



Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives

COLLECTION:
DIGITAL MODERN
LANGUAGES SECTION
LAUNCH ISSUE

ARTICLES – DIGITAL
MODERN LANGUAGES

CHRISTOF SCHÖCH

ROXANA PATRAS

TOMAŽ ERJAVEC

DIANA SANTOS

**Author affiliations can be found in the back matter of this article*



ABSTRACT

The aim of this contribution is to reflect on the process of building the multilingual *European Literary Text Collection* (ELTeC) that is being created in the framework of the networking project *Distant Reading for European Literary History* funded by COST (European Cooperation in Science and Technology). To provide some background, we briefly introduce the basic idea of ELTeC with a focus on the overall goals and intended usage scenarios. We then describe the collection composition principles that we have derived from the usage scenarios. In our discussion of the corpus-building process, we focus on collections of novels from four different literary traditions as components of ELTeC: French, Portuguese, Romanian, and Slovenian, selected from the more than twenty collections that are currently in preparation. For each collection, we describe some of the challenges we have encountered and the solutions developed while building ELTeC. In each case, the literary tradition, the history of the language, the current state of digitization of cultural heritage, the resources available locally, and the scholars' training level with regard to digitization and corpus building have been vastly different. How can we, in this context, hope to build comparable collections of novels that can usefully be integrated into a multilingual resource such as ELTeC and used in Distant Reading research? Based on our individual and collective experience with contributing to ELTeC, we end this contribution with some lessons learned regarding collaborative, multilingual corpus building.

CORRESPONDING AUTHOR:

Christof Schöch

Trier University, Germany

schoech@uni-trier.de

TO CITE THIS ARTICLE:

Schöch, C, Patras, R, Erjavec, T and Santos D 2021 Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, 2021(1): 25 pp. 1–19. DOI: <https://doi.org/10.3828/mlo.v0i0.364>

The creation, enrichment, curation, and publication of datasets for research is an essential element in Digital Humanities research. The existence of such machine-readable datasets is a precondition for the application of any computational method of analysis and for further investigation into the pertinence of quantitatively grounded concepts such as “Distant Reading”, “the great unread” or “data-rich literary history”.¹ However, the relationship between the composition and modelling of the dataset, on the one hand, and the methods of analysis that the dataset supports, on the other, is not always self-evident. In other words, the ways in which a given dataset has been designed directly influence the sorts of analyses that become possible and the kinds of literary concepts that may be addressed. Some datasets are designed for rather specific usage scenarios, while others aim to allow a much wider range of applications. As a consequence, it is important to reflect on the intended use-cases of a given dataset when setting out to create, enrich, and curate a dataset. This is true in Digital Humanities more generally, but also in the subfield of Digital Humanities that employs computational methods for the analysis of literary texts and which is sometimes called Distant Reading research.

Against this background, the aim of this contribution is to propose a reflection on the dataset design and creation process in the case of the *European Literary Text Collection* (ELTeC) that is being built in the framework of the COST Action *Distant Reading for European Literary History*. To provide some background, we briefly introduce the basic idea of ELTeC with a focus on the overall goals and the intended usage scenarios. We then describe the collection composition principles that we have derived from the usage scenarios. In our discussion of the corpus-building process, we focus on collections of novels from four different literary traditions as components of ELTeC: French, Portuguese, Romanian, and Slovenian, selected from more than twenty collections that are currently in preparation.² For each collection, we describe some of the challenges we have encountered and the solutions developed while building ELTeC. In each case, the literary tradition, the history of the language, the current state of digitization of cultural heritage, the resources available locally, and the scholars’ training level with regard to digitization and corpus building have been vastly different. How can we, in this context, hope to build comparable collections of novels that can usefully be integrated into a multilingual resource such as ELTeC and used in Distant Reading research? Based on our individual and collective experience with contributing to ELTeC, we end this contribution with some lessons learned vis-à-vis collaborative, multilingual corpus building.

In this way, we hope to contribute to a better understanding both of the relationship between corpus building and Distant Reading research and of the challenges inherent in multilingual corpus building.

THE EUROPEAN LITERARY TEXT COLLECTION (ELTEC)

While creating the *European Literary Text Collection* (ELTeC) is only one among several objectives of the COST Action *Distant Reading for European Literary History*, it is probably the project’s key output.³

DISTANT READING RESEARCH: USAGE SCENARIOS FOR ELTEC

ELTeC is intended to support a broad range of investigations into the multilingual tradition of the European novel using methods widely used in Distant Reading research, primarily computational methods of text analysis. However, ELTeC is also intended to support the development and evaluation of precisely these methods of research as well as to open new perspectives on how established notions in literary theory (such as periodization, authorship, genre, or canonization) might be reconceptualized given corpus-based evidence.

1 See Moretti, Bode, and Cohen, respectively.

2 The authors of this contribution are each among the editors of one of these collections.

3 This project is a pan-European networking initiative involving researchers from Computational Literary Studies, Computational Linguistics and Literary Theory and History. General information on the project can be found here: <https://www.distant-reading.net/>.

ELTeC is a multilingual collection of roughly comparable corpora each containing 100 novels from a given national (or rather: language-based) literary tradition. With this number of novels, we consider our corpora to be clearly useful for the usage scenarios outlined below. We focus on the novel because it was an important – probably the most important – literary genre from the mid-nineteenth to the early twentieth century in Europe (McKeon), with rich traditions in many European countries. Moreover, novels are usually quite long, resulting in a collection of considerable size, in terms of number of words. This strongly facilitates certain statistical approaches to textual data, like the ones typical of Distant Reading research that we intend to support. We focus on the period 1840 to 1920, mainly for pragmatic reasons: we begin in 1840, because starting at around this time, a number of novels that is sufficient for our purposes had been published in many European languages. And we end in 1920 because this allows us to focus exclusively on texts that are in the public domain.⁴ As a consequence, ELTeC can be shared freely and reused as widely as possible, without restrictions imposed by copyright law. We focus on novels from the European literary traditions because we are interested in the connections, mutual influences, and differences in development in the history of the novel, and expect that being able to analyse and compare a dense network of literary traditions will be relevant in this context. Our goal is to provide at least ten collections of 100 novels each.⁵

ELTeC is designed with several groups of users and application domains in mind, which have guided us when fixing the corpus composition criteria and encoding principles. While the usage scenarios below presuppose users with considerable technical expertise, we also plan to offer ELTeC collections pre-loaded in one or several web-based corpus-analysis systems (such as noSketchEngine, KonText, or Open Corpus Workbench) and in versions suitable for import into desktop corpus analysis tools (such as TXM or Antconc). Both approaches make the corpora available to users via friendly but powerful user interfaces for querying ELTeC.

On the one hand, we aim to provide ELTeC as a benchmarking corpus to researchers interested in evaluating existing methods of Distant Reading research, or in developing new methods and algorithms. ELTeC is meant to allow benchmarking studies across several languages in several domains: two examples are studies that evaluate algorithms and feature-sets for stylometric authorship attribution or measures of keyness.⁶ Another example is research aiming to develop methods to detect reported speech in narrative texts.⁷ Still other examples are the creation and analysis of character networks, and literary geographical mappings.⁸ There are also initiatives working on the ELTeC paratext such as titles, prefaces, and authors' notes.⁹ In most of these domains, evaluation studies using multilingual corpora exist, but they usually suffer from a lack of comparability between the corpora that limits the degree to which their results can be generalized. On the other hand, ELTeC is meant to allow the investigation of connections between, as well as shared or divergent developments in, several European literary traditions. For instance, ELTeC is intended to support investigations into mentions of names of places, authors, and publications across the various collections.

In order to facilitate such comparative investigations across several language-based collections, whether for benchmarking of methods or for questions of literary history, ELTeC requires all language-based collections to follow shared corpus-composition criteria. In addition, any document-level metadata or token-level linguistic annotations (part-of-speech and named entities, in particular) also need to be encoded in a comparable manner across collections. The next section provides additional details on these issues.

⁴ Since the 1993 Copyright Directive (Council Directive 93/98/EEC), the general rule in most European countries is that literary works remain in copyright for seventy years after the death of their author. We describe unusual cases below.

⁵ These core collections are being supplemented by extension collections providing additional novels from the same or an earlier time period. At the time of writing, the size of ELTeC stands at more than 1,300 novels in more than twenty collections. We provide an entry-point to information about ELTeC here: <https://www.distant-reading.net/eltec/>. See also Burnard et al.

⁶ Examples for such studies include Rybicki and Eder, and Evert et al. (for authorship attribution) as well as Schöch et al. "Burrows' Zeta" (for measures of keyness).

⁷ Examples for this area of research include Brunner, and Byszuk et al.

⁸ Such studies include Vala et al., Dekker et al., Jannidis et al., Santos and Freitas (character networks), and Piatto et al. (literary geography).

⁹ An example for such a study is Patras et al.

The guidelines for corpus building concern corpus composition as well as text encoding.¹⁰ The guidelines are meant to make sure that the resulting collection of corpora is indeed suitable for the range of use-cases we describe in the previous section.

With respect to corpus building, we distinguish eligibility criteria and corpus-composition criteria. The eligibility criteria offer a very simple, formal definition of the novel: for our purposes, a fictional narrative prose text of at least 10,000 words in length. And they specify the required time of first publication in a European country: 1840–1920. Novels also need to have originally been written in the language of the subcollection and must have been published in Europe during the period covered by ELTeC.

