Corpus del español del siglo XXI (CORPES)

Descripción del sistema de codificación
TEXTOS ORALES

Académico responsable

D. Guillermo Rojo

Equipo central CORPES XXI
Mercedes Sánchez Sánchez (coordinadora)

Julia Fernández Fernández
Mar Capilla Martín
Carme Lamela Villaravid

La descripción del sistema de codificación de los textos orales ha sido posible gracias a la colaboración del equipo de la Universidad de Santiago de Compostela dirigido por Victoria Vázquez y coordinado por Alba Sanmartín

CORPES XXI

TEXTOS ORALES

Descripción del sistema de transcripción y codificación -------

Versión 22/06/2018

Contenido

Generalidades

Procedencia de textos fuente para el componente oral del CORPES.

Transcripción y codificación:

- Convenio RTVE: audios correspondientes a diversos programas.
- Internet: YouTube, páginas de medios de comunicación con disponibilidad de audios, etc. Transcripción y codificación previas
- Textos procedentes de otros corpus Transcripción previa:
- Entrevistas en medios de comunicación, con disponibilidad de acudir al sonido para comprobar la transcripción.
- Materiales transcritos: Congreso, Senado, instituciones, etc.

Tipos de texto que podrán ser consultados en la aplicación de CORPES

- Alineados
- No alineados

1. Normas de transcripción

Ortografía

- En cuanto a la representación de las palabras, se siguen las convenciones ortográficas del español. Por tanto, no se reflejan fenómenos como:

- seseos, aspiraciones, rehilamientos, etc.
- "formas reducidas", como *[pa] para, [na] nada, [to] todo, [amás] además, [alante] adelante,* etc.
- contracciones no contempladas en la ortografía convencional: *[pal] para el, [deste] de este, [patrás] para atrás*.

- En cambio, se representan los elementos léxicos con acortamientos si son de uso general, estén o no recogidos en el DLE: *cole, boli, disco...*

- Las interjecciones se representan de acuerdo con la *Ortografía*, la *Gramática* y el *Diccionario* de la Real Academia Española.

- Además, se transcriben otros elementos funcionales que no aparecen en el DLE, tales como: *eeh, mmm, pff, sshh, eemm, buah, bueh, uff.* Para permitir su recuperación, se irá elaborando una lista que, dentro de la variedad existente, estandarice su modo de representación.

- Las secuencias numéricas: se transcriben en forma de palabra, siguiendo las normas ortográficas del español.

Uso de mayúsculas

- Solo se emplean mayúsculas para los nombres propios. En este sentido, se seguirán en lo posible las recomendaciones de la *Ortografía de la lengua española* (cap. IV).

- Puesto que no se usa la puntuación convencional, no habrá mayúsculas por razones de puntuación ni tampoco al principio de turno.

Puntuación

- Los únicos signos de puntuación convencionales que se utilizan en las transcripciones son:

  ¡! Para enunciados exclamativos
  ¿? Para enunciados interrogativos

- No se usan nunca coma, punto, punto y coma, dos puntos, puntos suspensivos, comillas ni guion.
- Para representar las pausas se emplean las siguientes marcas:

  / Pausas breves.
  // Pausas largas, de menos de un segundo.

- Tanto la pausa breve como la pausa se representan entre espacios.
- En caso de haber una pausa entre dos turnos, se transcribe al final del primero.
- Si hay una pausa tras un solapamiento, se transcribe después de la intervención del segundo hablante.
- Para pausas de más de un segundo se emplea la etiqueta *<silencio/>*.
- Si hay un silencio entre dos turnos, se transcribe al final del primer turno. Dentro de cada turno, las pausas y los silencios entre dos líneas de transcripción se transcriben siempre al final de la primera.

1. CODIFICACIÓN

<CORPES id="">

id="OR2003_0001">

OR: Oral
AAAA: Año de clasificación en CORPES. _0000: número correlativo de textos de ese año en Oral-CORPES.

<cabecera fecha_electrónica="">

fecha_electrónica="2014-12-11">

<título_principal autor_título_principal=""></título_principal>

autor_título_principal=""

- El autor es la cadena de radio / tv / internet / Subcorpus.
- La fuente para establecer el título será el título del programa en el caso de las grabaciones de radio/televisión. En el resto de los casos, será el título del texto fuente.

<edición procedencia="" subcorpus="" archivo_fuente_tipo="" archivo_fuente_localización="" lugar_grabación="" fecha_de_grabación="" fecha_de_emisión="" fecha_de_transcripción="" sonido_alineado=""/>

- Datos del archivo fuente. Condicionados por la procedencia.

procedencia=""

Valores:

Transcripción_y_codificación_previas:

- Texto previamente transcrito y codificado, procedente de otros corpus.

