CAPITEL

Corpus **A**notado del **P**lan de **I**mpulso de las **Te**cnologías del **L**enguaje

Centro de Estudios
de la
Real Academia Española

24 de septiembre de 2019

Índice

1. Marco de colaboración

1. Convenio Marco SESIAD-RAE (19 de mayo de 2016)
2. Adenda al Convenio (16 de enero de 2018)

↓

CAPITEL
(**C**orpus **A**notado del **P**lan de **I**mpulso de las **Te**cnologías del **L**enguaje)

2. Diseño y adquisición del corpus (1/2)

- Diseño
  - Sincrónico: textos producidos principalmente después de 2005
  - Tamaño: sujeto a los acuerdos con las fuentes proveedoras de datos
  - Organización de las noticias (anotaciones *stand-off*):
    - Metadatos: Año, fuente, título, autor, tema
    - Datos primarios
    - Anotación morfosintáctica
    - Entidades nombradas
    - Anotación sintáctica
- Codificación de datos primarios:
  - Usando estándares: UTF-8, XCES, TEI y XHTML.

Temas:

1. Ciencias y tecnología
2. Ciencias sociales, creencias y pensamiento
3. Política, economía y justicia
4. Artes cultura y espectáculos
5. Actualidad, ocio y vida cotidiana
6. Salud
7. Otros

Metadatos


```
<teiHeader xmlns="http://www.tei-c.org/ns/1.0" xmlns:xi="http://www.w3.org/2001/XInclude">
<fileDesc>
  <titleStmt>
    <title type="main">Falling Walls celebra su primer laboratorio de proyectos en España - EFE emprende</title>
    <author>EFEemprende | Madrid</author>
  </titleStmt>
  <extent resp="RAE"><measure type="tokens">344</measure></extent>
  <sourceDesc> <p>https://www.efeemprende.com/noticia/falling-walls-lab-espana/</p></sourceDesc>
</fileDesc>
<profileDesc>
  <creation><date>2015</date></creation>
  <langUsage><language ident="spa">Español</language></langUsage>
  <textClass>
    <classCode scheme="http://corpussesiad.rae.es/doc/clasif#pol">Política, economía y justicia</classCode>
    <classCode scheme="http://corpussesiad.rae.es/doc/src#EFEEMP">Agencia EFE. Portal temático: EFE emprende</classCode>
  </textClass>
</profileDesc>
<xenoData>
  <meta xmlns="" property="article:tag" content="innovacion"/>
  <meta xmlns="" property="article:tag" content="investigacion"/><meta xmlns=""property="article:section" content="Actualidad"/>
</xenoData>
<revisionDesc> <change who="RAE" when="2019-03-29">Creación</change></revisionDesc>
</teiHeader>
```

Datos primarios

```
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:xi="http://www.w3.org/2001/XInclude"
id="ELPAISW2017ESspa_1499378823_732351"
source="https://elpais.com/internacional/2017/07/07/actualidad/1499378823_732351.html">
<xi:include href="ELPAISW2017ESspa_1499378823_732351.hdr"/>
<text>
<p class="titulo"><hi>Emerge un nuevo vídeo del accidente aéreo en San Francisco</hi></p>
<p class="subtitulo"><hi>La toma de 47 minutos muestra el momento del impacto y como actúan los servicios de
emergencia para rescatar a los pasajeros</hi></p>
<p class="autor">Sandro Pozzi</p>
<p class="data">Nueva York</p>
<p class="data">7 JUL 2017 - 20:11 <abbr title="Central European Summer Time">CEST</abbr></p>
<p>Hace justo cuatro años, un avión operado por la compañía Asiana Airlines se estrellaba contra la pista
del aeropuerto de San Francisco en el momento en el que tomaba tierra tras cruzar el Pacífico. Las primeras
imágenes, con la aeronave en llamas, hicieron temer lo peor. En el avión viajaban 292 pasajeros y 16
miembros de la tripulación. Ahora emerge un vídeo en el que muestra el momento en el que los equipos de
emergencia acuden a la zona del siniestro para rescatar al mayor número de personas posible. Hubo solo tres
víctimas mortales.</p>
</text>
</TEI>
```

