Autónoma de Madrid).

• Corpus para fines específicos

Los corpus incluidos en esta categoría se presentan en la Tabla 4.4.

2. Corpus para el desarrollo de aplicaciones en Tecnologías del Habla

El objetivo de este tipo de corpus es desarrollar aplicaciones para el entrenamiento y evaluación de sistemas de reconocimiento.

• Corpus generales

Aquellos desarrollados sin ninguna finalidad específica dentro del ámbito del reconocimiento. Los corpus pertenecientes a esta categoría aparecen en la Tabla 4.5.

Proyecto	Organismos	Estado
EUROM.1	Universidad Politécnica Cataluña	Finalizado
ROARS (Robust Analytical Recognition System)	Universidad Politécnica Valencia	Finalizado
ALBAYZÍN	Universidad Politécnica Cataluña Universidad Autónoma Barcelona Universidad Politécnica Madrid Universidad Politécnica Valencia	En desarrollo

Tabla 4.5. Corpus generales para el desarrollo de aplicaciones en Tecnologías del Habla

Objetivo	Proyecto	Organismos	Tamaño actual	Tamaño final	Estado
Análisis léxico	DIES-RTP	Universidad de Alcalá de Henares	750.000 palabras		En desarrollo
Análisis de la conversación	ACUAH	Universidad de Alcalá de Henares		800 minutos de grabación	En desarrollo
	Corpus de conversación coloquial	Universidad de Valencia	300 horas de grabación		En desarrollo
Análisis del discurso	ADPA	Universidad de La Coruña	75 horas de grabación		En desarrollo
	Análisis del discurso oral	Universidad de Granada	100 horas de grabación		En desarrollo
Estudios de variación geográfica	ALMECOR	Universidad de Granada	30 horas de grabación[2]		En desarrollo
	FAE_Esp Can	Universidad de La Laguna	30 minutos de grabación[3]		En desarrollo
	ILSE	Universidad de Almería	75 horas de grabación		En desarrollo
	VUA	Universidad de Granada Universidad de Málaga	250 horas de grabación, 290 hablantes		En desarrollo
Estudios sobre el desarrollo del lenguaje	Adquisición, desarrollo y representación de categorías semánticas en niños de edad escolar	UNED	18.000-20.000 palabras 846 frases		En desarrollo
	Diferencias individuales en la adquisición del lenguaje	Universidad de Barcelona		10 hablantes en 10 sesiones por año (3-4 años)	Finalizado
	Corpus de habla infantil	CSIC-UNED	6 hablantes en 15-20 sesiones cada año desde 1991[4]		En desarrollo
	Disponibilidad léxica de los adolescentes	Universidad de Salamanca	200.000 palabras		En desarrollo

Tabla 4.4. Corpus orales para fines específicos

* ADPA : Análisis del discurso público actual

* FAE-Esp Can : Fonética acústica y experimental del español de Canarias

* ILSE : Investigaciones histórico-lingüísticas y de las hablas vivas del sudeste español

* VUA : Variedades urbanas andaluzas

[2] Horas de grabación por cada una de las 4 zonas estudiadas.

[3] Minutos de grabación por hablante.

[4] Sesiones de grabación de 45 minutos.

- Corpus para el desarrollo de aplicaciones específicas

La siguiente Tabla muestra los corpus que pertenecen a esta categoría:

Objetivo	Aplicación	Proyecto	Organismo
Reconocimiento del habla	Desarrollo de sistemas	TIC-0448/89	Universidad Politécnica de Valencia
		-	Microsoft
		-	Lernaut & Huspie
	Reconocimiento de dígitos	PA 85/86 – Corpus de dígitos	Universidad Politécnica de Valencia
	Reconocimiento de letras	PA 85/86 – Corpus de letras	Universidad Politécnica de Valencia
	Reconocimiento del habla en condiciones adversas	ROARS – Corpus de frases con efecto Lombard	Universidad Politécnica de Valencia
Reconocimiento del habla continua	Dictado automático	TANGORA	IBM España
		-	Universidad Politécnica de Madrid
Reconocimiento del habla para aplicaciones telefónicas	Desarrollo de sistemas	SPEECHDAT	Universidad Politécnica de Cataluña
		CEUDEX	Telefónica I+D
		-	Microsoft
	Reconocimiento de dígitos, números de teléfono y órdenes	NÚMERO	Universidad Politécnica de Cataluña
		TELÉMACO	Universidad Politécnica de Cataluña
		VESTEL	Telefónica I+D
	Información sobre vuelos	SPATIS	Telefónica I+D

