CREA. Corpus de Referencia del Español Actual

Mercedes Sánchez

Departamento de Banco de Datos de
la Real Academia Española

Introducción

El *CREA*, Corpus de Referencia del Español actual, es una muestra representativa y equilibrada de todas las variedades que presenta el español en nuestros días.

Está compuesto por una amplia variedad de textos escritos y orales completos producidos en todos los países de habla hispana desde 1975 hasta la actualidad.

Actualmente cuenta con unos 140 millones de formas, y está previsto que, a finales de 2004, cuente con 170 millones.

Para la realización de *CREA*, desde sus inicios hasta la fase actual, se han seguido las siguientes líneas de trabajo:

1. Determinar las dimensiones del corpus.
2. Componer el diseño del Corpus: qué textos se incorporarán y en qué porcentaje cronológico, espacial, temático, etc.
3. Adquisición de materiales: cómo hacernos con los textos.
4. Introducción y codificación de textos: cómo pasan a formar parte del Corpus y qué tipo de marcas se introducirán para permitir después recuperar la información, cómo se preparan los textos para las fases posteriores, anotación y explotación.

Dimensiones

CREA comenzó a realizarse en 1996; la primera fase, finalizada en diciembre de 2000, reúne textos editados entre 1975-1999. Es una fase ya terminada pero en continua revisión. Cuenta con 130 millones de formas, de modo que se cumplieron los objetivos marcados al inicio del proyecto.

La fase actual, 2000-2004, tiene previsto incorporar 37 Mill. y medio más, de manera que, a finales de 2004, contaremos con un corpus de referencia del español formado por unos 170 millones de formas.

Diseño

Los materiales que conforman el *CREA* se seleccionan de acuerdo con una serie de parámetros:

- *Medio:* El 90 % corresponde a la lengua escrita y el 10 %, a la lengua oral. De ese 90%, un 49% son libros, otro 49% es prensa y el 2% restante recoge los textos que denominamos miscelánea: folletos, prospectos.
- *Cronológico:* en períodos de cinco años: 1975-79; 1980-84; 1985-89; 1990-94; 1995-99. *CREA* siempre abarcará los últimos 25 años de la lengua. Los textos anteriores pasarán a *CORDE*, que da cuenta de la lengua española desde sus inicios hasta su frontera con *CREA*.
- *Origen: CREA* da cuenta, ya lo hemos dicho, de la lengua española en todo su ámbito. El 50% pertenece a España y el otro 50% son textos procedentes de Hispanoamérica. A su vez el 50% de Hispanoamérica se distribuye en zonas lingüísticas según el número de hablantes. Las zonas son: andina, caribeña, central, chilena, mexicana y rioplatense.
- *Tipo de texto: CREA* está formado por tres grandes bloques de materiales: libros y prensa, miscelánea, que conforman la parte escrita, y las transcripciones de la lengua hablada, el corpus oral.
- Para *Libros y Prensa* se establecieron dos grandes bloques: ficción y no ficción. Se subdivide cada uno a su vez, en siete grandes hipercampos formados por distintas áreas temáticas. En la fase de diseño también se señala el porcentaje, el número de formas correspondiente a cada uno de ellos. Los hipercampos son:

    1. Ciencias y tecnología
    2. Ciencias sociales, creencias y pensamiento
    3. Política, economía, comercio y finanzas.
    4. Artes
    5. Ocio y vida cotidiana
    6. Salud
    7. Ficción: novela, relatos, teatro

- Por su parte *miscelánea* se agrupa en dos bloques: impresa y no impresa. Cada bloque lo forman diferentes tipos textuales: prospectos, propaganda, correos electrónicos, páginas web...
- Para el Corpus Oral se tiene en cuenta el género radiofónico o televisivo del documento sonoro o si se trata de un discurso, una clase, una conversación espontánea, etc. Lo segundo se agrupa en el bloque "otras grabaciones" (2), pero nuestra labor está centrada, sobre todo, en textos procedentes de grabaciones de radio (1). A su vez el género 1, radiofónico, se clasifica en los siguientes subgéneros:

    1. Noticias
    2. Reportajes
    3. Entrevistas
    4. Debates
    5. Tertulias
    6. Documentales
    7. Retransmisiones deportivas
    8. Magacines
    9. Revistas deportivas
    10. Variedades
    11. Sorteos y concursos

Adquisición de materiales

En *CREA* se incorporan materiales diariamente. Los textos se seleccionan intentando mantener siempre el equilibrio establecido en la fase de diseño en todos sus parámetros. Para ello trabajamos con una base de datos que permite conocer de qué hipercampo o país o zona necesitamos incorporar más palabras o cuál debe ser nuestra próxima grabación sonora.

