Корпус *SrpELTeC*, његово анотирање и коришћење

проф. др Цветана Крстев
Универзитет у Београду, Филолошки факултет
Друштво за језичке ресурсе и технологије

Радионица „Увод у дигиталну хуманистику“
Филолошки факултет, Београд, 9. април 2024.

Нешто о називу корпуса

- ***ELTeC*** – вишејезични корпус проистекао из COST **акције CA16204 - *Distant Reading for European Literary History*** – Удаљено читање за европску књижевну историју (2017-2022) који садржи 12 комплетних подколекција (100 романа) и 9 непотпуних (<100 романа);
- ***SrpELTeC*** – српска подколекција од 100 романа
- ***SrpELTeC-ext*** – додатних 20 романа припремљених у оквиру **COST акције d-reading**
- ***SrpELTEC-plus*** – корпус у настајању који ће садржати SrpELTeC и SrpELTeC-ext али ће се ширити по разним димензијама: обухватаће друге временске периоде и друге текстуалне жанрове (приповетке, мемоаре, путописе итд.) – тренутно садржи додатна 44 текста.

Критеријуми за израду корпуса *ELTeC*

- Текстови који улазе у корпус треба да буду **романи** (или нешто што би се тако могло назвати)
  - долазе у обзир наративни текстови од најмање **10.000 речи**, књижевна дела која описују измишљене догађаје и људе (романи или приповетке, а не путописи, мемоари и сл.)
- Долазе у облик дела први пут објављена у **периоду 1840-1920.** (тиме се избегавају проблеми с ауторским правима приликом обраде, дистрибуције корпуса и публиковања резултата)
- За укључивање неког дела у одређену језичку подколекцију потребно је да је **оно оригинално написано на том језику**.

Балансираност корпуса

- Једна од карактеристика корпуса *ELTeC* је да се он састоји од више подколекција, тачно 21 за **21 европски језик**, а све ове подколекције поштују (колико је могуће) неке заједничке критеријуме.
  - Свака подколекција би требало да има тачно **100 романа** (таквих је 12);
  - Пожељна је једнака заступљеност **мушких и женских аутора**;
  - Пожељно је да романи **равномерно покривају изабрани период 1840-1920**. Требало би да буде подједнак број дела објављених у интервалима 1840-1859, 1860-1879, 1880-1899, 1900-1920.
  - Пожељно је да романи буду **разноврсни по питању дужине**: подједнак број кратих дела (до 50.000 речи), средње дугачких (више од 50.000 а мање од 100.000 речи) и дугачких (више од 150.000 речи)
  - Пожељно је да у свакој подколекцији буде **познатих и признатих дела** (дела која су у канону) као и оних **мање познатих или сасвим непознатих** (једном објављени)
  - Пожељна је **разноврсност аутора**: 10 аутора би требало да буде представљено са по 3 дела, док сви остали треба да буду представљени само с једним делом.

Да ли је лако саставити подколекцију која поштује ове критеријуме?

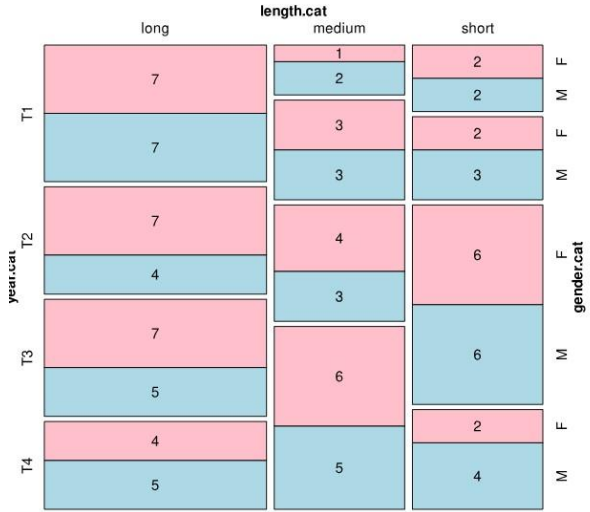
- Ни најмање, чак ни за „велике“ језике, а камоли за српски за који је роман у почетку назначеног периода био тек у настајању.
  - По окончању COST акције *Distant Reading* колекција *ELTeC* има **21 подколекција за 21 европски језик.**
  - Од ове 21 подколекције само **12 подколекција се састоји од 100 романа**.
    - међу њима је и **српски**.
  - Од ових 12 подколекција само **6 подколекција** има готово **савршено задовољене** све наведене **критеријуме**. То су:
    - **немачки,**
    - **енглески,**
    - **француски,**
    - **мађарски,**
    - **португалски,**
    - **шпански**.

