



# CORPUS MEsA 2.0

«Macrosintaxis del Español Actual»

GUÍA DE USUARIO

**Plan Estatal Excelencia – Proyectos I+D**

IP: Dra. Catalina Fuentes Rodríguez



## **CORPUS MEsA 2.0**

Este corpus es resultado del trabajo de recopilación de material llevado a cabo durante la vigencia de dos Proyectos de I+D correspondientes al Programa Estatal de Fomento de la Investigación Científica y Técnica de Excelencia, Subprograma Estatal de Generación del Conocimiento:

- «Macrosintaxis del Español Actual. El Enunciado: Estructura y Relaciones» (FFI2013-43205-P) (2014-2017)
- «De Construcciones Periféricas a Operadores Discursivos: un Estudio Macrosintáctico del Español Actual» (FFI2017-82898-P) (2018-2021)

 Licencia de Creative Commons.

Todos los documentos del Corpus MEsA 2.0 están sujetos a la licencia de Creative Commons Reconocimiento-NoComercial-SinObraDerivada 4.0 Internacional.

## ÍNDICE

※ Introducción .....	3
Nuevo MEsA 2.0 .....	3
¿Sintaxis del discurso? .....	3
¿Por qué Internet?.....	3
¿Para qué este corpus?.....	4
¿Qué puedo encontrar?.....	4
※ Organización de documentos .....	6
※ Constitución interna de los documentos.....	6
1. Organización interna .....	6
※ Sistema de compilación/transcripción.....	8
1. Pautas generales en la compilación de textos escritos .....	8
2. Pautas generales en las transcripciones de textos orales .....	9
3. Organización de los emisores en las interacciones .....	9
4. Características particulares de la compilación por fuente .....	11
※ Versión 2.0.....	12

## ✧ **Introducción**

El corpus de fuentes digitales «MEsA» es el resultado del esfuerzo de cuatro años de compilación de materiales lingüísticos procedentes de Internet. Su creación forma parte de los objetivos del proyecto de excelencia “Macrosintaxis del Español Actual” (MEsA), financiado por el Ministerio de Economía y Competitividad (FFI2013-43205-P) y dirigido por la Dra. Catalina Fuentes Rodríguez, catedrática de Lengua Española de la Universidad de Sevilla. Forman parte de este proyecto investigadores de la Universidad de Sevilla, León y Graz, y todos ellos se han volcado en el análisis y el estudio de la unidad básica de la sintaxis del discurso: el enunciado. El corpus recoge material público disponible en Internet y solo usado a efectos de investigación lingüística.

### **Nuevo MEsA 2.0**

La actualización y ampliación del corpus MEsA (MEsA 2.0) forma parte de los objetivos del proyecto de excelencia “De Construcciones Periféricas a Operadores Discursivos: un Estudio Macrosintáctico del Español Actual”, financiado por el Ministerio de Economía, Industria y Competitividad (FFI2017-82898-P) y dirigido también por la Dra. Catalina Fuentes Rodríguez.

En esta nueva actualización se recogen nuevos materiales comprendidos entre los años 2018 y 2021. Por otra parte, se ha realizado una exhaustiva revisión del material precedente con el fin de corregir distintos errores encontrados en la codificación, la cual se ha simplificado y unificado para una mayor homogeneidad del corpus. También se han añadido notas explicativas en los materiales que presentaban algunas particularidades. Por último, se han elaborado índices para un manejo más sencillo del corpus por parte del investigador y se ha reflejado la paginación de todos los documentos. Todos los detalles acerca de esta nueva actualización se describen en el último apartado de esta guía ([ver Versión 2.0](#)).

### **¿Sintaxis del discurso?**

Es una denominación que afirma que el discurso, más allá de los límites oracionales, posee su propia estructura y organización. Además, expresa que existen patrones y herramientas dentro del sistema lingüístico para fijar esta estructura en cualquier uso real de la lengua.

Si el discurso tiene su propio sistema de estructuración, debe poseer unidades mínimas que constituyan los bloques sobre los que se sustenta. Esto es el enunciado, que da sentido al subtítulo de este proyecto: El enunciado: estructura y relaciones. Uno de los objetivos principales de sus miembros es determinar los límites de esta unidad y ver las relaciones que se generan con el resto de componentes del discurso. Para lograrlo, es necesario poseer un material de realizaciones lingüísticas reales muy variado y que dé cuenta de contextos de uso muy diversos. De esta forma, se afina el análisis y garantiza que las conclusiones obtenidas se basan siempre en elementos efectivos y totalmente funcionales dentro del sistema lingüístico.