In the absence of exhaustive bibliographic records of novelistic production for most of the languages covered in ELTeC, no attempt at a randomly sampled, statistically representative corpus can reasonably be made. Instead, the corpus-composition criteria aim to ensure that the breadth and variety of novels produced during the period covered by ELTeC are well represented, while at the same time ensuring rough comparability across collections. We intend to achieve both aims by defining a number of relevant descriptive categories as well as criteria for their distribution in each collection. The categories retained are the following:

- The gender of the author, in three groups: female, male, mixed/diverse/undefined. Novels by female authors need to make up least 10%, ideally around 30%, with an upper bound of 50%, of the novels in each collection. In some literary traditions, such as in English, novels by female authors make up a significantly larger part of the overall production than 10%. In other traditions, however, the proportion of novels by female authors that we know of is clearly lower, for a number of reasons related to production, attribution practices, and conservation. However, in order to provide minimal support for comparisons between novels written by men and by women, 10% appears to be a lower bound. The upper bound ensures some level of comparability between collections in terms of author gender.
- The length of the novel in words, in three groups: short (10–50k words), medium (50–100k words), and long (more than 100k words). At least 20% of the novels should be short and 20% should be long. We realize that these three categories are, to some extent arbitrary. However, the principal goal of this criterion is to ensure that neither short, medium-long, or long novels dominate different collections to an excessive extent, something that would be a threat to the comparability between language collections.
- The time of publication of the novel, in four groups: 1840–1859, 1860–1879, 1880–1899, and 1900–1920. Each group of novels should comprise approximately one quarter of the collection. Defining precisely four time periods is an arbitrary choice, but it does support our goal of making sure that different time periods do not dominate different collections to an excessive extent, in a bid to ensure comparability between language collections.
- The reprint count of each novel, specifically in the period 1970–2010, as an operationalization of one aspect of canonicity, in two groups: low (0 or 1 reprints) and high (2 or more reprints). Neither group should be represented with less than 30% of the novels in a given collection, with an equal proportion deemed ideal. This criterion means that substantial numbers both of novels that can be considered part of our contemporary canon and of novels that have largely been forgotten today, are included in the corpus. Because digitization follows canonicity, demanding a certain balance in each collection leads to ELTeC containing many novels that have previously been unavailable in digital format. Further, this again supports comparisons between groups of novels with high and low reprint counts, both within a collection and comparability between individual collections.¹¹

¹⁰ The guidelines have been developed jointly by a large number of Action members in the Working Group devoted to “Scholarly Resources” and are laid out in detail in a white paper published by the group (WG Scholarly Resources, *Sampling Criteria for the ELTeC*).

¹¹ As far as possible, we also aim to respect the overall proportions specified by the different criteria within each time period.

- Finally, we aim to include 9–11 authors represented with three novels, while the remainder of the novels should have been written by a different author each. This criterion helps support the evaluation of stylometric methods of authorship attribution. For these studies, a certain number of authors represented with no less than three novels is a requirement, and 9–11 appears to be the absolute minimum. At the same time, requiring the remaining novels to be written by as many different authors as possible supports our goal of representing the largest possible variety of novels produced in the targeted time period.

The text encoding principles have also been developed jointly by members of the Action. They concern encoding of metadata as well as encoding of the text body and follow the Guidelines of the Text Encoding Initiative (TEI Consortium).

Indeed, the corpus-composition criteria described so far not only form the basis for the composition of the different collections, they are of course also documented, along with other metadata, in each individual text file, in a standardized manner. In fact, at least four kinds of metadata can be distinguished: (a) basic metadata serving to identify the work, such as its author and title, including authority file identifiers; (b) metadata describing each novel's characteristics in terms of the composition criteria mentioned above; (c) information on the provenance of the digital text, notably the digital source and the print edition it is based on as well as the first edition of the novel, as far as this information is known; (d) characteristics of the language of the text, like the alphabet used or whether the orthography has been modernized.

The project-specific text encoding principles fall into three closely connected levels, each of which is designed for a specific purpose and is documented in a machine-readable schema: Level 0 is meant as the target format for automatic conversions to XML-TEI when relatively poor input formats are available. Level 1 is somewhat richer in encoding, with textual phenomena such as foreign words, emphasis, or inserted quotations marked-up semantically. Level 2, finally, is designed for a version of our texts with linguistic annotation.¹² The intended application domains, the corpus composition criteria, and the encoding principles for text and metadata form a relatively flexible but nevertheless rather challenging set of constraints and represent high expectations with regard to availability of texts and metadata collection. For each of the collections we describe in the following sections, this has resulted in somewhat different challenges, depending on the literary tradition, the history of the language, the current state of digitization of cultural heritage, the resources available locally, and the level of training of the scholars in each of the countries involved.

BUILDING INDIVIDUAL COLLECTIONS

In the following, we describe the process of building several individual ELTeC collections to illustrate the corpus-building process as well as the specific challenges that each collection involved.

THE FRENCH COLLECTION (ELTEC-FRA)

The state of digitization for French literary texts for the period that concerns us is relatively comfortable, with respect to the availability of texts, when compared to most other European languages except German or English. As a consequence, the challenges for finding relevant digital texts when creating the French contribution to ELTeC are relatively minor, though certainly not absent.¹³ The multiplicity of platforms that offer French novels in digital formats is quite large. This is not a problem in itself—quite the contrary, especially as there is also an online search across multiple (but not all) platforms.¹⁴ However, it does create a number of difficulties both for corpus composition and for text encoding. In terms of corpus composition, the limited amounts of useful metadata that are available is a challenge, particularly for discovery of non-canonized texts. In terms of text encoding, the multiplicity of different platforms that offer digital texts and the many different formats that these platforms have on offer are the key challenge, with the added issue of the very variable quality of the available full text.

¹² For more details, see WG Scholarly Resources, *Encoding Guidelines for the ELTeC*.

¹³ For ELTeC-fra, see Schöch and Burnard and <https://github.com/COST-ELTeC/ELTeC-fra>.

¹⁴ See <http://noslivres.net/>.

Apropos corpus composition, the lack of rich and reliable metadata has been the major issue. Generally speaking, none of the available catalogues features information on more than one of our eligibility criteria (novels of a minimum length first published between 1840 and 1920) and key corpus composition criteria (author gender, length of the novels, year of publication, and reprint count). For example, the *Gallica* catalogue of the Bibliothèque nationale de France does not include any information about genre, so that it is hard to identify novels outside the usual suspects of canonized authors and outside those novels that have a subtitle making the genre explicit. Similarly, there is usually no indication of text length (beyond page numbers, which cannot be included as a search criterion), reprint count, and author gender. Conversely, *Ebooks libres et gratuits* does allow to search specifically for novels; however, this platform, surprisingly, includes neither data about the year of first publication nor, less surprisingly, about author gender, text length, or reprint count. Generally, this platform does not offer advanced search facilities. Not being able to search for any novel from a certain time period, or for novels specifically written by female authors, obviously makes it a lot more difficult to identify non-canonized novels and novels by female writers for inclusion into ELTeC. As a result of this limited discoverability of novels, the distribution of novels in terms of several metadata categories that we were able to achieve is not entirely balanced (*Figure 1*).

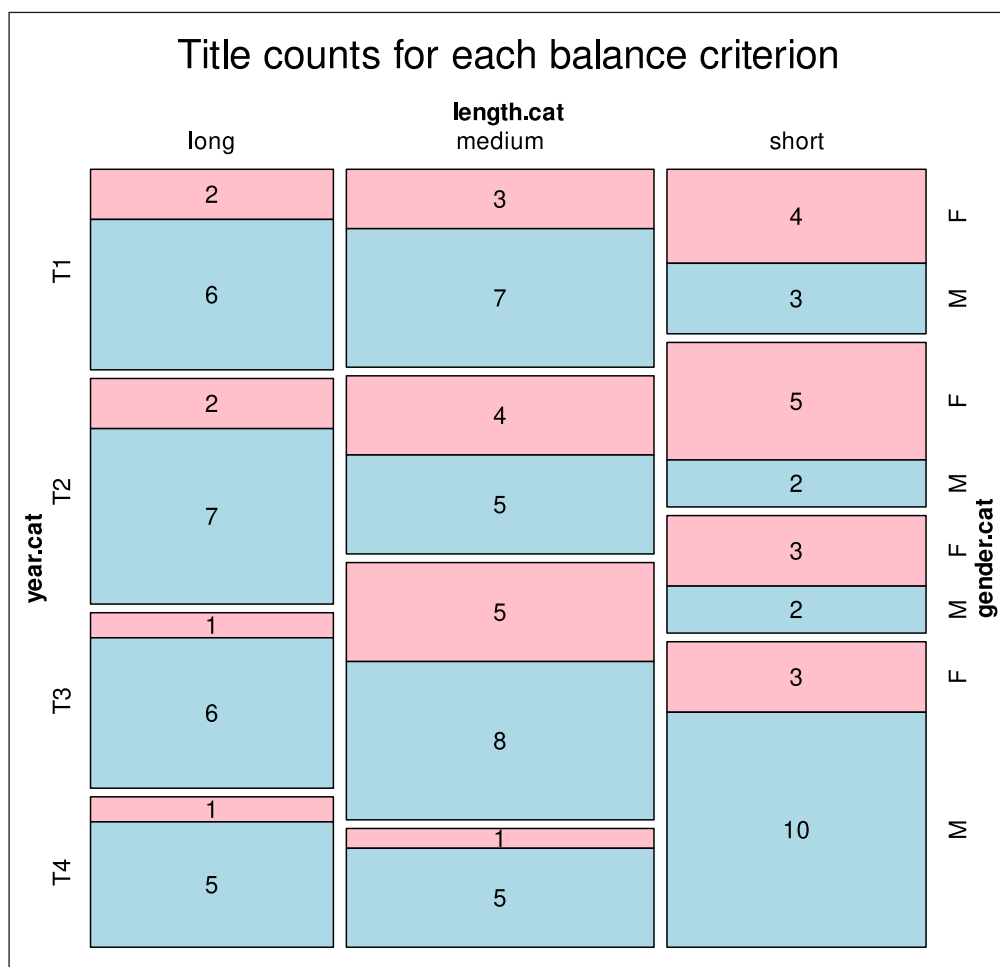


Figure 1 Mosaic plot showing the distribution of novels in terms of several metadata categories: time period, author gender, and novel length, for ELTeC-fra (24 April 2021). Image credit: Lou Burnard and Carolin Odebrecht, 2021. Source: <https://distantreading.github.io/ELTeC/fra/mosaic.svg>.