Изазови задовољавања критеријума балансираности

- Највећи изазови са српску подколекцију су:
- **проналажење потребног броја дела за почетне периоде**, пронађена су само 2 дела публикована у периоду 1840-1859. и 18 у периоду 1860-1879,
- **проналажење женских аутора**: уврштено је само 8 дела која су написала 4 различита женска аутора:
  - Исидора Секулић, Јелена Димитријевић, Драга Гавриловић и Милица Јанковић;
- **проналажење „дугачких дела“**: уврштено је само 6 која имају више од 100.000 речи
- **разноврсност аутора**: уместо 70 аутора представљених са по једним делом, у српској подколекцији је 48 аутора представљених са по једним делом, остали су представљени са 2, 3 или 5 дела:
  - Јаша Игњатовић

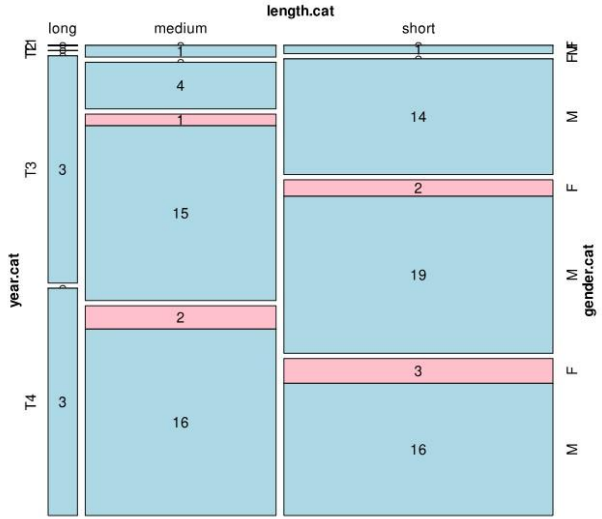
Визуелни приказ балансираности корпуса

ИДЕАЛНА БАЛАНСИРАНОСТ (ЕНГЛЕСКИ)



12.227.703 РЕЧИ

СРПСКИ ELTEC



4.931.503 РЕЧИ

Изазови дигитализације

- Највећи број дела која су уврштена у колекцију *SrpELTeC* није раније била дигитализована
  - или јесте, али није доступна или нема података како је извршена дигитализација ни које издање је коришћено;
- За дигитализацију су коришћена кад год је то могуће **прва издања** или најстарије доступно издање;
  - Први корак: **сканирање;**
  - Други корак: **OCR – оптичко препознавање карактера;**
  - Трећи корак: **корекција** (комбинација аутоматксе и „ручне“ корекције);
  - Четврти корак: **основна анотација** (поглавља, пасуси, фусноте, делови на страном језику, делови у курзиву или друкчије истакнути);
  - Пети корак: **метаподаци** о првом издању и о издању које је дигитализовано; опис дела (дужина, период, пол аутора, каноничност; везе са спољним изворима (COBISS, VIAF, Wikipedia, Wikidata); људи и организације који су допринели дигитализацији;
  - Шести корак: **аутоматска напредна анотација** (врсте речи, леме, именовани ентитети).

Значај и изазов сканирања првих издања

- Недостаци:
  - теже доступна;
  - резултат OCR-а је лошији;
  - стари правопис, стара транскрипција властитих имена, тежи за читање савременом читаоцу.
- Предности:
  - добија се оригинални, првобитни текст;
  - могућа је анализа језика, изучавање промена у језику у времену, и сл.;
  - могућа је израда савремених издања поштујући различите критеријуме.

Анотација – основна структура текста поштује TEI – *Text Encoding Initiavtive Guidelines* (1)

```
<text>
   <front>
      <div>
         ...
      </div>
   </front>
   <body>
      <div>
         ...
      </div>
   </body>
   <back>
      <div>
         ...
      </div>
   </back>
</text>
```

```
<text>
  <front>
    <div type="titlepage" xml:id="SRP18631_F1">
      <p>ПРИПОВЕТКЕ</p>
      <p>ВЛАДАНА ЂОРЂЕВИЋА</p>
      <p>КЊИГА ПРВА:</p>
      <p>КОЧИНА КРАЈИНА.</p>
      <p>У БЕОГРАДУ</p>
      <p>У ДРЖАВНОЈ ШТАМПАРИЈИ.</p>
    </div>
    …
    <div type="liminal" xml:id="SRP18631_F5">
      <p>Молим многопоштоване читаоце, да ми допусте рећи напред неколико речи, колико за роман толико и за ме самог.</p>
…
    </div>
  </front>
```

