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

Schedule

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

Schedule

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!

Introduction

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

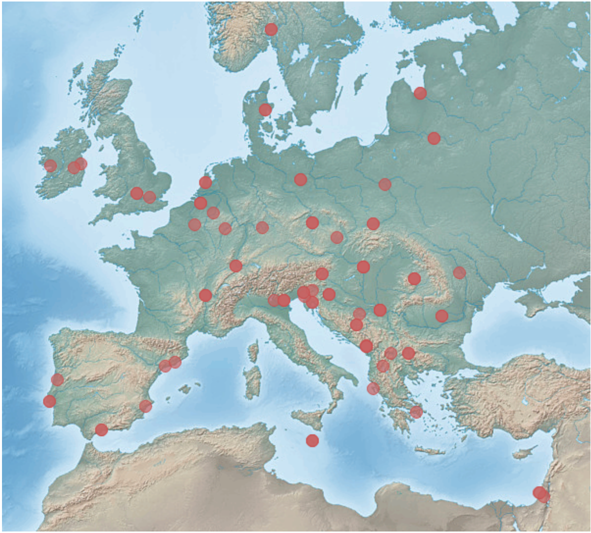
[1] www.cost.eu

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”[2]
- Working groups
  - WG 1: Scholarly Resources
  - WG 2: Methods and tools
  - WG 3: Literary Theory and History
  - WG 4: Dissemination

[2] www.distant-reading.net

COST Action Distant Reading

[3]

[3] https://www.distant-reading.net/about/network/

WG1 Scholarly Resources

- 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.

[4] https://www.distant-reading.net/wg-1/

ELTeC

- 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

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

ELTeC – current state

Overview on ELTeC Language Collections:
`https://distantreading.github.io/ELTeC/index.html`

Research data management for ELTeC

- "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[7]
- (meta) data should findable, accessible, interoperable and reusable (Wilkinson et al. 2016)

[7] http://www.dcc.ac.uk/digital-curation/glossary#R

Research data management for ELTeC

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

[8] https://github.com/COST-ELTeC

[9] ODD https://github.com/distantreading/WG1/ and schema https://github.com/COST-ELTeC/Schemas

[10] https://github.com/distantreading/WG1/wiki

[11] https://zenodo.org/communities/eltec/

[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

Encoding | Annotation

- 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

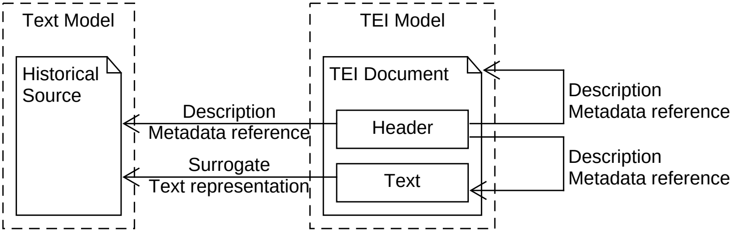
XML Extensible Markup Language

- 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

TEI document



- Consist of `teiHeader` and `text`
- Text contains e.g. `front`, `body`, `trailor`, `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 manually[13]
- data for tutorial on https://github.com/distantreading/WG1/tree/master/Training/2019-10-08-Sofia

[13] Many approaches on transformation processes, see e.g. Distant Reading Training School Budapest 2019 https://distantreading.github.io/Training/Budapest/#(2)

XML – post card example

```
Warm greetings
from Sofia
Carolin
```

XML – example

- Open start.xml
- Encode
  - line
  - place
  - name

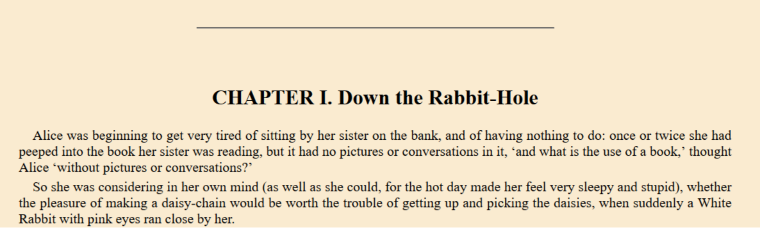
Hands on tutorial – data creation

- How do we encode texts for ELTeC?
- First steps: using TEI XML

→ Today's examples (taken from ELTeC)

  - *Why Paul Ferroll Killed His Wife* by Clive, Caroline, (1801-1873) Saunders, Otley, and Co. London 1860.
  - *Alice's Adventures in Wonderland*, by Lewis Carroll, (1832-1898).London: Macmillan 1865.

Alice

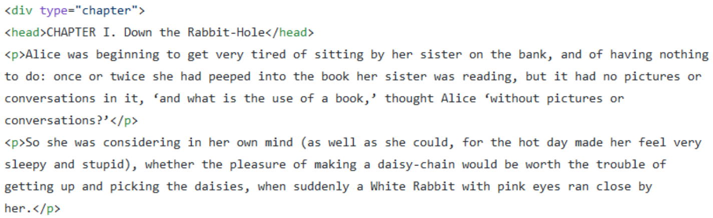


CHAPTER I. Down the Rabbit-Hole

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.

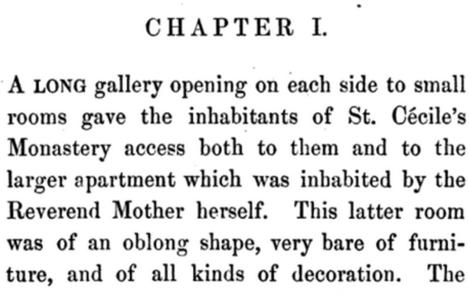
Figure: Gutenberg ebook Produced by Arthur DiBianca and David Widger. Gutenberg ebook



```
<div type="chapter">
<head>CHAPTER I. Down the Rabbit-Hole</head>
<p>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?’</p>
<p>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.</p>
```

Figure: ELTeC text in English language collection ELTeC version

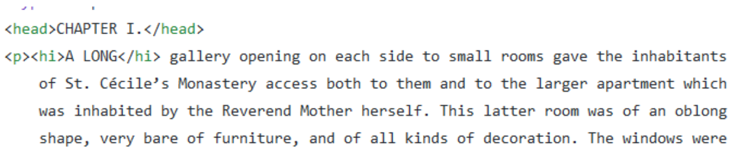
Paul Ferroll



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



```
<head>CHAPTER I.</head>
<p><hi>A LONG</hi> 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 windows were
```

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[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)

[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

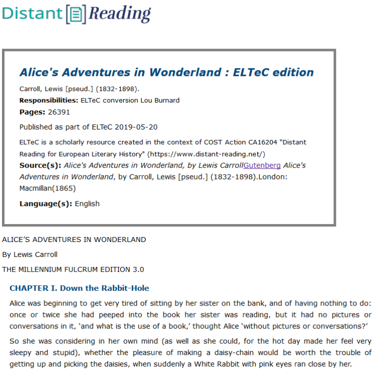
5. Using TEI encoded texts

6. References

Current state

ELTeC Language Collections:
`https://distantreading.github.io/ELTeC/index.html`

Uniform text display



Distant Reading

Alice's Adventures in Wonderland : ELTeC edition

Carroll, Lewis [pseud.] (1832-1898).

Responsibilities: ELTeC conversion Lou Burnard

Pages: 26391

Published as part of ELTeC 2019-05-20

ELTeC is a scholarly resource created in the context of COST Action CA16204 "Distant Reading for European Literary History" (https://www.distant-reading.net/)

Source(s): *Alice's Adventures in Wonderland, by Lewis Carroll*Gutenberg *Alice's Adventures in Wonderland*, by Carroll, Lewis [pseud.] (1832-1898).London: Macmillan(1865)

Language(s): English

ALICE'S ADVENTURES IN WONDERLAND

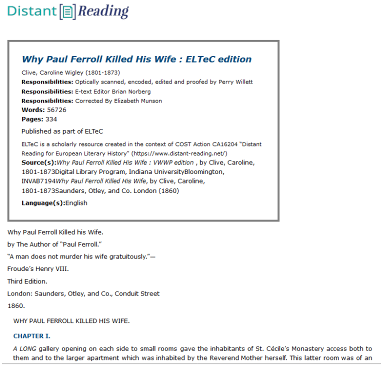
By Lewis Carroll

THE MILLENNIUM FULCRUM EDITION 3.0

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.



Distant Reading

Why Paul Ferroll Killed His Wife : ELTeC edition

Clive, Caroline Wigley (1801-1873)

Responsibilities: Optically scanned, encoded, edited and proofed by Perry Willett

Responsibilities: E-text Editor Brian Norberg

Responsibilities: Corrected By Elizabeth Munson

Words: 56726

Pages: 334

Published as part of ELTeC

ELTeC is a scholarly resource created in the context of COST Action CA16204 "Distant Reading for European Literary History" (https://www.distant-reading.net/)

Source(s):*Why Paul Ferroll Killed His Wife : VWWP edition* , by Clive, Caroline, 1801-1873Digital Library Program, Indiana UniversityBloomington, INVAB7194*Why Paul Ferroll Killed His Wife*, by Clive, Caroline, 1801-1873Saunders, Otley, and Co. London (1860)

Language(s): English

Why Paul Ferroll Killed his Wife.

by The Author of "Paul Ferroll."

"A man does not murder his wife gratuitously."—

Froude's Henry VIII.

Third Edition.

London: Saunders, Otley, and Co., Conduit Street

1860.

WHY PAUL FERROLL KILLED HIS WIFE.

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

Figure: Text display is based on TEI encoded files: HTML version for Alice, HTML version for Paul

Metadata composition plot

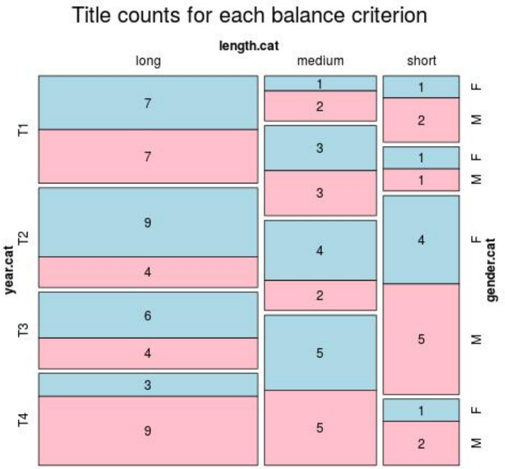


Figure: ELTeC-eng: Metadata in `teiheader` are parsed for each encoded file. Data is aggregated and visualized for corpus monitoring. Produced with ELTeC metadata and R package vcd by David Meyer [aut, cre], Achim Zeileis [aut], Kurt Hornik [aut], Florian Gerber [ctb], Michael Friendly [ctb][16].

[15] `https://cran.r-project.org/web/packages/vcd/index.html`

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 I

Algee-Hewitt, Mark and Mark McGurl (2018). *Between canon and corpus: six perspectives on 20th-century novels*. URL: https://litlab.stanford.edu/LiteraryLabPamphlet8.pdf.

Biber, Douglas (1993). "Representativeness in Corpus Design". In: *Literary and Linguistic Computing* (8), pp. 243–257.

Bode, Katherine (2018). *A World of Fiction - Digital Collections and the Future of Literary History*. eng. University of Michigan Press.

Burnard, Lou and Sebastian Rahtz (2004). *RelaxNG with Son of ODD*. Extreme Markup Languages. URL: http://www.tei-c.org/cms/Talks/extreme2004/paper.html (visited on 01/08/2017).

Caton, Paul (2013). "On the term text in digital humanities". In: *Literary and Linguistic Computing* 28.2, pp. 209–220.

Claridge, Claudia (2008). "Historical Corpora". In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 242–259.

De Rose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear (2002). "What is Text, Really?" In: *Journal of Computing in Higher Education* I(2), pp. 3–26. (Visited on 04/05/2016).

Herrmann, Leonhard (2011). "System? Kanon? Epoche?" In: *Kanon, Wertung und Vermittlung. Literatur in der Wissensgesellschaft*. Ed. by Claudia Stockinger Matthias Beilein and Simone Winko. Berlin: De Gruyter, pp. 59–75.

Hunston, Susan (2008). "Collection strategies and design decisions". In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 154–168.

Kytö, Merja (2011). "Corpora and historical linguistics". In: *Revista Brasileira de Linguistica Aplicada* 11, pp. 417–457.

References II

Lüdeling, Anke (2011). “Corpora in Linguistics. Sampling and Annotation”. In: *Going Digital. Evolutionary and Revolutionary Aspects of Digitization*. Ed. by Karl Grandin. Vol. 147. Nobel Symposium 147. New York: Science History Publications, pp. 220–243.

Lüdeling, Anke, Julia Ritz, Manfred Stede, and Amir Zeldes (2016). “Corpus Linguistics”. In: *OUP Handbook of Information Structure*. Ed. by Caroline Fery and Shinishiro Ishihara. Oxford: Oxford University Press, pp. 599–617.