### **¿Por qué Internet?**

Internet ha dejado de ser el futuro; Internet es el presente. Constituye una de las principales vías de comunicación en la sociedad actual y su presencia es constante en nuestra vida cotidiana: una publicación en Facebook, un tuit, una noticia en un periódico digital, un correo electrónico, una entrada de videoblog en YouTube, una reseña en TripAdvisor o un mensaje de WhatsApp. Desde la revolución de la Web 2.0, que permite a cualquier usuario generar contenidos en la Red, las formas de comunicación y de interacción lingüística entre hablantes de una lengua (o varias) han aumentado de forma exponencial, y creemos que los lingüistas tienen mucho que decir al respecto. Es parte de sus labores analizar y estudiar la comunicación verbal mediada por ordenador, con el objetivo de realizar una descripción fiel de este medio, así como potenciar y difundir las herramientas lingüísticas de comunicación verbal más efectivas a través de Internet.

### **¿Para qué este corpus?**

Con este corpus ofrecemos una herramienta a todos aquellos lingüistas o investigadores que quieran acercarse a la comunicación verbal en Internet, poniendo especial foco en los textos escritos (aunque no exclusivamente), ya que este canal sigue siendo el principal medio de expresión de la Red. Los materiales compilados se han modificado para facilitar las búsquedas y las comparaciones de resultados de distintas fuentes, pero respetando en la medida de lo posible el original. Además, todos los textos de este corpus aparecen exhaustivamente organizados bajo etiquetas que los clasifican por fuente, temática y fecha. De esta forma, se consigue realizar consultas de fenómenos lingüísticos más eficaces que la búsqueda directamente por Internet.

### **¿Qué puedo encontrar?**

En el corpus MEsA es posible encontrar textos en español, en un ámbito panhispánico, de una amplia gama de lugares en Internet: blogs, Facebook, foros, Instagram, Twitter, WhatsApp, YouTube y páginas web de distinto tipo. Mensajes escritos y transcripciones de vídeos constituyen la mayor parte del corpus y estos se complementan con descripciones de los elementos no verbales que rodean cada texto. Además, gracias al sistema de códigos que organizan los materiales, es posible vincular cada fragmento con la imagen de la fuente real en el momento de la recopilación.

La ortografía original de los textos recopilados se mantiene. Aunque esto puede dificultar la búsqueda de algunos fenómenos, consideramos que es fundamental respetar la forma del mensaje original para no perder ciertos significados o sentidos. Creemos que el buen analista es capaz de ver más allá de las faltas de ortografía y de expresión para desmembrar la estructura lingüística que hay detrás y el funcionamiento de la lengua en Internet, que está caracterizada prominentemente por la coloquialidad.

Este corpus puede ser una herramienta útil para el analista de la lengua. Puede abrir muchas líneas de investigación y dar cuenta de fenómenos lingüísticos que no se encuentren con tanta frecuencia en corpus de textos escritos convencionales o de conversaciones transcritas. Nuestra intención es seguir ampliando el material del corpus a medida que avanza el contenido de la

Red, y también seguir mejorando las herramientas de consulta del material recopilado facilitando el contacto entre el lingüista y su objeto de estudio.

**Directora del corpus e IP de los proyectos:** Catalina Fuentes Rodríguez.

**Edición y formato del corpus (primera versión):** María Soledad Padilla Herrada, Rocío Martínez Barcelona, Irene Mateo Nieto, Víctor Pérez Béjar y María Luz Díaz Domínguez.

**Edición y formato del corpus (versión 2.0):** Paula Gil y Martínez e Irene Martín del Barrio.

**Compiladores:** María Soledad Padilla Herrada, Rocío Martínez Barcelona, Irene Mateo Nieto, Víctor Pérez Béjar, María Luz Díaz Domínguez, Anabel Alfaro Romero, Giulia Maltese, Laura Ángeles Sánchez, Manuel Contreras, Vanessa Gómez Torres, Juan Luis Ramírez, Cristina Bermúdez, Guadalupe Luna, Marina Jiménez Rodríguez, Lydia Silvestre Roldán, Mariela Andrade Maureira, Irene Martín del Barrio y Paula Gil y Martínez.