(из SRP18631 – „Кочина крајина “ Владан Ђорђевић)

Анотација – основна структура текста (2)

```
<text>
   <front>
      <div>
         ...
      </div>
   </front>
   <body>
      <div>
         ...
      </div>
   </body>
   <back>
      <div>
         ...
      </div>
   </back>
</text>
```

```
<body>
  <pb n="1"/>
  <div type="chapter" xml:id="SRP18631_C1">
    <head>I</head>
    <p>Кроз стотине година од како је пропала српска царевина, кроз стотине година тешка српска робовања, било је много јада и чемера; србињска мајка пролила је море од суза, од њених уздаха за красним српским соколовима, што као жртве турскога ћеифа изгибоше, помрча плаво небо, помрча сјајно сунце; али опет — све те страхоте не беху ништа спрам оних које снађоше Србина у ово време, о ком хоћу да причам.</p>
    <p>Али како ћу да причам?... Каким речима да искажем оно, што се једва помислити даје?...</p>
    ...
    </div>
    </body>
```

(из SRP18631 – „Кочина крајина“ Владан Ђорђевић)

Анотација – основна структура текста (3)

```
<text>
   <front>
      <div>
         ...
      </div>
   </front>
   <body>
      <div>
         ...
      </div>
   </body>
   <back>
      <div>
         ...
      </div>
   </back>
</text>
```

```
<back>
  <div type="notes"  xml:id="SRP18631_B1">
    <note xml:id="SRP18631_N1">Мој господине.</note>
    <note xml:id="SRP18631_N2">Смрсити коме конце.</note>
    <note xml:id="SRP18631_N3">Играчице.</note>
    <note xml:id="SRP18631_N4">Соколе.</note>
    <note xml:id="SRP18631_N5">Пријатељ.</note>
    <note xml:id="SRP18631_N6">Ово је историјска истина.</note>
  </div>
</back>
</text>
```

(из SRP18631 – „Кочина крајина “ Владан Ђорђевић)

Метаподаци у TEI заглављу

```
<teiHeader>
  <fileDesc> <!– подаци о дигиталном издању
-->
    <titleStmt>
      <!– подаци о наслову -->
    </titleStmt>
    <extent>
      <!– величина текста у броју страна и
речима -->
    </extent>
    <publicationStmt>
      <!– подаци о издавању -- >
    </publicationStmt>
    <sourceDesc>
      <!--- подаци о извору -->
    </sourceDesc>
  </fileDesc>
  <profileDesc>
    <!– додатне описне информације -->
  </profileDesc>
  <revisionDesc>
    <!– подаци о изменама -->
  </revisionDesc>
</teiHeader>
```

Напредна анотација

- Напредна анотација је урађена аутоматски:
- први ниво анотације:
  - подела на пасусе (уз помоћ корекције акцијаша);
  - подела на реченице;
  - додела сваком токену његове врсте, за речи је то врста речи;
  - и леме (за речи).
- други ниво анотације:
  - препознавање и анотација 7 именованих ентитета:
    - **PERS** – особе
    - **LOC** – места, локације
    - **DEMO** – демоними
    - **ROLE** – улоге, професије, позиције људи
    - **ORG** – организације
    - **WORK** – називи ауторских дела
    - **EVENT** – догађаји

Пример анотације једне реченице

Бежите, јер Беганова мајка, кажу, да не плаче.

```
<s xml:id="s_SRP19181_2548">
        <w pos="VERB" lemma="бежати" xml:id="w_SRP19181_41166" join="right">Бежите</w>
        <pc pos="PUNCT" xml:id="w_SRP19181_41167">,</pc>
        <w pos="SCONJ" lemma="јер" xml:id="w_SRP19181_41168">јер</w>
        <w pos="ADJ" lemma="беганов" xml:id="w_SRP19181_41169">Беганова</w>
        <w pos="NOUN" lemma="мајка" xml:id="w_SRP19181_41170" join="right">мајка</w>
        <pc pos="PUNCT" xml:id="w_SRP19181_41171">,</pc>
        <pb n="100" />
        <w pos="VERB" lemma="казати" xml:id="w_SRP19181_41172" join="right">кажу</w>
        <pc pos="PUNCT" xml:id="w_SRP19181_41173">,</pc>
        <w pos="SCONJ" lemma="да" xml:id="w_SRP19181_41174">да</w>
        <w pos="PART" lemma="не" xml:id="w_SRP19181_41175">не</w>
        <w pos="VERB" lemma="плакати" xml:id="w_SRP19181_41176" join="right">плаче</w>
        <pc pos="PUNCT" xml:id="w_SRP19181_41177">.</pc>
</s>
```

