COST Action Distant Reading for European Literary History
Corpus Design Principles and Challenges

Distant Reading

Carolin Odebrecht (Humboldt-Universität zu Berlin) together with Christof Schöch, Lou Burnard, Borja Navarro-Colorado et al.

PARTHENOS Workshop for CEE countries, 2019-10-08
Outline

1. Schedule
2. COST Action Distant Reading and ELTeC
3. Corpus design
4. Introduction to XML and ELTeC TEI schema
5. Using TEI encoded texts
6. References
9:00-9:15  Introduction COST and ELTeC
9:15-9:30  Introduction Romanian novels / literary contexts
9:30-09:55 Corpus design
09:55-10:30 Romanian language collection
10:30-11:00 Break
11:00-12:00 Introduction to TEI XML and ELTeC schema
12:00-13:00 Transcribus demonstration
Goals of our sessions

- Present our research approach in Digital Humanities
- Concepts on corpus design and annotation model
- Language-specific contexts on text selection and balancing
- First steps encoding TEI XML
- First steps text digitization with Transcribus

→ Focus on data design and creation
→ Looking forward to discussing each part in the break out sessions!
Corpus linguistics
Every linguistic analysis is an interpretation of the data.
(Lüdeling 2011)

Digital literary studies
Two scholars can read the same dataset - like the same literary work - and derive different meanings.
(Bode 2018)
Outline

1. Schedule
2. COST Action Distant Reading and ELTeC
3. Corpus design
4. Introduction to XML and ELTeC TEI schema
5. Using TEI encoded texts
6. References
COST Actions are research networks\(^1\)

- for any scientific field,
- for workshops, conferences, working group meetings, training, schools, short-term scientific missions, and dissemination and communication activities,
- for fostering Inclusiveness Target Countries (ITC).

Each country member has a national supporting institution

\(^1\)www.cost.eu
Distant Reading

Christof Schöch, University of Trier

CA16204 will

“create a vibrant and diverse network of researchers jointly developing the resources and methods necessary to change the way European literary history is written”

“contribute to the development and distribution of methods, competencies, data, best practices, standards and tools relevant to Distant Reading research”

Working groups

WG 1: Scholarly Resources
WG 2: Methods and tools
WG 3: Literary Theory and History
WG 4: Dissemination

2www.distant-reading.net
https://www.distant-reading.net/about/network/
Creating an open source multi-lingual benchmark corpus for European literature: European Literary Text Collection (ELTeC)⁴

- (currently) 34 Members of 22 countries
- Main tasks are
  - defining corpus design,
  - developing basic encoding schemas,
  - developing workflows.

⁴https://www.distant-reading.net/wg-1/
Digitized and annotated European novels of the 19th century

Uniform sampling and balancing criteria

Uniform and consistent encoding schemas in TEI XML
  - Basic encoding to facilitate distant reading
  - Applicable for different languages
  - Currently working on English, German, French, Spanish, Italian, Romanian, Slovenian, Polish, Hungarian, Portuguese, Serbian, Greek, Norwegian, Czech
Outline

1. Schedule
2. COST Action Distant Reading and ELTeC
3. Corpus design
4. Introduction to XML and ELTeC TEI schema
5. Using TEI encoded texts
6. References
Corpus

Historical corpora (cf. for example Claridge 2008; Kytö 2011)

▶ Digitized and annotated (encoded) historical texts
▶ Resources with a complex publication history and often conflicting texts definitions (cf. for example Caton 2013; De Rose et al. 2002; van Zundert and Andrews 2017)
▶ Divers methods and approaches towards corpus creation in relation to corpus architecture, annotations etc.
▶ Widely-used complex subtype of text corpora in (digital) humanities
Corpus design defines two things (cf. a.o. Hunston 2008; Lüdeling et al. 2016):

▶ Candidates: Which text(s) can be included in the corpus? Which don’t?
▶ Proportion: How many texts with which characteristics should the corpus contain?
Corpus design – Action’s purpose

- Benchmark corpus for distant reading
- Methods for data creation and analysis, e.g.
  - Part-of-speech tagging
  - Lemmatization
  - Morphological information
  - Authorship attribution
  - Network analysis
  - Topic modelling
  - Sentiment analysis
Corpus design – challenges

- Different publication histories in Europe
- Different literary scholars and traditions
- Accessibility of information and resources

Is it possible to define criteria for selecting novels from all over Europe?
Corpus design – Action’s approach

Sampling and balancing criteria\(^5\) will

- not define what a novel is,
- follow a non-normative but metadata-based approach (not canon-based),
- aim to represent the variety of a population\(^6\),
- allow for a comparability of texts and individual sub-collections according to different metadata set(s).