**Agradecimientos:** Mario de la Fuente García.

## ※ Organización de documentos

- Todo el material de una misma fuente digital será integrado en un mismo documento de texto. Ej.: Twitter, Facebook, Blogs, etc.
- El título de cada documento nombra la fuente digital de su contenido.
- No hay límite de páginas y los documentos están sujetos a la ampliación de material.

## ※ Constitución interna de los documentos

### 1. Organización interna

El material se reúne, con carácter general, siguiendo estos criterios:

- Primer nivel: por año, de más antiguo a más reciente.
- Segundo nivel: de un mismo año, se ordenarán por orden alfabético según el código del perfil ([ver 1.3](#)).
- Tercer nivel: de un mismo perfil, se ordenarán por criterio cronológico (día, mes) o por orden de compilación ([ver 1.3](#)).

#### 1.1. Ficha técnica del documento

En cada documento se incluye una ficha técnica que posee la siguiente información, distribución y formato:

FUENTE DIGITAL - FICHA TÉCNICA	
<b>1. Datos de la recopilación:</b>	
<b>1.1. Tipo(s) de documento:</b>	Se refiere a la fuente original del material recopilado (Facebook, Twitter, foros, blogs...). En ocasiones se indican los diversos subtipos de que se recopilan.
<b>1.2. Material compilado:</b>	códigos que se compilan en el cuerpo del documento.
<b>1.3. Fechas:</b>	período que abarca el material recopilado.
<b>1.4. Distribución:</b>	el material se reúne teniendo en cuenta la fuente original siguiendo un criterio cronológico y alfabético.
<b>1.5. Recopiladores / Transcripores:</b>	el nombre de las personas que han compilado o transcrito el material lingüístico.
<b>2. Temas tratados:</b>	temáticas generales del material.
<b>3. Características de la comunicación:</b>	
<b>3.1. Emisores:</b>	nombre de los perfiles recogidos.
<b>3.2. Ámbito comunicativo:</b>	se especifica el alcance que pueden tener los mensajes y sus receptores potenciales (público, privado).

**3.3. Rasgos lingüísticos:** se indica el registro (formal, coloquial o los dos) en el que se enmarcan los mensajes, características ortográficas relevantes y otros fenómenos pertinentes relacionados con cualquier nivel o dimensión lingüística.

**3.4. Otros códigos implicados:** se indican los sistemas comunicativos implicados en la transmisión de los mensajes: texto, imágenes, vídeos, audio, etc. Así como la aparición de fenómenos de relación textual (intertextualidad y uso de hipervínculos).

**3.5. Elementos tipográficos:** se indican los elementos tipográficos utilizados en el texto original (negrita, cursiva, mayúscula, subrayado...) y cuáles aparecen reflejados en el documento.

**4. Otros factores:** se explica otro tipo de características que no hayan quedado reflejadas en los apartados anteriores.

## 1.2. Cabeceras

Al comienzo de cada apartado (el material que se engloba bajo un mismo código por tener el mismo origen y el mismo año de emisión) se incluirá un breve cuadro descriptivo, denominado cabecera, que contendrá la siguiente información:

**1<sup>1</sup>. TW 2016 ERR<sup>2</sup>**

**Cuenta de Twitter de Íñigo Errejón<sup>3</sup>**

[<http://www.twitter.com/ierrejon>]<sup>4</sup>

**Transcriptor / Recopilador<sup>5</sup>: M<sup>a</sup> Soledad Padilla Herrada**

1. Número, en orden, de la cabecera.
2. Código genérico del fragmento [fuente, año, código perfil/página]
3. Origen del material. En caso de páginas web, debe determinarse la sección.
4. Enlace al material original
5. Nombre de los recopiladores y transcriptores.

## 1.3. Entradas y códigos

Cada material se encuadra bajo una entrada que lo diferencia con un código único, en Times New Roman 16, negrita:

**TW<sup>1</sup> 2016<sup>2</sup> jul<sup>3</sup> ERR<sup>4</sup> 01<sup>5</sup>**

1. **Código de la fuente:** Blog (BL), Facebook (FB), Foros (FO), Página Web (PW), Twitter (TW), WhatsApp (WA), YouTube (YT) e Instagram (IG).
2. **Año** (cuatro dígitos).
3. **Mes:** ene (enero), feb (febrero), mar (marzo), abr (abril), may (mayo), jun (junio), jul (julio), ago (agosto), sep (septiembre), oct (octubre), nov (noviembre) o dic (diciembre).