In addition, the fact that very few of these platforms use a standard data format such as XML-TEI does create problems. In some cases, the conversion process to XML-TEI according to ELTeC standards was easy, because the texts from one of the original sources had already been transformed to XML-TEI and published by the CLiGS group in the “romandixneuf” collection.¹⁵ The texts currently included in ELTeC-fra come mainly from the sources shown in *Table 1*:

¹⁵ The Computational Literary Genre Stylistics (CLiGS) group was active from 2014 to 2020 at Würzburg University, Germany. The “romandixneuf” collection contains 365 French novels first published in the nineteenth century. See Schöch et al. “The CLiGS Textbox”, and <https://github.com/cligs/romandixneuf> or DOI: <https://doi.org/10.5281/zenodo.4279128>.

PLATFORM	ORIGINAL DATA FORMATS	NUMBER OF NOVELS INCLUDED IN ELTEC-FRA
Ebooks libres et gratuits, including Bibliothèque électronique du Québec	EPUB (always) as well as some others; usually impeccable text quality	44
Project Gutenberg	EPUB (almost always), with very good text quality	12
Wikisource	EPUB (whenever the text quality is sufficient)	11
<i>Gallica</i> (Bibliothèque nationale de France)	PDF (always), TXT (usually, often imperfect OCR), EPUB (sometimes), Daisy-XML (very rarely)	9

Table 1 Sources for ELTeC-fra.

Sources of a smaller number of texts include *Bibebook* (EPUB), the *Bibliothèque numérique romande* and *Éfèle* (who provide texts in BML, the Book Markup Language). For each of these sources, the editors created a workflow to facilitate the conversion to XML-TEI. However, even with such pipelines, there remains the need for a substantial amount of manual checking and improvements. This is true particularly because in order to respect the corpus-composition criteria as well as we could, we were not always able to select texts depending on the most convenient format, but often chose to include texts for which no rich data format (such as XML or even EPUB) was available. In several cases, we chose texts for which only error-ridden texts produced with substandard OCR methods or based on low-quality scans were available (for instance, Jehan Gilbert's *Vers le pôle en aéroplane* from 1912 or Auguste Maquet's *La Maison du baigneur* from 1857). On the upside, this also means that ELTeC-fra does include a substantial number of texts that have not been available in XML-TEI before as well a fair number of texts that we make available in a reliable transcription for the first time. In some cases, relevant novels could not be included because of copyright issues, as in the case of *Monsieur Vénus* by Rachilde (Marguerite Eymery), which although first published in 1884, is still under copyright because the author published it when quite young and died only in 1953, less than 70 years ago.

A particularly challenging metadata item is the reprint-count criterion, our operationalization of canonicity as part of the corpus-composition criteria. The necessary data have been retrieved using a dedicated Python script that scrapes Worldcat, the union catalogue run by the Online Computer Library Center (OCLC). Because OCLC rejected our application to access the database via their API, and because the web-accessible data cannot be filtered reliably, some manual counter-checks are necessary. In this way, the reprint count of each novel in each year since its publication has been established and the reprint-count class (high or low) could be determined. In order to achieve a good balance with respect to this criterion, a number of largely forgotten and poorly digitized texts had to be manually curated (the two novels mentioned above, by Gilbert and Maquet, being examples of this).

Finally, in terms of data formats, ELTeC-fra is also available in the linguistically annotated format defined within the COST Action (Level 2 encoding). However, beyond the XML-TEI and plain-text versions available from GitHub, we do provide a version of ELTeC-fra that has been annotated using TreeTagger with the modern French model and converted to the TXM corpus format for analysis using the TXM corpus analysis tool.¹⁶

THE PORTUGUESE COLLECTION (ELTEC-POR)

When starting the project, we were appalled to see how problematic the situation was for public-domain Portuguese literature, as far as full-text availability was concerned, but also as to the absence of reliable metadata for most collections.¹⁷ Therefore, the creation of the Portuguese ELTeC also included detective work and a lot of opportunistic (not necessarily mainly literarily motivated) decisions. One of the reasons for the situation may be that Portuguese literature and texts span many centuries and genres, and that institutions like the National Library of Portugal have several mandates and tasks. Moreover, their digitization efforts have often focused on important (canonical) authors, or on thematic issues, preferring depth over breadth. The same can be said of other digital humanities projects in Portugal, such as the edition of Fernando

¹⁶ The TXM version is available from Zenodo: <https://doi.org/10.5281/zenodo.3939542>.

¹⁷ The Portuguese collection is a joint effort by Diana Santos, Raquel Amaro, Isabel Araújo Branco, and Paulo da Silva Pereira. For details, see Santos, "Portuguese Novel Collection (ELTeC-por, v2.0.0)" and <https://github.com/COST-ELTeC/ELTeC-por>.

Pessoa's writings by Universidade de Coimbra,¹⁸ to mention one such project. As to recovering or reassessing forgotten authors and works, there have been some projects on, for example, Portuguese women writers, both nationally and internationally,¹⁹ but as far as we could find out, they have so far dealt only with metadata. In addition, some of them provide their results only in a printed version (Lousada and Cantarin). This means that to base the choice on such materials would still leave us with the problem of finding the works themselves. This turns out to be non-trivial: out of an initial suggestion of works by women writers by Paulo da Silva Pereira, almost half of them could not be found in physical form in any Portuguese library.

So, we had to follow the principle to include, within ELTeC-por, all available works in digital form, before seeking out additional physical works to bring into the collection, adhering to the limitation of one to three works per author. **Table 2** shows the final choice. For most works we had to painstakingly revise the OCR, and some of it was of very low quality.

PLATFORM	ORIGINAL DATA FORMAT	NUMBER OF NOVELS INCLUDED IN ELTEC-POR
Gutenberg project	corrected OCR, original orthography	33
Internet archive	uncorrected OCR, very low quality, original orthography	29
National Library of Portugal	uncorrected OCR, rather low quality, original orthography	17
Luso livros	full edited text, modernized orthography	11
Projecto Adamastor	full edited text, modernized orthography	2
Other sources	uncorrected OCR, very low quality, original orthography	8

Table 2 shows that the collection includes two kinds of works: those keeping the original orthography, and those that have been modernized. In reality there are even more orthographies, depending on the time of printing and on the date of modernization. This may be a problem for some studies but we had no other choice than to use what existed, hoping that, for literary purposes, the analyses would not be too dependent on philological issues.

Table 2 Sources for ELTeC-por.

As for the books we chose to digitize from the holdings of the National Library of Portugal, we were provided with a list of all works published in the period 1840–1920. Unfortunately, the metadata did not include the genre, which meant that we had to base our selection on the books that had the word *romance* (novel) in the subtitle (incidentally, two of them turned out to be too short to be used in ELTeC). It should probably be mentioned that, given that Portuguese literary scholarship traditionally distinguishes between *conto*, *novela*, and *romance* (Moniz and Paz), we used six *novelas* in ELTeC-por, given that the denomination *novel* in English seems to cover both.

Admittedly, it was not easy to respect the composition criteria that had been agreed in common by the COST Action. In particular, we only managed to get seventeen works by fifteen women, and only eighteen long novels. The longest novel we prepared, *Gomes Freire*, written by Rocha Martins and published in 1900, could not be included because it only falls into public domain in 2022. For the first period (1840–1859) we could only make use of twelve works. Measuring reprint count was hard, too, given that the information in Worldcat was very deficient for Portuguese.²⁰ Furthermore, it was tricky to find enough canonical authors in order to include thirty canonical novels.²¹ The overview according to the distribution features can be seen in **Figure 2** below.

¹⁸ More specifically, those that are part of *Livro do desassossego*, <https://ldod.uc.pt/>.

¹⁹ See Osório and Esteves; Silva; and *Women Writers in Portuguese Before 1900* at <http://www.escriptoras-em-portugues.eu/>.

²⁰ For example, fifteen texts were simply missing from the catalogue.

²¹ We remind the reader that only nine authors were represented with three novels, and not all of them were chosen to be canonical, which means that ELTeC-por's canonical novels correspond to fourteen different authors. It would not be hard to find more than thirty canonical novels from 1840–1920, but these novels correspond to a small number of canonical authors who wrote many works.

