Diseño, compilacion y anotacion de corpus para estudios literarios computacionales.

Borja Navarro Colorado
Universidad de Alicante
borja@dlsi.ua.es

Humanidades Digitales. Del corpus a la interpretación: Estilometría con R
Curso de verano Universidad de Burgos

Septiembre 2021

Indice

1. Definiciones.
2. Diseño del corpus.
3. Compilación, marcado y anotación.
4. Estudio de caso: el corpus ELTeC.

De niciones

- Estudios literarios computacionales
  - Análisis a gran escala (*Distant reading* (Moretti 2007)).
  - Un análisis literario apropiado requiere un corpus bien diseñado y compilado.
- ¿Qué es un corpus?

>Que es un corpus?

A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to ***represent****, as far as possible, a language or language variety as a source of data for linguistic research.*[1]

[1] Sinclair, J. 2005. "Corpus and Text - Basic Principles" in *Developing Linguistic Corpora: a Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books: 1-16. Available online from `http://ota.ox.ac.uk/documents/creating/dlc/` [Accessed 2021-08-16].

Fases de creacion de un corpus

1. Especificación objetivos de investigación.
   - Documentación.
   - Especificar la población (*target domain*).
2. Diseño: criterios de selección.
3. Compilación: obtención de los textos y limpieza.
4. Marcado y anotación.
5. Evaluación.
6. Publicación.
7. Revisión, mejora y ampliación.

Diseño del corpus: la representatividad

Corpus: fuente de datos.

Para que el análisis sea válido, el corpus debe ser **representativo** del fenómeno o hecho literario que se quiere estudiar.

Representativeness refers to the extent to which a sample includes the full range of variability in a population (Biber 1993)

Diseño del corpus: la representatividad

Dos tipos de "representatividad":

- Representatividad del campo objeto de estudio (*target domain*).
- Representatividad lingüística (del fenómeno lingüístico).

Ejemplo

"Estudio de las formas verbales en estilo directo libre en la novela española del siglo XX."

Diseño del corpus: la representatividad

Corpus aleatorios vs. no aleatorios (Egbert 2019)

Diseño del corpus: la representatividad

Corpus aleatorios vs. no aleatorios (Egbert 2019)

- Aleatorio: selección totalmente **aleatoria** de los textos a partir de la totalidad de la población.
  - Permite hacer generalizaciones a partir de la muestra.

Diseño del corpus: la representatividad

Corpus aleatorios vs. no aleatorios (Egbert 2019)

- Aleatorio: selección totalmente **aleatoria** de los textos a partir de la totalidad de la población.
  - Permite hacer generalizaciones a partir de la muestra.
- No aleatorios: selección de textos según la conveniencia del estudio.
  - Las conclusiones no son generalizables más allá del corpus.
  - **Corpus balanceados**: los textos se seleccionan en función de determinadas categorías procurando que la cantidad de textos por categoría esté compensada.

Diseño del corpus: la representatividad

Corpus aleatorios vs. no aleatorios (Egbert 2019)

- Aleatorio: selección totalmente **aleatoria** de los textos a partir de la totalidad de la población.
  - Permite hacer generalizaciones a partir de la muestra.
- No aleatorios: selección de textos según la conveniencia del estudio.
  - Las conclusiones no son generalizables más allá del corpus.
  - **Corpus balanceados**: los textos se seleccionan en función de determinadas categorías procurando que la cantidad de textos por categoría esté compensada.

Todo corpus se situa entre estos dos polos.

Ambos son válidos.
Depende del objeto de estudio: un corpus puede ser representativo para una cuestión, pero no serlo en absoluto para otra.

Corpus balanceados: criterios

- Definir qué categorías deben estar bien representadas en el corpus.
- Establecer la cantidad necesaria / suficiente de textos por categorías para que la representación de cada una quede compensada.
- Las categorías dependen del objeto de estudio:
  - Género literario: lírica, drama, épica...
  - Idioma
  - Rasgos del autor: sexo, año de nacimiento, edad...
  - Edición de la obra: primera, última supervisada por el autor...
  - Periodo, fechas de publicación..
  - Temas, subgéneros...
  - Determinado rasgo literario que sea de interés: métrica, estilo, metáforas y tropos, símbolos, motivos...

