RADNA AKCIJA DIGITALNA BIBLIOTEKA 100 SRPSKIH ROMANA

Cvetana Krstev

Jerteh, 22. XI 2018.

D-READING

Cost akcija: CA16204 - Distant Reading for European Literary History

Cost D-reading D-reading

- Trajanje akcije 2017-2021. - početak novembar 2017.
- Sastanci do sada:
  - Sastanak upravnog odbora i radnih grupa, Prag, februar 2018. (Cvetana)
  - Radna grupa W4 (diseminacija, veb strana, reklama) – Amsterdam, januar 2018.
  - Radna grupa W2 (automatskja obrada prikupljenih tekstova – Antverpen, oktobar 2018. (Ranka)
  - Radionica W1 (priprema tekstova za korpus) – Würzburg, april 2018. (Jelena Andonovski)
- Naredni sastanci:
  - Sastanak upravnog odbora i radnih grupa, Lisabon, januar 2019. (Cvetana, Ranka?)
  - Radionica W2 (obrada tekstova) – Galway, Irska, decembar 2019. (Mihailo Škorić)
- STSM misija
  - Rad na alatu TMX (tekstometrija) i uputstvu an engleski – Lion, Francuska, februar 2019. (Jelena Jaćimović)

CILJEVI AKCIJE (IZ MOU)

The Action will:

1. build a multilingual European Literary Text Collection (ELTeC), ultimately containing around 2,500 full-text novels in at least 10 different languages, permitting to test methods and compare results across national traditions;

- Koji su tih 10 jezika: bar Dutch, English, French, German, Modern Greek, Italian, Polish, Portuguese, Russian and Spanish (nema srpskog)

2. establish and share best practices and develop innovative computational methods of text analysis adapted to Europe´s multilingual literary traditions;

3. consider the consequences of such resources and methods for rethinking fundamental concepts in literary theory and history.

WG1 (IZ MOU)

WG 1, “Scholarly Resources”, will coordinate building and publishing the ELTeC (European Literary Text Collection). Its tasks will be to advise partners in structuring the subcollections, including linguistic annotation and metadata collection. Members of WG 1 will primarily come from participants active in computational and corpus linguistics. Expert Meetings will mainly concern discussion of the structure, content, data, annotation and metadata standards for the ELTeC.

- ELTeC je naziv korpusa koji se razvija;
- Korpus će biti slobodno dostupan jednom kad bude razvijen

SADRŽAJ KORPUSA ELTEC

Tokom sastanka u Pragu iskristalisani su jasniji kriterijumi izgradnje ovog korpusa, pre svega što se tiče njegovog sadržaja.

- Korpus se sastoji od više podkorpusa jezika (najmanje predviđenih deset, a verovatno i više);
- Korpus sadrži isključivo romane originalno napisane na jeziku podkorpusa (ili nešto što bi se moglo time nazvati, npr. duže pripovetke)
  - Ne sadrži putopise, eseje, biografije i autobiografije, prevode, istorijske spise ...
- Svi romani moraju da budu iz perioda 1840-1920. (pre svega zbog autorskih prava)
  - Podrazumeva se da su prvi put objavljeni kao zasebno delo u ovom periodu
- Prilikom odabira romana teži se raznovrsnosti:
  - Zadati vremenski period treba da bude ravnomerno pokriven;
  - Korpus treba da sadrži više dela nekih autora (zbog eksperimentisanja sa autorstvom);
  - Trebalo bi da podjedanko (ili bar...) budu zastupljeni muški i ženski autori;
  - Dužina dela u korpusu (merena brojem reči) treba da bude raznovrsna;
  - U korpusu treba da se nađu i kanonizovana dela (poznata, popularna, najkvalitetnija) i zaboravljena dela jednom objavljena (opet zbog raznih eksperimenata).

KRITERIJUMI ODABIRA (1)

Po godinama (idealno 25 u grupi)

- group A (1840-1859)
- group B (1860-1879)
- group C (1880-1899)
- group D (1900-~~1919~~) 1920

Broj izdanja (idealno pola-pola, ali bar 30 po grupi) – pojavile su se i druge ideje, npr. Objavljen je bar jednom 20 godina posle prvog objavljivanja.

- low: no reprints at all, or one reprint (30% of all novels in a language collection)
- high: reprinted more than once (30% of all novels in a language collection)
  - We will not include digitizations of texts in the reprint count.

KRITERIJUMI ODABIRA (2)

Pol autora (bar 10 ženskih):

- male
- female
- mixed (undefined or more than one author)

Dužina romana (bar 20 po grupi):

- short (10k-50k word tokens)
- medium (50k-200k word tokens)
- long (>200k word tokens)
  - I ovde ima drugih ideja; mnogi smatraju da romana sa više od 200K reči neće biti mnogo (ili uopšte) u mnogim književnostima.