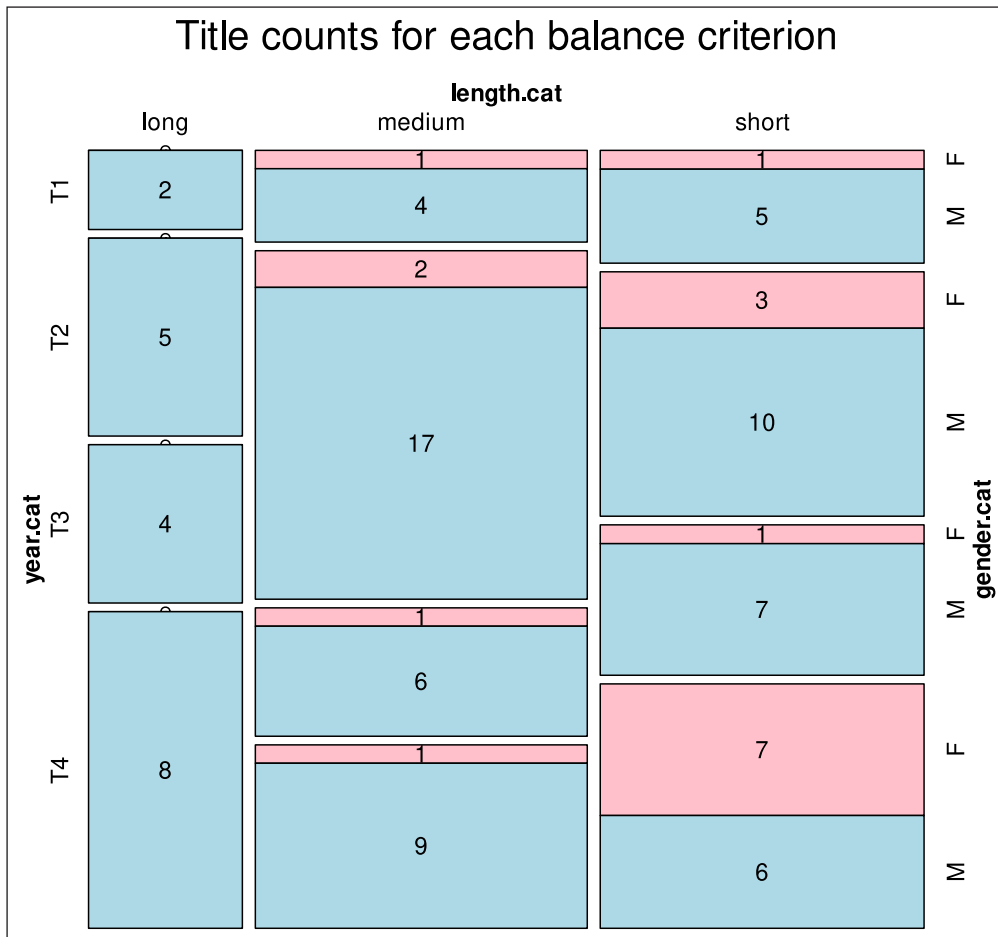


Figure 2 Mosaic plot showing the distribution of novels in terms of several metadata categories: time period, author gender, and novel length, for ELTeC-por (24 April 2021). Image credit: Lou Burnard and Carolin Odebrecht, 2021. Source: <https://distantreading.github.io/ELTeC/por/mosaic.svg>.

Just by considering this collection of 100 novels, some preliminary observations about Portuguese literary history can already be made:

- It appears that female authors had a tendency to write short or medium-length novels: it was not possible to find a single long novel written by a woman, and in fact most of the novels by Portuguese female authors are short.
- The historical novel (Marinho, Chaves) was very much in favour in the nineteenth century in Portugal: in the ELTeC-por collection, thirty-nine cases are historical novels, and they cover the whole “national” history (from pre-Roman times to the late nineteenth century).
- There are novels whose plot takes place in Africa and India, but by far the most frequent non-Portuguese main locations are in Brazil. Also common are novels relating the fates of people who went to Brazil to make a fortune and came back.²² As to the collection itself, in order to balance its size, most canonical authors appear with long novels (although they have also written medium or short ones); and, in order to minimize revision work, it is mainly canonical authors who feature three works (six out of nine authors).

The processing pipeline was as follows: first, the revised text would be included in the NOBRE corpus available from Linguateca, and it would be automatically transformed into the TEI XML format from there, followed by manual editing to fill in the TEI header, to reproduce the title page, and to guarantee that notes, poetry, and foreign citations were in place. NOBRE is a corpus of Portuguese literary works that did not appear in Vercial (a corpus of Portuguese canonical works), and it is part of Literateca, the subset of the AC/DC project corpora that contains literary texts.²³ In this manner, we ensured that the work became immediately available to the public, at the same time as increasing the material that is used for linguistic and NLP studies of

²² Traditionally called “Brazilians” in Portuguese literature, see Santos, “Portuguese Language Identity in the World: Adventures and Misadventures of an International Language”.

²³ See Santos, “Literature Studies in Literateca: Between Digital Humanities and Corpus Linguistics” for an early introduction to Literateca, and Santos, “Corpora at Linguateca: Vision and Roads Taken” for AC/DC.

Portuguese. All the works included in ELTeC-por are parsed and semantically analysed,²⁴ and can be queried from Literateca.²⁵ Although part of a larger corpus—Literateca—as explained above, they can also be selected as the ELTeC subset. By working this way, we were able to find problems in the ELTeC-por material that we subsequently corrected. And we were able to do further exploration studies in a larger collection, as illustrated in Santos et al.

In March 2021, we added named-entity annotation to the collection using PALAVRAS-NER (Bick) and converted the output to Level 2 format. In addition to “traditional” named entities (such as people and place-names), demonyms, professions, and titles of works were also marked, following the pilot project done on multilingual named-entity annotation of the ELTeC collection, described in Stankovic et al.

THE ROMANIAN COLLECTION (ELTEC-ROM)

Back in 2017, when *Distant Reading for European Literary History* started, the Romanian offer to contribute to ELTeC did not seem very promising. As both the Action’s timespan and its focus on networking rather than research did not seem enough to cope with so many challenges, any attempt at addressing them appeared an impossible task on account of the scarcity of literary texts that met the criteria for inclusion, as well as the lack of digitized texts available in Romanian. The most apparent challenges could be said to fall in the following categories: 1. *infrastructure* (underdeveloped library services and lack of librarian training); 2. *digitization policies* (scarce open access resources and available formats); 3. *text processing* (OCR and POS tagging suboptimal performance on diachronic varieties of Romanian); 4. *editing* (data on book-length available in page numbers but not in word count); 5. *cultural and literary tradition* (imbalance of the four time slots; the percentage of female-authored novels) and, last but not least, 6. *the Romanian team’s low training level and lack of experience* with XML, EPUB, and/or HTML formats, with collaborative work on platforms such as GitHub, with TEI markup and annotation in general, and with editing software (Patras et al.).

First, the policies concerning the digitization of national cultural heritage as well as the existing infrastructure and library services—for instance, none of the larger libraries provided scanning-on-request—would not support the creation of a fiction collection according to strict composition criteria. It was obvious that the few existing Romanian digitization initiatives had previously been focused mainly on non-fiction: press, public, and private archives, and even a few historical literary pieces proved that historians had been served better than literary scholars.²⁶ As far as Romanian literature was concerned, it was quite apparent that canonical rather than non-canonical novels and authors were chosen to be digitized first.

Second, data from the table of 13,726 scanned books that have kindly been made available by the Metropolitan Library of Bucharest demonstrate that from a potential ELTeC list (roughly 110 novels) only thirty-seven had been scanned. Setting aside the fact that some of them were hard to track down because library metadata did not make explicit the entries titled as “complete works”, “prose pieces”, and so on, a score of approximately 33% might have sounded like a good start for an enthusiastic team. Yet, from the thirty-seven items, almost all books issued before 1900 were printed in either a transition alphabet (Latin letters mixed with Cyrillic letters in the same word) or in a Latin-extended variety containing at least ten characters that are no longer used in contemporary standardized Romanian (e.g. “ș”, “é”, “ê”, “à”, “ó”, “ç”, “đ”, “ı”, “ü”, “Đ”, “É”, cedilla “ș” and “ț”), a fact that was evaluated as a potential hurdle for OCR.

Third, even if some digital libraries—such as Biblioteca Digitala BCU Cluj,²⁷ which currently has 3,019 items labelled as “literature”—could possibly grow and diversify their entries over time, it was obvious that the national digitization agenda would not keep up with the project’s timetable and milestones. All in all, except for a few PDF files, the Romanian team could not tap into any of the usual “fountains”: Wikisource, EPUBs, ebooks, or larger digitization projects.

24 The PALAVRAS parser (Bick, “PALAVRAS, a Constraint Grammar-Based Parsing System for Portuguese”) was used for syntactical annotation, and Linguatca tools were used for semantics.

25 See <https://www.linguatca.pt/aceso/corpus.php?corpus=LITERATECA>.

26 See <http://digibuc.ro/> and <http://digitool.bibnat.ro>.

27 See <http://dspace.bcucluj.ro/>.

While the aforementioned status of digital resources and appropriate tools was definitely challenging, a careful evaluation of the Romanian literary context and of the so-called emergent literary traditions in South, Eastern, and Central Europe (Cornis-Pope and Neubauer) flared our hopes that for the timespan 1840–1920, the ELTeC composition criteria—as described above—might be easier to comply with in Romania’s case (Munteanu, Drace-Francis, Metzeltin). Such an assumption was supported not only by various literary histories and cultural studies (most recently on the “factories” of “national writers” and “national geniuses” (see Tudurachi and Thiesse), but also by a very rough testing of ELTeC’s composition criteria against generally accepted notions of periodization, canonicity, literary history, genre, style, and authorship. More specifically, metadata provided by *Dicționarul cronologic al romanului românesc* (Istrate et al.) indicated a convenient overlapping of Romanian traditional periodization and ELTeC periodization principles, and a comfortable distribution of Romanian female novelists between 1840 and 1920.²⁸ Consequently, challenges #4 and #5, concerning digital editing and local literary tradition, did not seem insurmountable but somehow negotiable according to each novel’s particular “story”. In most cases, chiefly for popular fiction items—as well as for others, labelled by Romanian literary historians and critics as “aesthetically faulty” pieces and thus neglected—*princeps* editions had to be digitized. This implied, of course, establishing a set of editing principles and deciding, in some of the borderline cases, whether a work published in several volumes represents several novels or actually only one (see *Mysterele căsătoriei* by C.D. Aricescu), or if a novel is really a translation or a second version engendered through co-authorship (see *Haiducul* by Bucura Dumbravă). For copyright reasons, we did not include the youth novels of the most important Romanian novelist. Indeed, if we could have chosen among Mihail Sadoveanu’s *Șoimii* (1904), *Floare ofilită* (1906), *Însemnările lui Neculai Manea* (1907), *Apa morților* (1911), and *Neamul Șoimăreștilor* (1915), then our selection of authors of three novels could have served as a useful ground not only for stylometric investigation on authorship attribution but also for reconsidering the relation between canonicity and quantity in general.