\(^5\)https://distantreading.github.io/sampling_proposal.html

\(^6\)Cf. for discussion of representativeness Biber (1993) and canonicity (Herrmann 2011) and corpus design Algee-Hewitt and McGurl (2018), Bode (2018), Hunston (2008), and Lüdeling et al. (2016).
ELTeC – sampling criteria

- **language**: European languages, no translations
- **prose**: narrative fictional prose
- **period**: 1840-1920
- **length**: min. 10,000 words
- **publication**: prefer books over novels published in serial publications
- **access**: only freely available digitizations
ELTeC – balancing criteria

- 100 texts per language (language collection)
- Period: distribution over time
  - T1: 1840-1859
  - T2: 1860-1879
  - T3: 1880-1899
  - T4: 1900-1920
- Gender: min. 10% and max. 50% have been written by female authors for the language subcollection
- Authorship: 9 - 11 authors with exact three novels
- Length: min. 20% are short novels (10-50k word tokens), min. 20% are long novels (>200k word tokens).
- Reprint: min. 30% are highly canonized novels, min. 30% should be non-canonized novels, based reprint counts within the period 1970-2009
Overview on ELTeC Language Collections:
https://distantreading.github.io/ELTeC/index.html
“Research data management is an explicit process covering the creation and stewardship of research materials to enable their use for as long as they retain value.” Whyte, A. and Rans, J., Glossary of Digital Curation Center

(meta) data should findable, accessible, interoperable and reusable (Wilkinson et al. 2016)
Research data management for ELTeC

- Data creation and update on GitHub\textsuperscript{8}
- Encoding schema developed and documented with TEI ODD\textsuperscript{9}
- Data and workflow documentation on GitHub\textsuperscript{10}
- Persistentreferencing and archiving on Zenodo\textsuperscript{11}
- Free licence to foster re-usability: CC-BY 4.0\textsuperscript{12}
- Further dissemination strategies are currently evaluated

\textsuperscript{8}https://github.com/COST-ELTeC
\textsuperscript{9}ODD https://github.com/distantreading/WG1/ and schema
https://github.com/COST-ELTeC/Schemas
\textsuperscript{10}https://github.com/distantreading/WG1/wiki
\textsuperscript{11}https://zenodo.org/communities/eltec/
\textsuperscript{12}https://creativecommons.org/licenses/by/4.0/
Outline

1. Schedule

2. COST Action Distant Reading and ELTeC

3. Corpus design

4. Introduction to XML and ELTeC TEI schema

5. Using TEI encoded texts

6. References
Corpus data

- Different starting points for data creation, e.g.:
  - Exemplar of the book
  - Digitized book
  - Plain text
  - Previously encoded data set

- Metadata describe the digital or/and analogue source(s) of the data set
  - Library catalogues
  - Online databases for texts, ebooks, corpora
Annotation: explicit assignment of categories to one or more exponents in a corpus, always interpretation (c.f. a.o. Lüdeling 2011; McEnery and Hardie 2012; Zinsmeister et al. 2008)

Tag set and guidelines: defining categories (and values) and formulate guidelines on how and when to assign them

Motivation: Research question/context!
XML  Extensible Markup Language

- W3C standard since 1998
- For structuring and organizing information.
- Metalanguage for defining domain-specific XML vocabularies
  → TEI XML
Text Encoding Initiative

- Encoding standard and guidelines for the representation of texts for humanities
- TEI consortium (since 1987) for developing and maintaining the standard (TEI-Consortium 2019)
- For various texts, e.g. manuscripts and prints, books, letters, poems, and dictionaries
  - Text-internal categories, text-external categories
  - Mark up, text structure(s) and divisions, content and references
- Guidelines provide ca. 500 elements and various specifying attributes
Consist of teiHeader and text

Text contains e.g. front, body, trailer, back

Customization TEI for domain-specific purpose (e.g. select elements and attributes, building subsets, defining new elements)

Validation mechanisms

Customization, documentation and validation via ODD (Burnard and Rahtz 2004)
TEI XML

- **start tag:**
  - `<title>`

- **end tag:**
  - `</title>`

- **single composite tag:**
  - `<lb/>`

- **attributes**
  - `type="..."`

→ Hierarchy of XML elements
Encoding XML

- We start with plain text (e.g. transcribed, OCR)
- We will encode manually¹³

---

¹³Many approaches on transformation processes, see e.g. Distant Reading Training School Budapest 2019
https://distantreading.github.io/Training/Budapest/#(2)
Warm greetings
from Sofia
Carolin
XML – example

- Open start.xml
- Encode
  - line
  - place
  - name
Hands on tutorial – data creation

How do we encode texts for ELTeC?
First steps: using TEI XML
Today’s examples (taken from ELTeC)

CHAPTER I. Down the Rabbit-Hole

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversations?’