Tabla 4.6. Corpus orales para el desarrollo de aplicaciones en tecnologías del habla

4.2 CORPUS DE REFERENCIA DEL ESPAÑOL ACTUAL (CREA)

La Real Academia Española está elaborando un corpus de referencia general para el español: el CREA. Pretende ser una representación de la lengua española de los últimos 25 años (1975-1999). Debido a la magnitud e importancia de este proyecto hemos creído conveniente dedicar un apartado para profundizar en él.

4.2.1 DISEÑO Y ESTRUCTURA DEL CREA

El diseño del CREA pretende ofrecer una muestra representativa y equilibrada del español estándar que se utiliza actualmente en el mundo con el fin de permitir la mayor flexibilidad posible en la obtención de datos, el CREA está estructurado en diferentes módulos, de manera que las consultas vayan referidas a la totalidad de los textos o bien únicamente a aquellas que poseen unas determinadas características:

- Cronológicos : últimos 25 años (desde 1975 a 1999).
- Geográficos : textos españoles y americanos distribuidos al 50%.
- Medio : textos publicados en libros, revistas, periódicos, textos orales.
- Temas generales : ciencia, política, vida cotidiana, economía, ficción, etc.

El tamaño del CREA al final de su segunda fase (diciembre del año 2000) será de 125 millones de formas. El 90% de esa cantidad procederá de textos escritos y el 10%, de textos orales.

1975-1979	10 %
1980-1984	15 %
1985-1989	20 %
1990-1994	25 %
1995-1999	30 %

Tabla 4.7. Distribución temporal de los textos del CREA

1. CIENCIA Y TECNOLOGÍA	10,125 %
2. CIENCIAS SOCIALES, CREENCIAS, PENSAMIENTO	13,5 %
3. POLÍTICA Y ECONOMÍA	13,5 %
4. ARTES	10,125 %
5. OCIO Y VIDA COTIDIANA	10,125 %
6. SALUD	10,125 %
7. FICCIÓN	22,5 %

Tabla 4.8. Distribución de los textos del CREA por grandes áreas temáticas (porcentajes sobre el total)

Se han incorporado al CREA los siguientes textos procedentes de otros corpus del español:

- Entrevis
- Corpus oral de referencia del español
- Proyecto Dies-RTV (España, Puerto Rico, Uruguay)
- Macrocorpus de ALFAL
- Corpus conversacional de Alcalá
- Archivo de textos hispánicos de la Universidad de Santiago

1. CODIFICACIÓN DEL CREA

- Introducción

El CREA se compone de un conjunto de textos en formato electrónico enriquecido con una serie de informaciones adicionales.

Hay tres aspectos diferenciados en el concepto de *codificación*, que pueden entenderse también como tres etapas en la elaboración de un corpus codificado:

1. La codificación consiste en la elección de un lenguaje de marcas o etiquetas que permiten representar la información añadida y definir un esquema jerarquizado de marcas, con una sintaxis precisa.

2. Proceso de introducción de marcas en el texto.
3. Conjunto de marcas que pertenecen a un nivel distinto al textual y que aparecen asociadas a los textos.

- Esquemas de codificación del CREA

El lenguaje de codificación escogido para el proyecto CREA es el que se ha impuesto como estándar en los últimos años: SGML (Standard Generalized Markup Language). Se han seguido, en la medida de lo posible, las recomendaciones de la TEI (Text Encoding Initiative) y del CES (Corpus Encoding Standard).