Diseño del corpus: casos

Caso 1. Corpus representativo de la novela realista española del s. XIX

- B. Pérez Galdós: 80 novelas aprox.
- Leopoldo Alas: 2 novelas.

Diseño del corpus: casos

Caso 1. Corpus representativo de la novela realista española del s. XIX

- B. Pérez Galdós: 80 novelas aprox.
- Leopoldo Alas: 2 novelas.

Caso 2. Corpus representativo de la poesía del Siglo de Oro.

- Garcilaso de la Vega: 38 sonetos conocidos.
- Lope de Vega: 1382 sonetos aprox.

Diseño del corpus: tamaño

¿El tamaño importa?

- ¡Sí!: "Mo'better"
  - Las técnicas de análisis computacional (como las disponibles en *stylo()*) funcionan con frecuencias altas, necesitan muestras recurrentes.
  - Muestra amplia de textos para estar seguros que el fenómeno a estudiar está suficientemente representado en toda su variedad.
- Pero el tamaño no lo es todo: en un corpus amplio mal diseñado los errores (falta de representatividad) también se amplían.
- Un corpus puede ser "pequeño" (Egbert 2019):
  - Áreas ("domain") específicas (como el literario).
  - Fenómenos lingüísticos poco frecuentes.

Compilacion

Búsqueda, recolección y almacenamiento de los textos.

Existe una edicion digital de la obra

- Calidad del texto: solo fuentes fiables. A ser posible, edición crítica digital.
- Problemas legales: cuidado con obras o ediciones modernas con derechos de autor/editor.
- Limpieza del fichero. Os será muy útil saber algo de expresiones regulares.

Compilacion

Búsqueda, recolección y almacenamiento de los textos.

No se dispone de una edicion digital de la obra

- Digitalización del texto con OCR y corrección de errores:
  Transkribus Tesseract OCR4all Kraken eScriptorium
- ¿Modernización del texto? ¿Edición crítica digital?

Algunas fuentes ables

- Biblioteca Virtual Miguel de Cervantes
- Biblioteca Digital Hispánica
- Colección Clásicos hispánicos
- Internet Archive (?)
- Project Gutenberg
- Oxford Text Archive
- Y muchísimas más...

Compilacion

Durante todo el proceso, guardad los metadatos de las obras (hoja de cálculo):

- Título
- Autor
- URL de descarga y fecha
- Responsable de la edición digital (si se conoce)
- Edición impresas original
- etc.

Almacenamiento

- Directorio propio
- Modularidad: un fichero por obra.
- Nombres de fichero descriptivo. Evitad tildes y eñes.
- Formato simple. Extensión ".txt" o similar.
- Número de identificación

Marcado y anotacion

Los textos se pueden presentar en tres niveles según la información incluida:

- Texto puro ("plain text").
- Texto marcado: metadatos y estructura.
- Texto anotado: información lingüística, literaria, ecdótica, etc.

Marcado

Lenguajes de marcado: lenguaje formal para codificar un documento mediante etiquetas.
Tipos:

- Basado en SGML (*Standard Generalized Markup Language*)
  - Etiquetas `<...>`.
  - Ej. `<b>casa</b>`
  - Lenguajes: HTML y XML.

Marcado

Lenguajes de marcado: lenguaje formal para codificar un documento mediante etiquetas.
Tipos:

- Basado en SGML (*Standard Generalized Markup Language*)
  - Etiquetas `<...>`.
  - Ej. `<b>casa</b>`
  - Lenguajes: HTML y XML.
- No basados en SGML.
  - LaTeX, MarkDown, Wikitextos, etc.

Marcado

Los corpus se suelen marcar con el lenguaje XML.