So she was considering in her own mind (as well as she could, for the hot day made her feel very sleepy and stupid), whether the pleasure of making a daisy-chain would be worth the trouble of getting up and picking the daisies, when suddenly a White Rabbit with pink eyes ran close by her.
CHAPTER I.

A LONG gallery opening on each side to small rooms gave the inhabitants of St. Cécile’s Monastery access both to them and to the larger apartment which was inhabited by the Reverend Mother herself. This latter room was of an oblong shape, very bare of furniture, and of all kinds of decoration. The

Figure: Google’s digitization of the novel. Google book

Figure: ELTeC text in English language collection ELTeC version
Defining encoding schema

Brain storming:
Which text features can to be encoded for analysing European novels?
ELTeC metadata

- teiHeader
  - Bibliographic information
  - Balancing information
  - Data processing information
ELTeC encoding

- text
  - paragraphs
  - highlighted
  - head
  - division
  - chapter
  - page breaks
  - ...

ELTeC encoding schemas

- Not to represent texts in all their original complexity\textsuperscript{14}
- Not aiming for duplicating the work of scholarly editors
- Aim to facilitate a richer and better-informed distant reading than a transcription of lexical content alone would permit
- Encoding levels (via ODD chaining)
  - level0: basic encoding
  - level1: richer encoding
  - level2: tokenization and linguistic annotation (work in progress)

\textsuperscript{14}cf. contribution to TEI Conference 2020 Burnard, Schöch and Odebrecht
http://gams.uni-graz.at/context:tei2019
Outline

1. Schedule

2. COST Action Distant Reading and ELTeC

3. Corpus design

4. Introduction to XML and ELTeC TEI schema

5. Using TEI encoded texts

6. References
Current state

ELTeC Language Collections:
https://distantreading.github.io/ELTeC/index.html
Figure: Text display is based on TEI encoded files: HTML version for Alice, HTML version for Paul
Metadata composition plot

**Figure:** ELTeC-eng: Metadata in `teiheader` are parsed for each encoded file. Data is aggregated and visualized for corpus monitoring. Produced with ELTeC metadata and R package vcd by David Meyer [aut, cre], Achim Zeileis [aut], Kurt Hornik [aut], Florian Gerber [ctb], Michael Friendly [ctb].

---

15 https://cran.r-project.org/web/packages/vcd/index.html
16
Outline

1. Schedule

2. COST Action Distant Reading and ELTeC

3. Corpus design

4. Introduction to XML and ELTeC TEI schema

5. Using TEI encoded texts

6. References
References COST Action Distant Reading

- COST Action Distant Reading homepage
  https://www.distant-reading.net/
- Documentation on https://distantreading.github.io/
  ▶ Corpus design
  ▶ Encoding guidelines
  ▶ Working Group
  ▶ Training schools
- ELTeC on https://github.com/COST-ELTeC
- ELTeC releases on Zenodo
  https://zenodo.org/communities/eltec/
References for introductions to XML and TEI

- Lou Burnard’s *Introduction to oxygen*
- Martina Scholger’s *Introduction to TEI XML*
- DARIAH’s *Training Digital Editions*
- Lou Burnard’s book on *What is the Text Encoding Initiative*
- Customizing TEI with ODD
References


09:15-09:30: Roxana Patras

Senior Researcher II/ Cercetator stiintific gr. II (CS II)
Institute for Interdisciplinary Research
“Alexandru Ioan Cuza” University of Iasi