En todo esquema de codificación hay que considerar los aspectos formales y de contenido. En cuanto a los primeros, hay que tener en cuenta que el intercambio de textos puede sufrir modificaciones si no se utiliza el código ASCII plano. La ortografía española utiliza muchos símbolos que no pertenecen a dicho código, como los acentos o las eñes. Lo mismo puede decirse para el resalte tipográfico. Para conseguir que un texto no pierda los signos ortográficos o tipográficos que no pertenecen al código ASCII se le somete a una conversión de los códigos no aceptados a ciertas marcas de carácter estándar que pueden llegar a verse en pantalla como los signos originales.

Otro tipo de marcas son aquellas que añaden informaciones relativas al contenido. Son las más interesantes para la explotación del texto. El espectro que abarcan las marcas de contenido es muy amplio: marcas de carácter bibliográfico, de la anotación lingüística, estructurales, intratextuales no estructurales.

Se han definido tres esquemas de codificación distintos para el CREA:

- Esquema de nivel 1 : para textos escritos. Está formado por marcas mínimas fácilmente automatizables.
- Esquema de nivel 2 : para textos escritos. Añade información intratextual muy diversa.
- Esquema de nivel 3 : para textos orales. Tiene un grado de complejidad comparable al de la codificación de nivel 2 de textos escritos.

La distinción de dos niveles en la codificación de textos escritos se debe a que el trabajo en dos fases resulta más rápido que la marcación completa de los textos escritos en una sola fase.