Transcripción_y_codificación_CORPES:

- Textos *genuinos* de CORPES, transcritos y codificados desde un archivo de sonido.

Transcripción_previa:

- Textos que se incorporan a CORPES pero que no proceden de un corpus: transcripciones de entrevistas a un político…

subcorpus="" Valor: denominación del corpus del que procede en caso de que así sea: PRESEEA, CORALES.

archivo_fuente_tipo=""

Valores: audio | vídeo | texto

archivo_fuente_localización=""

Url del audio o el vídeo.

- Si procede de otro subcorpus, el nombre del archivo en ese subcorpus.

lugar_grabación=""

- Consignar según el archivo fuente en el caso de los subcorpus incorporados.

fecha_de_grabación=""

- Solo si tenemos el dato, claro, independientemente del texto fuente.

fecha_de_emisión="" o
Lo mismo.

fecha_de_transcripción="" o La que aparece en el archivo fuente si procede de otro subcorpus o la que consigne el transcriptor de CORPES.

sonido_alineado=""

Sí: lleva marcas de alineamiento con el texto. No:
no está alineado con el archivo de audio.

<numpal n=""/> n=""
Número de palabras del texto.
<duración minutos="" segundos=""/>
Duración del archivo en minutos y segundos

<criterio_clasificación_CORPES criterio="" año=""/>

Datación del texto en CORPES de acuerdo con el criterio seleccionado.

criterio=""

Fecha_de_grabación
Fecha_de_emisión
Fecha_de_transcripción

año=""

Consignado AAAA.

<clasificación_textual medio_difusión="" tipología=""/>

medio_difusión=""

Radio | Televisión | Internet | Otros corpus

tipología=""

Conversación | Debate | Discurso | Entrevista | Entrevista_semidirigida |Magacines_y_variedades | Noticia | Publicidad Reportajes_y_documentales | Retransmisiones_deportivas | Sorteos_y_concursos | Tertulia | Otros |

<hablante hb="" nombre="" sexo="" grupo_edad="" edad="" nivel_edu="" estudios="" profesión="" ciudad_origen="" país="" zona="" origen="" otros_datos="" papel=""/>

- Es en el hablante donde se concentran realmente los datos *importantes*. Hay que entenderlo como una especie de microtexto sobre el que irá la carga de la recuperación de la información, y debe integrarse, por tanto, con el resto de textos de CORPES (escritos).

- Es muy probable que no sea posible consignar algunos datos como nombre, estudios, profesión, etc. Para esos casos utilizaremos 'No_indicado'.

hb=""

Código del hablante, formado por un número correlativo '001'…

En caso de un hablante colectivo del tipo 'varios' o 'todos', el valor del atributo 'hb' podrá ser ese, pero solo llevará los atributos comunes: por ejemplo, si tenemos la certeza de que todos los participantes son españoles, lo haremos constar en el parámetro geográfico, pero no será posible añadir los valores de edad, etc… es decir, valores individuales.

Si se trata de dos hablantes de distintas nacionalidades, consignaremos 'No_indicado'. nombre=""

Nombre del hablante, en formato Apellidos, Nombre. Si no lo conocemos, 'No_indicado'.

Si conocemos algún dato identificativo (solo nombre, solo apellido, apodo…) se hará constar igualmente.

sexo=""

mujer | hombre | 'No_indicado'

grupo_edad=""

0-14 | 15-19 | 20-34 | 35-54 | 55_adelante | No_indicado

edad=""

Si no consta, 'No_indicado'

nivel_edu=""

bajo | medio | superior | No_indicado

estudios=""

Si no consta, 'No_indicado'

profesión=""

Si no consta, 'No_indicado'

ciudad_origen=""

Si no consta, 'No_indicado'

país=""

Argentina | Bolivia | Chile | Colombia | Costa_Rica | Cuba | Ecuador | El_Salvador | España | Estados_Unidos | Filipinas | Guatemala | Guinea_Ecuatorial | Honduras | México | Nicaragua | Panamá | Paraguay | Perú | Puerto_Rico | República_Dominicana | Uruguay | Venezuela | No_identificado | No_nativo

No_identificado: Hablante cuyo origen no es posible identificar. No_nativo: Hablante de español pero no nativo (italiano, francés… Se complementa en otros_datos.

zona=""