Introduction to Romanian novels/literary context
Relevant to the Action’s goal (e.g. focus on previously non preferred periods)
<table>
<thead>
<tr>
<th>File Path</th>
<th>Action</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROM001_Annonim_MCP_ROSCAN_HAIIDUCUL.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM002_BalanescuSimion_BLESTEMUL.xml</td>
<td>reapply fix</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM003_BalanescuSimion_SFARISITUL_BLESTEMUL.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM004_DumbraBarbu_HAIIDUCUL.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM005_MaciPanait_GHITA_CATANUTA.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM006_PopescuND_IancuLianuZapciu.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM007_PopescuND_IancuLianuCapitan.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM008_PopescuND_BUJOR_HAIIDUCUL.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM009_MacedonskiA_Thalassa.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM010_VlahutaA_Dan.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM011_GrandeaF_Fulga.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM012_GrandeaF_Vlasiu.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM013_IonescuR_LeGureSobei.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM014_IonescuR_CatastichiuAmoraului.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM015_IonescuR_DonJuaniiBucuresti.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM016_BolintineanuD_Manoli.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM017_BolintineanuD_Elena.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM018_BolintineanuD_DoritoriiNebuni.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM019_VDemetrius_MateiDumbarau.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM020_VDemetrius_OrasulBucuresti.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM021_VDemetrius_PacatulRabirului.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM022_PanaitMaciu_HaiduculTandura.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM023_Igehei_TilharulFilger.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM024_DumbraBarbu_Pandurului.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM025_Bujoreanu_MistereDinBucuresti.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
<tr>
<td>ROM026_MaciP_Bostan.xml</td>
<td>Add files via upload</td>
<td>17 days ago</td>
</tr>
</tbody>
</table>
ELTeC Summary Page

As well as the following summary statistics, this page provides links to human-readable versions of each text currently included in the European Literary Text Collection (ELTeC). Click on a language code in the table below to see a list of texts now available in that language. Then click on the identifier of a text to see a simple rendering of the text as produced by CETEiClean. The original source files are stored in a GitHub repository at COST-ELTeC, and may be downloaded freely from there.

Please note: this is a work in progress! Comments and reports of any problems are much appreciated: send them to the [WG1 Issue Tracker](https://github.com/ELTeC-COST/WG1-Issue-Tracker).

<table>
<thead>
<tr>
<th>Language</th>
<th>Texts</th>
<th>Words</th>
<th>Male</th>
<th>Female</th>
<th>Short</th>
<th>Medium</th>
<th>Long</th>
<th>1840-59</th>
<th>1860-79</th>
<th>1880-99</th>
<th>1900-20</th>
</tr>
</thead>
<tbody>
<tr>
<td>cze</td>
<td>23</td>
<td>692936</td>
<td>21</td>
<td>2</td>
<td>20</td>
<td>3</td>
<td>0</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>deu</td>
<td>44</td>
<td>8422194</td>
<td>27</td>
<td>17</td>
<td>13</td>
<td>16</td>
<td>15</td>
<td>13</td>
<td>8</td>
<td>14</td>
<td>9</td>
</tr>
<tr>
<td>eng</td>
<td>91</td>
<td>11990064</td>
<td>46</td>
<td>45</td>
<td>17</td>
<td>25</td>
<td>49</td>
<td>20</td>
<td>21</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>fra</td>
<td>92</td>
<td>6939456</td>
<td>57</td>
<td>35</td>
<td>25</td>
<td>45</td>
<td>22</td>
<td>16</td>
<td>31</td>
<td>32</td>
<td>13</td>
</tr>
<tr>
<td>gre</td>
<td>11</td>
<td>42524</td>
<td>10</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>hun</td>
<td>102</td>
<td>7641508</td>
<td>86</td>
<td>16</td>
<td>46</td>
<td>33</td>
<td>23</td>
<td>24</td>
<td>24</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>ita</td>
<td>34</td>
<td>3328244</td>
<td>32</td>
<td>2</td>
<td>13</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>12</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td>nor</td>
<td>27</td>
<td>1114092</td>
<td>22</td>
<td>5</td>
<td>18</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>2</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>por</td>
<td>57</td>
<td>3685825</td>
<td>46</td>
<td>11</td>
<td>27</td>
<td>20</td>
<td>10</td>
<td>8</td>
<td>18</td>
<td>13</td>
<td>18</td>
</tr>
<tr>
<td>rom</td>
<td>26</td>
<td>1196258</td>
<td>23</td>
<td>2</td>
<td>17</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>sly</td>
<td>72</td>
<td>3894267</td>
<td>65</td>
<td>7</td>
<td>37</td>
<td>32</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>30</td>
<td>38</td>
</tr>
<tr>
<td>spa</td>
<td>19</td>
<td>2022041</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>11</td>
<td>5</td>
<td>10</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>srp</td>
<td>29</td>
<td>1087455</td>
<td>25</td>
<td>4</td>
<td>20</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>

Last updated: 2019-10-03
<head>CATASTHIUL AMORULUI</head>

Scena se petrece în capul servitorei dumneavoastră.

La dreapta și la stingă paradoxe, reflecțiiunii, proiecte, fragmente începutul, suveniri, speranțe, căințe.

În scurt niște mobile de simpli creeri, dar de ajuns.

Așezat înaintea unei mese, eu privesc cu un aer multumit un splendid soare care se joacă printre sticlele ferestrei mele.