McEnery, Tony and Andrew Hardie (2012). *Corpus Linguistics. Method, Theory and Practice*. Cambridge Textbooks in Linguistics. Cambride [u.a.]: Cambridge University Press.

TEI-Consortium (2019). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. URL: `http://www.tei-c.org/Guidelines/P5/` (visited on 06/21/2019).

van Zundert, Joris and Tara L. Andrews (2017). “Qu’est-ce qu’un texte numérique? A new rationale for the digital representation of text”. In: *Digital Scholarship in the Humanities* 32, pp. 78–88.

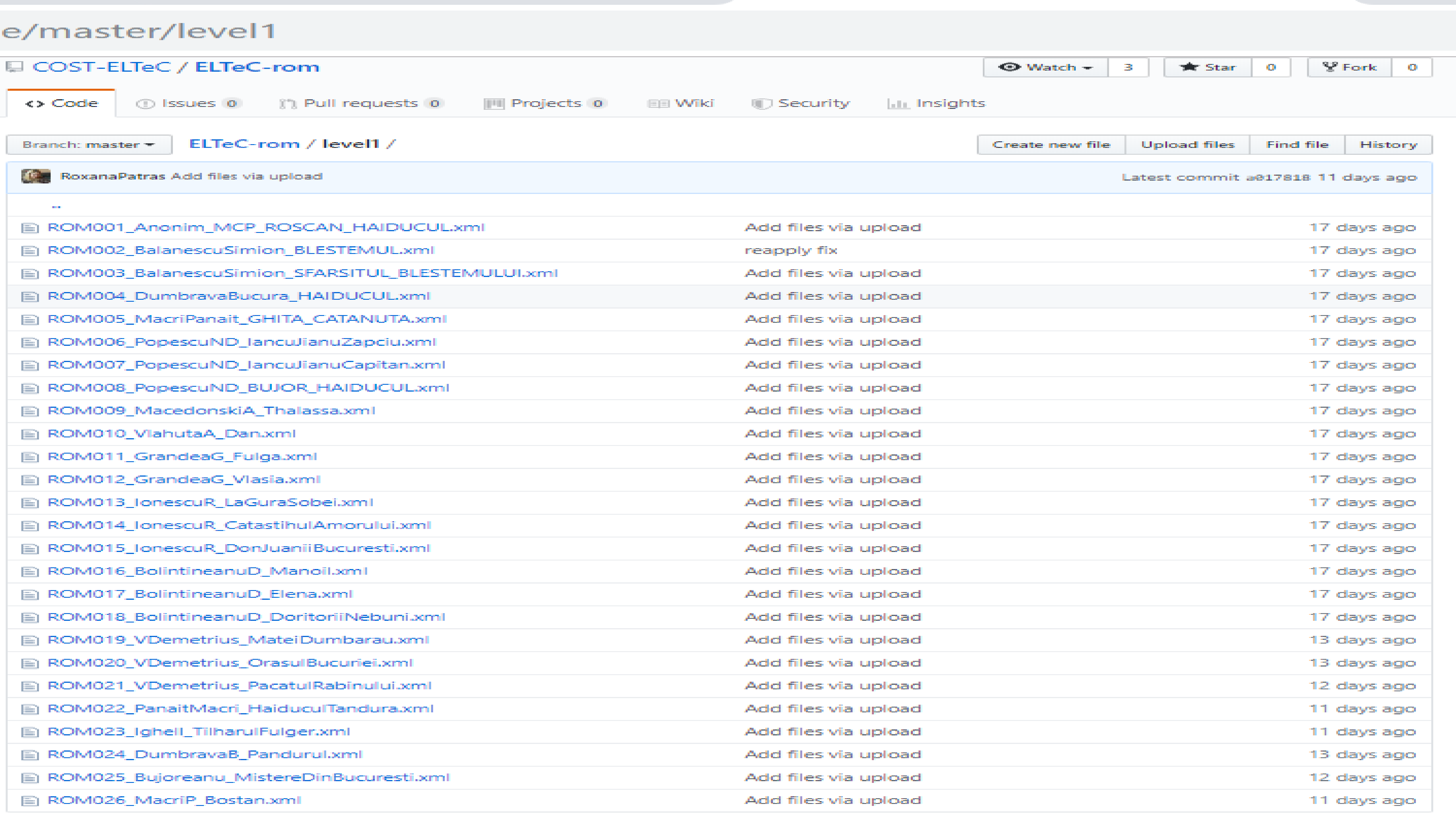
Wilkinson, Mark D., Michel Dumontier, Aalbersberg, I Jsbrand Jan, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, da Silva Santos, Luiz Bonino, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, et al. (2016). “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific data* 3, p. 160018.

Zinsmeister, Heike, Erhard W. Hinrichs, Sandra Kübler, and Andreas Witt (2008). “Linguistically annotated corpora. Quality assurance, reusability and sustainability”. In: *Corpus Linguistics. An International Handbook*. Ed. by Anke Lüdeling and Merja Kytö. Vol. 1. 2 vols. 1. Berlin: De Gruyter, pp. 759–776.

- **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)

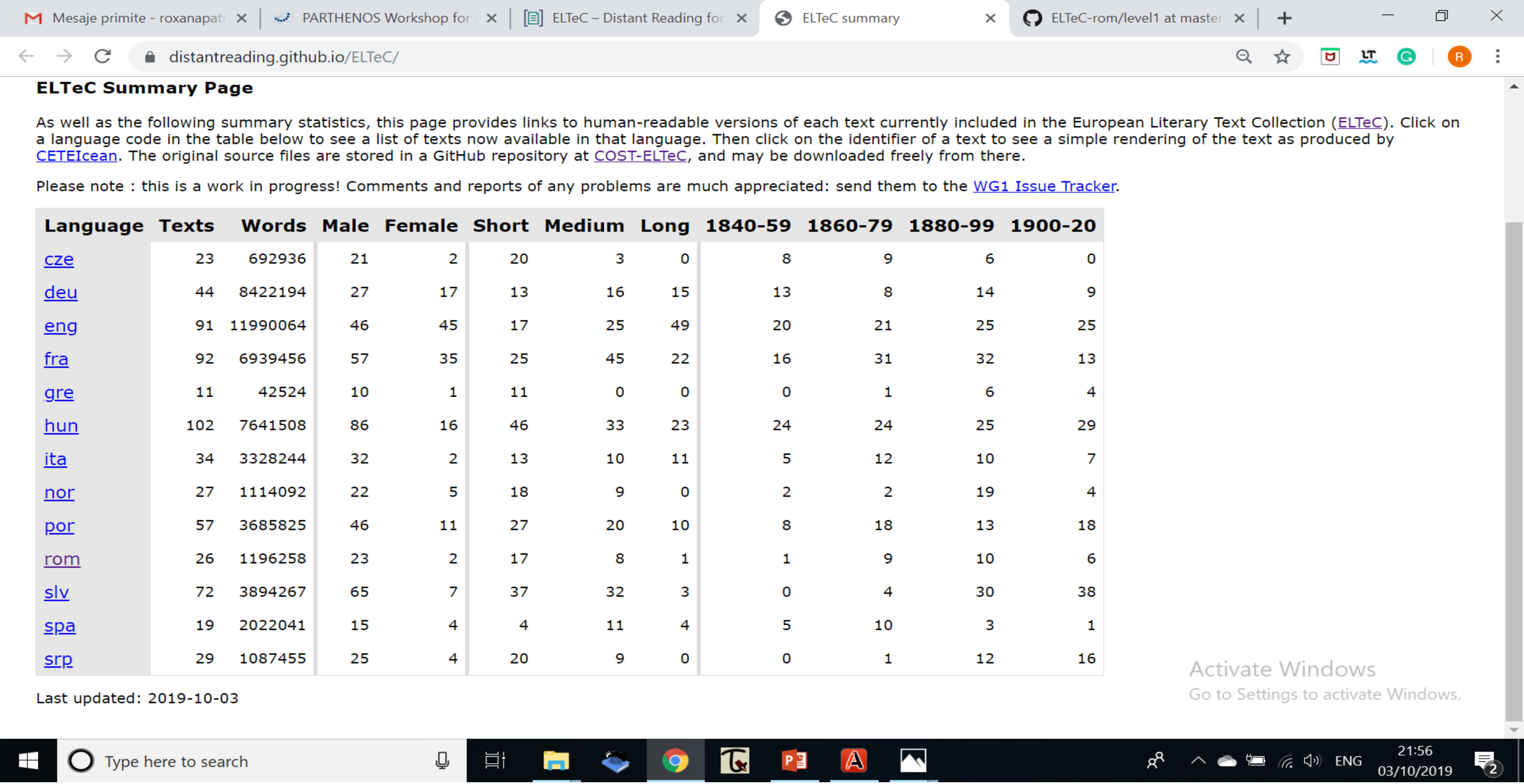


ELTeC-rom / level1 /

RoxanaPatras Add files via upload — Latest commit a817818 11 days ago

..

ROM001_Anonim_MCP_ROSCAN_HAIDUCUL.xml	Add files via upload	17 days ago
ROM002_BalanescuSimion_BLESTEMUL.xml	reapply fix	17 days ago
ROM003_BalanescuSimion_SFARSITUL_BLESTEMULUI.xml	Add files via upload	17 days ago
ROM004_DumbravaBucura_HAIDUCUL.xml	Add files via upload	17 days ago
ROM005_MacriPanait_GHITA_CATANUTA.xml	Add files via upload	17 days ago
ROM006_PopescuND_IancuJianuZapciu.xml	Add files via upload	17 days ago
ROM007_PopescuND_IancuJianuCapitan.xml	Add files via upload	17 days ago
ROM008_PopescuND_BUJOR_HAIDUCUL.xml	Add files via upload	17 days ago
ROM009_MacedonskiA_Thalassa.xml	Add files via upload	17 days ago
ROM010_VlahutaA_Dan.xml	Add files via upload	17 days ago
ROM011_GrandeaG_Fulga.xml	Add files via upload	17 days ago
ROM012_GrandeaG_Vlasia.xml	Add files via upload	17 days ago
ROM013_IonescuR_LaGuraSobei.xml	Add files via upload	17 days ago
ROM014_IonescuR_CatastihulAmorului.xml	Add files via upload	17 days ago
ROM015_IonescuR_DonJuaniiBucuresti.xml	Add files via upload	17 days ago
ROM016_BolintineanuD_Manoil.xml	Add files via upload	17 days ago
ROM017_BolintineanuD_Elena.xml	Add files via upload	17 days ago
ROM018_BolintineanuD_DoritoriiNebuni.xml	Add files via upload	17 days ago
ROM019_VDemetrius_MateiDumbarau.xml	Add files via upload	13 days ago
ROM020_VDemetrius_OrasulBucuriei.xml	Add files via upload	13 days ago
ROM021_VDemetrius_PacatulRabinului.xml	Add files via upload	12 days ago
ROM022_PanaitMacri_HaiduculTandura.xml	Add files via upload	11 days ago
ROM023_Ighell_TilharulFulger.xml	Add files via upload	11 days ago
ROM024_DumbravaB_Pandurul.xml	Add files via upload	13 days ago
ROM025_Bujoreanu_MistereDinBucuresti.xml	Add files via upload	12 days ago
ROM026_MacriP_Bostan.xml	Add files via upload	11 days ago


Mesaje primite - roxanapat
PARTHENOS Workshop for
ELTeC – Distant Reading for
ELTeC summary
ELTeC-rom/level1 at master
distantreading.github.io/ELTeC/
Activate Windows
Go to Settings to activate Windows.
Type here to search
ENG
21:56
03/10/2019