**4. Perfil de usuario/empresa/periódico/etc.:** Se usarán las tres primeras letras del título. En el caso de que la primera palabra pueda tener coincidencias con otros contextos, se usarán las tres primeras letras de la palabra principal (por ej., se usará "ALF" para referirnos al "Hotel Alfonso XIII", pues "hotel" puede corresponder a muchas otras entradas). Se evitarán también los artículos que introduzcan sustantivos. Ejs: +Risto Mejide: RIS +El Corte Inglés: COR.

**5. Día/orden de recopilación:** el código preferido siempre será el que se corresponda con el día de la publicación. En aquellos casos en los que no se pueda determinar el día (por ej. en las entradas de las páginas webs), se introducirá bajo la cabecera una nota explicativa entre corchetes que dé cuenta de esta condición.

Por otra parte, en aquellos casos en los que coincida el día (por ej., cuando se recogen varios tuits en Twitter de la misma cuenta y del mismo día), se añadirá la numeración correspondiente al orden de aparición del material tras un punto: TW 2016 jul ERR 01.1, TW 2016 jul ERR 01.2, TW 2016 jul ERR 01.3, etc.

Los códigos que introducen cada fragmento del corpus irán separados del comienzo de este por un salto de línea. El final de un fragmento con el código que inicia el siguiente estará separado por otro salto de línea. Ej.:

## **TW 2015 abr EVO 01**

**Jordi Évole:**

Otro conflicto entre familias español: los Sánchez y los Díaz.

## **TW 2015 abr EVO 02**

**Usuario 1 (no identificado):**

Los díaz son el poder en la sombra

## **✂ Sistema de compilación/transcripción**

### **1. Pautas generales en la compilación de textos escritos**

- Por norma general, se mantiene la ortografía original de los textos escritos de fuentes digitales.
- La fuente será adaptada siempre a Times New Roman 12.
- Se respetarán las tabulaciones y la distribución general del texto en párrafos.
- Los emoticonos reconocidos por el editor de texto aparecerán en su posición original sin ningún tipo de modificación. Los que no son reconocidos, se simplifican con caracteres más sencillos ( :) // :( // ;) // ¬\_¬ // XD // :P) o se describen entre paréntesis.

- Las imágenes que acompañan al texto original se describen entre corchetes introducidos en la posición en la que aparece en el mensaje real. En los casos en los que sea necesario, será indicada la localización exacta de la imagen.
- En el caso de que las imágenes incluyan texto, este aparecerá enmarcado por comillas dentro de la descripción entre corchetes.

## 2. Pautas generales en las transcripciones de textos orales

- Las transcripciones de textos orales siguen, como norma general, las pautas de transcripción del grupo Val.Es.Co<sup>1</sup>.
- El discurso referido, principalmente el directo, se marcará en cursiva. También se usará esta marca para cambios evidentes de enunciadores: “Me dijo *no me hables más*”.
- Los neologismos de uso no general y las palabras en lenguas extranjeras también se marcan en cursiva: “*Love wins*”.
- En las transcripciones de los vídeos, podrán aparecer los siguientes tipos de información entre corchetes: cortes en la edición del vídeo, sonidos o música de fondo y contenido de los textos escritos que aparezcan durante la reproducción del vídeo.
- En las transcripciones de los vídeos, se pondrán notas explicativas a pie de página para los siguientes casos: descripción de gestos o movimientos del emisor y otros elementos visuales que sean relevantes para la comprensión del mensaje.

## 3. Organización de los emisores en las interacciones

- En la recopilación del material se diferencia entre personajes públicos y usuarios anónimos. Los personajes públicos son identificados explícitamente. Los anónimos, con el objeto de proteger su identidad, son identificados solamente con la etiqueta "Usuario". Esta etiqueta es seguida de un número que refleja el orden de intervención de los mismos. Ej.: Usuario 1, Usuario 2, Usuario 3, Usuario 4, etc.
- El sexo de los participantes anónimos se refleja entre paréntesis detrás del número del usuario. Las etiquetas para marcar el sexo son "hombre", "mujer" y "no identificado". Esta última designación es utilizada en los casos en los que no es posible identificar el sexo del emisor. Ejs.: Usuario 2 (mujer), Usuario 6 (hombre), Usuario 9 (no identificado).
- El nombre propio o la etiqueta "usuario" seguida del sexo del participante usará la fuente Times New Roman 12 en negrita. Irá seguida de dos puntos como forma de introducir su mensaje. Ej.: **Jordi Évole: // Usuario 5 (hombre):**

---

<sup>1</sup> El sistema de transcripción del grupo Val.Es.Co. puede consultarse en el siguiente enlace <http://www.uv.es/valesco/sistema.pdf>. [Consultado el 28/12/2021]

- El texto que cada usuario emite irá en una línea inmediatamente posterior al nombre propio del personaje público o de la etiqueta "usuario". Ejs.:

**Jordi Évole:**

El sorpasso nos tiene a todos sobrepasados.