ODABIR SRPSKIH 100 ROMANA

Težak zadatak zbog izabranog perioda, da je period pomeren bar do 1930. ili 1940. bilo bi mnogo lakše.

Ipak, uspeli smo za sada da sačinimo listu od 87 (potencijalnih) dela

- To su uradili dr Aleksandra Trtovac i dr Vasilije Milnović iz Univerzitetske biblioteke.

Lista se nalazi ovde: 100 romana

Jedino što mi možemo da uradimo, a to je da poštujemo „obavezne kriterijume“:

- Roman ili novela
- Dužina > 10K
- Period 1840-1920

KAKO STOJI SRPSKA LISTA PO KRITERIJUMIMA

Periodi (25 po grupi) – total ≠ 87 jer za neka dela još nismo sigurni oko godine, bićemo sigurni tek kad delo „uzmemo u ruke“.

Grupa	Od	Do	ukupno	%
A	1840	1859	2	2.5
B	1860	1869	9	11.25
C	1880	1899	32	39.02
D	1900	1920	39	47.56
total			82	

KAKO STOJI SRPSKA LISTA PO KRITERIJUMIMA

Muški/ženski autori

Grupa	Broj	%
Male	81	93.1
Female	6	6.9
Mixed	0	0
	87	

KAKO STOJI SRPSKA LISTA PO KRITERIJUMIMA

Dužina dela - srpski romani (iz ovog perioda) nisu suviše dugački

Dužina nije poznata za mnoge romane – nju možemo znati tek kada se delo skanira i oceruje (!?)

Grupa	Broj	ukupno	%
short	(10k,50k)	17	62.96
medium	(50K,200K)	11	39.29
long	>200K	0	0
total		28	

ŠTA TREBA URADITI S ODABRANIM ROMANIMA

Ovako izgleda tok rada:

1. Da li delo postoji u digitalnom obliku? Ako postoji, prelazi se na korak 7 (ako su korektno uređeni)
2. Ako ne postoji, treba pronaći štampano delo – srećom najveći deo postoji u fondu Univerzitetske biblioteke, ostalo se može naći u drugim bibliotekama preko međubibliotečke pozajmice (SANU, MS)
3. Delo treba da se skanira (najveći deo će biti urađen u Univerzitetskoj biblioteci)
4. Treba uraditi OCR da bi se iz slike dobio tekst (Cvetana)
5. Treba obaviti automatsku korekciju teksta, koliko je moguće (Cvetana)
6. Treba obaviti ostale korekcije čitanjem (akcijaši) i ponovnom kontrolom (Cvetana)
7. Tekst treba na adekvatan način obeležiti, u skladu s preporukama ove akcije, ujednačeno za sve tekstove korpusa ELTeC (varijanta TEI) (delom akcijaši, delom automatski, delom Branislava)
8. Svaki tekst treba da dobije zaglavlje s metapodacima, koje je takođe u skladu s TEI i ELTeC-om (Jelena Andonovski)
9. Posle slede svakojake obrade, ali o tome potom...

ŠTA ZNAČI OBELEŽAVANJE TEKSTA U SKLADU S TEI

To znači da treba obeležiti samo osnovne strukturne elemente i osnovne tekstualne elemente.

Pomenuću samo one koji se tiču akcijaša:

Strukturni elementi:

- Najosnovniji je element (division) koji će se naći na početku svakog poglavlja, a ako je roman podeljen i u delove, onda i na početku svakog dela ;
- Ako se u romanu pojavljuje nešto što nalikuje na pesmu – u vidu zasebnih grupa redova – koriste se obeležja <lg> (line group) za celu „pesmu“ ili „strofu“ i <l> (line) za pojedinačne redove;
- Oznaka mesta na kojem počinje nova stranica u štampanom delu (<pb n="55"/>);
- Podela u pasuse <p> će biti obavljena automatski.