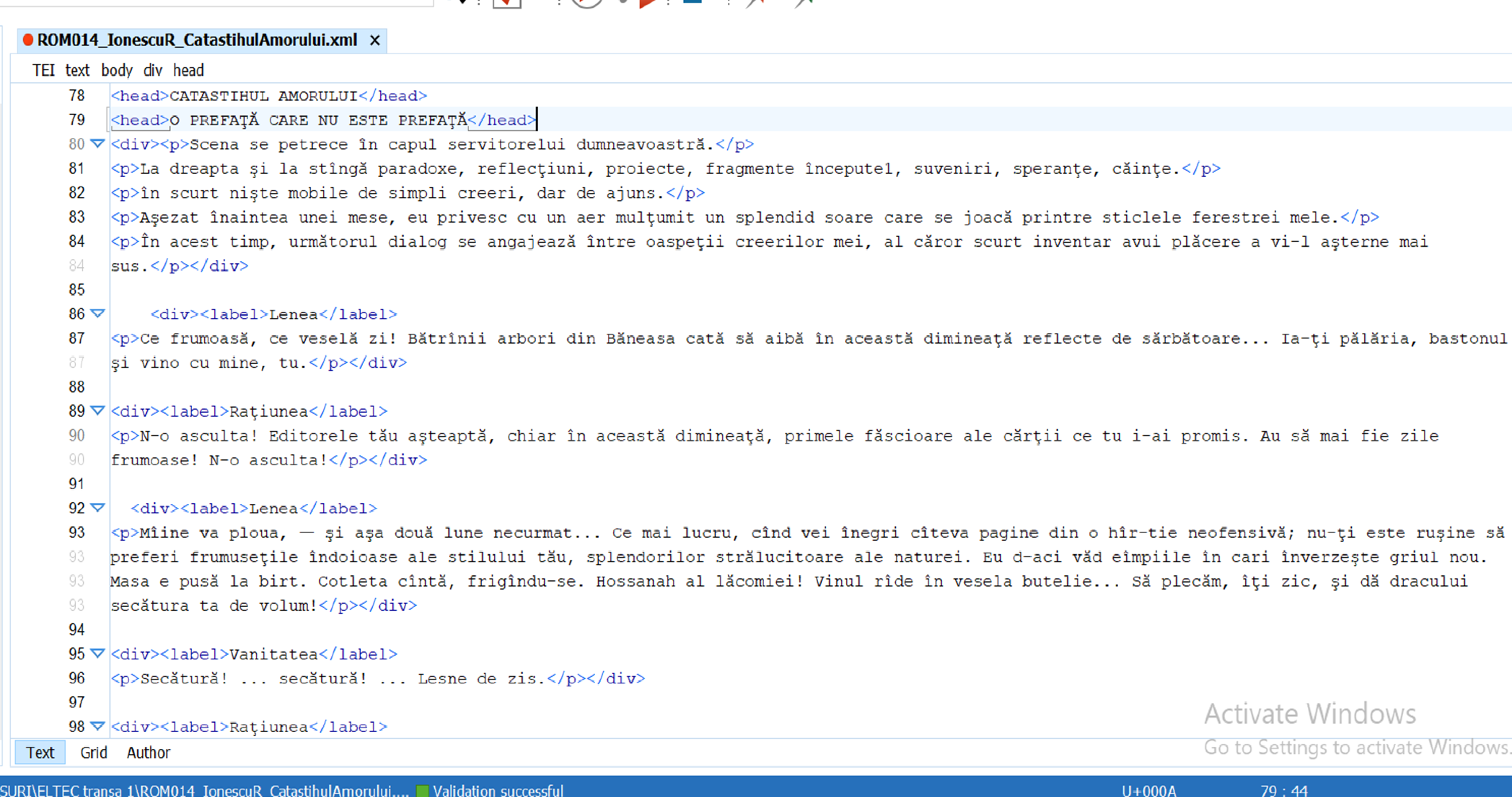
ELTeC Summary Page

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 CETEIcean. 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.

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

Last updated: 2019-10-03


ROM014_IonescuR_CatastihulAmorului.xml ×
TEI text body div head
Text Grid Author
Activate Windows
Go to Settings to activate Windows.
SURI\ELTEC transa 1\ROM014_IonescuR_CatastihulAmorului.... Validation successful
U+000A
79 : 44

```
<head>CATASTIHUL AMORULUI</head>
<head>O PREFAŢĂ CARE NU ESTE PREFAŢĂ</head>
<div><p>Scena se petrece în capul servitorelui dumneavoastră.</p>
<p>La dreapta şi la stîngă paradoxe, reflecţiuni, proiecte, fragmente începute1, suveniri, speranţe, căinţe.</p>
<p>în scurt nişte mobile de simpli creeri, dar de ajuns.</p>
<p>Aşezat înaintea unei mese, eu privesc cu un aer mulţumit un splendid soare care se joacă printre sticlele ferestrei mele.</p>
<p>În acest timp, următorul dialog se angajează între oaspeţii creerilor mei, al căror scurt inventar avui plăcere a vi-l aşterne mai sus.</p></div>

    <div><label>Lenea</label>
<p>Ce frumoasă, ce veselă zi! Bătrînii arbori din Băneasa cată să aibă în această dimineaţă reflecte de sărbătoare... Ia-ţi pălăria, bastonul şi vino cu mine, tu.</p></div>

<div><label>Raţiunea</label>
<p>N-o asculta! Editorele tău aşteaptă, chiar în această dimineaţă, primele făscioare ale cărţii ce tu i-ai promis. Au să mai fie zile frumoase! N-o asculta!</p></div>

  <div><label>Lenea</label>
<p>Mîine va ploua, — şi aşa două lune necurmat... Ce mai lucru, cînd vei înegri cîteva pagine din o hîr-tie neofensivă; nu-ţi este ruşine să preferi frumuseţile îndoioase ale stilului tău, splendorilor strălucitoare ale naturei. Eu d-aci văd eîmpiile în cari înverzeşte griul nou. Masa e pusă la birt. Cotleta cîntă, frigîndu-se. Hossanah al lăcomiei! Vinul rîde în vesela butelie... Să plecăm, îţi zic, şi dă dracului secătura ta de volum!</p></div>

<div><label>Vanitatea</label>
<p>Secătură! ... secătură! ... Lesne de zis.</p></div>

<div><label>Raţiunea</label>
```

Cultural context (1840-1920): traditional periodization vs. ELTeC sampling principles

1840-1859 (48 revolution)	1860-1866 (post-48 revolution) 1866-1880 (Junimea age)	1880-1900 (Literatorul + Contemporanul)	1900-1918 (Samanatorul + Viata Romaneasca) 1905-1920 (Belle Epoque/Simbolism)
ELTeC t1 = 1840-1859	ELTeC t2=1860-1879	ELTeC t3=1880-1899	ELTeC t4=1900-1920
-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)

1840-1859 (48 revolution) ELTeC t1 = 1840-1859	1860-1866 (post-48 revolution) 1866-1880 (Junimea age) ELTeC t2=1860-1879	1880-1900 (Literatorul + Contemporanul) ELTeC t3=1880-1899	1900-1918 (Samanatorul + Viata Romaneasca) 1905-1920 (Belle Epoque/Simbolism) ELTeC t4=1900-1920
Greek influence vs. intelligentsia educated in France and Germany brought in progressist views which culminated in the 1848 Revolution (late Enlightenment + Romanticism)	-Imports from Western literary traditions (forms, styles, manners) -The obsession of the language's *genius* + national literature (original) -aestheticism	-a socially-oriented literature -modernism	-a nationally oriented literature -aesthecization of daily life

1840-1859 (48 revolution) ELTeC t1 = 1840-1859	1860-1866 (post-48 revolution) 1866-1880 (Junimea age) ELTeC t2=1860-1879	1880-1900 (Literatorul + Contemporanul) ELTeC t3=1880-1899	1900-1918 (Samanatorul + Viata Romaneasca) 1905-1920 (Belle Epoque/Simbolism) ELTeC t4=1900-1920
Print culture: religious and lay texts in Romanian but printed in the Cyrilic alphabet; transition alphabet or Latin alphabet	Print culture: dominated by lay texts in Romanian but printed in the transition alphabet or Latin alphabet	Print culture: lay texts printed in non-standardized variants of Romanian (lack of consistent norms)	Print culture: lay texts printed in non-standardized variants of Romanian (norms)

“Birth certificates” of Romanian novels

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

VOLUME: *Elvira sau amorul fără de sfârşit. 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

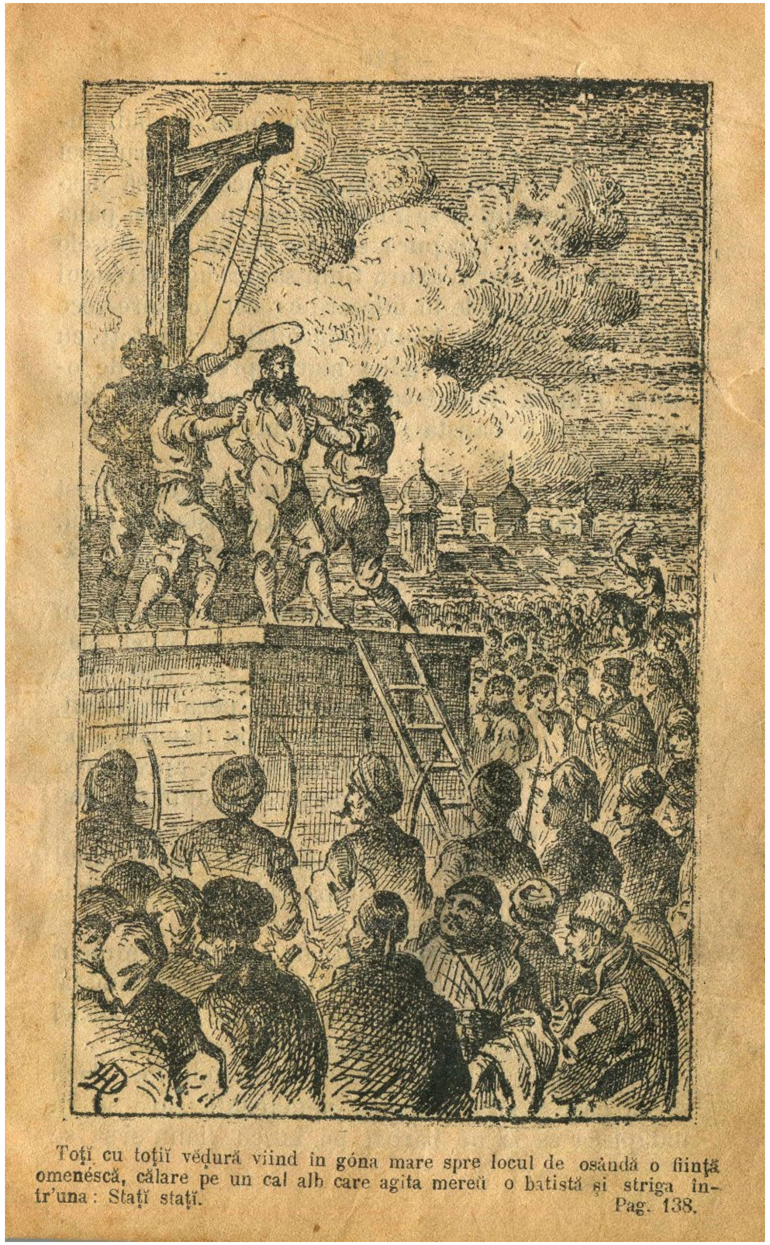
How to... Build a National novel collection?

Drawback 1: DIGITIZATION from scratch

The Romanian Collection of ELTeC starts from facsimiles:

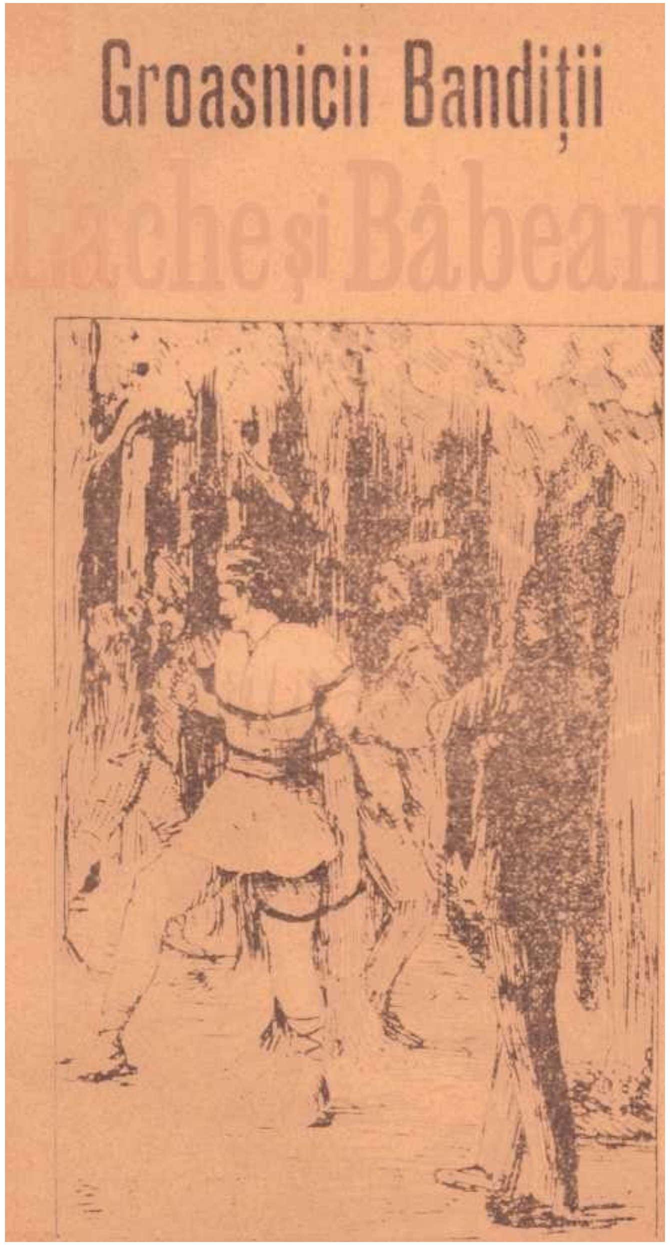
- ➢we scan (very fragile) books
- ➢OCR (when possible)
- ➢convert them into xml (automatically when scripts are available, manually when scripts aren’t really reliable)
- ➢A toilsome process of manual cleaning
- ➢Encode files into TEI

