



SERBIAN IN ELTEC

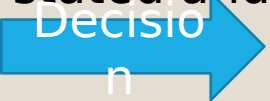
Cvetana Krstev
Faculty of Philology
University of Belgrade

ELTEC ZOOM PARTY, 7 December 2020

Lou's questions

- how was the project organized ? (personnel, management...)
- what is the current state of the collection ? (E5C, encoding, open issues, etc.)
- how were texts selected? (source/s used; balancing methods etc,)
- how did you learn how to use TEI ? (self-study, training, previous knowledge?)
- how were texts processed? (software used etc.)
- what problems did you encounter?

What problems did you encounter?

- Only a limited number of Serbian novels written and published 1840-1920 were digitized;
- Most of the digitized novels are well-known novels, with many editions;
- The big problem: the source for digitization is rarely explicitly stated and almost never first editions were used (it is more  Decision
- Do everything from the scratch

How were texts selected? (sources used; balancing methods etc)

- The problem in compiling the Serbian collection was to **retrieve** novels, not to **select** them.
- That means that it was necessary to compile the list of novels 1840-1920 that would contain:
 - **all novels** written by **women**;
 - **all novels** written in time slots **T1**=1840-1859 and **T2**=1860-1879;
 - **all long** novels (>100,000 words).
- In order to produce such a list we consulted:
 - literary encyclopedias and text books (which talk mostly about established authors);
 - library catalogues (not everything is OPAC, category „novel“ not always attributed or errors, unknown publishing year e.g. 19??);
 - advertisement lists of books published by same publisher that were attached usually at the end of some „old novels“ (already digitized).

How was the project organized? (personnel, management...)

- The work is being done by many volunteers.
 - some volunteers were „volunteers“ – PhD and master students.
- The head of the project: Cvetana
- digitization:
 - **three major libraries** (University Library „Svetozar Marković“, National Library of Serbia, Library of Matica srpska) – they were very helpful; we retrieved first editions for many novels
 - private library Cvetana & Dusko
- OCR: Cvetana & Dusko
- Correction and simple encoding: volunteer readers (25 readers that read from one to 26 novels);
 - level 1 encoding - front matter, chapters, headings, page numbers, footnotes, foreign words, highlighted words, paragraphs, separate lines.
- Final correction, TEI header: Cvetana

How were texts processed? (software used etc.)

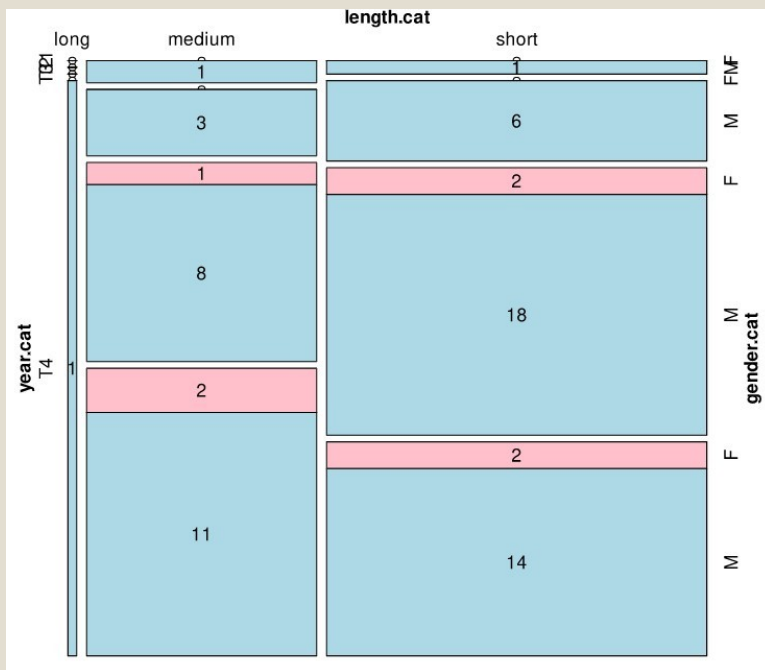
- For OCR we used Abby Finereader (home copy)
- For XML & TEI – Oxygen
- For OCR correction:
 - the results of OCR varies from excellent (rarely) to unusable – in the later case we had to find another copy or edition or we had to remove the novel from the list;
 - first step: we used the self-produced system for OCR correction implemented in Unitex (open source corpus processing suite) that relies on Serbian morphological dictionaries;
 - This procedure can correct up to 90% of errors, but in most of the cases 50-60%;
 - second step: a proofreader-volunteer reads the whole novel and corrects remaining errors (along with structure tagging)