- Permite definir etiquetas propias (DTD o Schema).
- Etiquetas simples:

```
<title>La Celestina</title>
```

- Etiquetas complejas (atributo - valor)

```
<verso type=‘‘endecasilabo’’>Un soneto me manda hacer
Violante</verso>
```

Marcado

Estándar TEI (*Text Encoding Initiative*)

`https://tei-c.org/`

Estructura general de un fichero TEI:

- Encabezado (`<teiHeader>`):

  Metadatos como título, autor, datos bibliográficos, codificación, historial de revisiones, etc.

- Cuerpo (`<text>`).

  Estructura de la obra: volúmenes, capítulos, párrafos, etc.
  Citas, versos, salto de página, notas, cambio de idioma, énfasis, etc.

Marcado

Estándar TEI (*Text Encoding Initiative*)

`https://tei-c.org/`

Estructura general de un fichero TEI:

- Encabezado (`<teiHeader>`):

  Metadatos como título, autor, datos bibliográficos, codificación, historial de revisiones, etc.

- Cuerpo (`<text>`).

  Estructura de la obra: volúmenes, capítulos, párrafos, etc.
  Citas, versos, salto de página, notas, cambio de idioma, énfasis, etc.

Más info:

`https://tei-c.org/Guidelines/P5/`

`https://tthub.io/aprende/`

`http://www.teibyexample.org/`

Anotacion

Cualquier tipo de información que se quiera hacer explícita en el texto:

- Literaria: personajes, estilo indirecto libre, referencias mitológicas, métrica, etc.
- Lingüística: categorías gramaticales, lemas, papeles semánticos, metáforas, etc.
- Ecdótica: testimonios, variantes, etc.
- etc.

Anotacion

- Etiquetas XML.

  Pero no todo estandarizado en TEI.
- Trabajo complejo y costoso que requiere anotación por pares para asegurar la consistencia de la anotación.
- Guía de anotación: documento donde se especifica y justifica qué anotar, con qué etiquetas, el proceso, qué hacer en casos complejos o dudosos, etc.
- Base de sistemas de aprendizaje automático (*Machine Learning*).

Anotacion - Ejemplo

```
<text>
 <body>
  <head>
   <title>-I-</title>
  </head>
  <lg type="cuarteto">
   <l n="1" met="---+---+-+-">Cuando me paro a contemplar mi estado,</l>
   <l n="2" met="-+-+---+-+-">y a ver los pasos por do me ha traído,</l>
   <l n="3" met="+--+--+--+-">hallo, según por do anduve perdido,</l>
   <l n="4" met="--++-+-+-+-">que a mayor mal pudiera haber llegado;</l>
  </lg> (...)
  <lg type="terceto">
   <l n="9" met="+--+---+-+-">Yo acabaré, que me entregué sin arte</l>
   <l n="10" met="---+-+---+-">a quien sabrá perderme y acabarme</l>
   <l n="11" met="+--+---+-+-">si ella quisiere, y aun sabrá querello;</l>
  </lg>
 </body>
</text>
```

https://github.com/bncolorado/CorpusSonetosSigloDeOro/blob/master/GarcilasoDeLaVega/GarcilasoDeLaVega_01.xml

Evaluacion

Se debe demostrar que el corpus está bien hecho.

- Representatividad:

  La colección de textos es una muestra representativa de la población (el campo de estudio).
  El fenómeno lingüístico o literario a estudiar está presente en variedad suficiente.

- Anotación:
  - Consistencia: ante un mismo texto, dos anotadores anotan lo mismo.
  - Acuerdo entre anotadores.

Publicacion

El corpus es para utilizarlo, no para esconderlo: ¡Compártelo!

- Alojar en algún repositorio que asegure mantenimiento, y permita su descarga y consulta:

  `https://github.com` (u otros GIT).
  `https://zenodo.org/`
  `https://teipublisher.com/index.html`
  `https://textgrid.de/en/web/guest/home`
  `http://gams.uni-graz.at/`

Publicacion

El corpus es para utilizarlo, no para esconderlo: ¡Compártelo!