Samples I: printing popular books in 19th-century Romania



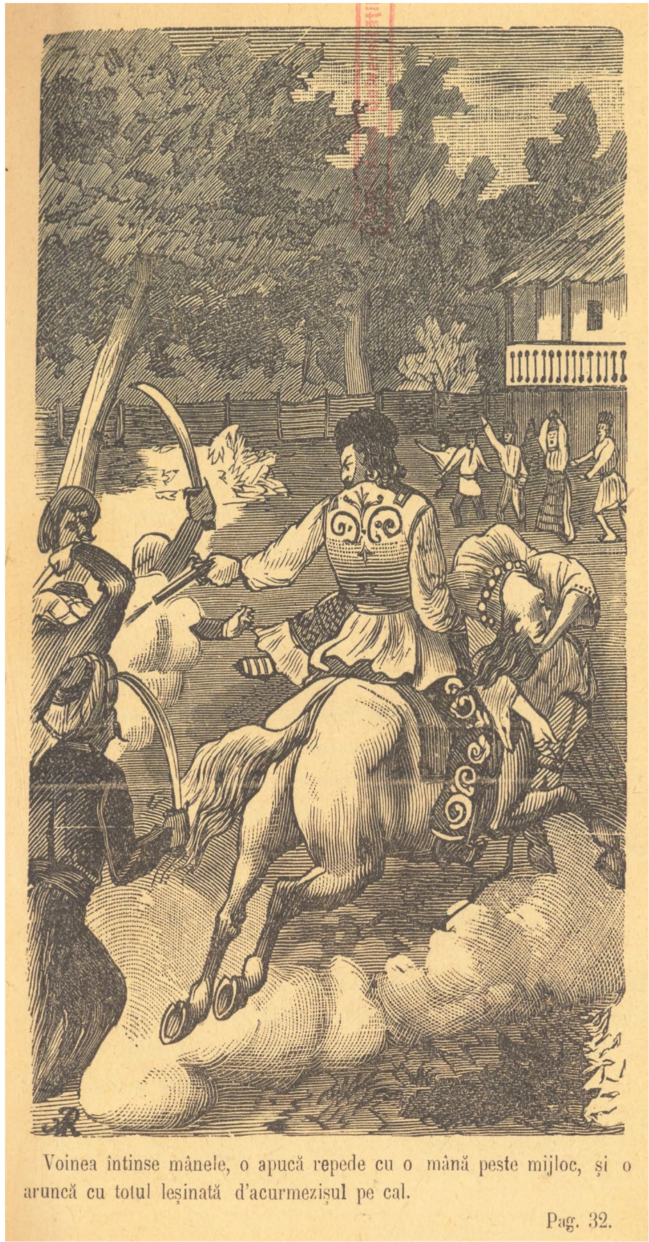
Toţĭ cu toţiĭ vĕdură viind in gόna mare spre locul de osândă o fiinţă omenéscă, călare pe un cal alb care agita mereŭ o batistă şi striga într'una: Staţĭ staţĭ.

Pag. 138.



Groasnicii Banditţii

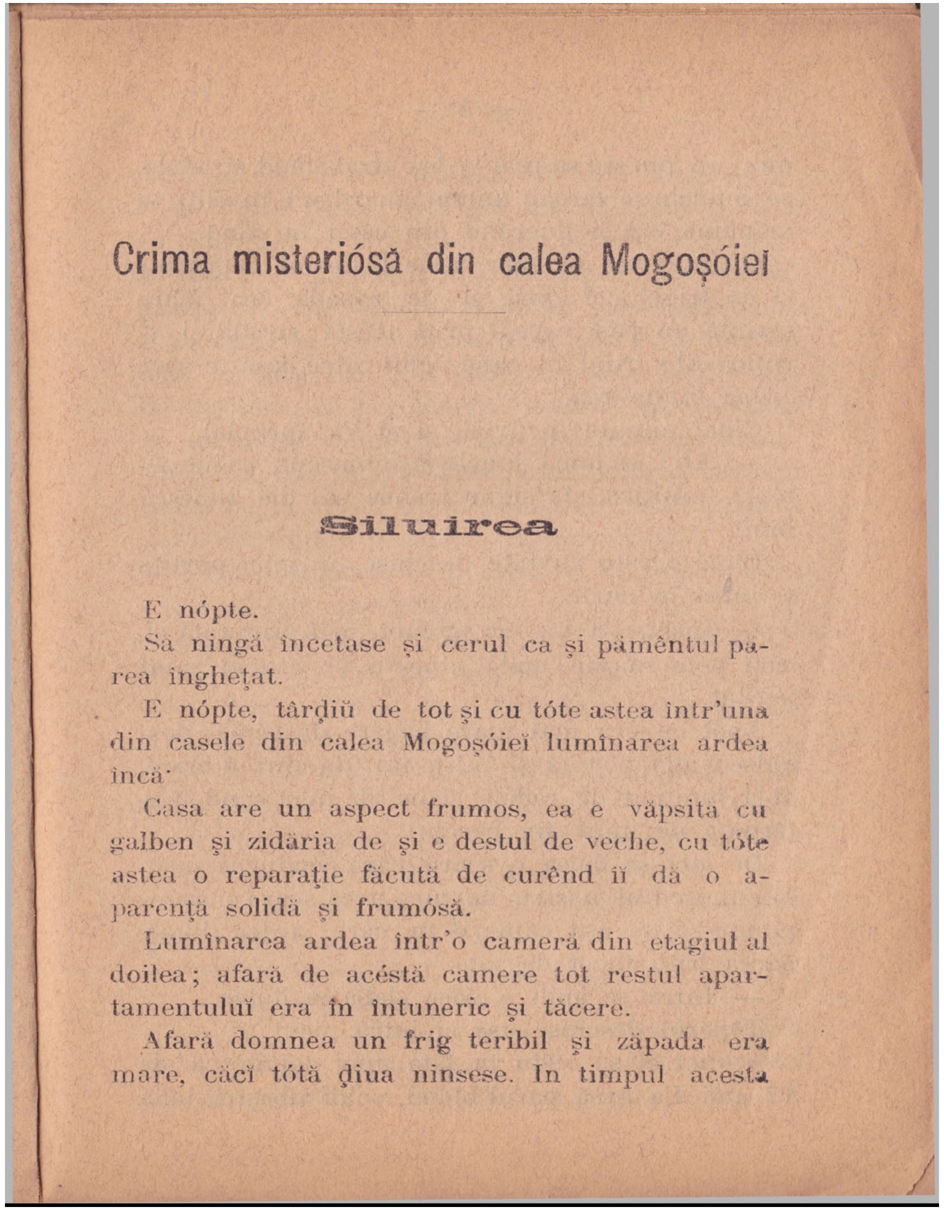
Lache şi Bâbean



Voinea intinse mânele, o apucă repede cu o mână peste mijloc, şi o aruncă cu totul leşinată d'acurmezişul pe cal.

Pag. 32.

Samples II: take a close look at the glyphs
How does OCR perform on this type of print?



Crima misteriósă din calea Mogoșóieĭ

Siluirea

E nópte.

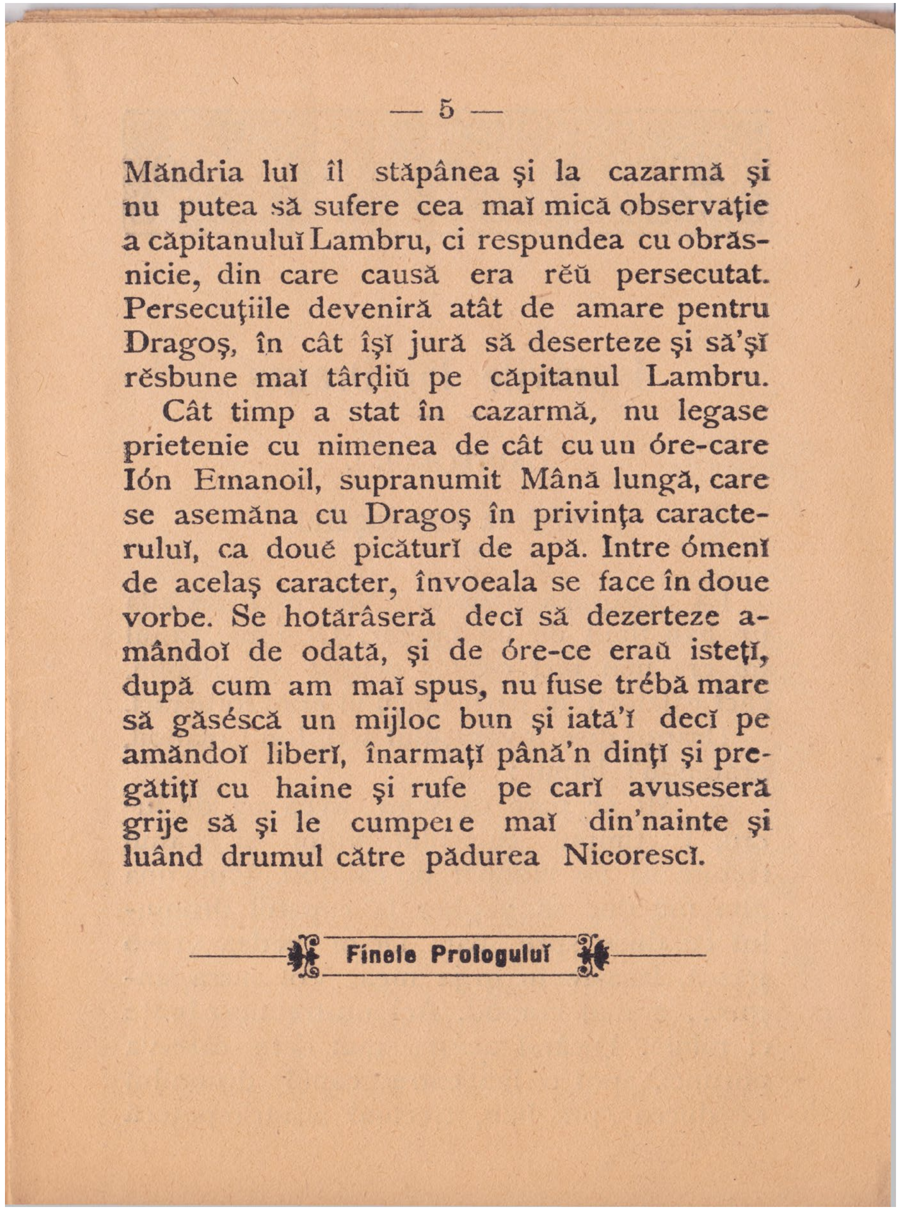
Să ningă încetase și cerul ca și pămêntul parea înghețat.

E nópte, târḑiŭ de tot și cu tóte astea într'una din casele din calea Mogoșóieĭ lumînarea ardea încă·

Casa are un aspect frumos, ea e văpsita cu galben și zidăria de și e destul de veche, cu tóte astea o reparație făcută de curênd iĭ dă o aparență solidă și frumósă.

Lumînarea ardea într'o cameră din etagiul al doilea; afară de acéstă camere tot restul apartamentuluĭ era în întuneric și tăcere.