We therefore decided to adopt “a collector’s strategy”, to scan 77% of the items from the provisional list and to find solutions for proper conversion of PDF files into editable formats for each particular case. At the same time, two of the project members contacted the heads of several libraries to arrange protocols of collaboration and ensure, where sites were no longer able to ensure maintenance (for instance, *dacoromanica*), the transfer of archives. As, in 2017–2018, not many of the Romanian libraries manifested a keen interest in digitizing nineteenth-century Romanian literature, we decided to create an ad hoc library by uploading PDF files to a Zenodo community and by assigning them DOIs.²⁹ Currently, 135 items are available there, and in the near future we plan to reorganize them in an ELTeC repository on the site of the Digital Humanities Laboratory (UAIC).

As already mentioned, between 1840 and 1920, there were at least three varieties of Romanian in print: a. *transition alphabet* (1840–1865); b. *Latin-extended alphabet* (1865–1900/1910); c. *quasi-contemporary alphabet* (1910–). While the first two varieties called for special treatment, the editions issued after 1900/1910 as well as the scholarly editions (mostly normalized) did not seem to raise problems. For novels printed in the transition alphabet and which had no other subsequent edition, we trained a HTR model on Transkribus available as RTA2 (*Romanian Transition Alphabet*), which performed very well, so that we managed to transcribe eight items automatically, among which three are already included in the Romanian collection.³⁰

The novels printed in the Latin-extended alphabet were converted with Abby FineReader 15, which proved to be a time-consuming solution: the output was so untidy that our volunteer students needed manually to add all ten characters for the “extended” Latin set. The positive part of this tiresome operation consisted in the fact that word about ELTeC has spread and that, since the summer session 2020, several bachelor’s and master’s dissertations in Romanian literature (supervised by professors of the “G. Ibrăileanu” Chair of “Alexandru Ioan Cuza” University of Iași) have been built on the ELTeC Romanian collection and rely on the digitization experience that students gained while contributing to it.

28 More consideration on the total number of Romanian novels issued between 1840 and 1920, on volume and instalment publications, and other concerns can be found in the materials used for the Parthenos workshop in Sofia; see <https://www.clarin.eu/event/2019/parthenos-workshop-cee-countries>.

29 See https://zenodo.org/communities/romanian_novel_library.

30 See Kahle et al. and <https://transkribus.eu/>. The RTA2 model was curated by Roxana Patras (April 2019).

Recently, the Romanian collection of ELTeC has reached 100 novels encoded at Level (see [Figure 3](#)).³¹ In order to evaluate whether Level 2 encoding is feasible in this particular case, we have tested four samples with UDPipe, a tool for the linguistic annotation of texts that is able to perform tokenization, lemmatization, part-of-speech tagging, and dependency parsing for multiple languages (see Straka). The output will be available with release 2.0.0 of ELTeC-rom on GitHub.

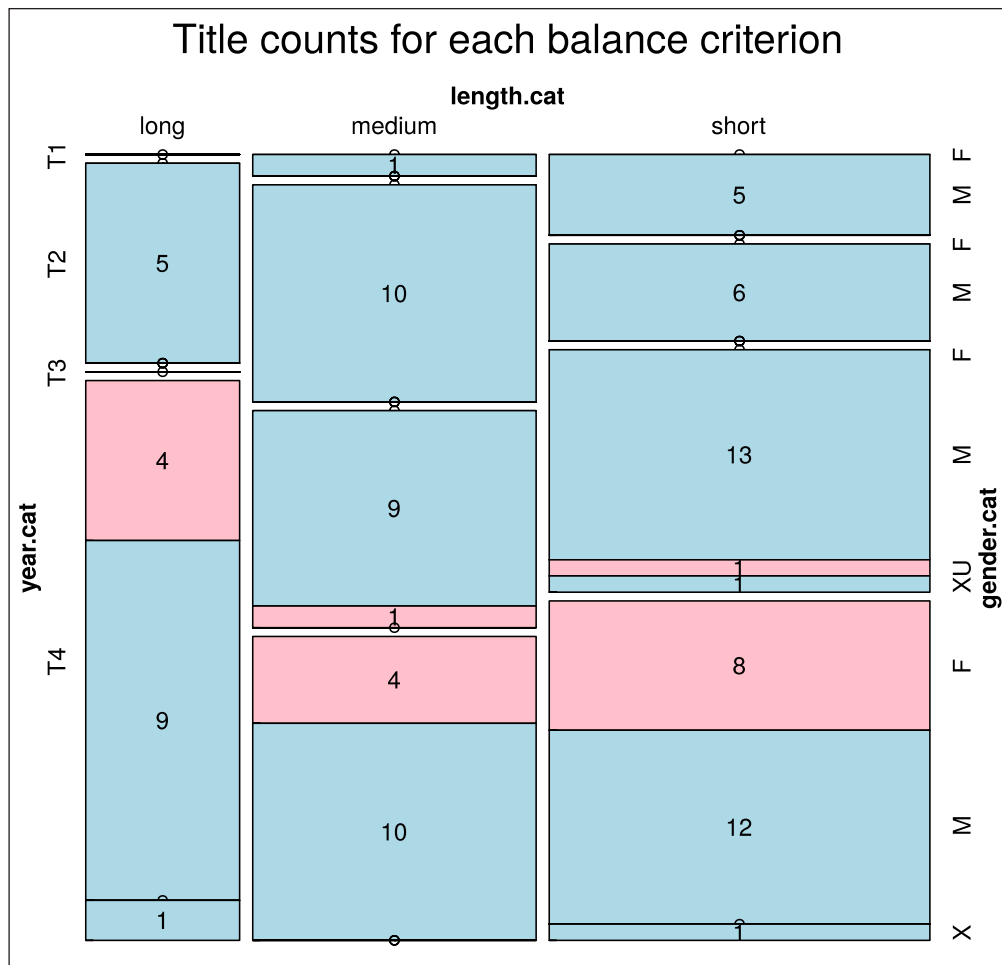


Figure 3 Mosaic plot showing the distribution of novels in terms of several metadata categories: time period, author gender, and novel length, in ELTeC-rom (20 October 2021). Image credit: Lou Burnard and Carolin Odebrecht, 2021. Source: <https://distantreading.github.io/ELTeC/rom/mosaic.svg>.

THE SLOVENIAN COLLECTION (ELTEC-SLV)

The Slovenian ELTeC collection is almost exclusively based on the novels available in the *Slovenian Literary Classics* project on WikiSource.³² This initiative, whereby all out-of-copyright Slovenian literature would be made available on Wikisource, was initiated over fifteen years ago by Miran Hladnik from the Slovenian language department of Ljubljana University. Each year, the Ministry of Science, Education, and Sport provides funds to the department that are used for student work to proofread the OCR of selected works, these typically being already available as facsimile PDF and OCRed HTML in the dLib digital library of the National and University Library of Slovenia.³³

In 2015, 500 of the books then available on Wikisource had already been converted to TEI in the scope of the EU IMPACT project, where they had been carefully structured, page-aligned with their facsimile, linguistically annotated, and compiled into the so-called IMP digital library and corpus, which is available from the CLARIN.SI repository.³⁴ The first phase of compiling the Slovenian ELTeC corpus involved selecting the 100 novels that would meet, as far as possible, the ELTeC composition criteria. This turned out to be a difficult process, for several reasons.

³¹ For ELTeC-rom, see Patras and <https://github.com/COST-ELTeC/ELTeC-rom>.

³² See https://sl.wikisource.org/wiki/Wikivir:Slovenska_leposlovna_klasika.

³³ See <http://www.dlib.si/>.

³⁴ See <http://nl.ijs.si/imp/index-en.html>, Erjavec, “The IMP Historical Slovene Language Resources” and Erjavec, *Digital Library and Corpus of Historical Slovene IMP 1.1*.

First, most older Slovenian novels are rather short, so there was a general dearth of novels in the “long” category. Second, Slovenian novels from the earlier periods are quite scarce, with those written by female writers especially so. Third, the Wikisource project, naturally, concentrates on key Slovenian books, so there was a lack of non-canonical novels available.

The final selection for the ELTeC corpus was a compromise between availability and the reality of published novels from Slovenian writers with their distribution by source and format (see *Table 3*).³⁵

PLATFORM	DATA FORMAT	NUMBER OF NOVELS INCLUDED IN ELTEC-SLV
IMP digital library	TEI P5, in richer encoding than required by ELTeC	65
Previously existing in “Slovenian literary Classics” at Wikisource	(Idiosyncratic) Markdown format	29
Proofread and made available in “Slovenian literary Classics” at Wikisource in the scope of ELTeC	(Idiosyncratic) Markdown format	5
eZISS digital library ³⁵	TEI P3, in richer encoding than required by ELTeC	1

The overview according to the distribution of features can be seen in *Figure 4*.

Table 3 Sources for ELTeC-slv.