Chilena | Río_de_la_Plata | Andina | Caribe_continental | México_y_Centroamérica | Antillas | Estados_Unidos | España | Filipinas | Guinea_Ecuatorial | No_identificado | No_nativo

No_identificado: Hablante cuyo origen no es posible identificar.
No_nativo: Hablante de español pero no nativo (italiano, francés... Se complementa en otros_datos.

origen=""

E | A | F | G | No_identificado | No_nativo

España, América, Filipinas, Guinea_Ecuatorial

No_identificado: Hablante cuyo origen no es posible identificar.
No_nativo: Hablante de español pero no nativo (italiano, francés… Se complementa en otros_datos.

otros_datos=""

Incluye alguna información adicional sobre el hablante. Ejemplo:

otros_datos="el hablante es italiano y habla español como segunda lengua".

papel=""

Si no consta, 'No_indicado'

Se irá preparando una lista de posibles opciones. Por ahora:

Presentador | Concursante | Entrevistado | Entrevistador | Circunstancial | Audiencia participante | Participante telefónico | Participante puntual

<codificación equipo_codificación="" persona_codificación="" fecha_codificación=""/>

Datos del responsable primero del texto: transcripción y codificación.

<validación valor_validación="" persona_validación="" fecha_validación=""/>

valor_validación=""

1 | 2 | 3

Identificación del responsable de la validación del texto.

<revisión_RAE valor_revisión_RAE="" persona_revisión_RAE="" fecha_revisión_RAE=""/>

valor_revisión_RAE=""

1 | 2 | 3 Datos sobre los responsables de revisión en la RAE.

<notas></notas>

Se añaden datos pertinentes relativos a todo el proceso y procedencia del texto.

Texto

<turno hb="" seg="">

Marca cada turno de hablante. Cada turno está compuesto por una sola intervención.

hb=""

Su valor está definido en la cabecera, en los datos del hablante. Recibe un número correlativo.

seg=""

Segundo en el que comienza la intervención del hablante. **Solo para los textos alineados**. Sincronización con el archivo de audio alineado.

Formato de las etiquetas

- No hay espacio entre la forma etiquetada y la etiqueta. Las únicas etiquetas que van entre espacios son las que no se refieren a ninguna forma, sino que marcan elementos no lingüísticos (risas, ruidos etc.), o sustituyen segmentos no transcritos (ininteligible, vacilación).

- No hay espacios antes ni después del signo = que contienen algunas etiquetas.

Marcación del tiempo

<tiempo seg=""/>

Sincronización con el archivo de audio. Se introduce en el momento en el que el hablante realice una pausa, larga o breve, de acuerdo con el criterio del transcriptor y codificador.

Lleva un atributo seg=”” cuyo valor es el del segundo en el que se marca el segundo exacto del alineamiento.

Solo para los textos alineados.

Marcación de fenómenos

<aplausos/>

<apoyo></apoyo>

Para segmentos que, para mantener la coherencia, es necesario incluir en la transcripción, pero que no deberían salir en los recuentos. Por ejemplo, hay entrevistas en las que las preguntas no se formulan oralmente sino que aparecen escritas en pantalla.

<cita></cita>

Para transcribir estilo directo; se utiliza en lugar de las comillas.

<fático hb=""/>

Sustituye a ciertas señales fónicas, no transcribibles con la ortografía convencional, usadas a menudo por los hablantes como señal de que están prestando atención al interlocutor.

Después del signo = se especifica el hablante que emite el fático.

<ininteligible/>

Sustituye un fragmento no transcrito por resultar ininteligible.

<música/>

Música no simultánea con el parlamento de un hablante.

<música_de_fondo></música_de_fondo>

Encierra un fragmento de texto con música de fondo perceptible.

<nrp/>

Sustituye segmentos no transcritos por cualquier motivo, como parlamentos en otro idioma, fragmentos de películas… Si se considera pertinente, puede ir acompañada de una <observación_complementaria desc=””> para especificar qué es lo que se ha eliminado.

<observación_complementaria desc=""/>

Para añadir cualquier tipo de información, en el valor del atributo ‘desc’ que no se puede codificar mediante las etiquetas existentes: explicaciones sobre la situación, silencios largos en los que ocurre algo relevante para la interacción, gestos especialmente significativos… También se puede emplear para introducir aclaraciones a otras etiquetas que lo necesiten, como <nrp/> o <apoyo></apoyo>.

<palabra_cortada></palabra_cortada>

Etiqueta fragmentos de palabras no concluidas.