**Usuario 7 (mujer):**

Jajaja qué más da

- Las intervenciones se recogerán según el orden de aparición en la fuente original. En el caso de que la plataforma permita la ordenación de los comentarios, se preferirá el cronológico, de más antiguo a más reciente.
- Entre cada intervención se dejará una línea en blanco, tal y como se muestra en el ejemplo anterior y en los sucesivos.
- Cuando una publicación o intervención genera una línea conversacional propia o una sucesión de enunciados reactivos se seguirán las siguientes pautas:

- El comentario inicial será situado sin tabulación.

- Los comentarios reactivos se marcarán con una tabulación en 1 cm.

- Otros niveles conversacionales, como las líneas temáticas dentro de un comentario reactivo se marcarán con tabulaciones sucesivas: 2 cm, 3 cm, 4 cm... Ej.:

**Jordi Évole:** .....*Tab. 0 cm*  
Otro conflicto entre familias español: los Sánchez y los Díaz.

**Usuario 1 (no identificado):** .....*Tab. 1 cm*  
Ls díaz son el poder en la sombra

**Usuario 2 (mujer):** .....*Tab. 2 cm*  
La Susana sombra?? Esa brilla más que los farolillos de la feria

**Usuario 3 (no identificado):** .....*Tab. 2 cm*  
La smbra es el sobako de los politicos

**Usuario 4 (hombre):** .....*Tab. 1 cm*  
Los Sánchez caerán a manos de los árboles comandados por los Díaz.

**Usuario 5 (mujer):** .....*Tab. 1 cm*  
Los Évole no son también mafia familiar?

- Las intervenciones de cualquiera de estos niveles de conversación que tengan más de una línea tendrán la misma sangría en la primera que en las líneas sucesivas. Ej.:

**Usuario 1 (mujer):**

Hola a todos, vendo un reloj de pulsera de mujer de segunda mano. Es Rolex y apenas lo he utilizado. Lo tengo desde hace dos años y para lo único que me servía era para estar en un cajón. Lo vendo por 100 €. Si alguien está interesado, le envío una foto por MP.

**Usuario 2 (hombre):**

Podría ser un buen regalo para mi mujer. Si te confirmo ahora, me harías una rebaja?

**Usuario 1 (mujer):**

[mención al Usuario 2] Creo que ya he hecho una rebaja bastante grande con los 100 €. Lo tengo desde hace dos años y está nuevo, y no sé si sabes cuánto te puede costar un buen Rolex. Además, se vende con la caja original y una correa de repuesto.

**Usuario 2 (hombre):**

De acuerdo, lo pienso y te escribo MP.

**Usuario 3 (no identificado):**

Me interesa. Si aún no lo has vendido, MP!!!!

## **4. Características particulares de la compilación por fuente**

### *4.1. Blogs*

El material compilado de los blogs consiste en publicaciones del autor del blog y sus respectivos comentarios de usuarios anónimos. Se identifican tres tipos de blogs: personales (de autores particulares), profesionales (con un perfil de trabajo y una imagen pública) y corporativos (gestionados por empresas), todos ellos abiertos al público. Se recogen publicaciones desde 2009 hasta 2017. Por motivos que desconocemos, muchos blogs y/o entradas de blogs han sido eliminados, por lo que la información es irrecuperable en su versión original.

### *5.2. Facebook*

Se recopilan únicamente publicaciones de páginas públicas y sus comentarios. Las páginas son de diversa naturaleza: celebridades, programas de televisión o distintos tipos de entidades. Se compilan publicaciones desde 2012 hasta 2020. Por motivos que desconocemos, muchos perfiles han sido eliminados, por lo que la información es irrecuperable en su versión original.

### *5.3. Foros*

Se recopilan hilos de conversación de foros públicos de dos tipos: generales (constituidos por usuarios anónimos al mismo nivel) y corporativos (con usuarios que representan a la empresa y usuarios que son clientes). Se compilan conversaciones desde 2004 hasta 2018. Por motivos que desconocemos, muchos foros y/o entradas de foros han sido eliminadas, por lo que la información es irrecuperable en su versión original.