2. Diseño y adquisición del corpus (1/2)

- Adquisición: Convenios de colaboración y textos con licencia CC
- Organización en colecciones según fuente (231.314/129.057.760):
  - EFE (231.314/83.601.678)
    - EFE hackatón (6.587/27.470.551)
    - Efe Agro (6.020/2.998.433), Efe Doc Análisis (11.337/5.564.749), Efe Emprende (8.949/3.654.851), Efe Empresas (38.873/14.243.776), Efe Escuela (1.654/715.548), Efe Estilo (4.631/2.928.494), Efe Futuro (14.430/6.541.983), Efe Motor (3.308/1.515.630), Efe Practico deporte (4.870/2.024.644), Efe Salud (6.318/3.876.912), Efe Tur (3.959/2.642.982), Efe Verde (20.765/9.331.283), Efeminista (114/91.842)
    - Hemeroteca (?/?)
  - Eldiario.es (1.870/1.605.769)
  - Elimparcial.es (46.929/27.228.355)
  - La Vanguardia (33.325/13.219.257)
  - La Voz de Galicia (17.375/3.402.701)
  - ? (?/?)

LA VANGUARDIA

eldiario.es

EFE: Agencia EFE

La Voz de Galicia

EL IMPARCIAL

3. Segmentación de oraciones, *tokens* y palabras

- Segmentación de oraciones ortotipográficas:
  - Utiliza información sobre caja tipográfica, abreviaciones y signos de puntuación
  - Basado en reglas (implementadas como transductores de estados finitos)
- Segmentación en dos niveles: *tokens*-palabras
  - Locuciones (aquellas que no admiten análisis sintagmático)
  - Formas verbales con enclíticos (‘*invitarle’* → *‘invitar’* + *‘le’*)
  - Tiempos compuestos (‘*ha terminado’* → *‘ha_terminado’*)
  - Comitativos (‘*conmigo’* → *‘con’* + *‘mí’*)
  - Contracciones (‘*del’* → *‘de’* + *‘el’*)
  - Otros: palabras guionadas,...
- Formato: ISO MAF
- Manual: *Esquema de segmentación de palabras en el CAPITEL*

4. Anotación morfosintáctica y lematización (1/2)

- Etiquetario morfosintáctico:
  - Neutralidad teórica, enfoque descriptivo
  - Compatibilidad y comparabilidad con otros de amplia difusión (UD, FreeLing)
  - Las etiquetas constan de categoría y rasgos gramaticales, referenciados al DCR
  - Se ha limitado/eliminado el uso de rasgos léxicos o particulares (p. ej. tipo adverbial)
- Lematización:
  - Se añade un lema (del DLE o forma de citación para *OOV*)
  - Lexicón de análisis morfosintáctico:
    - DLE++ + apreciativos + adv. en *'-mente'* + superlativos + prefijación + locuciones

4. Anotación morfosintáctica y lematización (2/2)

- Manual: *Desambiguación morfosintáctica del CAPITEL*
  - Contiene indicaciones prácticas y posibilistas más que finura lingüística
  - Se anota forma, no función (leísmo, cortesía, etc.)
  - Tiene en cuenta manuales de otros corpus
- Formato: ISO MAF
  - Textos incluyen etiquetas compactas desarrolladas para el CORPES
  - Hay previstas traducciones entre formatos (CORPES, UD, FS), en ocasiones con pérdidas