În acest timp, următorul dialog se angajează între oaspeții creerilor mei, al căror scurt inventar avui plăcere a vi-l așterne mai sus."

Ce frumoasă, ce veselă zi! Bătrinii arbori din Băneasa cadă să aibă în această dimineață reflecte de sărbătoare... Ia-ți pălăria, bastonul și vino cu mine, tu.

N-o asculta! Editorele tău așteaptă, chiar în această dimineață, primele fășcioare ale cărții ce tu i-ai promis. Au să mai fie zile frumoase! N-o asculta!

Milene va ploua, și așa două lune necurmat... Ce mai lucrui, cind vei înegri citvea pagine din o hir-tie neofensivă; nu-ți este rușine să preferi frumusețile îndoloase ale stilului tău, splendorilor strălucitoare ale naturii. Eu d-acui văd eimpile în cari înverzește griul nou. Masa e pusă la birt. Cotleta cintă, frigindu-se. Hossanah al lăcomiei! Vinul ride în vesela butelie... Să plecăm, și zic, și dă dracului secătura ta de volum!

Secătura! ... secătura! ... Lesne de zis.
## Cultural context (1840-1920): traditional periodization vs. ELTeC sampling principles

<table>
<thead>
<tr>
<th>Period</th>
<th>ELTeC t1 = 1840-1859</th>
<th>ELTeC t2 = 1860-1879</th>
<th>ELTeC t3 = 1880-1899</th>
<th>ELTeC t4 = 1900-1920</th>
</tr>
</thead>
<tbody>
<tr>
<td>1840-1859</td>
<td>(48 revolution)</td>
<td>1860-1866 (post-48 revolution)</td>
<td>1880-1900 (Literatorul + Contemporanul)</td>
<td>1900-1918 (Samanatorul + Viata Romaneasca)</td>
</tr>
<tr>
<td>1866-1880</td>
<td>(Junimea age)</td>
<td></td>
<td></td>
<td>1905-1920 (Belle Epoque/Simbolism)</td>
</tr>
<tr>
<td>1880-1900</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1900-1918</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **National, unionist movement:** Wallachia and Moldavia which were semi-autonomous provinces (Ottoman Empire and other empires interested in the geo-strategic position of the 2 principalities)
- **UNION of 1859** (national culture and literature = Romanian language)
- **Romanian prince (1859-1866)**
- **German prince (starting in 1866)**
- **parliamentarianism; state modernization; rural to city life**
- **The war of independence in 1877**
- **proclamation of the Romanian kingdom (German dynasty)**
- **rise of socialism**
- **the peasants’ riot in 1907**
- **national movement (union with Transylvania)**
<table>
<thead>
<tr>
<th>Period</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1840-1859</td>
<td>(48 revolution) Greek influence vs. intelligentsia educated in France and Germany brought in progressist views which culminated in the 1848 Revolution (late Enlightenment + Romanticism)</td>
</tr>
<tr>
<td>1860-1866</td>
<td>(post-48 revolution) 1866-1880 (Junimea age) Imports from Western literary traditions (forms, styles, manners) The obsession of the language’s genius + national literature (original) -aestheticism</td>
</tr>
<tr>
<td>1880-1900</td>
<td>(Literatorul + Contemporanul) a socially-oriented literature -modernism</td>
</tr>
<tr>
<td>1900-1918</td>
<td>(Samanatorul + Viata Romaneasca) 1905-1920 (Belle Epoque/Simbolism) a nationally oriented literature -aestheticization of daily life</td>
</tr>
<tr>
<td>Time Period</td>
<td>Description</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1840-1859 (48 revolution)</td>
<td>Print culture: religious and lay texts in Romanian but printed in the Cyrilic alphabet; transition alphabet or Latin alphabet</td>
</tr>
<tr>
<td>1860-1866 (post-48 revolution) 1866-1880 (Junimea age)</td>
<td>Print culture: dominated by lay texts in Romanian but printed in the transition alphabet or Latin alphabet</td>
</tr>
<tr>
<td>1880-1900 (Literatorul + Contemporanul)</td>
<td>Print culture: lay texts printed in non-standardized variants of Romanian (lack of consistent norms)</td>
</tr>
<tr>
<td>1900-1918 (Samanatorul + Viata Romaneasca) 1905-1920 (Belle Epoque/Simbolism)</td>
<td>Print culture: lay texts printed in non-standardized variants of Romanian (norms)</td>
</tr>
</tbody>
</table>
“Birth certificates” of Romanian novels

INSTALMENT: *Pustnicul* by GRAF Valberg [Mihail KOGĂLNICEANU], 1844, incomplete novel

VOLUME: *Elvira sau amorul fără de sfârșit. Romans original* by D.F.B., 85 p, 1845, complete but too short novel
The Romanian Novel in figures relevant for ELTeC sampling (1840-1920)

- 350 novels published in volumes
- 279 novels published in instalments

Note that current data on Romanian novels also include items written in other languages (French, German) by “Romanian” authors, then imported into Romanian as a new version or translation.