Los libros se desencuadernan, se escanean y se interpretan por un programa de reconocimiento óptico de caracteres.

La prensa se obtiene actualmente a través de Internet y con ayuda de algunos programas off-line. En la fase 1975-1999 se utilizó también un programa de reconocimiento de voz para introducir los ejemplares más antiguos.

Los textos de miscelánea se recogen por los miembros del equipo, también desde Internet, y avisando en ocasiones a amigos y conocidos que viajan a hispanoamérica.

Por lo que respecta a las transcripciones de la lengua hablada, en la fase 1975-1999 se incorporaron textos procedentes de otros corpus orales que se adecuaron a *CREA*, y también grabaciones de radio y televisión obtenidas a través de convenios con RTVE o la Cadena Ser. Las grabaciones se recibían en cinta de audio y se transcribían con ayuda de una grabadora/reproductora.

En la fase actual, y aunque continúan vigentes todos estos convenios, hemos optado por la grabación de la radio desde Internet, que permite mayor versatilidad y rapidez, además de la digitalización directa del sonido y su posterior alineamiento con la transcripción. Para ambas tareas se utilizan programas específicos de libre distribución en Internet.

Introducción y codificación de textos escritos

Una vez escaneados los libros, cada codificador lee atentamente el texto y corrige los posibles errores cometidos por el OCR.

Simultáneamente introduce marcas de codificación SGML, el estándar recomendado, que permite la posterior recuperación de la información. Utilizamos el procesador de textos *Word*, que hemos ido adaptando a nuestras propias necesidades: hemos creado *macros* que permiten una rápida inserción de las marcas, así como la adecuación de los ficheros de prensa, marcados *HTML*, al sistema de codificación de *CREA*. Cada codificador revisa y corrige la actuación de la macro y asigna, además, un hipercampo y área temática a los artículos periodísticos.

En la parte escrita de *CREA* se introducen dos tipos de marcas:

- Estructurales: párrafo, oración, número de página:

```
<p><s>— Te aseguro que no habría el menor riesgo.
<pb n=256> <p><s>— Valentín, te lo ruego. <s>No quiero saber una
palabra de nada. <s>Ni donde están, ni quienes son, nada.
```

- De resalte tipográfico: negritas, cursivas, texto entrecomillado:

```
<p><s>— <hi rend="cdob">"Querido, vuelvo otra vez a conversar
contigo... La noche, trae un silencio que me invita a hablarte...
los sueños tristes de este amor extraño... te juro, que el alma
```

mía será toda tuya, mis pensamientos y mi vida tuyos, como es tan tuyo... este dolor..."</hi> o <hi rend="curs">*este penar*</hi>.