Segmentación, lematización y anotación morfosintáctica

```
<maf addressing="byte_offset">
<s id="s1" start="301" end="361">
<token from="301" to="307" class="tok" id="s1_t1">Emerge</token>
<token from="308" to="310" class="tok" id="s1_t2">un</token>
<token from="311" to="316" class="tok" id="s1_t3">nuevo</token>
<token from="317" to="323" class="tok" id="s1_t4">vídeo</token>
<token from="324" to="327" class="tok" id="s1_t5">del</token>
<token from="328" to="337" class="tok" id="s1_t6">accidente</token>
<token from="338" to="344" class="tok" id="s1_t7">aéreo</token>
<token from="345" to="347" class="tok" id="s1_t8">en</token>
<token from="348" to="351" class="tok" id="s1_t9">San</token>
<token from="352" to="361" class="tok" id="s1_t10">Francisco</token>
<wordForm tokens="s1_t1" lemma="emerger" tag="Vis-3p0n" form="Emerge" id="s1_w1"/>
<wordForm tokens="s1_t2" lemma="uno" tag="Qms---dn" form="un" id="s1_w2"/>
<wordForm tokens="s1_t3" lemma="nuevo" tag="Amsq--pn" form="nuevo" id="s1_w3"/>
<wordForm tokens="s1_t4" lemma="vídeo" tag="Nmsc---n" form="vídeo" id="s1_w4"/>
<wordForm tokens="s1_t5" lemma="de" tag="P------n" form="de" id="s1_w5"/>
<wordForm tokens="s1_t5" lemma="el" tag="Tms----n" form="el" id="s1_w6"/>
<wordForm tokens="s1_t6" lemma="accidente" tag="Nmsc---n" form="accidente" id="s1_w7"/>
<wordForm tokens="s1_t7" lemma="aéreo" tag="Amsq--pn" form="aéreo" id="s1_w8"/>
<wordForm tokens="s1_t8" lemma="en" tag="P------n" form="en" id="s1_w9"/>
<wordForm tokens="s1_t9 s1_t10" lemma="San Francisco" tag="N--p---f(#P)" form="San Francisco" id="s1_w10"/>
</s>...
```

5. Entidades nombradas

- Tipos: PER(SON), LOC(ATION), ORG(ANIZATION), OTH(ER)
- Clasificación contextual: ‘*Madrid’* puede ser PER, LOC o ORG
- Manual: *Esquema de anotación de entidades nombradas en el CAPITEL*

Entidades nombradas

```
<NEC>
  <NamedEntity id="s1_e1" type="LOC" wordForms="s1_w10">San Francisco</NamedEntity>
  <NamedEntity id="s3_e2" type="PER" wordForms="s3_w1">Sandro Pozzi</NamedEntity>
  <NamedEntity id="s4_e3" type="LOC" wordForms="s4_w1">Nueva York</NamedEntity>
  <NamedEntity id="s5_e4" type="OTH" wordForms="s5_w4">CEST</NamedEntity>
  <NamedEntity id="s6_e5" type="ORG" wordForms="s6_w12 s6_w13">Asiana Airlines</NamedEntity>
  <NamedEntity id="s6_e6" type="ORG" wordForms="s6_w13">Airlines</NamedEntity>
  <NamedEntity id="s6_e7" type="LOC" wordForms="s6_w23">San Francisco</NamedEntity>
  <NamedEntity id="s6_e8" type="LOC" wordForms="s6_w35">Pacífico</NamedEntity>
  <NamedEntity id="s13_e9" type="OTH" wordForms="s13_w12">YouTube</NamedEntity>
  <NamedEntity id="s14_e10" type="LOC" wordForms="s14_w13">Oeste</NamedEntity>
  <NamedEntity id="s15_e11" type="LOC" wordForms="s15_w10 s15_w11">Asiana Airlines</NamedEntity>
  <NamedEntity id="s16_e13" type="PER" wordForms="s16_w6">B777</NamedEntity>
  <NamedEntity id="s16_e14" type="LOC" wordForms="s16_w16">Boeing</NamedEntity>
  <NamedEntity id="s20_e15" type="LOC" wordForms="s20_w21">Corea del Sur</NamedEntity>
  <NamedEntity id="s22_e16" type="PER" wordForms="s22_w2">B777</NamedEntity>
</NEC>
```

6. Anotación sintáctica

- Sigue el modelo de *Universal Dependencies v2*
- Formato CONLL-U
- Manual: *Esquema de anotación sintáctica del CAPITEL*