- Proceso de codificación de los textos

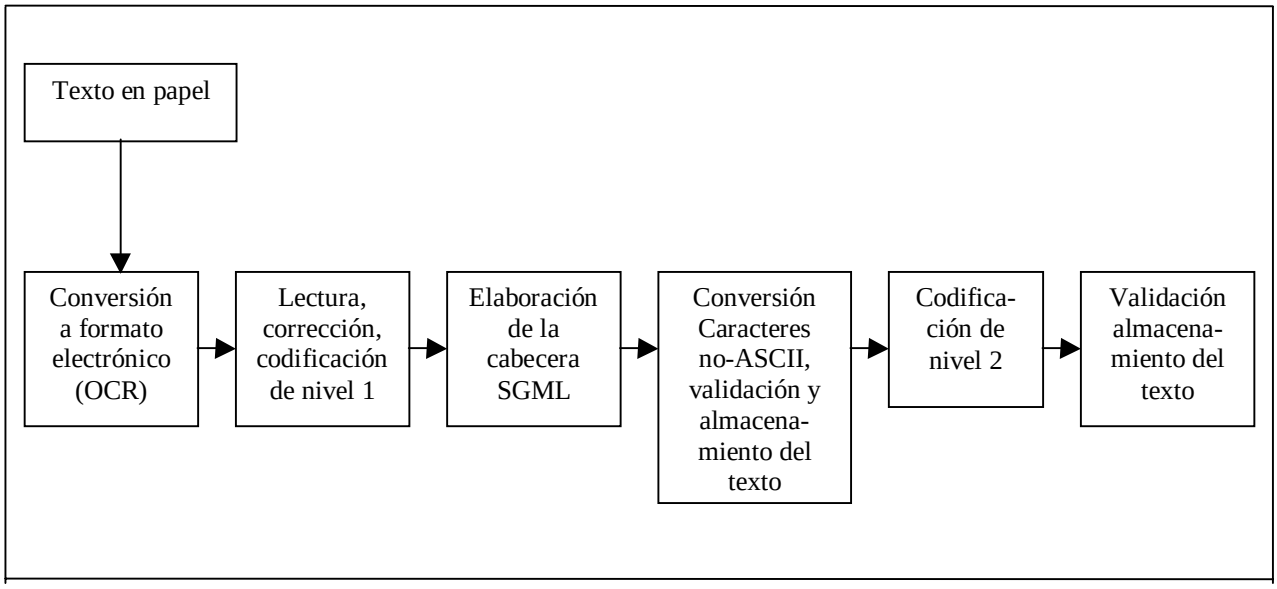

Figura 4.1. Proceso de codificación de textos escritos

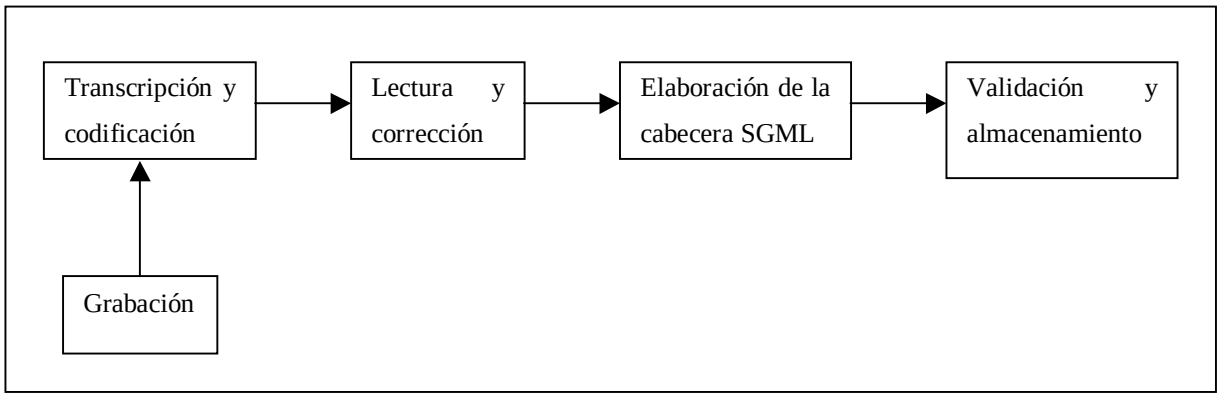

Figura 4.2. Proceso de transcripción y codificación de textos orales

4.3 CORPUS DE ENTRENAMIENTO Y EVALUACIÓN

4.3.1 CORPUS DE ENTRENAMIENTO

Nuestro Corpus de Entrenamiento está constituido por textos del periódico español EL MUNDO publicados en los años 1994 y 1995, disponibles en CD-ROM. Cada CD

contiene los artículos publicados durante seis meses. Así, por ejemplo, al año 1994 le corresponden dos CD's, el primero contendrá los artículos publicados desde el 1 de enero al 30 de junio y el segundo irá desde el 1 de julio al 31 de diciembre. Estos CD's son de dominio público, disponibles como producto comercial, si bien su precio es asequible.

- Tamaño

Nuestro Corpus de Entrenamiento contiene más de 54 millones de palabras, entendiendo por palabra cualquier cadena de caracteres entre dos separadores (espacios en blanco, comas, puntos y comas, dos puntos, signos de interrogación o admiración, retornos o tabuladores) consecutivos. Los signos de puntuación (cadenas con un solo carácter) también se consideran palabras. Otro dato interesante es el número de frases, entendiendo como frase la cadena de palabras, la primera de ellas comenzando con letra mayúscula, comprendida entre dos puntos. Nuestro corpus contiene 2.120.328 frases en un total de 96.529 textos. A continuación presentamos una tabla donde se recogen todos estos datos de una manera más detallada:

Año	Palabras	Frases	Textos
1994	2.236.627	85.653	4086
1995	2.264.445	91.041	3958
Media	2.250.536	88.347	4022

Tabla 4.9. Datos medios mensuales

Palabras	54.012.863
Frases	2.120.328
Textos	96.529

Tabla 4.10. Datos totales (1994-95)

Estos datos pueden contrastarse con los que ofrece el Laboratorio de Lingüística Informática de la Universidad Autónoma de Barcelona [Subirats 99], cuyo corpus tiene

91.505.114 palabras y 6.826.000 frases. Dicho corpus está integrado por textos periodísticos de la prensa española (85 %) y latinoamericana (1 %), publicados durante la década de los 90, y por textos de ensayos de filosofía, antropología, psicología, etc. (14%).

- Formato

Los textos se distribuyen por meses, en orden cronológico. Todos ellos están estructurados de la siguiente forma:

REGISTRO
FECHA
SECCIÓN
(CABECERA)
EDICIÓN
COLUMNA
PÁGINA
TIPO
LEAD
TEXTO
(PIE)
(FIRMA)

Entre paréntesis figuran los campos que aparecen de forma esporádica.