- Alojar en algún repositorio que asegure mantenimiento, y permita su descarga y consulta:

  `https://github.com` (u otros GIT).
  `https://zenodo.org/`
  `https://teipublisher.com/index.html`
  `https://textgrid.de/en/web/guest/home`
  `http://gams.uni-graz.at/`

- Publicación científica: un buen corpus suele venir acompañado de una publicación científica de referencia:
  - Congresos sobre recurso como LREC
    `http://www.elra.info/en/lrec/`
  - Revistas sobre recursos como LRE journal
    `https://www.springer.com/journal/10579`
  - Revistas y congresos de Humanidades Digitales.

Recapitulacion

Aspectos a tener en cuenta en el diseño y compilación de corpus para estudios literarios computacionales:

- Criterios de selección: representatividad y balanceado.
- Fiabilidad de la fuente de los textos digitales. Necesidad de ediciones críticas digitales.
- Codificación de los textos.
- Saber usar expresiones regulares para limpieza de texto, corrección de errores OCR, modernización...
- Conocer XML y TEI para marcado y anotación. Guía de anotación y evaluación.
- Publicar en repositorios fiable y escribir un artículo de referencia.

Algunos ejemplos de corpus literarios literarios[2]

- Góngora *Soledades*, edición crítica de A. Rojas Castro:
  `https://github.com/arojascastro/soledades`
- Corpus de sonetos del Siglo de Oro (con anotación métrica):
  `https://github.com/bncolorado/CorpusSonetosSigloDeOro`
- DISCO: Diachronic Spanish Sonnet Corpus
  `https://github.com/pruizf/disco`
- Biblioteca Electrónica Textual del Teatro en Español (1868-1936)
  `https://github.com/GHEDI/BETTE`
- The CLiGS textbox (varios)
  `https://github.com/cligs/textbox`
- ELTeC corpus
  `https://github.com/COST-ELTeC`
- Más...
  `https://tthub.io/recursos/ejemplos-tei/`

[2] listos para descargar y procesar, sin contar bibliotecas, colecciones, bases de datos, etc.

Estudio de caso: el corpus ELTeC

ELTeC: *European Literary Text Collection*

- Corpus de novela europea (1840-1920)
- Actualmente en desarrollo. Primera versión inicios de 2022.

  En estos momentos, más de 1200 novelas en 17 idiomas y creciendo
- Proyecto *Distant Reading for European Literary History* (COST Action CA16204) 2017-2022.

Contexto

Grupo de trabajo 1 `https://www.distant-reading.net/wg-1/`

Carolin Odebrecht (leader)

Lou Burnard

Borja Navarro Colorado

Martina Scholger

Objetivo

... build a multilingual European Literary Text Collection (ELTeC), (...) containing around 2,500 full-text novels in at least 10 different languages, ***permitting to test methods and compare results across national traditions.***[a]

[a] Memorandum of Understanding

- Evitar otro corpus de novela del XIX como los que ya existen.
- El corpus debe permitir la comparación entre idiomas y tradiciones culturales.

Representatividad y criterios de seleccion

Corpus balanceado (sin selección aleatoria).

[3] https://distantreading.github.io/sampling_proposal.html

Representatividad y criterios de seleccion

Corpus balanceado (sin selección aleatoria).
Criterios de selección y balanceo:[3]

- Periodos: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)

[3] `https://distantreading.github.io/sampling_proposal.html`

Representatividad y criterios de seleccion

Corpus balanceado (sin selección aleatoria).
Criterios de selección y balanceo:[3]

- Periodos: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)
- Tamaño: al menos 20 %
  - *short* (10k⇀50k word tokens)
  - *medium* (50k⇀100k word tokens)
  - *long* (>100k word tokens)

[3] `https://distantreading.github.io/sampling_proposal.html`

Representatividad y criterios de seleccion

Corpus balanceado (sin selección aleatoria).
Criterios de selección y balanceo:[3]

- Periodos: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)
- Tamaño: al menos 20 %
  - *short* (10k∽50k word tokens)
  - *medium* (50k∽100k word tokens)
  - *long* (>100k word tokens)
- Sexo/género autor: 10 % ∽ 50 % mujeres.