Afară domnea un frig teribil și zăpada era mare, căcĭ tótă ḑiua ninsese. In timpul acesta

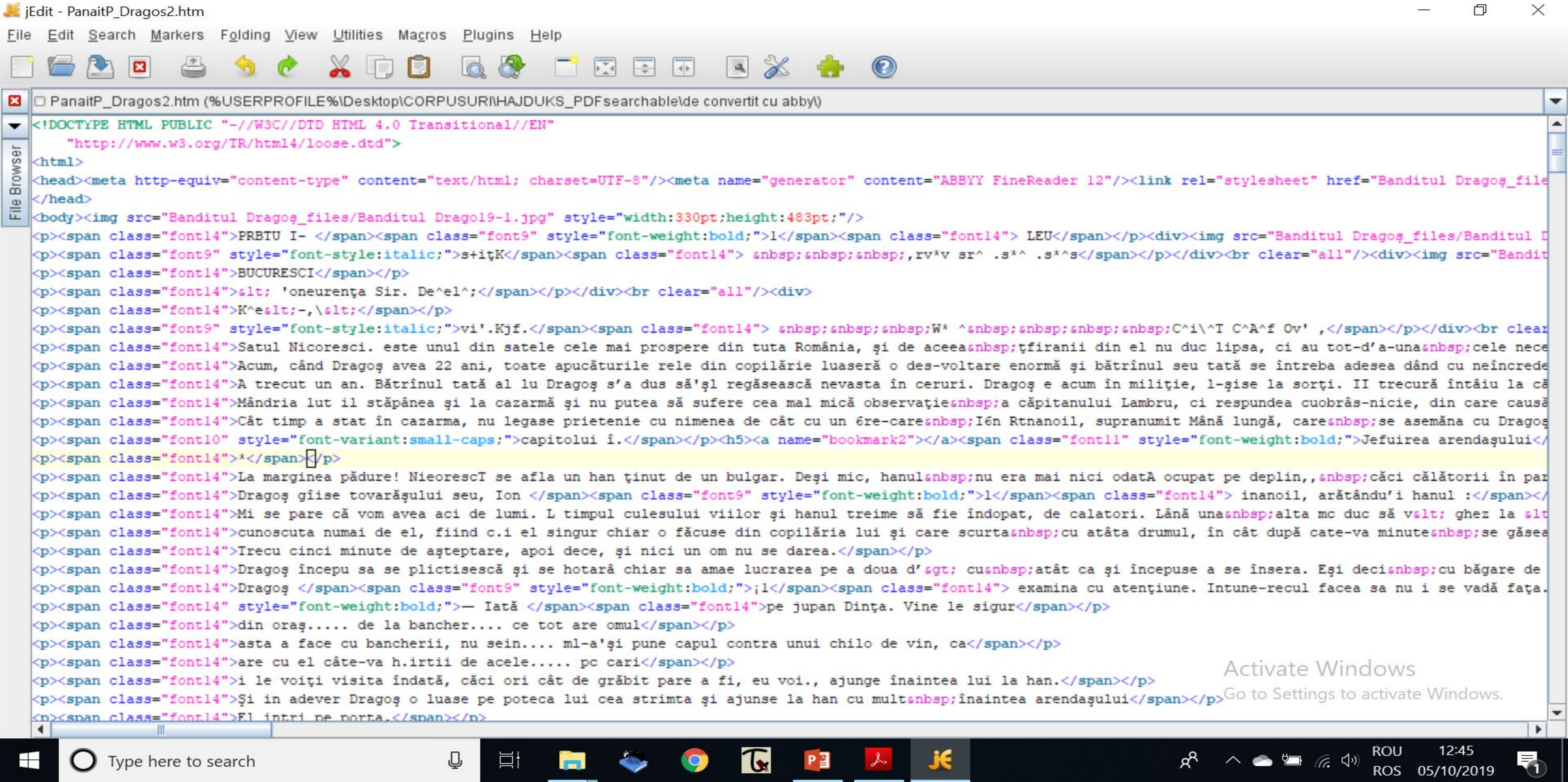


Măndria luĭ îl stăpânea şi la cazarmă şi nu putea să sufere cea maĭ mică observaţie a căpitanuluĭ Lambru, ci respundea cu obrăsnicie, din care causă era rĕŭ persecutat. Persecuţiile deveniră atât de amare pentru Dragoş, în cât îşĭ jură să deserteze şi să'şĭ rĕsbune maĭ târḑiŭ pe căpitanul Lambru.

Cât timp a stat în cazarmă, nu legase prietenie cu nimenea de cât cu un óre-care Ión Emanoil, supranumit Mână lungă, care se asemăna cu Dragoş în privinţa caracteruluĭ, ca doué picăturĭ de apă. Intre ómenĭ de acelaş caracter, învoeala se face în doue vorbe. Se hotărâseră decĭ să dezerteze amândoĭ de odată, şi de óre-ce eraŭ isteţĭ, după cum am maĭ spus, nu fuse trébă mare să găséscă un mijloc bun şi iată'ĭ decĭ pe amăndoĭ liberĭ, înarmaţĭ până'n dinţĭ şi pregătiţĭ cu haine şi rufe pe carĭ avuseseră grije să şi le cumpere maĭ din'nainte şi luând drumul către pădurea Nicorescĭ.

Finele Prologuluĭ

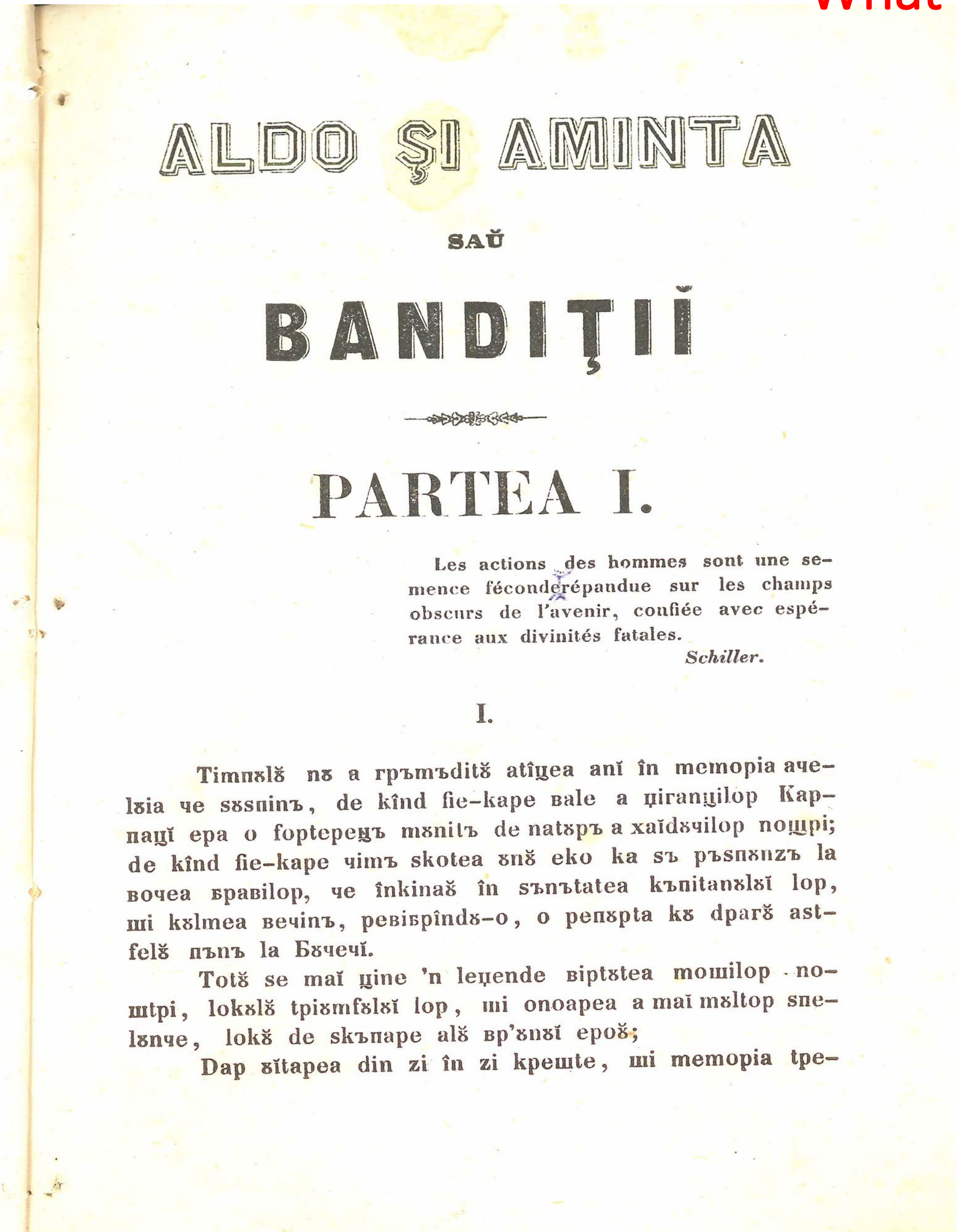
OCR output: really untidy



jEdit - PanaitP_Dragos2.htm
File Edit Search Markers Folding View Utilities Macros Plugins Help
PanaitP_Dragos2.htm (%USERPROFILE%\Desktop\CORPUSURI\HAJDUKS_PDFsearchable\de convertit cu abby\)
File Browser
Activate Windows
Go to Settings to activate Windows.
Type here to search
ROU
ROS
12:45
05/10/2019