- Tipología

Los textos periodísticos que componen el Corpus de Entrenamiento tratan temas muy variados, si bien pueden clasificarse atendiendo a la **Sección** a la que pertenecen:

1. PORTADA
2. OPINIÓN
3. NACIONAL
4. INTERNACIONAL
5. SOCIEDAD
6. COMUNICACIÓN
7. MADRID
8. CRÓNICA
9. EXTRA
10. ECONOMÍA
11. BOLSA
12. CAMPUS
13. MOTOR
14. DEPORTES
15. CULTURA
16. TELEVISIÓN
17. CINE
18. ESCENA
19. TOROS
20. SALUD
21. DOCUMENTOS
22. ESFERA
23. SERVICIOS
24. UVE
25. METROPOLI
26. ÚLTIMA
27. 7DÍAS

Dentro de las secciones nos encontramos artículos que son Críticas, Entrevistas o Encuestas. Algunas de las secciones son diarias (**Opinión**, **Nacional**, **Economía**, **Deportes**) y otras varían en función de la época del año, como por ejemplo **UVE** (*Un Verano Europeo*) que es un suplemento veraniego, o del día de la semana, como es el caso de **7Días**, que es dominical.

La distribución de los textos por secciones se presenta a continuación. Se han considerado las secciones fijas y las más relevantes, las demás se han englobado bajo el término **Restantes**.

PORTADA	3.11%	BOLSA	0.67%
OPINIÓN	11.02%	DEPORTES	10.60%
NACIONAL	13.32%	CULTURA	6.06%
INTERNACIONAL	10.05%	TELEVISIÓN	3.30%
SOCIEDAD	7.47%	CINE	1.02%
COMUNICACIÓN	1.30%	ÚLTIMA	1.47%
MADRID	11.42%	7DIAS	1.21%
ECONOMÍA	8.89%	RESTANTES	9.09%

Tabla 4.11 Distribución de los textos por secciones
(porcentajes sobre el total)