<risa hb=""/>

Para risas que no acompañan al hablante. Debe especificarse el hablante que se ríe, tanto si está en posesión de la palabra como si no.
Cuando existan dos o más hablantes que se puedan identificar, se introduce un elemento <risa../> por cada uno de ellos:
<risa hb="001"/><risa hb="002"/>
Si se ríe más de un participante sin identificar, entonces la indicación será <risa hb="varios"/> y 'varios' debe estar recogido en los datos de la cabecera, en el elemento <hablante…>

<risas_inicio hb=""/>…<risas_fin hb=""/>

Para risas que se extienden a lo largo del parlamento de un hablante. Por tanto, se sitúa rodeando un fragmento de texto.
Hay que especificar el hablante que se ríe, tanto si está en posesión de la palabra como si no.
Si se ríe más de un participante, entonces la indicación será <risas hb="varios">…</risas> y 'varios' debe estar recogido en los datos de la cabecera, en el elemento <hablante…>

<ruido desc=""/>

Para un ruido que no se solapa con el habla de ningún participante. No se utiliza como etiqueta de apertura y cierre. Se especifica el tipo de ruido en el valor del atributo 'desc'. Por ejemplo, "chasquido boca" o "resoplido".

<ruido desc="chasquido de boca"/>

<ruido_de_fondo></ruido_de_fondo>

Encierra un fragmento emitido simultáneamente con cualquier tipo de ruido claramente perceptible y que pueda resultar relevante para interpretar la transcripción. No se especifica de qué ruido se trata.

<sic></sic>

Para errores de pronunciación llamativos que se considere oportuno reflejar en la transcripción (no acortamientos, contracciones u otros fenómenos propios de la oralidad), y que puedan ser confundidos con un error del transcriptor.

<siglas desc=""></siglas>

Se transcriben convencionalmente, pero dentro de las comillas, se representan tal y como las ha pronunciado el informante, aunque utilizando las normas ortográficas convencionales.

<silencio/>

Para silencios de más de un segundo.
Las pausas no se marcan con etiquetas sino con barra / (breve) y doble barra // (larga)

<simultáneo></simultáneo>

Para segmentos de habla solapada.

<traducción></traducción>

Para segmentos traducidos. Por ejemplo: casos en los que se oye el parlamento original en otro idioma de fondo y al mismo tiempo la traducción simultánea.

<transcripción_dudosa></transcripción_dudosa>

Para fragmentos en los que la transcripción no está del todo clara pero que se entienden lo suficiente como para que no parezca adecuado emplear <ininteligible/>.

<vacilación/>

Sustituye a fragmentos similares a palabras cortadas pero imposibles de transcribir.

Archivo oral: jerarquización de elementos

```
<CORPES id="">
  <cabecera fecha_electrónica="">
    <título_principal autor_título_principal=""></título_principal>
    <edición procedencia="" subcorpus="" archivo_fuente_tipo="" archivo_fuente_localización="" lugar_grabación="" fecha_de_grabación="" fecha_de_emisión="" fecha_de_transcripción="" sonido_alineado=""/>  <numpal n=""/>  <duración minutos="" segundos=""/>
    <criterio_clasificación_CORPES criterio="" año=""/>
    <clasificación_textual medio_difusión="" tipología=""/>
    <hablante hb="" nombre="" sexo="" grupo_edad="" edad="" nivel_edu="" estudios="" profesión="" ciudad_origen="" país="" zona="" origen="" otros_datos="" papel=""/>
    <codificación equipo_codificación="" persona_codificación="" fecha_codificación=""/>
    <validación valor_validación="" persona_validación="" fecha_validación=""/>
    <revisión_RAE valor_revisión_RAE="" persona_revisión_RAE="" fecha_revisión_RAE=""/>
<notas></notas>
  </cabecera>
  <texto>
    <turno hb="" seg="">
      <aplausos/>
      <apoyo></apoyo>
      <cita></cita>
      <fático hb=""/>
      <ininteligible/>
      <música/>
      <música_de_fondo></música_de_fondo>
      <nrp/>
      <observación_complementaria desc=""/>
      <palabra_cortada></palabra_cortada>
      <risa hb=""/>
      <risas_inicio hb=""/>…<risas_fin hb=""/>
      <ruido desc=""/>
      <ruido_de_fondo> </ruido_de_fondo>
```

```
        <sic></sic>
        <siglas desc=""></siglas>
        <silencio/>
        <simultáneo></simultáneo>
        <tiempo seg=""/>
        <traducción></traducción>
        <transcripción_dudosa></transcripción_dudosa>
        <vacilación/>
      </turno>
  </texto>
</CORPES>
```