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"
    "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head><meta http-equiv="content-type" content="text/html; charset=UTF-8"/><meta name="generator" content="ABBYY FineReader 12"/><link rel="stylesheet" href="Banditul Dragoș_file
</head>
<body><img src="Banditul Dragoș_files/Banditul Dragol9-1.jpg" style="width:330pt;height:483pt;"/>
<p><span class="font14">PRBTU I- </span><span class="font9" style="font-weight:bold;">l</span><span class="font14"> LEU</span></p><div><img src="Banditul Dragoș_files/Banditul D
<p><span class="font9" style="font-style:italic;">s+iţK</span><span class="font14"> &nbsp;&nbsp;&nbsp;,rv*v sr^ .s*^ .s*^s</span></p></div><br clear="all"/><div><img src="Bandit
<p><span class="font14">BUCURESCI</span></p>
<p><span class="font14">&lt; 'oneurenţa Sir. De^el^;</span></p></div><br clear="all"/><div>
<p><span class="font14">K^e&lt;-,\&lt;</span></p>
<p><span class="font9" style="font-style:italic;">vi'.Kjf.</span><span class="font14"> &nbsp;&nbsp;&nbsp;W* ^&nbsp;&nbsp;&nbsp;&nbsp;C^i\^T C^A^f Ov' ,</span></p></div><br clear
<p><span class="font14">Satul Nicoresci. este unul din satele cele mai prospere din tuta România, şi de aceea&nbsp;ţfiranii din el nu duc lipsa, ci au tot-d'a-una&nbsp;cele nece
<p><span class="font14">Acum, când Dragoş avea 22 ani, toate apucăturile rele din copilărie luaseră o des-voltare enormă şi bătrînul seu tată se întreba adesea dând cu neîncrede
<p><span class="font14">A trecut un an. Bătrînul tată al lu Dragoş s'a dus să'şl regăsească nevasta în ceruri. Dragoş e acum în miliţie, l-şise la sorţi. II trecură întâiu la că
<p><span class="font14">Mândria lut il stăpânea şi la cazarmă şi nu putea să sufere cea mal mică observaţie&nbsp;a căpitanului Lambru, ci respundea cuobrâs-nicie, din care causă
<p><span class="font14">Cât timp a stat în cazarma, nu legase prietenie cu nimenea de cât cu un 6re-care&nbsp;I6n Rtnanoil, supranumit Mână lungă, care&nbsp;se asemăna cu Dragoş
<p><span class="font10" style="font-variant:small-caps;">capitolui î.</span></p><h5><a name="bookmark2"></a><span class="font11" style="font-weight:bold;">Jefuirea arendaşului</
<p><span class="font14">*</span></p>
<p><span class="font14">La marginea pădure! NieorescT se afla un han ţinut de un bulgar. Deşi mic, hanul&nbsp;nu era mai nici odatA ocupat pe deplin,,&nbsp;căci călătorii în par
<p><span class="font14">Dragoş gîise tovarăşului seu, Ion </span><span class="font9" style="font-weight:bold;">l</span><span class="font14"> inanoil, arătându'i hanul :</span></
<p><span class="font14">Mi se pare că vom avea aci de lumi. L timpul culesului viilor şi hanul treime să fie îndopat, de calatori. Lână una&nbsp;alta mc duc să v&lt; ghez la &lt
<p><span class="font14">cunoscuta numai de el, fiind c.i el singur chiar o făcuse din copilăria lui şi care scurta&nbsp;cu atâta drumul, în cât după cate-va minute&nbsp;se găsea
<p><span class="font14">Trecu cinci minute de aşteptare, apoi dece, şi nici un om nu se darea.</span></p>
<p><span class="font14">Dragoş începu sa se plictisescă şi se hotarâ chiar sa amae lucrarea pe a doua d'&gt; cu&nbsp;atât ca şi începuse a se însera. Eşi deci&nbsp;cu băgare de
<p><span class="font14">Dragoş </span><span class="font9" style="font-weight:bold;">¡l</span><span class="font14"> examina cu atenţiune. Intune-recul facea sa nu i se vadă faţa.
<p><span class="font14" style="font-weight:bold;">— Iată </span><span class="font14">pe jupan Dinţa. Vine le sigur</span></p>
<p><span class="font14">din oraş..... de la bancher.... ce tot are omul</span></p>
<p><span class="font14">asta a face cu bancherii, nu sein.... ml-a'şi pune capul contra unui chilo de vin, ca</span></p>
<p><span class="font14">are cu el câte-va h.irtii de acele..... pc cari</span></p>
<p><span class="font14">i le voiţi visita îndată, căci ori cât de grăbit pare a fi, eu voi., ajunge înaintea lui la han.</span></p>
<p><span class="font14">Şi in adever Dragoş o luase pe poteca lui cea strimta şi ajunse la han cu mult&nbsp;înaintea arendaşului</span></p>
<p><span class="font14">El intri pe porta.</span></p>
```

Samples III: take a close look at the glyphs
What is this alphabet?



ALDO ŞI AMINTA

SAŬ

BANDIŢIĬ

PARTEA I.

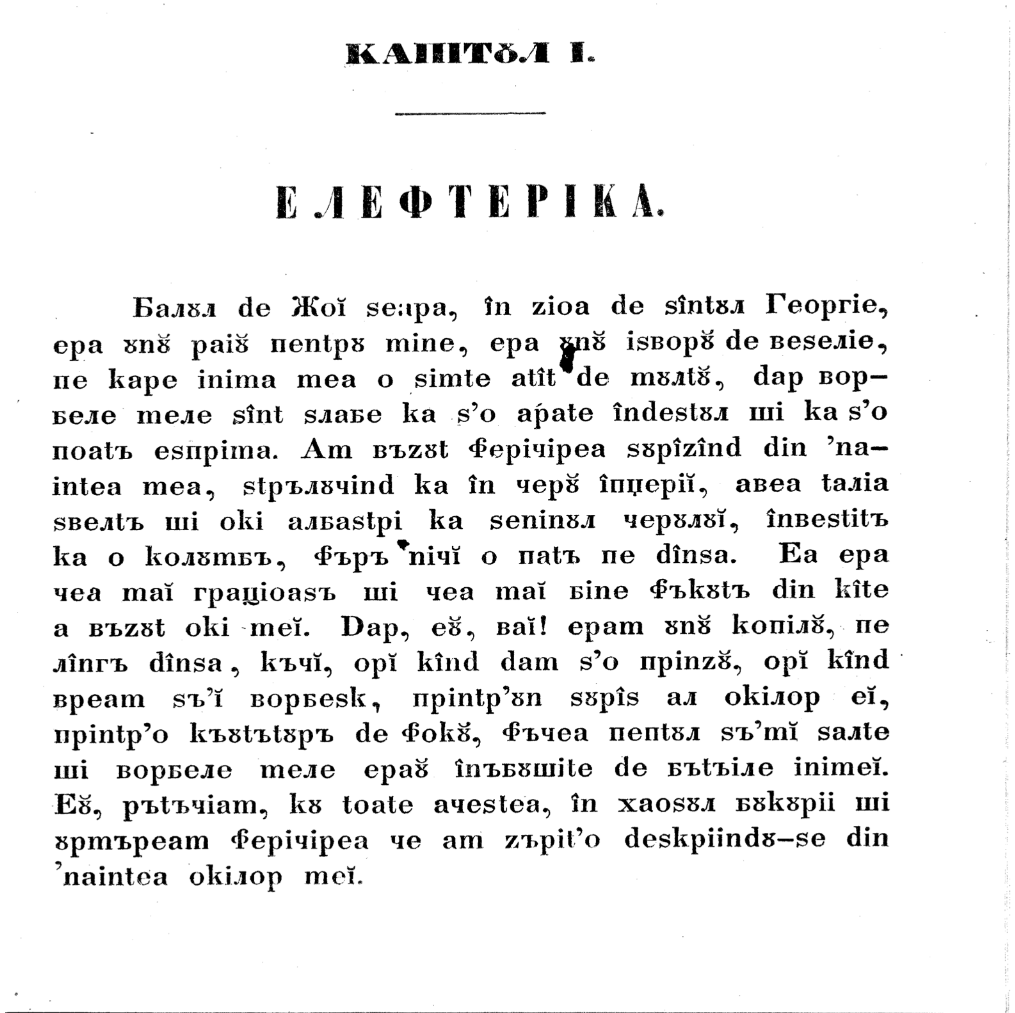
> Les actions des hommes sont une semence fécondérépandue sur les champs obscurs de l'avenir, confiée avec espérance aux divinités fatales.
>
> *Schiller.*

I.

Timпȣlȣ̆ пȣ a гръmъditȣ̆ atîџea anĭ în memoрia ачелȣia че sȣsпinъ, de kînd fie-kaрe вale a џiranџiloр Каpпаџĭ eрa o foрteреџъ mȣnitъ de natȣръ a xaĭdȣчiloр noшрi; de kînd fie-kaрe чimъ skotea ȣnȣ̆ eko ka sъ ръsпȣnzъ la вочea враviloр, че înkinaȣ̆ în sъпъtatea къпitanȣlȣĭ loр, шi kȣlmea вечinъ, реviвpîndȣ-o, o репȣрta kȣ dрarȣ̆ astfelȣ̆ пъпъ la Бȣчечĭ.

Totȣ̆ se maĭ џine 'n leџende viрtȣtea moшiloр noшtрi, lokȣlȣ̆ tрiȣmfȣlȣĭ loр, шi onoaрea a maĭ mȣltoр sпeлȣnче, lokȣ̆ de skъпaрe alȣ̆ вр'ȣnȣĭ eроȣ̆;

Dap ȣĭtaрea din zi în zi kрeшte, шi memoрia tре-

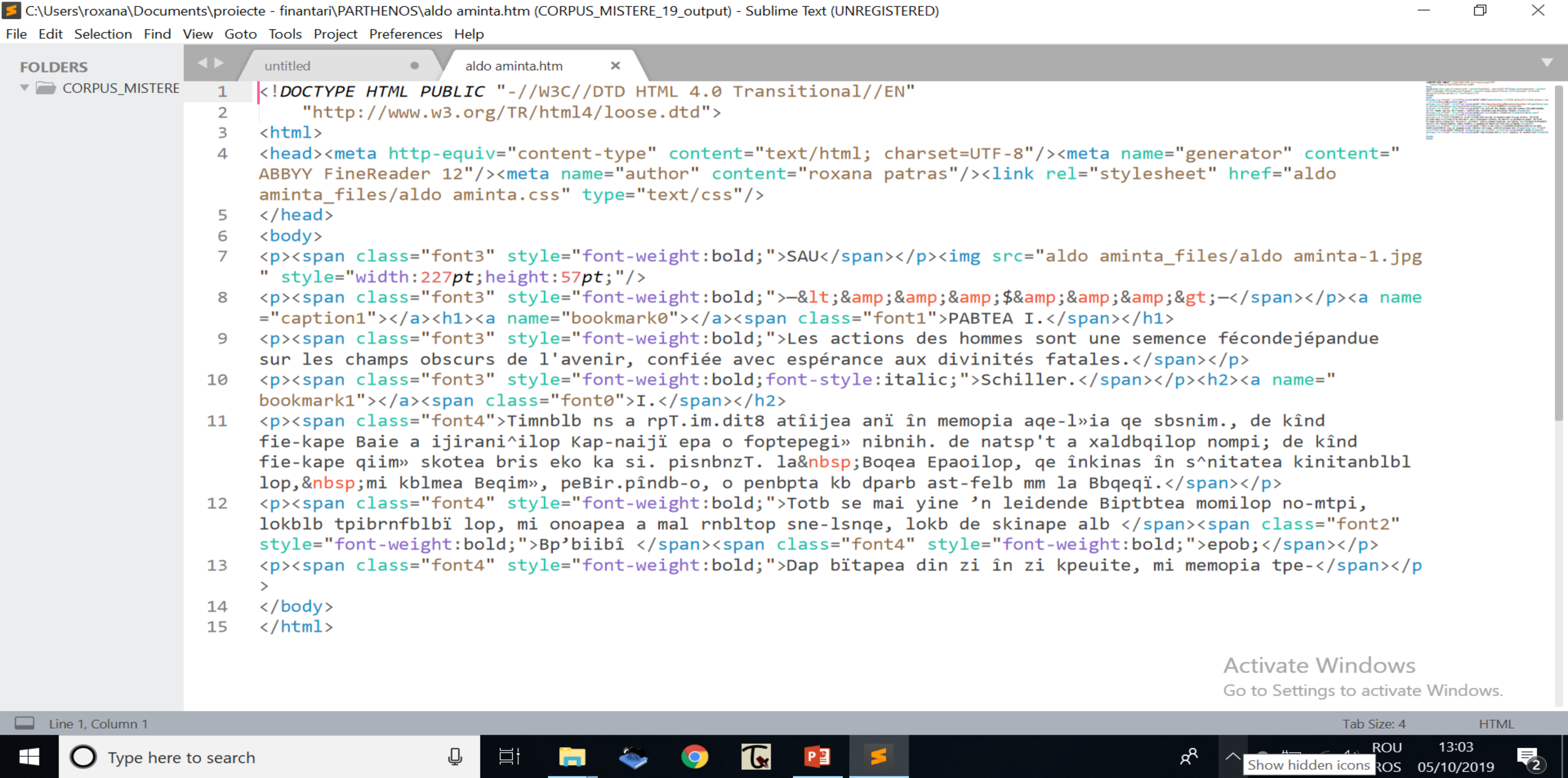


КАПІТȣ̆Л І.

ЕЛЕФТЕРІКА.

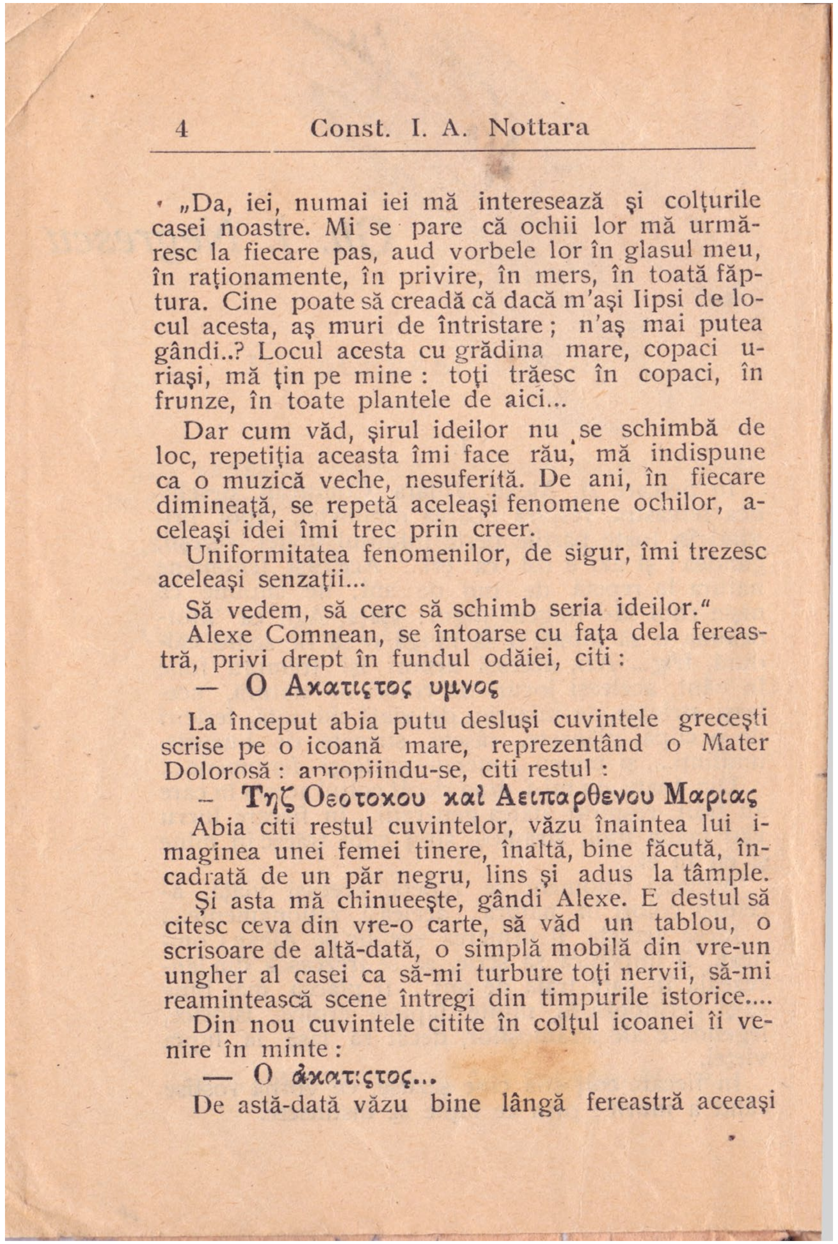
Балȣл de Жоĭ seaрa, în zioa de sîntȣл Георгie, eрa ȣnȣ̆ рaiȣ̆ пentрȣ mine, eрa ȣnȣ̆ isвoрȣ̆ de вeseлie, пе kaрe inima mea o simte atît de mȣлtȣ̆, dap вoрвеле меле sînt sлаве ka s'o aрate îndestȣл шi ka s'o поatъ esпрima. Am въzȣt фeрiчiрea sȣрîzînd din 'naintea mea, stрълȣчind ka în чеpȣ̆ înџeрiĭ, aвea taлia sвeлtъ шi oki aлвastрi ka seninȣл чеpȣлȣĭ, înвestitъ ka o koлȣmвъ, фъръ nічĭ o пatъ пе dînsa. Ea eрa чеа maĭ граџioasъ шi чеа maĭ віne фъkȣtъ din kîte a въzȣt oki meĭ. Dap, eȣ̆, вaĭ! eрam ȣnȣ̆ koпiлȣ̆, пе лîngъ dînsa, къчĭ, oрĭ kînd dam s'o пріnzȣ̆, oрĭ kînd вреam sъ'ĭ вoрвesk, пріntр'ȣn sȣрîs aл okiлoр eĭ, пріntр'o къȣtъtȣръ de фokȣ̆, фъчеа пentȣл sъ'mĭ saлte шi вoрвеле меле eрaȣ̆ înъвȣшite de вътъiле inimeĭ. Eȣ̆, ръtъчiam, kȣ toate aчestea, în xaosȣл вȣkȣрii шi ȣрmъреam фeрiчiрea че am zърit'o deskрiindȣ-se din 'naintea okiлoр meĭ.

OCR output: absolutely unreliable



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Transitional//EN"
    "http://www.w3.org/TR/html4/loose.dtd">
<html>
<head><meta http-equiv="content-type" content="text/html; charset=UTF-8"/><meta name="generator" content="ABBYY FineReader 12"/><meta name="author" content="roxana patras"/><link rel="stylesheet" href="aldo aminta_files/aldo aminta.css" type="text/css"/>
</head>
<body>
<p><span class="font3" style="font-weight:bold;">SAU</span></p><img src="aldo aminta_files/aldo aminta-1.jpg" style="width:227pt;height:57pt;"/>
<p><span class="font3" style="font-weight:bold;">—&lt;&amp;&amp;&amp;$&amp;&amp;&amp;&gt;—</span></p><a name="caption1"></a><h1><a name="bookmark0"></a><span class="font1">PABTEA I.</span></h1>
<p><span class="font3" style="font-weight:bold;">Les actions des hommes sont une semence fécondejépandue sur les champs obscurs de l'avenir, confiée avec espérance aux divinités fatales.</span></p>
<p><span class="font3" style="font-weight:bold;font-style:italic;">Schiller.</span></p><h2><a name="bookmark1"></a><span class="font0">I.</span></h2>
<p><span class="font4">Timnblb ns a rpT.im.dit8 atîijea anï în memopia aqe-l»ia qe sbsnim., de kînd fie-kape Baie a ijirani^ilop Kap-naijï epa o foptepegi» nibnih. de natsp't a xaldbqilop nompi; de kînd fie-kape qiim» skotea bris eko ka si. pisnbnzT. la&nbsp;Boqea Epaoilop, qe înkinas în s^nitatea kinitanblbl lop,&nbsp;mi kblmea Beqim», peBir.pîndb-o, o penbpta kb dparb ast-felb mm la Bbqeqï.</span></p>
<p><span class="font4" style="font-weight:bold;">Totb se mai yine ’n leidende Biptbtea momilop no-mtpi, lokblb tpibrnfblbï lop, mi onoapea a mal rnbltop sne-lsnqe, lokb de skinape alb </span><span class="font2" style="font-weight:bold;">Bp’biibî </span><span class="font4" style="font-weight:bold;">epob;</span></p>
<p><span class="font4" style="font-weight:bold;">Dap bïtapea din zi în zi kpeuite, mi memopia tpe-</span></p>
</body>
</html>
```