### *5.4. Instagram*

Se compilan los comentarios de una determinada imagen publicada desde cuentas abiertas. Los tipos de cuentas originales pertenecen a personajes públicos, programas de televisión y distintas entidades, pero los comentarios son, principalmente, de usuarios anónimos. En cada

publicación, los comentarios se ordenan cronológicamente del más antiguo al más reciente. Se compilan materiales desde 2012 hasta 2021.

### *5.5. Páginas web*

Aunque “página web” es un hiperónimo de cualquier fuente en la Red, reunimos bajo esta etiqueta los dominios web corporativos como páginas de periódicos, revistas culturales o empresas de diversa naturaleza. Por este motivo, los materiales recopilados provienen de publicaciones de distinto tipo: noticias, reseñas, textos divulgativos... y sus comentarios. Se recopilan materiales desde 2009 hasta 2018.

### *5.6. Twitter*

Se compilan tuits de personajes públicos o entidades de diversa naturaleza y todos aquellos que conforman sus hilos conversacionales. Estos últimos son emitidos, principalmente, por usuarios anónimos. Los tuits que conforman un hilo conversacional se reúnen bajo la misma fuente (mismo código) y se ordenan respetando la organización original (criterio cronológico). Se compilan tuits desde octubre de 2015 hasta 2021.

### *5.7. YouTube*

Se transcriben los vídeos de diversos canales públicos y se compilan los comentarios de cada vídeo. Los canales pertenecen a usuarios anónimos, personajes públicos, programas de televisión y varias entidades. Los comentarios en cada vídeo publicado se ordenan respetando la distribución original (ordenados por relevancia, por orden cronológico y por hilos conversacionales). Se recopilan publicaciones de vídeos desde 2012 hasta 2018.

### *5.8. WhatsApp*

Se recopilan conversaciones privadas entre dos personas o de chats grupales. Los usuarios se identifican según el género (H: hombre, M: mujer) y el orden de aparición (M1, M2, M3... H1, H2, H3...).

El material se reúne en torno a la conversación a la que pertenezca (mismo código) y las intervenciones según un estricto criterio cronológico. Se especifica la fecha y hora de cada comentario según aparece en el historial de conversaciones. Las conversaciones pueden abarcar grandes períodos de tiempo. Se recopilan chats desde 2015 hasta 2018.

## **✂ Versión 2.0**

En este apartado detallaremos el trabajo llevado a cabo en esta actualización y las mejoras producidas:

- Se ha reemplazado el índice de códigos de cada documento por un único catálogo que recoge, por fuente y en orden cronológico, la relación de todos los códigos incorporados al corpus.

- Se ha hecho un reajuste y unificación del sistema de codificación. En la versión anterior se observaron distintos solapamientos y/o discordancias en los códigos: algunas fuentes compartían un mismo código y, en otros casos, una misma fuente aparecía con distinta codificación. De este modo, conseguimos una relación 1 código – 1 fuente.
- Reducción de los códigos compuestos por cuatro letras a tres.
- Corrección de codificaciones erróneas.
- Reordenación cronológica de todo el material.
- Corrección y unificación de fuentes tipográficas: Times New Roman.
- Unificación de la etiqueta [Comentarios]: en la versión anterior se observaron distintas formas de reflejar la relación de interacciones de los usuarios anónimos, por lo que se decidió establecer un criterio único.
- Elaboración de indicaciones explicativas en las cabeceras. Ante la problemática de diferenciar si el código de algunas entradas se correspondía con la fecha o el orden de recopilación, se comprobaron las fuentes de origen y se añadieron las explicaciones correspondientes bajo las cabeceras, entre corchetes.
- Elaboración de índices en cada documento con hipervínculos para permitir un manejo más cómodo del corpus por parte del investigador.
- Revisión, edición e incorporación de nuevo material al corpus. En concreto, en cada corpus, ha supuesto el siguiente incremento:
  - MEsA Blogs: de 859 a 1060 páginas.
  - MEsA Facebook: de 772 a 928 páginas.
  - MEsA Foros: de 737 a 929 páginas.
  - MEsA Instagram: de 1288 a 1540 páginas.
  - MEsA Páginas Web: de 743 a 1280 páginas.
  - MEsA Twitter: de 1012 a 2680 páginas.
  - MEsA YouTube: de 871 a 1194 páginas.