From 350 volume-novels:
- 37 are written by women (10.5%)
- Between 20 and 30 have been subject to constant reprint/ canonization (5.6%-8.5%)
- Around 100 novels are adventure novels (hajduk + city mysteries), thus short novels (28%)
1/3 of the Romanian novels published in volumes should go into ELTeC (100 novels per each collection)

Is that possible?
09:50-10:20: Roxana Patras

Romanian Language collection.
Specific characteristics and challenges for selecting and digitizing novels
How to... Build a National novel collection?
Drawback 1: DIGITIZATION from scratch

The Romanian Collection of ELTeC starts from facsimiles:

- we scan (very fragile) books
- OCR (when possible)
- convert them into xml (automatically when scripts are available, manually when scripts aren’t really reliable)
- A toilsome process of manual cleaning
- Encode files into TEI
Samples I: printing popular books in 19th-century Romania
Samples II: take a close look at the glyphs
How does OCR perform on this type of print?

Crina misterioasă din calea Mogoşoasă

Siluirea

E noapte.
Sa ninge incetase si cerul ca si pamantul parea ingheatat.
E noapte, tarzii de tot si cu toate asta intr'una din casele din calea Mogoşoasă luminaea ardea inca.

Casa are un aspect frumos, ca o vâpsita cu galben si zidaria de si de destul de veche, cu toate astea o reparaţie facută de curând ii dă o aparentă solidă şi frumoasă.

Luminaea ardea într'o cameră din etajul al doilea; afara de acesta camere tot restul apartamentului era în întuneric si tacere.

Afara dormeau un frig teribil si zăpadă era mare, caci tota gîna ninsește. In timpul acesta

Mândria lui il stăpanea și la cazarmă și nu putea să sufere cea mai mica observație a căpitanului Lambu, ci respundea cu obrăsnice, din care cauză era rea persecutat. Persecuțiile deveniră atât de amare pentru Dragoș, în cât își jură să deserteze și să rămâne mai târziu pe căpitanul Lambru.

Cât timp a stat în cazarmă, nu legase prietenie cu nimeni de cât cu un lărmă-căr bomboanele, supranumit Mâna lungă, care se asemena cu Dragoș în privința caracterului, ca două picături de apa. Intre omen de acelaș caracter, invoala se face în două vorbe. Se hotărăseră deci să dezerteze amândoi de odată, și de oarece erau isteți, după cum am mai spus, nu fuseseră mere sau găseșcă un mijloc bun și iată’ deci pe amândoi libert, înarmați până din manș și pregătiți cu haine și rufe pe cari avuseseră grijă să și le cumpere mai din nașinte și luând drumul către pădurea Nicolai.
really untidy
Samples III: take a close look at the glyphs
What is this alphabet?

**ALDO ŞI AMINTA**

**BANDIŢII**

**PARTEA I.**

Les actions des hommes sont une se-

meure fécondément sur les champs

obscurs de l'avenir, confiée avec espé-

rance aux divinités fatales.

Schiller.

I.

Timmiă ne a țărădită atigea an în memória ave-

lvia ce sănătate, de câte făceșe să se înno-

zări pe fortăreața să vină în culcare și

de câte făceșe vremea scoatea și să-e ca că ștă toate la

covea făcărilor, ce înkină în starea ca țări la

mi câlma nevină, pesisînd-o, o pensa că străză ast-

felă să că la București.

Totși se mai gine 'n legende sipră a musulmonilor,

no-

mii, locului trînsenă ală, mi o opoare a mai maltașe

că se avea, locului de cămăril a și apă șpă;-

Dap starea din zi în zi crește, mi memória tre-

**ΚΑΠΙΤΛΟ I.**

**ΕΛΕΦΤΕΡΙΚΑ.**

Bazalt de șold scară, în zioa de șintila Cooprie,

era șpă păiș nentre mine, era șpă izbor șă dăselește;

ne Kapă inima mea o simțe altă de malăș,

dar vor-

bele mele șint slabe ka s'o șapte înseși și ka s'o

nuașe esplăma. Am năzăt șeșișurea săpăză din 'nă-

intea mea, străzăvindă ka în ceră șeșei, șeva tuia ași

svlăză și șe alăstră ka senină căzășă, șinășii țuia

ka o doare, șeră înăi o șapte ne dința. Ea era

che mai găzioasă șii chea mai eșă șeșii din kită

a năzăt oșă mei. Dar, ea, sâă, eșă eșă conuișă,

șe dința, că, opți și șapte s'o prință, opți șinășii

vreme să șili vorbesă, pirnită săpă ală oță mei,

pirnită să șătără de fokă, șeșia nentâ s'mă săme

șii vorbele mele șapte înțeșii de șeșii inimă.