Some normal and, of course, normalized texts
Books published between 1920-1945 & scholarly editions published between 1950-1989



4 Const. I. A. Nottara

„Da, iei, numai iei mă interesează şi colţurile casei noastre. Mi se pare că ochii lor mă urmăresc la fiecare pas, aud vorbele lor în glasul meu, în raţionamente, în privire, în mers, în toată făptura. Cine poate să creadă că dacă m'aşi Iipsi de locul acesta, aş muri de întristare; n'aş mai putea gândi..? Locul acesta cu grădina mare, copaci uriaşi, mă ţin pe mine: toţi trăesc în copaci, în frunze, în toate plantele de aici...

Dar cum văd, şirul ideilor nu se schimbă de loc, repetiţia aceasta îmi face rău, mă indispune ca o muzică veche, nesuferită. De ani, în fiecare dimineaţă, se repetă aceleaşi fenomene ochilor, aceleaşi idei îmi trec prin creer.

Uniformitatea fenomenilor, de sigur, îmi trezesc aceleaşi senzaţii...

Să vedem, să cerc să schimb seria ideilor."

Alexe Comnean, se întoarse cu faţa dela fereastră, privi drept în fundul odăiei, citi:

— Ο Ακατιςτος υμνος

La început abia putu desluşi cuvintele greceşti scrise pe o icoană mare, reprezentând o Mater Dolorosă: anropiindu-se, citi restul:

– Τηζ Οεοτοκου καὶ Αειπαρθενου Μαριας

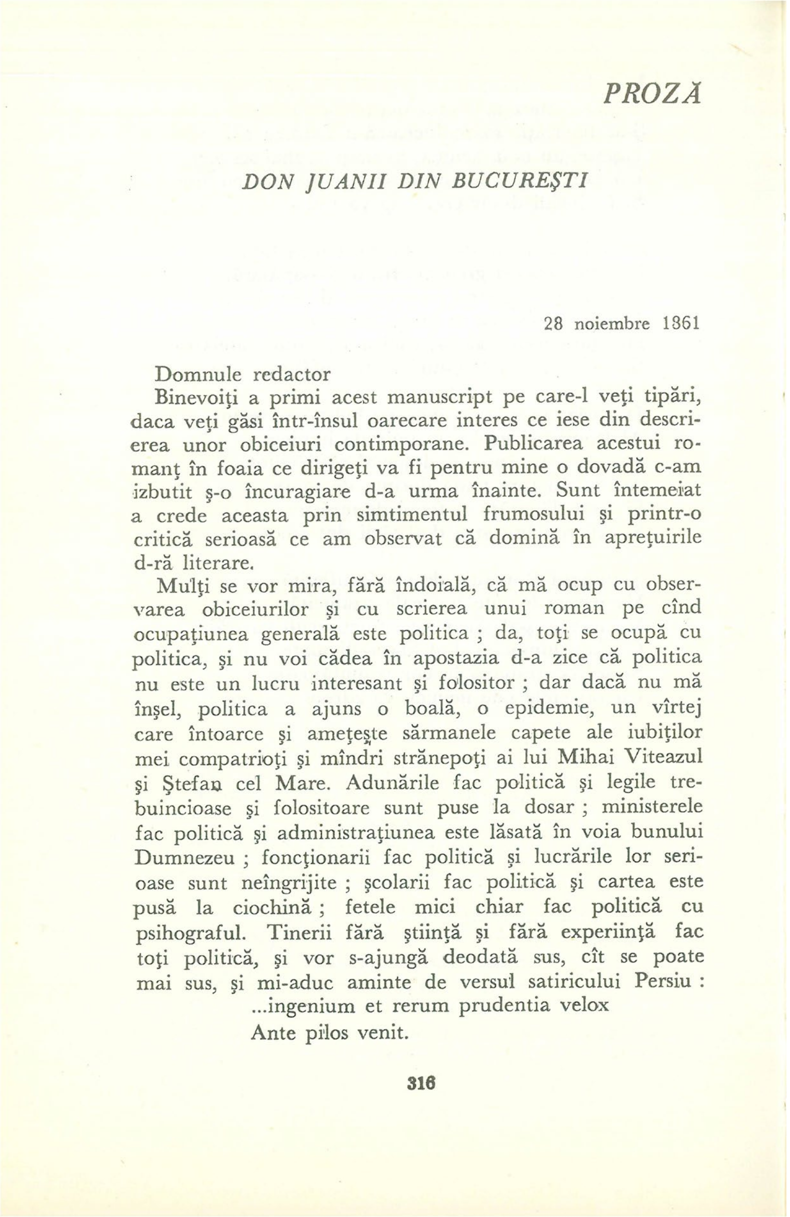
Abia citi restul cuvintelor, văzu înaintea lui imaginea unei femei tinere, înaltă, bine făcută, încadrată de un păr negru, lins şi adus la tâmple.

Şi asta mă chinueeşte, gândi Alexe. E destul să citesc ceva din vre-o carte, să văd un tablou, o scrisoare de altă-dată, o simplă mobilă din vre-un ungher al casei ca să-mi turbure toţi nervii, să-mi reamintească scene întregi din timpurile istorice....

Din nou cuvintele citite în colţul icoanei îi venire în minte:

— Ο ἀκατιςτος...

De astă-dată văzu bine lângă fereastră aceeaşi



PROZĂ

DON JUANII DIN BUCUREŞTI

28 noiembre 1861

Domnule redactor

Binevoiţi a primi acest manuscript pe care-l veţi tipări, daca veţi găsi într-însul oarecare interes ce iese din descrierea unor obiceiuri contimporane. Publicarea acestui romanţ în foaia ce dirigeţi va fi pentru mine o dovadă c-am izbutit ş-o încuragiare d-a urma înainte. Sunt întemeiat a crede aceasta prin simtimentul frumosului şi printr-o critică serioasă ce am observat că domină în apreţuirile d-ră literare.

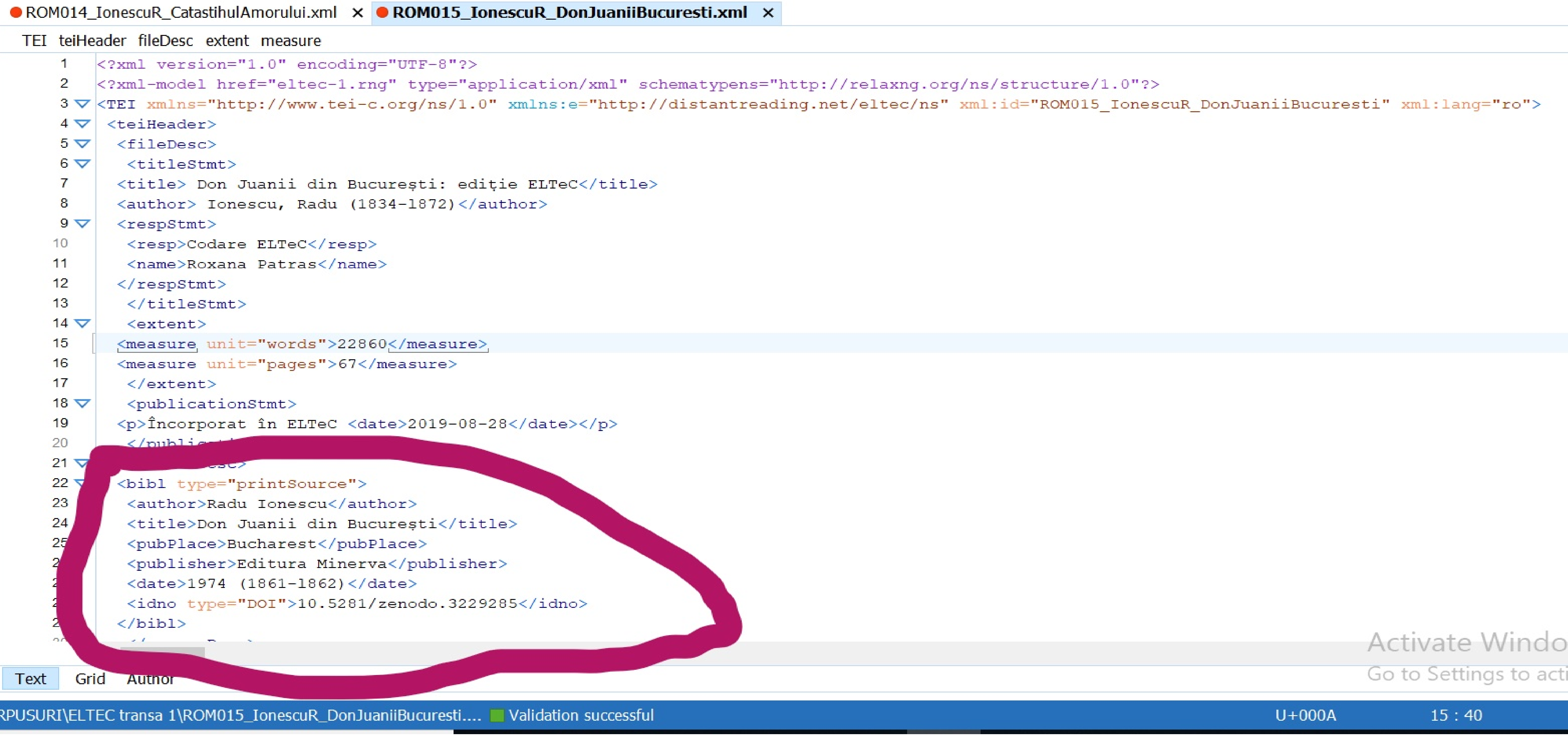
Mulţi se vor mira, fără îndoială, că mă ocup cu observarea obiceiurilor şi cu scrierea unui roman pe cînd ocupaţiunea generală este politica ; da, toţi se ocupă cu politica, şi nu voi cădea în apostazia d-a zice că politica nu este un lucru interesant şi folositor ; dar dacă nu mă înşel, politica a ajuns o boală, o epidemie, un vîrtej care întoarce şi ameţeşte sărmanele capete ale iubiţilor mei compatrioţi şi mîndri strănepoţi ai lui Mihai Viteazul şi Ştefan cel Mare. Adunările fac politică şi legile trebuincioase şi folositoare sunt puse la dosar ; ministerele fac politică şi administraţiunea este lăsată în voia bunului Dumnezeu ; foncţionarii fac politică şi lucrările lor serioase sunt neîngrijite ; şcolarii fac politică şi cartea este pusă la ciochină ; fetele mici chiar fac politică cu psihograful. Tinerii fără ştiinţă şi fără experiinţă fac toţi politică, şi vor s-ajungă deodată sus, cît se poate mai sus, şi mi-aduc aminte de versul satiricului Persiu :

...ingenium et rerum prudentia velox
Ante pilos venit.