Se añade después la cabecera del texto, que contiene los datos bibliográficos: título, autor, año de edición, país, hipercampo y área temática, revisiones, etc., con datos diferentes para prensa, libros o efímera.

```
<TEI.2 ID="CRL701316">
<TEIHEADER ID="thair001" TYPE="text" STATUS="new"
DATE.CREATED="14/05/2002">
<FILEDESC ID="fdair001">
<TITLESTMT>
<TITLE>Los aires dif&iacute;ciles: transcripci&oacute;n
electr&oacute;nica</TITLE>
</TITLESTMT>
<EDITIONSTMT N="1.0">
<EDITION>Primera versi&oacute;n.
<DATE>14/05/2002</DATE>
</EDITION>
[…]
<EXTENT>263066 palabras, 1.857.390 b.</EXTENT>
<SOURCEDESC ID="sdair001">
<BIBLSTRUCT>
<MONOGR>
<AUTHOR>Grandes, Almudena</AUTHOR>
<TITLE LEVEL="m">Los aires dif&iacute;ciles</TITLE>
<EDITION>Primera edici&oacute;n</EDITION>
<IMPRINT>
<PUBPLACE>Barcelona</PUBPLACE>
<PUBLISHER>Tusquets</PUBLISHER>
<DATE N="1.0">2002</DATE>
</IMPRINT>
<BIBLSCOPE TYPE="pages">13-593</BIBLSCOPE>
</MONOGR>
<SERIES>
<TITLE LEVEL="s">Andanzas</TITLE>
<BIBLSCOPE TYPE="volume">466</BIBLSCOPE>
</SERIES>
<IDNO TYPE="isbn">84-8310-195-5</IDNO>
<IDNO TYPE="dl">B.879-2002</IDNO>
<IDNO TYPE="ccorde">co1221</IDNO>
<IDNO TYPE="origin">E</IDNO>
<IDNO TYPE="country">Espa&ntilde;a</IDNO>
<IDNO TYPE="sex">M</IDNO>
</BIBLSTRUCT>
</SOURCEDESC>
</FILEDESC>
[…]
<PROFILEDESC ID="pdair001">
<CREATION>Primera edici&oacute;n
<DATE>2002</DATE>
</CREATION>
<TEXTCLASS>
<CATREF TARGET="cr701" SCHEME="crea">
```

```
<CATREF TARGET="L" SCHEME="medio">
<KEYWORDS>aires</KEYWORDS>
</TEXTCLASS>
</PROFILEDESC>
</TEIHEADER>
<TEXT N="air001" DECLS="thair001 fdair001 sdair001 edair001
pdair001">
<HEAD>Los aires dif&iacute;ciles</HEAD>
<HEAD>Almudena Grandes</HEAD>
```

Finalmente se valida como texto *SGML* y se genera la copia en formato texto, que se integrará en la totalidad del corpus.

CREA-Oral

En la fase 1975-1999 se incorporaron casi nueve millones de registros procedentes de transcripciones de la lengua hablada, que ya están disponibles a través de la página web.

La incorporación de este tipo de textos es mucho más lenta y laboriosa: en la parte escrita, un codificador puede incorporar unas 90 mil palabras a la semana, mientras que en oral solo puede llegarse a unas 10 mil palabras semanales por persona. Ahora bien, el resultado es la obtención por escrito de una muestra real de la lengua hablada.

La transcripción debe estar acompañada de una codificación específica que, en ocasiones, ayude incluso a comprender la lectura de un texto hablado.

Pero es que además debe estar acompañada de su correspondiente correlato sonoro, es decir, es necesario que el corpus oral esté alineado, que se pueda acceder tanto a la transcripción como a la audición del texto. Para ello es necesaria también la introducción de marcas de sincronización. Es así como estamos trabajando en la fase actual, 2000-2004, del corpus oral. Lo hacemos a través de Transcriber, un programa especialmente diseñado para la codificación, alineamiento y transcripción simultáneos de textos orales. Pero recordemos, brevemente, las características de la fase 1975-1999 del Corpus Oral que actualmente se puede consutar en la aplicación del Banco de Datos en la página web de la Real Academia Española. (http://www.rae.es)

1975-1999: Fenómenos codificados

- Turnos de palabra de cada hablante y tipo de transición entre turnos
- Superposiciones entre hablantes
- Pausas
- Fenómenos no vocales y no comunicativos
- Fenómenos no vocales comunicativos
- Expresiones léxicas y semiléxicas
- Cambios de intensidad en el enunciado
- Discurso directo
- Texto leído
- Citas y refranes
- Palabras extranjeras
- Fórmulas
- Nombres propios
- Números
- Palabras deletreadas
- Abreviaturas
- Fragmentos poco claros en la grabación
- Errores de producción del hablante
- Palabras fragmentarias o truncadas y repeticiones de palabras

1975-1999: Información contenida en la cabecera

- Título y subtítulo del documento
- Responsable de la transcripción y codificación del texto
- Información sobre la edición electrónica del texto
- Información sobre la extensión del texto
- Localización del texto en la grabación
- Procedencia española o hispanoamericana del texto y país de procedencia
- Descripción de la codificación
- Clasificación del texto
- Descripción de los hablantes: código de identificación, nombre, papel, sexo, edad, lengua materna, variante dialectal, origen geográfico, país de procedencia, clase social y nivel de estudios.
- Información sobre la revisión del texto

1975-1999: Nº total de palabras: 8.790.971

Introducción y codificación de textos orales, 2000-2004

La construcción de la parte oral del *CREA* en la fase 2000-2004 se basa en la creación de un corpus representativo de la lengua hablada, **transcrito**, pero también **audible**, de los distintos **usos y variaciones de la comunidad hispanohablante** que pueda utilizarse como recurso lingüístico.