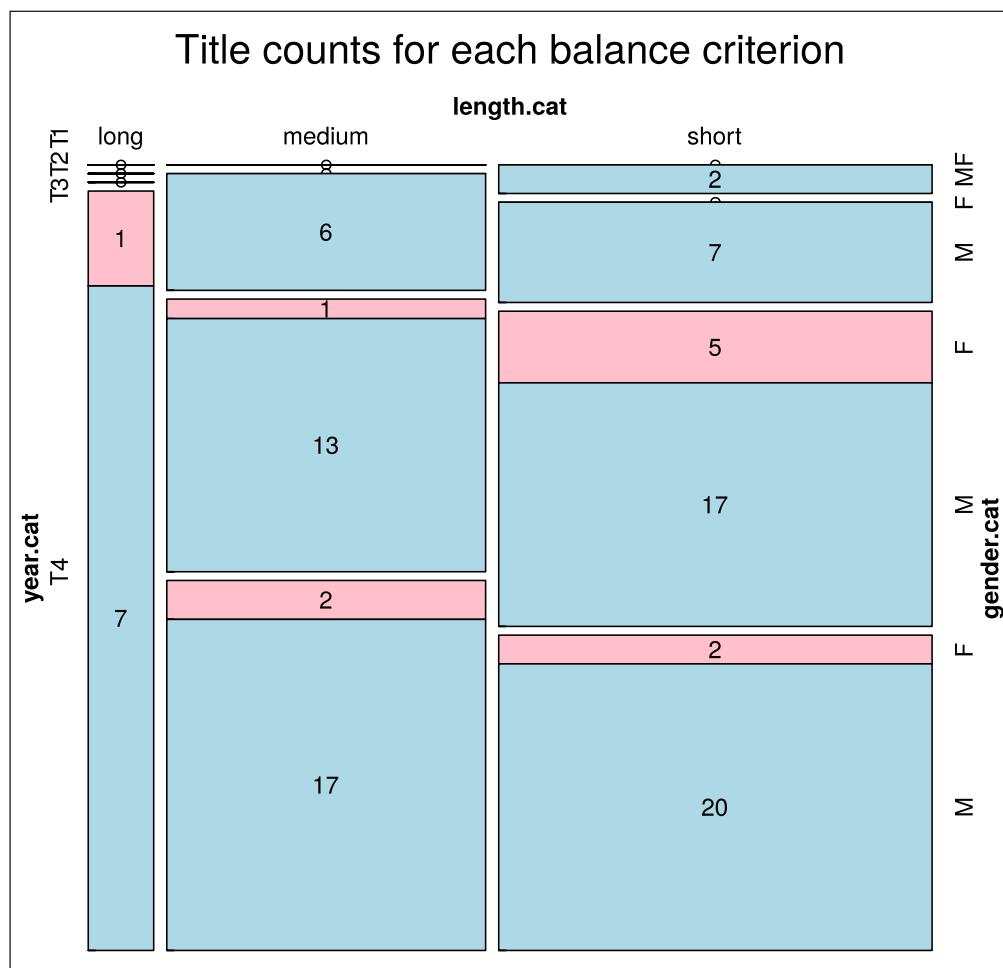


Figure 4 Mosaic plot showing the distribution of Slovenian ELTeC novels in terms of several metadata categories: time period, author gender, and novel length (24 April 2021). Image credit: Lou Burnard and Carolin Odebrecht, 2021. Source: <https://distantreading.github.io/ELTeC/slv/mosaic.svg>.

The list of selected novels was compiled in a spreadsheet, in which the metadata for each work were also entered: its ELTeC identifier (i.e. filename), title, author, year of first publication, size, canonicity, date, and the URL from where the novel can be retrieved (all are open source). Additionally, we made a table giving the VIAF and Wikipedia link, where these exist, for each author.

The conversion to ELTeC encoding proceeded for each of the three sources separately, but always included the following steps: 1. download and rename the novel; 2. convert the source encoding to the ELTeC TEI; 3. add novel and author metadata from the tables. Conversion

³⁵ See <http://nl.ijs.si/e-zrc/index-en.html>.

from the novels that were previously already encoded in TEI was performed with dedicated and fairly simple XSLT scripts; it mostly consisted of stripping out encoding and content not relevant for ELTeC, and some renaming of elements. More challenging was the conversion of Wikisource novels, as these are encoded in a Wikipedia-specific flavour of Markdown, as well as using various mark-up conventions idiosyncratic to the Slovenian literature collection. Here the conversion first fixed some character-level issues (e.g. removing soft hyphens, changing two successive commas to double quotation marks), removed excess text (e.g. the Wikisource metadata block) and formatting (e.g. hanging capitals), and transformed the input into standard Markdown. We then used the TEI Markdown to TEI XSLT stylesheet to convert the novels into “generic” TEI, and, finally, a dedicated XSLT script to convert this to ELTeC TEI.

The 100 Slovenian ELTeC novels are stored, as with other ELTeC languages, on the GitHub ELTeC repository, under the ELTeC-slv project. We also tried to include the complete conversion pipeline, including the metadata tables and download scripts, into the Git project. So, given the required locally installed programs (wget, Perl, Java, Saxon) anyone can recreate (and possibly modify) the Slovenian ELTeC Level 1 collection.³⁶ We have also produced a pipeline with which we have linguistically annotated the corpus, thus arriving at the Slovenian ELTeC Level 2 collection. For tokenisation and sentence segmentation we used the rule-based ReLDI tokeniser.³⁷ This tokeniser takes into account some specifics of the Slovenian language—in particular its abbreviations—and produces output compatible with the CoNLL-U tabular format, used in the Universal Dependencies project, a framework for linguistic annotation that is consistent and comparable across multiple languages.³⁸

In the second step, we modernize the spelling of the words in each novel, for which we use the character-based statistical machine translation tool cSTMTiser (Scherrer and Ljubešić) trained on the goo300k manually annotated corpus of historical Slovenian (Erjavec) for its translation model, and on the literary portion of the Gigafida reference corpus of contemporary standard Slovenian for its target language model.³⁹ The CSMTiser output is integrated into the CoNLL-U format, but with the original tokens (where they differ from the modernized ones) stored in the column with local features, while the modernized word replaces the original token. This enables further processing steps, which had been trained on contemporary Slovenian, to use the modernized words, thus leading to much better annotation results.

For part-of-speech tagging and lemmatization we used CLASSLA-StanfordNLP,⁴⁰ a branch of the well-known StanfordNLP library.⁴¹ As opposed to StanfordNLP, the CLASSLA-StanfordNLP strand introduces certain extensions that result in better quality annotation, such as using an external dictionary while performing lemmatization. The part-of-speech and morphological feature annotations are in the Universal Dependencies formalism for Slovenian (Dobrovoljc et al.), as well as containing the morphosyntactic descriptions from the MULTEXT-East schema for Slovenian.⁴² The corpus was also lemmatized, which is important for Slovenian, as it is a highly inflecting language.

The corpus was then annotated for named entities, using the Janes-NER tool,⁴³ which, given its training data for Slovenian, classifies named entities into persons, adjectives derived from person names, locations, organizations, and “other”. Finally, the CoNLL-U files with all the annotations were merged with the original TEI files, with the annotations encoded as specified by the ELTeC Level 2 schema, e.g. `<w xml:id="SLV20001.w274" pos="ADJ" msd="Case=Nom | Degree=Cmp | Gender=Fem | Number=Sing | ModernForm=večja | XPOS=Agcfsn" lemma="velik">veča</w>`. The Level 2 version of the ELTeC-slv corpus has also been converted to vertical format and mounted on the CLARIN.SI concordancers,⁴⁴ so that it is available for linguistic analysis.

36 For ELTeC-slv, see Erjavec et al. and <https://github.com/COST-ELTeC/ELTeC-slv>.

37 See <https://github.com/clarinsi/reldi-tokeniser>.

38 See <https://universaldependencies.org/>.

39 On Gigafida, see <https://viri.cjvt.si/gigafida/>.

40 See Ljubešić and Dobrovoljc and <https://github.com/clarinsi/classla-stanfordnlp>.

41 See Qi et al. and <https://stanfordnlp.github.io/stanfordnlp/>.

42 See Erjavec, “MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages” and <http://nl.ijs.si/ME/>.

43 See <https://www.github.com/clarinsi/janes-ner>.

44 See https://www.clarin.si/noske/run.cgi/corp_info?corpname=eltec_slv&struct_attr_stats=1.

Based on discussion of the challenges of pulling together several individual ELTeC collections, we can now reflect on the major shared challenges encountered when building our ELTeC collections, as well as on some lessons learned in collaborative, multilingual, multicultural corpus building.

Regarding corpus building, it has become increasingly clear that equally strict adherence to all corpus-composition principles in all collections is almost impossible to achieve. On the one hand, the corpus-composition criteria have turned out to be very demanding relative to the actual production in many literary traditions during the period 1840–1920. On the other hand, there remains a tension between representativeness, composition principles, and comparability of collections. For example, in the case of the French collection, it proved very difficult to identify long novels written by women in the period 1900–1920. As for the Slovenian collection, no long novels from the period 1840–1859 could be identified.⁴⁵ In many cases, these difficulties reflect a real rarity in the production of such novels; in others, it may be the result of a coincidence of digitization and discovery and/or an effect of processes of canonization or judgements of value more generally. Such cases, of which we encountered several, raise the question of whether or not it is acceptable to under-represent one category of novels in a given time period, if this reflects a probable historical reality, even if it negatively impacts comparability across collections. Within the Working Group creating ELTeC, a strategy to deal with differing levels of compliance with the corpus-composition criteria has been developed as a reaction to these difficulties. This strategy aims to make the level of compliance with the corpus-composition principles transparent by quantifying, for each ELTeC collection, the level of compliance with each corpus-composition criterion and by summarizing the results in an easily interpretable numerical score.⁴⁶

More generally, it has become clear that the corpus-composition criteria are a double-edged sword. On the one hand, they provide a definite incentive to avoid repeating the biases encountered in many corpus-building endeavours—we have made sure, for instance, that we include a certain minimum proportion of novels by female authors and of non-canonical novels, and the fact that we have undertaken this effort makes ELTeC unique. The higher the bars are raised for this (in terms of the minimum percentages of novels from a given, usually under-represented group), the stronger this effect will be. On the other hand, respecting the corpus-composition criteria is clearly more challenging for some languages than for others. Strict criteria favour better-resourced languages (such as English, German, and French), where finding 100 novels respecting all criteria is possible; it disadvantages lesser-resourced languages, where more challenges present themselves to reach the full size of 100 novels while including, for example, a sufficiently high proportion of novels by female authors and of low canonicity. This, in turn, means the lesser-resourced languages, which again are precisely what the COST Action aims to support and which make ELTeC unique, are penalized by the composition criteria. This “diversity paradox” also results in a conflict of targets: fostering the inclusion of collections of novels in lesser-resourced languages in ELTeC as a whole, while also fostering the inclusion of marginalized categories of novels in each collection within ELTeC. This is exacerbated by a third target, namely that of maintaining the comparability of the collections.