316

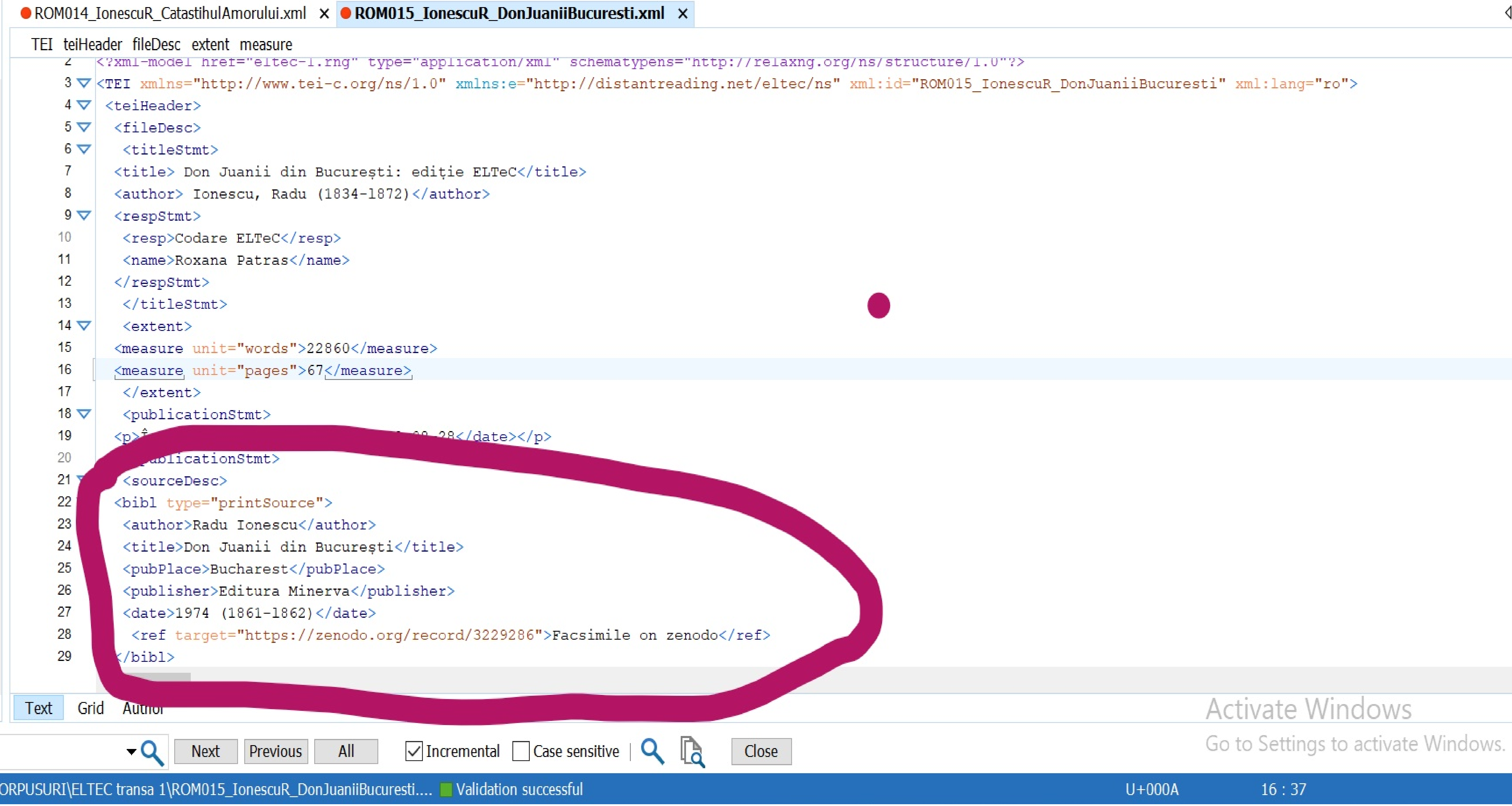
Drawback II: TEI HEADER – particularities

Referencing facsimiles on zenodo https://zenodo.org/record/3229284


ROM014_IonescuR_CatastihulAmorului.xml ✕
ROM015_IonescuR_DonJuaniiBucuresti.xml ✕
TEI teiHeader fileDesc extent measure
Text
Grid
Author

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-model href="eltec-1.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:e="http://distantreading.net/eltec/ns" xml:id="ROM015_IonescuR_DonJuaniiBucuresti" xml:lang="ro">
 <teiHeader>
  <fileDesc>
   <titleStmt>
  <title> Don Juanii din București: ediție ELTeC</title>
  <author> Ionescu, Radu (1834-1872)</author>
  <respStmt>
   <resp>Codare ELTeC</resp>
   <name>Roxana Patras</name>
  </respStmt>
   </titleStmt>
   <extent>
  <measure unit="words">22860</measure>
  <measure unit="pages">67</measure>
   </extent>
   <publicationStmt>
  <p>Încorporat în ELTeC <date>2019-08-28</date></p>
   </publicationStmt>
   <sourceDesc>
  <bibl type="printSource">
   <author>Radu Ionescu</author>
   <title>Don Juanii din București</title>
   <pubPlace>Bucharest</pubPlace>
   <publisher>Editura Minerva</publisher>
   <date>1974 (1861-1862)</date>
   <idno type="DOI">10.5281/zenodo.3229285</idno>
  </bibl>
  </sourceDesc>
```

RPUSURI\ELTEC transa 1\ROM015_IonescuR_DonJuaniiBucuresti.... 🟩 Validation successful U+000A 15 : 40



ROM014_IonescuR_CatastihulAmorului.xml ×
ROM015_IonescuR_DonJuaniiBucuresti.xml ×
TEI teiHeader fileDesc extent measure

```
<?xml-model href="eltec-1.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xmlns:e="http://distantreading.net/eltec/ns" xml:id="ROM015_IonescuR_DonJuaniiBucuresti" xml:lang="ro">
 <teiHeader>
  <fileDesc>
   <titleStmt>
  <title> Don Juanii din București: ediție ELTeC</title>
  <author> Ionescu, Radu (1834-1872)</author>
  <respStmt>
   <resp>Codare ELTeC</resp>
   <name>Roxana Patras</name>
  </respStmt>
   </titleStmt>
   <extent>
  <measure unit="words">22860</measure>
  <measure unit="pages">67</measure>
   </extent>
   <publicationStmt>
  <p>Î[illegible]-09-28</date></p>
   </publicationStmt>
   <sourceDesc>
  <bibl type="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>
  </bibl>
```

Text
Grid
Author
Activate Windows
Go to Settings to activate Windows.
Next
Previous
All
☑ Incremental
□ Case sensitive
Close
ORPUSURI\ELTEC transa 1\ROM015_IonescuR_DonJuaniiBucuresti....
Validation successful
U+000A
16 : 37

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: allign 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

Output of Romanian novels vs. ELTeC groups (Time Slots): culturally determined unbalance of a literary corpus

	T1 (1840-1859)	T2 (1860-1879)	T3 (1880-1899)	T4 (1900-1920)
Total no	11	38	120	181
Transition Alphabet	4	4	0	0
Incomplete	2	2	1	-
Under 10,000 words (around 100p)	4	5	19	27
Posthumous		2	1	-
Other languages		1 (French)	6 (French and German)	2 (German)
CANDIDATES for ELTeC	**1**	**24**	**94**	**152**

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 marterial 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

Dicționarul cronologic al romanului românesc (2003). Vol. 1. Bucharest: Editura Academiei Române.

Dicționarul romanului românesc tradus (2005). Vol. 1. Bucharest: Editura Academiei Române.

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

Drăgan, I. (2001). Romanul popular in România. Literar si paraliterar. Cluj: Casa Cartii de Stiinta.

Barbu, M. (2003). Romanul de mistere în literatura română. Craiova: Fundatia Scrisul românesc.

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

***, New Literary History. Theorizing Genres, Vol. 34, no. 2, spring 2003.

Arthur, P & Bode, Katherine, Advancing Digital Humanities: Research, Methods, Theories, Pagrave Macmillan, 2014.

Barbu Mititelu, V. et al., Corpus of Contemporary Romanian. Architecture, Annotation Levels and Analysis Tools, in Helga Bogdan Oprea et alii (eds.), Lingvistică românească, lingvistică Romanică, Editura Universității din București, 2017, pp. 13-20.

Bernard, M.; Bohet B. (2017). Littérométrie. Outils numériques pour l'analyse des textes littéraires, Paris, Presses de la Sorbonne-Nouvelle.

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.

Burnard, Lou, Qu’est-ce que la Text Encoding Initiative ?, Nouvelle édition [en ligne], Marseille, OpenEdition Press, 2015.

Chivu, Gh. et al. (éd.), Studii de istorie a limbii române. Morfosintaxa limbii literare în secolele al XIX-lea şi al XX-lea, Bucureşti, Editura Academiei Române, 2015

Eder, M. ; Rybicki, J. ; Kestemont, M., Stylometry with R: a package for computational text analysis, in R Journal, 8 (1), 2015, pp. 107–121.

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.

Gheţie, I. ; Mareş, Al., De când se scrie româneşte?, Bucureşti, Univers Enciclopedic, 2001.

Heiden, S., The TXM Platform : Building Open-Source Textual Analysis Software Compatible with the TEI Encoding Scheme, In K. I. Ryo Otoguro (Ed.), 24th Pacific Asia Conference on Language, Information and Computation (p. 389-398). Institute for Digital Enhancement of Cognitive Development, Waseda University, 2010.

Hettinger, Lena et alii, Significance Testing for the Classification of Literary Subgenres, DH conference paper, 2016, http://www.dhd2016.de/abstracts/vortr%C3%A4ge-049.html.

Hoorn, Johan F, How is a Genre Created? Five Combinatory Hypotheses, CLCWeb: Comparative Literature and Culture 2.2, 2000, https://pdfs.semanticscholar.org/0dea/af47fa52207388fbef4bf9f769d2e26a2a9b.pdf.

Ivănescu, G., Istoria limbii române, Iaşi, Junimea, 2000.

Jockers, M., Macronanalysis. Digital methods and literary history, Urbana, Chicago and Springfield, University of Illinois Press, 2013.

Jockers, M. L., Detecting and Characterizing National Style in the 19th Century Novel, Paper presented at Digital Humanities Conference 2011, Stanford University, 2011, pp. 159-160.

Jockers, Matthew L., Computing and Visualizing the 19th Century Literary Genome, Paper presented at Digital Humanities Conference 2012, University of Hamburg, 2012, p. 242-244.

Juola, P., Authorship Attribution, in Foundations and Trends in Information Retrieval. 1 (3): 3, 2006.

Kenny, A., The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities, Oxford: Pergamon Press, 1982.

Mancaş, M, Limbajul artistic românesc modern. Schiţă de evoluţie, Editura Universităţii din Bucureşti, 2005.

Moretti, F., Distant reading, Londres et New York, Verso, 2013.

Novakova, Iva, Siepmann, Dirk (Eds.), Phraseology and Style in Subgenres of the Novel: A Synthesis of Corpus and Literary Perspectives, Palgrave-Macmillan, 2019.

Pană Dindelegan, Gabriela (ed.)., The Syntax of Old Romanian, Oxford, Oxford University Press, 2016.

Rhody, L., Topic Modeling and Figurative Language, in Journal of Digital Humanities, 1/ 2012.

Schöch, Christof, Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama, in Digital Humanities Quarterly, 11, no. 2, 2017, p. 1-53.

Tufiș, D., CoRoLa Primul corpus computațional de referință pentru limba română contemporană, in Market Watch, no. 205, 2018, p. 28-29;

Van Holland, S.; Hengchen, S.; Gillet, F.; De Wilde, M., Introduction aux humanités numériques, De Boeck, 2016.

Villalva, A.; Williams, G. (eds.), The Landscape of Lexicography, John Benjamins, 2019.

Williams, G., Collocational Networks: Interlocking Patterns of Lexis in a Corpusof Plant Biology Research Articles, in International Journal of Corpus Linguistics, vol. 3, issue 1, 1998, p. 151-171.

Zafiu, R., Criterii estetice în normarea limbii române, în C. Stan, R. Zafiu, Al. Nicolae (coord.), Studii lingvistice. Omagiu profesoarei Gabriela Pană Dindelegan, la aniversare, Bucureşti, Editura Universităţii din Bucureşti, 2007, p. 467-473

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/

<p> <emph>THANKS!</emph></p>

</body>
<back></back>
</text>