Eșă, rătăciat, ka toate acestea, în șavoasă șe șeiă

șapte șeță șeșișurea șe am șăpță să se din șațea oță mei.
<p>OCR output: absolutely unreliable</p>
Some normal and, of course, normalized texts

Books published between 1920-1945 & scholarly editions published between 1950-1989
Don Juanii din București: ediție ELTeC

Author: Ionescu, Radu (1834-1872)

Resp: Codare ELTeC

Name: Roxana Patras

Extent: 22860 words

PublicationStmt:

<sourceDesc>
<printSource>
<author>Radu Ionescu</author>
<title>Don Juanii din București</title>
<pubPlace>Bucharest</pubPlace>
<publisher>Editura Minerva</publisher>
<date>1974 (1861-1862)</date>
<ref target="https://zenodo.org/record/3229286">Facsimile on zenodo</ref>
</sourceDesc>
Drawback III: The Splendors and Miseries of Digital Literary Studies in Romania in 2019
or
A Harlot High and Low
• treatment of metadata rather than data: a plethora of recent studies strongly rely on a basic counting of metadata provided by the entries in various dictionaries and lexicons (e.g. Transylvanian Review, vol XXXVIII, supplement no. 1, 2019, • Romanian Literature in the Digital Age)

• strong focus on post-45 Romanian literature and on non-literary corpora:
  Corpora such as ROMBAC, RODICA, and so forth
• low digitization of both resources and library metadata: e.g. partially digitized catalogue of the Library of the Romanian Academy;
• dacoromanica platform not really functional;
• shy initiatives started by university libraries, but no interoperable formats (xml, txt, epub)
• incomplete references of this literary span (1840-1920) in Worldcat & others
• Romanian corpora haven’t included enough literary texts and if they have, texts were selected from those published after 1945 (standardized Romanian)
• books indexed in older printed catalogues may not have survived (books to be scanned are fragile and not suited for regular scanning)
• no lemmatizers and POS taggers for diachronic varieties of Romanian (thus automatic analysis functions suboptimally)
Briefly...
Creating such a collection feels like writing fixed verse: align resources to sampling principles

• at least 10%-50% have been written by female authors for the language subcollection.
• 9 to 11 authors are represented with exact three novels.
• at least 30% are highly canonized novels, at least 30% should be non-canonized novels, based on the following reprint groups: reprinted not at all, reprinted once, reprinted more than once within the period 1980-2000
• at least 20% are short novels (10-50k word tokens), at least 20% are long novels (>100k word tokens).

Date: 1840 to 1920 (first iteration)

We will divide into four groups
• group A (1840-1859): code T1
• group B (1860-1879): code T2
• group C (1880-1899): code T3
• group D (1900-1920): code T4
<table>
<thead>
<tr>
<th>Category</th>
<th>T1 (1840-1859)</th>
<th>T2 (1860-1879)</th>
<th>T3 (1880-1899)</th>
<th>T4 (1900-1920)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total no</td>
<td>11</td>
<td>38</td>
<td>120</td>
<td>181</td>
</tr>
<tr>
<td>Transition Alphabet</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Incomplete</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Under 10,000 words (around 100p)</td>
<td>4</td>
<td>5</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>Posthumous</td>
<td>2</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Other languages</td>
<td>1 (French)</td>
<td>6 (French and German)</td>
<td>2 (German)</td>
<td></td>
</tr>
<tr>
<td>CANDIDATES for ELTeC</td>
<td>1</td>
<td>24</td>
<td>94</td>
<td>152</td>
</tr>
</tbody>
</table>
How many of them actually available in Romanian libraries after

• 2 World Wars
• Communist expurgation of undesirable books published during the interbellum period or even during the 19th-century
• The 89 Revolution
Female authors: bettering the scores
From 37 novels authored by females:
- 5 are written in French and German
- 4 are around 100p (possibly under 10,000 words)

If the 9 problematic items are taken out, in real terms we’ve got only 8% female-authored novels

Among the 28 novels that would fit as candidates:

2 authored 3 novels (I.G. Lecca, S. Nadejde)
2 authored 2 novels (Smara, E.I. De Reus, Olteo)
6 authored 1 novel (E. Bacaloglu, M. Miller Verghi, C. Hodos, A. Xenopol, S. Cassvan, V. Ermali)
1 authored 3 novels but 2 of them are too short (E. Tailler is left with only 1)
Sampling 9-11 authors with exactly 3 books

Who published at least 3 books between 1840-1920?