Durante el proceso de transcripción de textos orales se introducen de manera simultánea las marcas de codificación, XML en lugar de SGML, utilizado en el resto del *CREA*.

Se establecen los siguientes niveles de segmentación:

1. Delimitación del documento XML → <Trans>
2. Datos sobre el archivo (programa de radio/televisión) → <Episode>
3. Sección del programa (publicidad, noticia, entrevista) → <Section>
4. Turno de hablante (formado por 1 o más hablantes) → <Turn>
5. Elemento mínimo de sincronización → <Sync>

2000-2004: Información contenida en la cabecera

Localización del texto

<Trans scribe="Corpus de Referencia del Español Actual (CREA), Real Academia Española" audio_filename="DA011201" version="3" version_date="011218" xml:lang="Perú">

Clasificación

<Topic id="to1" desc="publicidad"/>
<Topic id="to2" desc="noticias 101"/>

Descripción de los hablantes

<Speakers>
<Speaker id="spk1" name="Arturo Muñoz" check="yes" type="male" dialect="native" accent="Peruano" scope="local"/>
<Speaker id="spk2" name="Arturo Valderrama Chávez" check="yes" type="male" dialect="native" accent="Peruano" scope="local"/>
</Speakers>

Datos del archivo: título y fecha

<Episode program="Debate: funciones del Congreso y de los Congresistas, 1/4, 01/12/01, CNR" air_date="2001">

Elementos estructurales

- **Secciones**

– <Section type="report" startTime="0" endTime="708.416" topic="to1">

- **Turnos**

– <Turn speaker="spk1" mode="spontaneous" fidelity="medium" channel="studio" startTime="5.231" endTime="32.48">

- **Sincronización:**

– <Sync time="17.512"/>

- **Turnos de un solo hablante**

```
<Turn speaker="spk5" mode="spontaneous" fidelity="high"
startTime="895.632" endTime="899.379">
<Sync time="895.632"/>
```
El partido se juega en las dos canchas,
```
<Sync time="897.684"/>
```
en la cancha de Chile y en la de Colombia .
```
</Turn>
```

Unturno.wav

```
<Turn speaker="spk7" mode="spontaneous" fidelity="high"
startTime="899.379" endTime="906281">
<Sync time="899.379"/>
```
Cinco minutos treinta va cobrándose en este momento.
```
<Sync time="901.629"/>
```
Levanta Jaime Riveros.
```
<Sync time="902.839"/>
```
De nuevo la pelota en las manos de Óscar Córdoba.
```
<Sync time="904.869"/>
¡Qué bien la para el arquero colombiano!
</Turn>
```

- **Turnos de dos hablantes:**


Dobleturno.wav

```
<Turn speaker="spk1 spk2" startTime="142.153"
endTime="142.966">
<Sync time="142.153"/>
<Who nb="1"/>
¿no es cierto?
<Who nb="2"/>
Ya.
</Turn>
```

Fenómenos codificados

Pausas


pause.wav

```
Ahora, estaba pensando,
<Sync time="28.308"/>
recién estamos a jueves ya,
<Sync time="29.52"/>
durante toda la semana no
<Event desc="pause" type="lexical"
extent="instantaneous"/>
 todavía nunca un chivito, o sea que
```
<Event desc="pause" type="lexical" extent="instantaneous"/>

Discurso directo


directo.wav

```
 al hacer la pregunta, usted dice:
 <Sync time="1219.463"/>
 <Event desc="q" type="lexical" extent="begin"/>
 yo no soy responsable
 <Event desc="q" type="lexical" extent="end"/>
 <Sync time="1220.406"/>
 Pero, alguien tiene que hacerse responsable
 </Turn>
```

Errores de producción: *sic*


sic.wav

```
Bien, muchas gracias,
<Event desc="sic" type="pronounce" extent="begin"/>
dondora
<Event desc="sic" type="pronounce" extent="end"/>
<Event desc="pause" type="lexical"
extent="instantaneous"/>
```
doctora Dora Núñez.