Anotación sintáctica

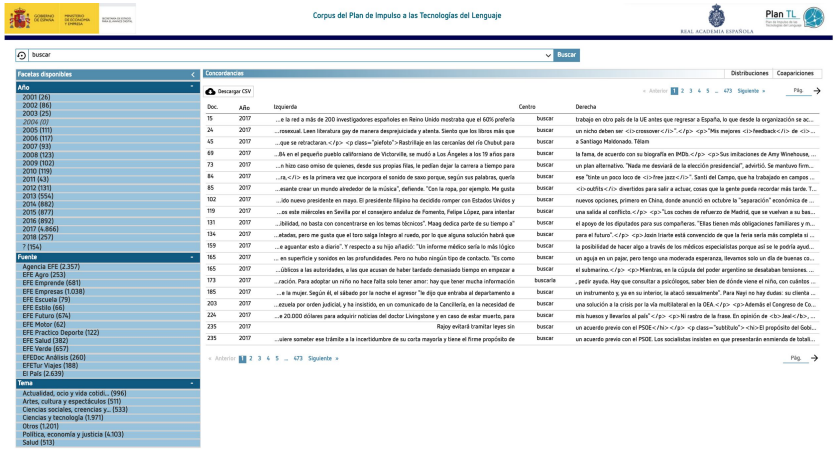
```
# Mario_Vargas_Llosa vuelve a ejercer como reportero en Israel
1	Mario_Vargas_Llosa	Mario_Vargas_Llosa	PROPN	_	_	4	nsubj	_	_
2	vuelve	volver	VERB	_	Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin	4	aux	_	_
3	a	a	ADP	_	_	4	mark	_	_
4	ejercer	ejercer	VERB	_	VerbForm=Inf	0	root	_	_
5	como	como	SCONJ	_	_	6	mark	_	_
6	reportero	reportero	NOUN	_	Gender=Masc|Number=Sing	4	obl	_	_
7	en	en	ADP	_	_	8	case	_	_
8	Israel	Israel	PROPN	_	_	4	obl	_	_

# LIBRO PARA CONMEMORAR LOS 50 AÑOS DE EL ESTADO
1	LIBRO	libro	NOUN	_	Gender=Masc|Number=Sing	0	root	_	_
2	PARA	para	ADP	_	_	3	mark	_	_
3	CONMEMORAR	conmemorar	VERB	_	VerbForm=Inf	1	nmod	_	_
4	LOS	el	DET	_	Gender=Masc|Number=Plur|PronType=Art	6	det	_	_
5	50	50	NUM	_	Gender=Masc|Number=Plur|NumType=Card	6	nummod	_	_
6	AÑOS	año	NOUN	_	Gender=Masc|Number=Plur	3	obj	_	_
7	DE	de	ADP	_	_	9	case	_	_
8	EL	el	DET	_	Gender=Masc|Number=Sing|PronType=Art	9	det	_	_
9	ESTADO	estado	NOUN	_	Gender=Masc|Number=Sing	6	nmod	_	_
```

7. Subcorpus revisado

- Anotación morfosintáctica, lematización y entidades nombradas:
  - PoS Tagging automático (ACC: 96-97 %) + NERC (ACC: 83 %)
  - Revisión manual (~1.000.000 p.)
  - Reentrenamiento del PoS Tagger (ACC: ?) y el NERC (ACC: ?)
- Anotación sintáctica:
  - Análisis sintáctico automático (UAS: ?, LAS: ?)
  - Revisión manual (160.000 p.)
  - Entrenamiento de un nuevo analizador (UAS: 86-87 %, LAS: 80-81 %)

8. Explotación de corpus

- Interfaz web
  - Búsquedas por forma, lema y categoría
  - Búsquedas facetadas
  - Concordancias
  - Distribuciones
  - Coapariciones
- API REST (swagger)



9. Documentación

1. Parámetros de equilibrio y representatividad del CAPITEL (2 pp.)
2. Esquema de codificación del CAPITEL (8 pp.)
3. Ejemplo de adquisición del CAPITEL: textos de la Agencia EFE (10 pp.)
4. Metodología y control de calidad del CAPITEL (2 pp.)
5. Esquema de segmentación de palabras del CAPITEL (18 pp.)
6. Esquema de anotación morfosintáctica del CAPITEL (14 pp. + apéndices)
7. Esquema de anotación de entidades nombradas del CAPITEL (36 pp.)
8. Guía de revisión manual de la anotación del CAPITEL (1 pp.)
9. Esquema de anotación sintáctica del CAPITEL (15 pp.)
10. Explotación del CAPITEL (2 pp.)
11. Licencias para la distribución de CAPITEL (x pp.)

¡FIN!