1. D. Bolintineanu
2. R. Ionescu
3. N.D. Popescu
4. P. Macri
5. C. Sandu-Aldea
6. N. Rădulescu-Niger
7. I. Pop-Florantin
8. A. Theochari
9. M. Sadoveanu
10. Duiliu Zamfirescu
11. I. Slavici
12. R. Rossetti
13. V. Demetrius
14. G. Baronzi
15. Al. Pelimon
16. C. D. Aricescu
17. S. Nadejde
18. I.G. Lecca
19. A. I. Alexandrescu
20. C. Oeconomu
21. D.C. Moruzi

Which one would you chose?

Debatable choices from the viewpoint of established Romanian literary historians and theorists
Open questions

How can we get more novels in a specific group?

• Pioneer editing of problematic novels (e.g. novels in the transition alphabet can be recovered by training a HTR model on Transkribus)

• Open discussion about the status of translations or versions (into national languages) that have been authorized by the authors themselves

What is the best way to get/produce useful resources?

• customized, project-oriented libraries (https://zenodo.org/login/?next=%2Fdeposit%3Fpage%3D1%26size%3D20)

• larger digitization projects with a strong and well defined research focus: e.g. literary genres (novel, poetry, theatre), periods, themes, and other types of topical interests
• 12:00-13:00    Roxana Patras

Experiece with building a Romanian Language collection

Focus on typefaces, special characteristics on the text material etc.

Hands-on session (transliteration)
Introduction to Transkribus

TRANSKRIBUS enables you to:

- register in the platform
- download an expert tool specifically designed for your needs
- upload your own documents/images
- manage your own private collection (no one else has access to your documents!)
- segment the images into blocks, lines/baselines and words with the support of layout analysis tools
- link the text with the image which increases the value of your transcription significantly (actually you cannot transcribe text without linking!)
- transcribe text in any language and with any character set (load your own virtual keyboard)
- export your documents at every time in several formats such as TEI, RTF, PDF, XML.
This sounds interesting, but it starts to get really exciting if you consider that

• once you have properly transcribed e.g. 100 images you may inform us and we will train an HTR engine from the Computational Intelligence Technology Lab (CITlab) of the University of Rostock on your documents and

• you will be able to transcribe further pages of your documents with the support of automatically produced handwritten text.
Installation

• Register
• Download Transkribus
• Try out some test documents with automatically produced full-text in the TranskribusCloud collection

Or

• Try to do it intuitively by starting from the documents available in the google drive
Samples from 4 Romanian typesets

1. C.A. Aricescu, Misterele Casatoriei. Barbatul desilusionat, vol 3, 1866
2. G. Baronzi, Fontana zanelor, 1896
3. Ilie Ighel, Banditul Simion Licinski, 1890
4. Th. A. Myller, In Iassy, 1871
5. G. Baronzi, Misterele Bucurestilor, vol 1, 1862
People that are not familiar with language and contents are better in transliteration:

What they see is what they retrieve

So

You won’t learn Romanian in 1h!
But at least you’ll train your eyes and patience for other situations when you need to treat a print as a handwritten text
References


Bibliografia românească modernă (1831-1918), https://biblacad.ro/bnr/brm.php


Transylvanian Review, vol XXXVIII, supplement no. 1, 2019, Romanian Literature in the Digital Age


Bode, Katherine, Reading by numbers: recalibrating the literary field, Anthem Press, 2012.

Bode, Katherine, The equivalence of “close” and “distant” reading; or, toward a new object for data-rich literary history, in Modern Language Quarterly, 78 (1), 2017, pp. 77-106.


Galleron, I, Conceptualisation of theatrical characters in the digital paradigm: needs, problems and foreseen solutions, Human and Social studies, De Gruyter, vol. 6, issue 1, 2017.


Tufiș, D., CoRoLa Primul corpus computațional de referință pentru limba română contemporană, in Market Watch, no. 205, 2018, p. 28-29;
If you have suggestions for ELTeC in general, for the national collections of ELTeC or anything else related to this project please drop me a line at

CS II dr. Roxana Patras
roxana.patras@uaic.ro
https://proiectulbrancusihairo.wordpress.com/
THANKS!