How did you learn how to use TEI? (self-study, training, previous knowledge?)

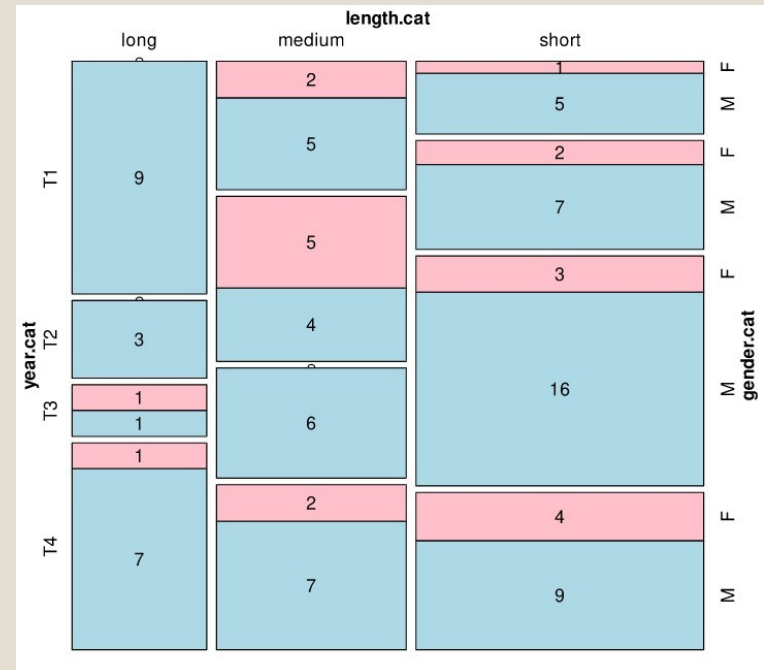
- Previous knowledge
- Back in 1997 in my PhD thesis I used TEI for one project;
- I was teaching my students about TEI (TEI header in more details) for many years;
- I am not a big expert in TEI but I can manage quite well (also my former students in libraries).

What is the current state of the collection ? (E5C, encoding, open issues, etc.)

Serbian collection - 70 novels, E5C=70.77

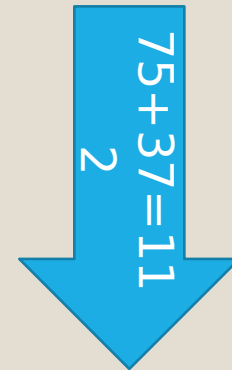


Hungarian collection - 100 novels, E5C=100.00



The present status of work on ELTeC-srp collection

- 75 novels in ELTeC-srp collection
- 3 novels: final correction pending
- 6 novels: volunteers are proof-reading
- 17 novels: scanned
- 11 novels: on the list but not scanned yet.
- We will arrive at the number 100, but:
 - **no more female** authors;
 - **no more** novels from time slot **T1**=1840-1839:
 - one or two novels from time slot **T2**=1860-1879:
 - one or two long novels.
 - more „new authors“, more forgotten (never-heard) novels.



Why all this work?

- Because I love novels
- Also, this will be a unique corpus
- Ultimately it will contain all novels (novellas and similar) written in Serbian 1840-1920
- It will be used to improve morphosyntactic tagging and NER for Serbian, leading to level 2 encoding (at least for some novels);
- Potential usage is practically unlimited:
 - actual reading
 - close reading
 - distant reading.