As a consequence of the diverse historical practices encountered in each language tradition, comparability between corpora will necessarily remain imperfect. There are several additional factors interfering with comparability, like the fact that we do not account for narrative perspective (e.g. first-person vs. third-person narrative) or for literary subgenres (such as adventure novels, crime fiction novels, or *Bildungsromane*) in our composition criteria. In terms of these factors, our collections cannot be comparable, strictly speaking, if only because neither the meaning of individual subgenre categories, depending on their sources, nor their function in each literary system, can be assumed to be equivalent in the different literary traditions ELTeC covers. However, we do attempt to document these factors in each novel’s metadata, so that analyses can account for them post hoc. Just as importantly, several factors—such as

⁴⁵ In addition, it should be kept in mind that creating classes of novels based on fixed word count limits is clearly a simplification, given differences between languages with regard to both language typology and cultural context.

⁴⁶ See the “ESC” column in the ELTeC overview table: <https://distantreading.github.io/ELTeC/>.

the shared eligibility criteria, identical metadata available for each text, shared text-encoding principles, and a shared linguistic annotation scheme—do create considerable potential for cross-linguistic analyses even in the absence of strict comparability. In addition, based on the metadata, more strictly comparable subcorpora can be created depending on the selection of languages that are considered for a given cross-linguistic analysis. This contributes to a certain level of comparability and is crucial for facilitating cross-linguistic analyses. Again, rather than exerting excessive pressure towards uniformity, the solution we have adopted is to set precise but flexible guidelines and to foster the largest degree of transparency we possibly can.

In terms of the text encoding, we have consistently been dealing with the difficult balance between a simple encoding scheme and the many requirements and possible phenomena one might want to encode. Given the multiplicity of formats we have encountered when collecting the novels, providing our collections in one shared, standardized, and expressive format has certainly proven its value and justified the considerable effort required. We started out with the idea to keep it very simple, and “less is more” has remained one of the principles when making encoding-related decisions. However, the demands from various Action participants and the requirements deriving from the materials as well as from the use-cases we envision, have led to a noticeable increase of complexity of the encoding scheme. For the most part, though, this has been limited to the *teiHeader* rather than the body of the text itself. Again, our three-level encoding scheme (described above) offers flexibility without reducing clarity and guidance.

A challenge that remains, and on which intense work is in progress within the COST Action at the time of writing, is to provide a shared and comparable set of linguistic annotations (particularly with regard to tokenization, lemmatization, part-of-speech tagging, and named-entity recognition) across all collections in a way that does not compromise the accuracy of the annotations with regard to the particularities of each language.⁴⁷ As mentioned, the Slovenian and the Portuguese collections have already made substantial progress in this direction. In addition to these fundamental layers of annotation, work is also in progress on sentiment analysis of the novels as well as the automatic identification and annotation of direct speech in a wide range of languages (Byszuk et al.).

Finally, ELTeC certainly fulfils the role of preserving and making available a specific part of the European cultural heritage. More importantly, and uniquely, however, ELTeC also functions like an *interface* for a very engaged and growing community of users. Matters such as multilingual usage, recalibration of focus according to broader or more narrow research agendas, and collection distribution and accessibility are issues that will continue to occupy us in the future. The latter issue, in particular, is not solved entirely by the open access policy we have adopted in the project because—in the absence of informed guidance on translation or transliteration—collections in lesser-resourced languages and collections lacking linguistic annotation (Level 2) may in some cases be excluded from some of the cross-linguistic explorations we seek to facilitate.

All of these challenges are to be expected in a corpus-building endeavour spanning such a large variety of literary traditions, linguistic practices, and scholarly communities, especially given that COST Actions, as networking projects, rely first and foremost on the intrinsic motivation and voluntary commitment of their members. The key question is how to provide reliable guidelines in the context of progressive collection building and how to foster their understanding and acceptance across a rather large group of participants. On the one hand, clear and stable guidelines based on a shared understanding of the aims and possibilities of the project are required for efficient collection building. On the other, our understanding of the implications of these guidelines in the context of each literary tradition constantly evolves as we are building the collections, in many cases producing a desire to revise the guidelines. Taking legitimate concerns and useful improvements into account must be balanced against the need not to invalidate work that has already been done, create additional workload, or question the overall objectives of the project, requiring a constant process of communication and negotiation, but also of training.⁴⁸

⁴⁷ This work is happening in the Working Group on “Methods and Tools”; see e.g. Cinková et al. and Stankovic et al.

⁴⁸ For this latter aspect, the Action’s Training Schools and Short-Term Scientific Mission programme have been instrumental.


Ultimately, we consider that ELTeC has been a challenge and a learning opportunity for everyone involved in creating it. We hope to have shown that corpus design is a scholarly endeavour in its own right, shaped by scholarly goals and dependent on shared methods and principles. We believe that ELTeC is already becoming a gathering point for a community of collection builders and users, both within and beyond our project. And we hope that it is opening up multiple perspectives for research in its own right, despite what is, compared to the temporal scope and breadth of production, a collection of limited size. However, ELTeC also serves to pave the way to curating larger sets of works in each language, so that Distant Reading at a larger scale, involving very significant parts of literary traditions, or even entire literary traditions, will one day be possible. Therefore, it is important to regard ELTeC as a seed and an inspiration for much larger endeavours that will form the future of Distant Reading research in Europe.

ACKNOWLEDGEMENTS

The COST Action *Distant Reading for European Literary History* (CA16204) is funded through the COST Association in the framework of the Horizon 2020 Framework Programme of the EU.

AUTHOR AFFILIATIONS

Christof Schöch  orcid.org/0000-0002-4557-2753
Trier University, DE

Roxana Patras  orcid.org/0000-0003-2534-8663
Universitatea Alexandru Ioan Cuza, RO

Tomaž Erjavec  orcid.org/0000-0002-1560-4099
Jožef Stefan Institute, SI

Diana Santos  orcid.org/0000-0002-3108-7706
University of Oslo, Norway and Linguatca, NO

REFERENCES

- Bick, Eckhard. "Functional Aspects in Portuguese NER." *Computational Processing of the Portuguese Language*, edited by Renata Vieira et al., Springer, 2006, pp. 80–89. DOI: https://doi.org/10.1007/11751984_9
- . "PALAVRAS, a Constraint Grammar-Based Parsing System for Portuguese." *Working with Portuguese Corpora*, edited by Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira, Bloomsbury Academic, 2014, pp. 279–302.
- Bode, Katherine. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press, 2018. DOI: <https://doi.org/10.3998/mpub.8784777>
- Brunner, Annelen. "Automatic Recognition of Speech, Thought, and Writing Representation in German Narrative Texts." *Literary and Linguistic Computing*, vol. 28, no. 4, 2013, pp. 563–575. DOI: <https://doi.org/10.1093/lc/fqt024>
- Burnard, Lou, et al. "In Search of Comity: TEI for Distant Reading." *Journal of the Text Encoding Initiative*, submitted, 2020. DOI: <https://doi.org/10.4000/jtei.3500>
- Byszuk, Joanna, et al. "Detecting Direct Speech in Multilingual Collection of 19th-Century Novels." *Proceedings of LT4HALA 2020 – 1st Workshop on Language Technologies for Historical and Ancient Languages*, European Language Resources Association (ELRA), 2020, pp. 100–04, <https://www.aclweb.org/anthology/2020.lt4hala-1.15>.
- Chaves, Castelo Branco. *O Romance Histórico no Romantismo Português*. Instituto de Cultura Portuguesa, Ministério da Cultura e da Ciência, Secretaria de Estado da Cultura, 1979.
- Cinková, Silvie, et al. "Evaluation of Taggers for 19th-Century Fiction." *DH_Budapest_2019*, edited by Gábor Pálko, ELTE, 2019, http://elte-dh.hu/dh_budapest_2019-abstract-booklet/.
- Cohen, Margaret. "Narratology in the Archive of Literature." *Representations*, vol. 108, no. 1, 2009, pp. 51–75. DOI: <https://doi.org/10.1525/rep.2009.108.1.51>
- Cornis-Pope, Marcel, and John Neubauer. *History of the Literary Cultures of East-Central Europe: Junctures and Disjunctures in the 19th and 20th Centuries*. John Benjamins, 2004. DOI: <https://doi.org/10.1075/chlel.xix>
- Dekker, Niels, et al. "Evaluating Named Entity Recognition Tools for Extracting Social Networks from Novels." *PeerJ Computer Science*, vol. 5, 2019, p. e189. DOI: <https://doi.org/10.7717/peerj-cs.189>
- Dobrovoljc, Kaja, et al. "The Universal Dependencies Treebank for Slovenian." *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, 2017, pp. 33–38. DOI: <https://doi.org/10.18653/v1/W17-1406>