(из SRP19181 – „Војник Стојан : недовршен ратни роман“ Драгомир С. Петровић)

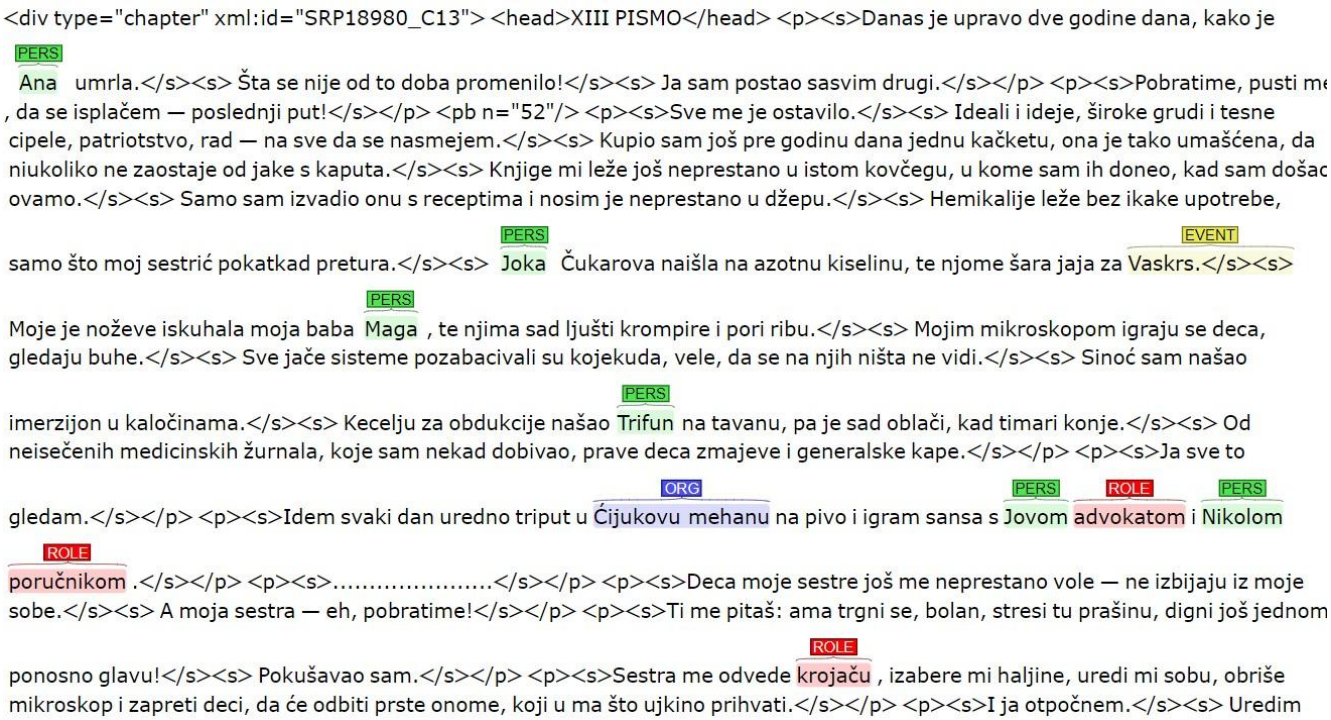
Анотације једне реченице са именованим ентитетима

Ићи ћемо о Илину-дне на Орид.

```
<s xml:id="s_SRP19181_766">
        <w pos="VERB" lemma="ићи" xml:id="w_SRP19181_12236">Ићи</w>
        <w pos="AUX" lemma="хтети" xml:id="w_SRP19181_12237">ћемо</w>
        <w pos="ADP" lemma="о" xml:id="w_SRP19181_12238">о</w>
        <rs type="EVENT">
          <w pos="PROPN" lemma="Илин-дан" xml:id="w_SRP19181_12239">Илину-дне</w>
        </rs>
        <w pos="ADP" lemma="на" xml:id="w_SRP19181_12240">на</w>
        <rs type="LOC">
          <w pos="PROPN" lemma="Орид" xml:id="w_SRP19181_12241" join="right">Орид</w>
        </rs>
        <pc pos="PUNCT" xml:id="w_SRP19181_12242">.</pc>
</s>
```

(из SRP19181 – „Војник Стојан : недовршен ратни роман“ Драгомир С. Петровић)

Визуелна презентација анотација у текст



(из SRP18980 – „Швабица “ Лаза Лазаревић)

Које су разлике а које предности *ELTeC*-а у односу на друге дигиталне репозиторијуме

- Сам одабир текстова;
- Припрема текстова (корекције и анотације);
- Конзистентност – сви текстови су припремљни на исти начин;
- Упоредивост – резултати се могу поредити са другим jezicimaјер су њихови репозиторијуми припремљени на исти начин;
- Све ***ELTeC*** колекције су слободно доступне.