Fragmentos poco claros o ininteligibles


inintelli.wav

¿están cumpliendo con sus
<Event desc="unclear" type="pronounce"
extent="instantaneous"/>
<Sync time="299.281"/>
o de repente solamente fue un engaña muchachos,
<Sync time="302.241"/>
como nos han tenido durante mucho tiempo?

Palabras fragmentarias o truncadas


truncation.wav

están diciendo,
<Sync time="704.759"/>
ya los últimos tres
<Event desc="T" type="pronounce"
extent="instantaneous"/>
par
<Event desc="pause" type="lexical"
extent="instantaneous"/>
cuatro partidos,

Repeticiones de palabras


repe.wav

Yo creo que las cosas se pueden arreglar
<Sync time="233.211"/>
a partir
<Event desc="repe2" type="pronounce"
extent="begin"/>
del <Event desc="repe2" type="pronounce"
extent="end"/> momento más álgido y
<Event desc="repe1" type="pronounce"
extent="begin"/>
más <Event desc="repe1" type="pronounce"
extent="end"/> terrible, ¿no?

Fenómenos vocales semiléxicos comunicativos


voca1.wav

```
<Event desc="pensa e..." type="noise"
extent="instantaneous"/>
```
Otro aspecto que también quería tocar,


voca2.wav

```
<Event desc="asen mhm mhm" type="noise"
extent="instantaneous"/> .
```


voca3.wav

salí del cine abriendo la puerta a estilo kárate
```
<Event desc="onomat ¡zas!" type="noise"
extent="instantaneous"/>
```


voca4.wav

Bueno, le cuento que Daniel nos está
```
<Event desc="risa" type="noise"
extent="instantaneous"/>
```
contando que dentro de poco va a actuar con la cantante nacional Miriam Quiñones,

Fenómenos no vocales no comunicativos


coment.wav

Radio Reloj.
```
<Sync time="179.884"/>
<Comment desc="pitido"/>
```
Seis cuarenta y tres minutos.