- Drace-Francis, Alex. *Geneza culturii române moderne. Instituțiile scrisului și dezvoltarea identității naționale (1700–1900)*. Polirom, 2016.
- Ervajec, Tomaž. *Digital Library and Corpus of Historical Slovene IMP 1.1*. 2014. <http://hdl.handle.net/11356/1031>.
- . “MULTeXt-East: Morphosyntactic Resources for Central and Eastern European Languages.” *Language Resources and Evaluation*, vol. 46, no. 1, 2012, pp. 35–57. DOI: <https://doi.org/10.1007/s10579-011-9174-8>
- . *Reference Corpus of Historical Slovene goa300k 1.2*. Slovenian language resource repository CLARIN.SI, 2015. <http://hdl.handle.net/11356/1025>.
- . “Slovenian Novel Collection (ELTeC-slv).” *European Literary Text Collection (ELTeC)*, edited by Carolin Odebrecht et al., COST Action Distant Reading for European Literary History, 2020. <https://github.com/COST-ELTeC/ELTeC-slv>. DOI: <https://doi.org/10.5281/zenodo.3518108>
- . “The IMP Historical Slovene Language Resources.” *Language Resources and Evaluation*, vol. 49, no. 3, 2015, pp. 753–75. DOI: <https://doi.org/10.1007/s10579-015-9294-7>
- Evert, Stefan, et al. “Understanding and Explaining Distance Measures for Authorship Attribution.” *Digital Scholarship in the Humanities*, 2017. DOI: <https://doi.org/10.1093/lc/fqx023>
- Istrate, Ion, et al. *Dicționarul cronologic al romanului românesc*. Editura Academiei Române, 2004.
- Jannidis, Fotis, et al. *Description of a Corpus of Character References in German Novels: DROC [Deutsches Roman Corpus]*. 2018.
- Kahle, Philip, et al. “Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents.” *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 4, 2017, pp. 19–24. DOI: <https://doi.org/10.1109/ICDAR.2017.307>
- Ljubešić, Nikola, and Kaja Dobrovoljc. “What Does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian.” *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, Association for Computational Linguistics, 2019, pp. 29–34. DOI: <https://doi.org/10.18653/v1/W19-3704>
- Lousada, Isabel Cruz, and Márcio Matias Cantarin. *Do feminino plural ou a singularidade pela voz e a escrita de mulheres*. CLEPUL/UNESP, 2016.
- Marinho, Fátima. *O Romance Histórico em Portugal*. Campo das Letras, 1999.
- McKeon, Michael. “Introduction.” *Theory of the Novel: A Historical Approach*, Johns Hopkins Univ. Press, 2000, pp. xiii–xviii.
- Metzeltin, Michael. *România: Stat, Națiune, Limbă*. Univers Enciclopedic, 2002.
- Moniz, António, and Olegário Paz. *Dicionário breve de termos literários*. Presença, 1997.
- Moretti, Franco. “Conjectures on World Literature.” *The New Left Review*, no. 1, 2000. <http://newleftreview.org/A2094>.
- Munteanu, Basil. *Panorama de la littérature roumaine contemporaine*. Éditions du Sagittaire, 1938.
- Osório, Zília, and João Esteves. *Dicionário no Feminino (séculos XIX–XX)*. Livros Horizonte, 2005.
- Patras, Roxana. “Romanian Novel Collection (ELTeC-rom, v0.9.0).” *European Literary Text Collection (ELTeC)*, edited by Carolin Odebrecht et al., COST Action Distant Reading for European Literary History, 2021. <https://github.com/COST-ELTeC/ELTeC-rom/>. DOI: <https://doi.org/10.5281/zenodo.4662599>.
- . “The Splendors and Mist(eries) of Romanian Digital Literary Studies: A State-of-the-Art Just before Horizons 2020 Closes Off.” *Hermeneia*, no. 23, 2019, pp. 207–22.
- Patras, Roxana, et al. “Thresholds to the ‘Great Unread’: Titling Practices in Eleven ELTeC Collections.” *Interférences Littéraires/Littéraire Interferentias*, 2021. DOI: <https://doi.org/10.5281/zenodo.4664715>.
- Piatti, Barbara, et al. “Mapping Literature: Towards a Geography of Fiction.” *Cartography and Art*, Springer, 2009, pp. 1–16. DOI: https://doi.org/10.1007/978-3-540-68569-2_15
- Qi, Peng, et al. “Universal Dependency Parsing from Scratch.” *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Association for Computational Linguistics, 2018, pp. 160–70. <https://nlp.stanford.edu/pubs/qi2018universal.pdf>.
- Rybicki, Jan, and Maciej Eder. “Deeper Delta across Genres and Languages: Do We Really Need the Most Frequent Words?” *Literary and Linguistic Computing*, vol. 26, no. 3, July 2011, pp. 315–21. DOI: <https://doi.org/10.1093/lc/fqr031>
- Santos, Diana. “Corpora at Linguatca: Vision and Roads Taken.” *Working with Portuguese Corpora*, edited by Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira, Bloomsbury, 2014, pp. 219–36.
- . “Literature Studies in Literateca: Between Digital Humanities and Corpus Linguistics.” *Humanists and the Digital Toolbox: In Honour of Christian-Emil Smith Ore*, edited by Martin Doerr et al., Novus forlag, 2019, pp. 89–109.
- . “Portuguese Language Identity in the World: Adventures and Misadventures of an International Language.” *Language – Nation – Identity: The Question of the Lingua in an Italian and Non-Italian Context*, edited by Elizaveta Khachatryan, Cambridge Scholars Publishing, 2015, pp. 31–54.
- . “Portuguese Novel Collection (ELTeC-por, v2.0.0).” *European Literary Text Collection (ELTeC)*, edited by Carolin Odebrecht et al., COST Action Distant Reading for European Literary History, 2021. <https://github.com/COST-ELTeC/ELTeC-por>. DOI: <https://doi.org/10.5281/zenodo.4288235>.

- Santos, Diana, and Cláudia Freitas. "Estudando personagens na literatura lusófona." *STIL-Symposium in Information and Human Language Technology*, 2019.
- Santos, Diana et al. "Periodização automática: Estudos linguístico-estatísticos de literatura lusófona." *Linguamática*, vol. 12, no. 1, 2020, pp. 80–95. DOI: <https://doi.org/10.21814/lm.12.1.314>
- Scherrer, Yves, and Nikola Ljubešić. "Automatic Normalisation of the Swiss German ArchiMob Corpus Using Character-Level Machine Translation." *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 2016, pp. 248–55.
- Schöch, Christof et al. "Burrows' Zeta: Exploring and Evaluating Variants and Parameters." *Book of Abstracts of the Digital Humanities Conference, ADHO*, 2018. <https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/>.
- Schöch, Christof et al. "The CLiGS Textbox: Building and Using Collections of Literary Texts in Romance Languages Encoded in XML-TEI." *The Journal of the Text Encoding Initiative*, no. rolling issue, 2019. <https://journals.openedition.org/jtei/2085>. DOI: <https://doi.org/10.4000/jtei.2085>
- Schöch, Christof, and Lou Burnard. "French Novel Collection (ELTeC-fra v1.0.1)." *European Literary Text Collection (ELTeC)*, edited by Carolin Odebrecht et al., COST Action Distant Reading for European Literary History, 2021. <https://github.com/COST-ELTeC/ELTeC-fra>. DOI: <https://doi.org/10.5281/zenodo.4662433>
- Silva, Fabio Mario da. *A autoria feminina na literatura portuguesa: Reflexões sobre as teorias do Cânone*. Edições Colibri, 2014.
- Stankovic, Ranka, et al. "Named Entity Recognition for Distant Reading in Several European Literatures." *DH_Budapest_2019*, edited by Gábor Pálko, ELTE, 2019. http://elte-dh.hu/dh_budapest_2019-abstract-booklet/.
- Straka, Milan. "UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task." *Proceedings of CoNLL 2018: The SIGNLL Conference on Computational Natural Language Learning*. ACL, 2018, pp. 197–207. <https://www.aclweb.org/anthology/K18-2020/>.
- TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. TEI, 2020. <https://tei-c.org/guidelines/P5/>.
- Thiesse, Anne-Marie. *La fabrique de l'écrivain national. Entre littérature et politique*. Gallimard, 2019.
- Tudurachi, Adrian. *Fabrica de geniu. Nașterea unei mitologii a productivității literare în cultura română (1825–1875)*. Institutul European, 2016.
- Vala, Hardik, et al. "Mr. Bennet, His Coachman, and the Archbishop Walk into a Bar but Only One of Them Gets Recognized: On the Difficulty of Detecting Characters in Literary Texts." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 769–74. DOI: <https://doi.org/10.18653/v1/D15-1088>
- WG Scholarly Resources. *Encoding Guidelines for the ELTeC*. 2018. https://distantreading.github.io/encoding_proposal.html.
- . *Sampling Criteria for the ELTeC*. 2018. https://distantreading.github.io/sampling_proposal.html.

TO CITE THIS ARTICLE:

Schöch, C, Patras, R, Erjavec, T and Santos D 2021 Creating the European Literary Text Collection (ELTeC): Challenges and Perspectives. *Modern Languages Open*, 2021(1): 25 pp. 1–19. DOI: <https://doi.org/10.3828/mlo.v0i0.364>

Published: 17 December 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Modern Languages Open is a peer-reviewed open access journal published by Liverpool University Press.