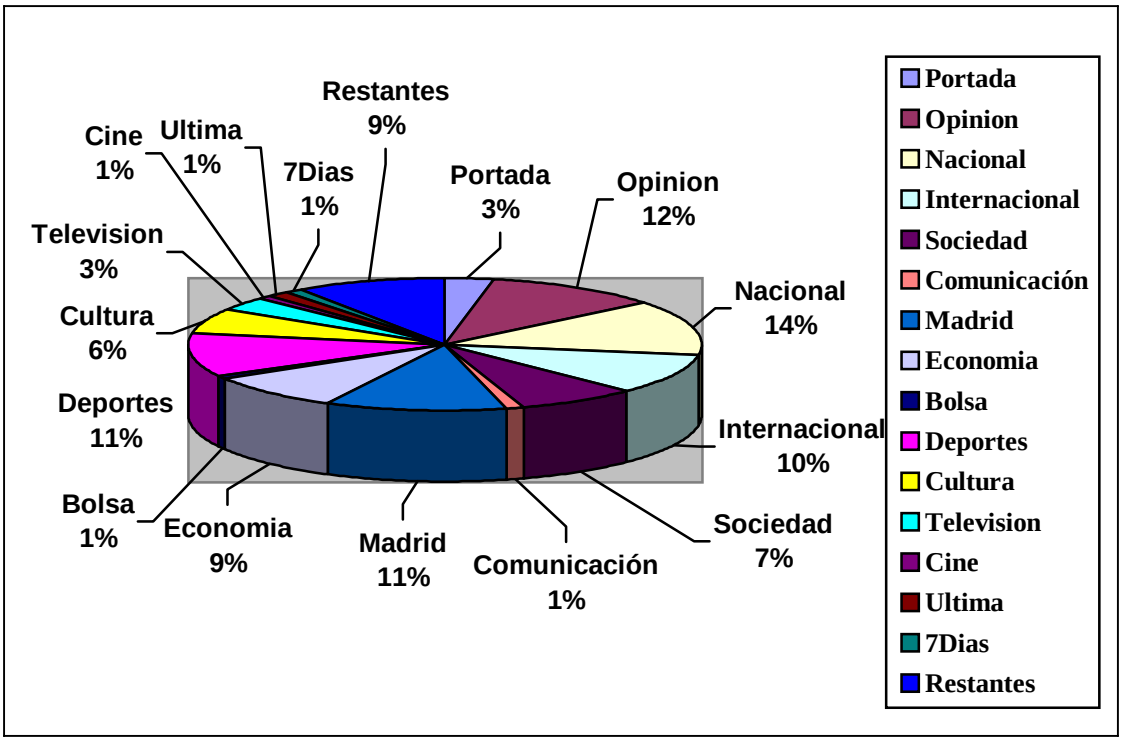

Figura 4.3 Distribución de los textos de El Mundo por secciones
(Porcentajes sobre el total)

4.3.2 CORPUS DE EVALUACIÓN

El Corpus de Evaluación está formado por textos periodísticos de *EL MUNDO* correspondientes al año 1996. Como puede observarse es un corpus más pequeño que el utilizado en la fase de entrenamiento.

Por otra parte, para evaluar el categorizador hemos utilizado un corpus que contiene textos depurados, que no necesitan pre-procesamiento: los *"Textos 860"*. Estos textos se pueden dividir en tres tipos:

1. Textos jurídicos y de legislación: **EEC**.
2. Textos de la Comunidad Europea: **CEE**.
3. Textos periodísticos: **TEXTSPA**.

Existe un cuarto tipo de textos dedicados a Discapacidades, pero no los vamos a tratar. Este cuarto grupo se incluyó en un proyecto posterior.

Los tres tipos de textos considerados presentan el mismo formato: frases consecutivas escritas en forma de columna, de manera que cada palabra va acompañada de la categoría gramatical[5] que le corresponde según el contexto en el que se encuentra.

	Nº textos total	Porcentaje sobre el total
EEC	39	30 %
CEE	82	63.08 %
TEXTSPA	9	6.92 %

Tabla 4.12 Clasificación de los Textos 860

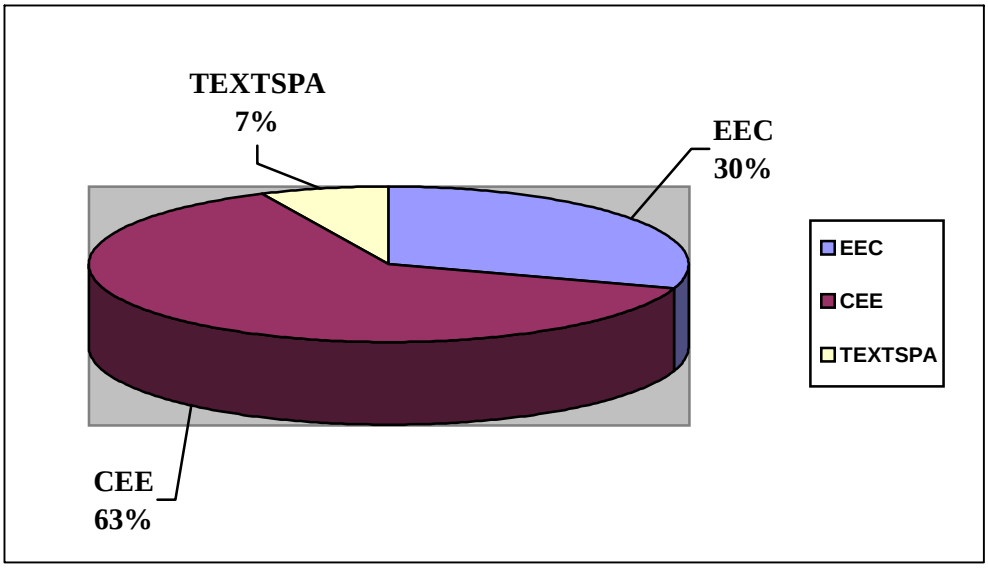

Figura 4.4 Distribución de los Textos 860 por temas
(Porcentajes sobre el total)

[5] La lista de categorías que utiliza el programa se da en el Anexo A.