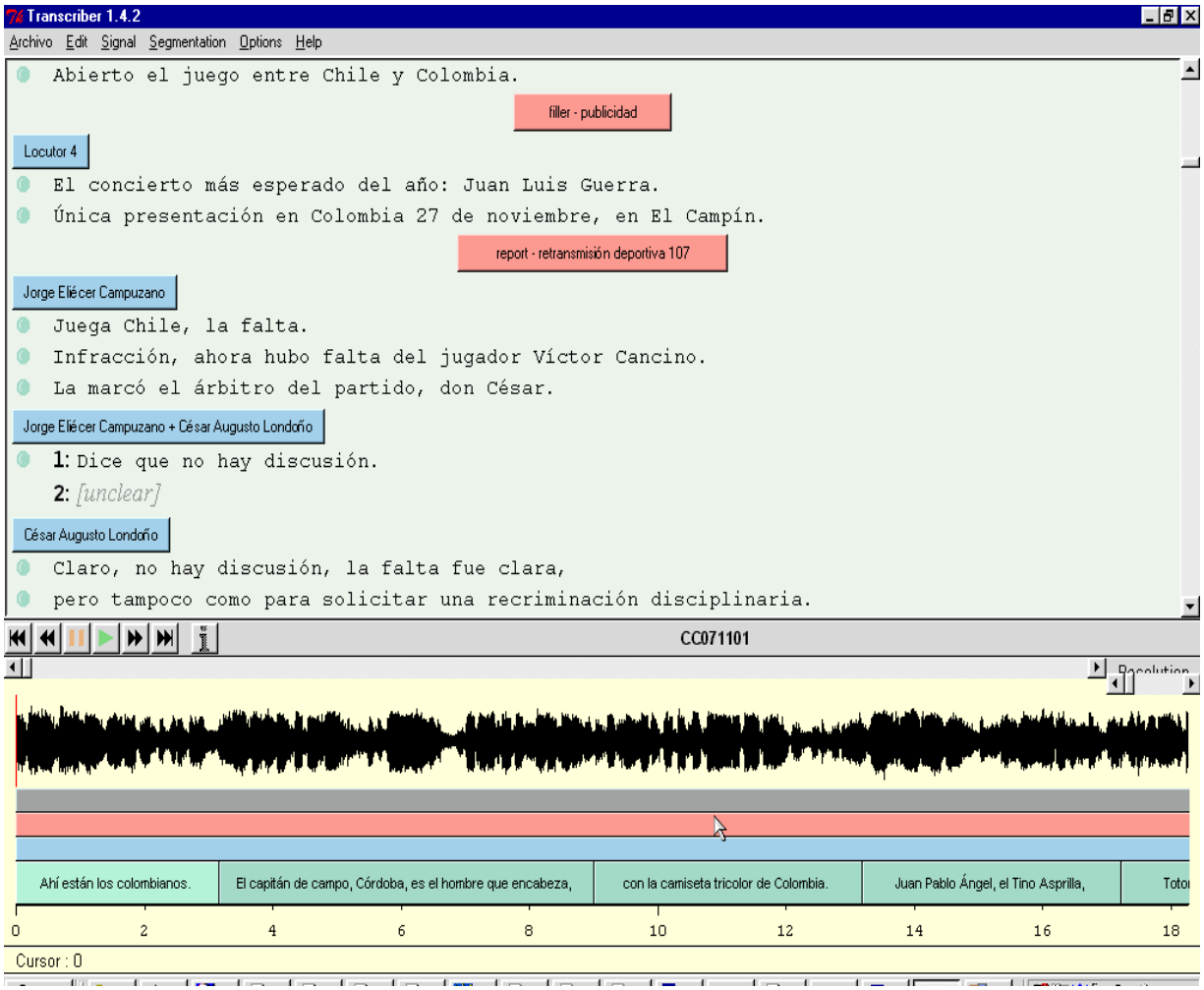
Palabras extranjeras

Los más sabrosos y jugosos
```
<Event desc="eng" type="language"
extent="begin"/>
```

foreign.wav

steaks
```
<Event desc="eng" type="language" extent="end"/>
```
para su
```
<Event desc="eng" type="language"
extent="begin"/>
```
barbecue
```
<Event desc="eng" type="language" extent="end"/>
```

Pantalla de codificación: Transcriber



Fichero de texto

```
<Sync time="797.83"/>
Abierto el juego entre Chile y Colombia.
</Turn>
</Section>
<Section type="filler" topic="to2" startTime="800.03" endTime="806.555">
<Turn speaker="spk8" mode="spontaneous" fidelity="high" startTime="800.03" endTime="806.555">
<Sync time="800.03"/>
El concierto más esperado del año: Juan Luis Guerra.
<Sync time="802.26"/>
Única presentación en Colombia 27 de noviembre, en El Campín.
</Turn>
</Section>
<Section type="report" topic="to1" startTime="806.555" endTime="866.754">
<Turn speaker="spk7" mode="spontaneous" fidelity="high" startTime="806.555" endTime="815.53">
<Sync time="806.555"/>
Juega Chile, la falta.
<Sync time="808.724"/>
Infracción, ahora hubo falta del jugador Víctor Cancino.
<Sync time="812.815"/>
La marcó el árbitro del partido, don César.
</Turn>
<Turn speaker="spk7 spk2" mode="spontaneous" fidelity="high" startTime="815.53" endTime="816.83">
<Sync time="815.53"/>
<Who nb="1"/>
Dice que no hay discusión.
<Who nb="2"/>

<Event desc="unclear" type="pronounce" extent="instantaneous"/>

</Turn>
<Turn speaker="spk2" mode="spontaneous" fidelity="high" startTime="816.83" endTime="826.233">
<Sync time="816.83"/>
Claro, no hay discusión, la falta fue clara,
<Sync time="818.34"/>
pero tampoco como para solicitar una recriminación disciplinaria.
```

2000-2004: Nº total de palabras CORPUS ALINEADO: 704454