[3] `https://distantreading.github.io/sampling_proposal.html`

Representatividad y criterios de seleccion

Corpus balanceado (sin selección aleatoria).
Criterios de selección y balanceo:[3]

- Periodos: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)
- Tamaño: al menos 20 %
  - *short* (10k∽50k word tokens)
  - *medium* (50k∽100k word tokens)
  - *long* (>100k word tokens)
- Sexo/género autor: 10 % ∽ 50 % mujeres.
- Cantidad de reimpresiones: al menos 30 % “high” y 30 % “low”.

[3] `https://distantreading.github.io/sampling_proposal.html`

Representatividad y criterios de seleccion

Otros criterios

- Prosa narrativa ficcional
- A ser posible, primera edición en libro entre 1840 y 1920.
- Publicado en Europa (geográfico).
- No traducciones: escrito en la lengua de la colección.
- Una novela por autor. Solo 9$\backsim$11 pueden estar representados por tres novelas.

Anotacion

Cada colección se organiza en tres niveles:[4]

- Nivel 0: *plain text*.
- Nivels 1: XML-TEI.
  - TEI Header:
    - ★ Autor, título, responsables.
    - ★ Fuente bibliográfica.
    - ★ Idiomas
    - ★ Criterios de selección.
  - Estructura.
  - Etiquetas en el texto para *code switching*, títulos, énfasis, versos, citas.
- Nivel 2: lemas y categorías gramaticales.

Ejemplo: Gómez de Avellaneda *Sab* 1841.

[4] https://distantreading.github.io/Schema/eltec-1.html

Estado actual

				AUTHORSHIP				LENGTH			TIME SLOT					REPRINT COUNT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99	1900-20	range	Frequent	Rare	E5C
cze	2021-04-09	100	5621667	88	12	62	6	43	49	8	12	21	39	28	27	1	19	80.00
deu	2021-04-09	100	12738842	67	33	35	9	20	37	43	25	25	25	25	0	48	46	96.92
eng	2021-04-09	100	12227703	49	51	70	10	27	27	46	21	22	31	26	10	32	68	100.00
fra	2021-04-09	100	8712219	66	34	58	10	32	38	30	25	25	25	25	0	44	56	101.54
hrv	2021-03-22	21	1440018	21	0	4	0	6	12	3	6	12	2	1	11	1	0	23.08
hun	2021-04-09	100	6948590	79	21	71	9	47	31	22	22	21	27	30	9	32	67	100.00
ita	2019-11-21	34	3328244	32	2	19	3	13	10	11	5	12	10	7	7	12	0	55.97
lav	2020-12-20	2	106045	2	0	2	0	0	2	0	0	0	1	1	1	0	1	21.54
lit	2020-08-20	25	636132	18	7	16	1	19	3	2	5	3	3	14	11	6	18	55.38
nor	2021-04-09	53	3432676	38	15	18	11	26	18	9	4	2	30	17	28	32	21	67.69
pol	2021-04-09	100	8500172	58	42	1	33	33	35	32	8	11	35	46	38	39	61	80.00
por	2021-04-09	100	6799385	83	17	73	9	40	41	19	13	37	19	31	24	26	60	94.62
rom	2021-04-09	95	5558961	74	16	54	9	46	31	18	4	17	25	49	45	24	71	80.00
slv	2021-04-09	100	5682120	89	11	26	5	53	39	8	2	13	36	49	47	48	52	78.46
spa	2021-04-09	84	7147890	67	17	45	5	31	28	25	17	17	25	25	8	42	42	92.31
srp	2021-04-09	90	3961405	82	8	35	11	56	32	2	2	12	38	38	36	32	58	73.68
swe	2021-04-09	58	4960085	29	28	18	8	16	24	18	15	3	20	20	17	17	41	76.92
ukr	2021-04-09	50	1840062	37	13	23	7	34	13	3	5	10	11	24	19	30	20	70.77

https://distantreading.github.io/ELTeC/index.html

ELTeC-SPA

Situación actual:

- 84 novelas (7147890 tokens).
- Selección balanceada:
  - Periodos: 17, 17, 25, 25.
  - Autores: 67 hombres y 17 mujeres.
  - Tamaño: 31 *short*, 28 *medium* y 25 *long*.
  - Reimpresiones: 42, 42.

Lista de novelas

ELTeC corpus - Descarga y consulta

- Versión desarrollo: `https://github.com/COST-ELTeC`
- Versión estable:
  - **Oficial:** `https://zenodo.org/communities/eltec`
  - TEIpublisher: `https://teipublisher.com/exist/apps/eltec/index.html`
  - GAMS: `http://glossa.uni-graz.at/context:eltec`
  - TextGRID (test): `https://dev.textgridrep.org/browse/3thgt.0`

¡UTILIZADLO!

Bibliografía

- Biber (1993) "Representativeness in corpus desing" *Literary and Linguistic Computing* 19, 219-241.
- Egbert, Jesse (2019) "Corpus Design and Representativeness" en Berber Sardinha, Tony y Veirano Pinto, Marcia *Multi-Dimensional Analysis*, Londres, Nueva York, Bloomsbury Academics.
- Moretti, Franco *La literatura vista desde lejos*. Barcelona: Marbot Ediciones, 2007.
- Borja Navarro Colorado, María Ribes Lafoz and Noelia Sánchez (2016) "Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation", *Proceedings of the 10th edition of the Language Resources and Evaluation Conference*, 23-28 May 2016, Portorož (Slovenia)
- Odebrecht, Carolin; Burnard, Lou; Navarro Colorado, Borja; Eder, Maciej; Schöch, Christof (2019) "The European Literary Text Collection (ELTeC)." *Digital Humanities Conference*, Utrecht.

Apendices

Algunos aspectos más...

\Electronic form" o el problema de la codi cacion.

La representación digital de cada carácter textual es un número binario de 7 u 8 dígitos:

ASCII	Carácter
1100001	a
1100010	b
1100011	c
1100100	d
etc.	

Otros códigos de caracteres: Latin1 o ISO 8859-1, ISO 8859-5 (cirílico), ISO 8859-6 (árabe), ISO 8859-7 (griego), etc ... y UNICODE.

\Electronic form" o el problema de la codi cacion.

La representación digital de cada carácter textual es un número binario de 7 u 8 dígitos:

ASCII	Carácter
1100001	a
1100010	b
1100011	c
1100100	d
etc.	

Otros códigos de caracteres: Latin1 o ISO 8859-1, ISO 8859-5 (cirílico), ISO 8859-6 (árabe), ISO 8859-7 (griego), etc ... y UNICODE.

Problema

En ocasiones la máquina no sabe qué código aplicar la abrir/procesar un texto.

En la medida de lo posible, **¡utilizad UTF-8 (UNICODE)!**

ELTeC-SPA

Fuentes de los textos:

- Biblioteca Virtual Miguel de Cervantes (Universidad de Alicante);
- CLIGS corpus (Universidad de Würzburg);
- Biblioteca digital hispánica.

ELTeC-SPA - Situacion actual

Problemas:

- Compilación: pocas novelas digitalizadas del periodo 1840-1859.
- Anotación: difícil encontrar los cambios de idioma en el texto

Tareas pendientes:

- ELTeC-CAT y ELTeC-EUS comienzan este mes, pero no hay nadie para desarrollar ELTeC-GLG (¿Algún voluntario?)
- Anotación del nivel 2.