The *European Literary Text Collection* (ELTeC)

A multilingual corpus of European novels (1840-1920)

Borja Navarro Colorado

University of Alicante

June 2021

Context

Working Group 1. COST Action Distant Reading for European Literary History (CA16204)

Carolin Odebrecht (leader)

Lou Burnard

Borja Navarro Colorado

Martina Scholger

Objectives

Objective

... 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

- Avoid "Yet Another 19th-Century Novels corpus"
- Allow comparison between languages and cultural traditions.

Representativeness and balance

Non random corpus, but balanced.

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

Representativeness and balance

Non random corpus, but balanced.
Balance criteria:[1]

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

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

Representativeness and balance

Non random corpus, but balanced.
Balance criteria:[1]

- Date: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)
- Length: at least 20 %
  - short (10k∽50k word tokens)
  - medium (50k∽100k word tokens)
  - long (>100k word tokens)

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

Representativeness and balance

Non random corpus, but balanced.
Balance criteria:[1]

- Date: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)
- Length: at least 20 %
  - short (10k∽50k word tokens)
  - medium (50k∽100k word tokens)
  - long (>100k word tokens)
- Author gender: 10 ∽ 50 % female authors.

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

Representativeness and balance

Non random corpus, but balanced.
Balance criteria:[1]

- Date: 1840 - 1920
  - 1840-1859 (T1)
  - 1860-1879 (T2)
  - 1880-1899 (T3)
  - 1900-1920 (T4)
- Length: at least 20 %
  - short (10k∽50k word tokens)
  - medium (50k∽100k word tokens)
  - long (>100k word tokens)
- Author gender: 10 ∽ 50 % female authors.
- Reprint count: at least 30 % "high" and 30 % "low".

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

Representativeness and balance

- Fictional prose narrative
- First edition as book between 1840 and 1920s.
- Published in a European country.
- No translations: originally written in the language of the collection.
- One novel per author. Only 9∽11 authors should be represented by three novels.

Annotation

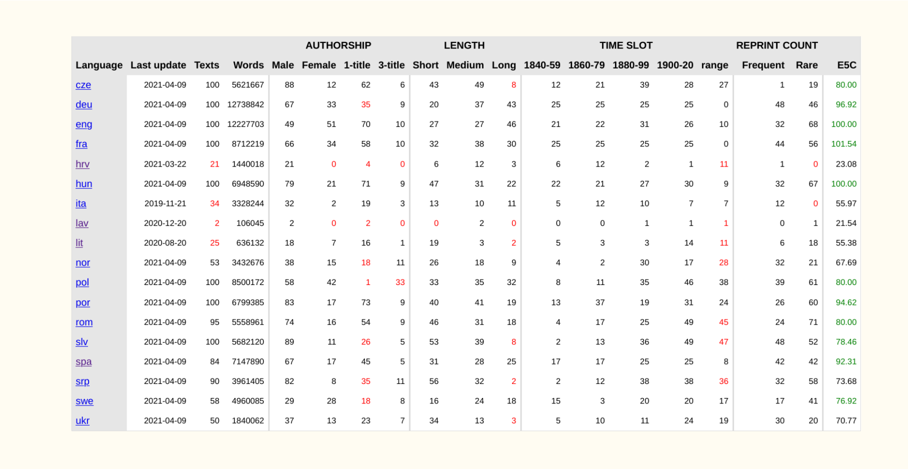
Each collection is organized in three levels:[2]

- Level 0: plain text.
- Level 1: XML-TEI.
  - TEI Header:
    - Author, title, extent.
    - Source
    - Profile: languages
    - Balance criteria
  - Structure.
  - Text tags as code switching, titles, “emphs”, verse lines or quotes.
- Level 2: Lemmas and PoS.

Example: Gómez de Avellaneda *Sab* 1841.

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

Current state



				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

Current situation:

- 84 novels (7147890 tokens).
- Balancing criteria satisfied:
    - Time slots: 17, 17, 25, 25.
    - Genre: 67 male and 17 female authors.
    - Size: 31 Short, 28 medium and 25 long.
    - Reprints: 42, 42.

List of novels

ELTeC-SPA

Text sources:

- Miguel de Cervantes Virtual Library (University of Alicante);
- CLIGS corpus (University of Würzburg), mainly for the two last periods;
- Hispanic Digital Library (Biblioteca digital hispánica) from the Spanish National Library (Biblioteca Nacional).

ELTeC-SPA - Development team

I have practically done all the selection, review and annotation work by myself, but I have had help from:

- Pilar Escobar, Gustavo Candela and other colleagues from the BVMC; as well as José Calvo from CLIGS, share with me novels annotated in XML-TEI.
- Lou Burnard provided me several scripts to transform between TEI formats: from BVMC or CLIGS format to ELTeC format.
- During the annotation process, I have had help from my students of the *Master of Literary Studies* at the University of Alicante.

Many thanks to all of them!

ELTeC-SPA - Problems

Compilation problems:

- It has been hard to find novels from the first period 1840-1859. There are few novels digitalized from this period.

Annotation problems:

- Perhaps the annotation of code switching:
  - sometimes it was difficult to find CS in the texts;
  - time consuming task.
- Other aspects were annotated with regular expressions and some patience.

Open issues

Languages: What about Catalan, Basque and Galician?

Open issues

Languages: What about Catalan, Basque and Galician?

- ELTeC-CAT: two novels under development:
  - Vayreda, Marià (1904) *La punyalada. Novela montanyenca*
  - Bosch de la Trinxeria, Carles (1889) *L'Heréu Noradell: estudi de família catalana*
- ELTeC-EUS :-(
- ELTeC-GLG :-(

Volunteers? Join the Action:

`https://www.distant-reading.net/about/participate/`

Open issues

Level 2 annotation:

- Part of Speech and lemma for each token.
- Universal dependencies.
- Automatic annotation with human validation of some fragments.

ELTeC corpus

- Development version: `https://github.com/COST-ELTeC`
- Stable versions:
  - Official `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`