Захвалност

- Библиотекама које су пронашле дела у својим фондовима и сканирали их (за корпус ***SrpELTeC*** и ***SrpELTeC-ext***):
  - Универзитетска библиотека „Светозар Марковић“ (82)
  - Народна библиотека Србије (11)
  - Библиотека Матице српске (3)
  - Библиотека САНУ (2)
  - Дигитална библиотека Народне библиотеке Крушевац (1) (прекуцано)
  - Приватна библиотека Д. Витас и Ц. Крстев (17)
  - Антологија српске књижевности, Учитељски факултет, Универзитет у Београду (4) (преузето)

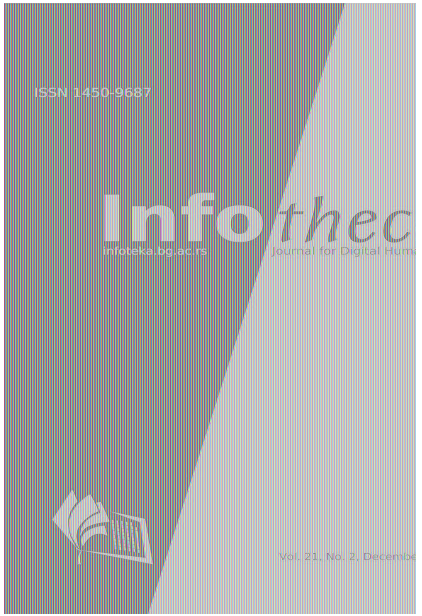
Акцијаши који су кориговали и анотирали текст – 27 учесника

Чланови Друштва за језичке технологије и ресурсе JePTex и
мастер студенти и докторанди ЕУ

акцијаш	бр. р	страна	речи	акцијаш	бр. р	страна	речи
Цветана Крстев	41	8246	1863281	Тања Ћулафић	2	296	68481
Душко Витас	15	2912	647920	Андреа Адамовић	2	503	101640
Иван Обрадовић	13	3021	673724	Александра Јовановић	1	80	10095
Ранка Станковић	7	1192	255718	Александра Марковић	1	204	59480
Бранислава Шандрих	4	805	189575	Аница Милановић	1	125	44760
Оливера Китановић	4	559	130493	Вања Радуловић	1	92	20414
УФАСК	4		223155	Ђорђе Стакић	1	276	51288
Александра Томашевић	3	595	111913	Марина Ђорђевић	1	39	8625
Биљана Лазић	3	549	120890	Милена Михајловић	1	400	71824
Александра Давидовић	2	517	113385	Милица Антић	1	125	29543
Јелена Андоновски	2	213	60262	Милош Утвић	1	249	81827
Јована Димитријевић	2	361	92327	Ненад Зекавица	1	296	60464
Милица Иконић Нешић	2	575	115708	Сергеј Адамов	1	122	22904
Стефан Степановић	2	77	19128	Тамара Радак	1	52	14247
					120	**22481**	**5263071**

Желите да сазнате више о овом корпусу

- Специјалан број часописа за дигиталну библиотеку Инфотека
- Издавачи:
  - Универзитетска библиотека „Светозар Марковић“
  - Филолошки факултет, Универзитет у Београду;
  - Заједница библиотека универзитета у Србији.
- Сви бројеви су слободно доступни:
  - https://infoteka.bg.ac.rs/ojs/index.php/Infoteka/issue/view/21



Шта доноси корпус ELTeC-plus?

- Доста нових романа, посебно допуна опуса аутора:
  - **Јаша Игњатовић, Лазар Комарчић, Пера Тодоровић**;
- Још неки романи/новеле женских аутора:
  - **Милица Јанковић, Анђелија Лазаревић, Милка Гргурова**;
- Путописи:
  - **Љубомир Ненадовић, Чедомиљ Мијатовић, Владан Ђорђевић, Јелена Димитријевић, Марко Цар, Исидора Секулић, Станиша Станишић, Растко Петровић, Богобој Атанацковић;**
- Монографије:
  - **Вук С. Караџић, Коста Н. Костић;**
- Рад се наставља...