ŠTA ZNAČI OBELEŽAVANJE TEKSTA U SKLADU S TEI

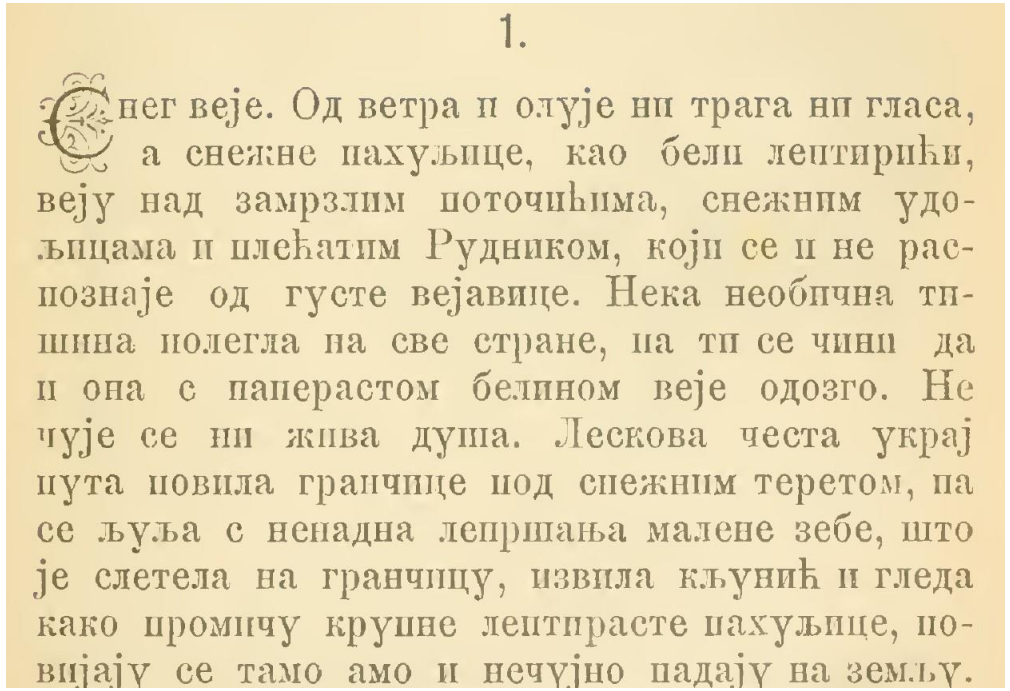
Tekstualni elementi (koji se tiču akcijaša)

- U tekstu se pominje neki naslov (najčešće je dat u kurzivu) – naslov novina, knjige, pozorišne predstave... – koristi se etiketa <title>;
- U tekstu se pojavljuje tekst na stranom jeziku (i on može biti u kurzivu, ali ne mora) – koristi se etiketa <foreign> kojoj se mora dodati atribut jezika <foreign lang="FR">;
- Nešto što je u tekstu nekako istaknuto (kurziv, masna slova, podvučeno, povećan font usred teksta...) a nije ništa od prethodnog, koristi se etiketa <hi> (highlighted);
- Ako se u tekstu pojavljuju autorske fusnote treba i njih obeležiti – kako? Biće objašnjeno kasnije.

Zašto ovo treba da rade akcijaši?

Jer će oni čitati ceo tekst, i neće im biti veliki problem da to urade. Kasnije vraćanje na sve to oduzima jako puno vremena.

KAKO IZGLEDA REZULTAT KORAKA SKAN I OCR *HADŽI-ĐERA* DRAGUTIN ILIĆ

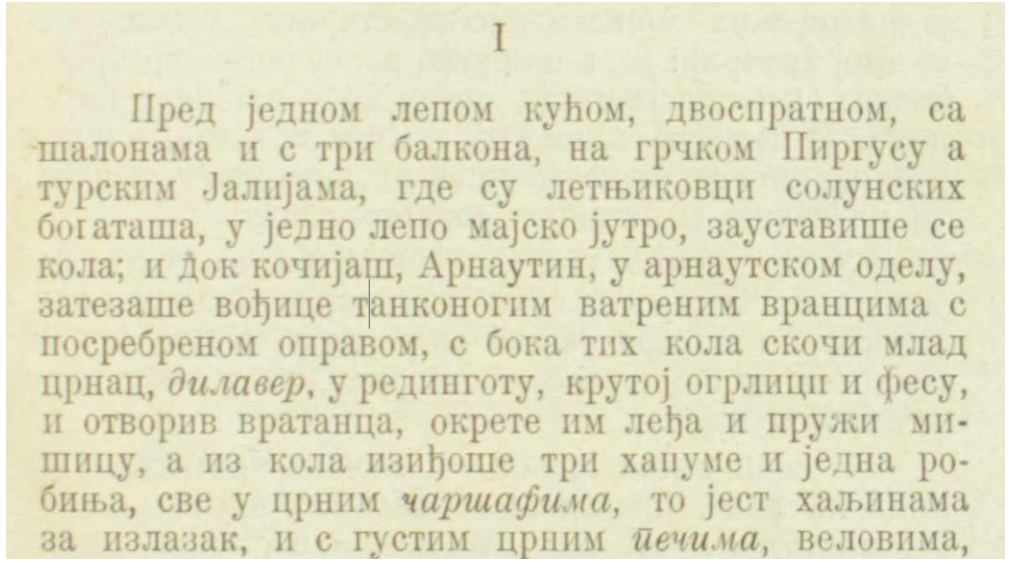


1.

Снег веје. Од ветра и олује ни трага ни гласа, а снежне пахуљице, као бели лептирићи, вију над замрзлим поточићима, снежним удољицама и плећатим Рудником, који се и не распознаје од густе вејавице. Нека необична тишина полегла на све стране, па ти се чини да и она с паперастом белином веје одозго. Не чује се ни жива душа. Лескова чета украј пута повила гранчице под снежним теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, извила кљунић и гледа како промичу крупне лептирасте пахуљице, повијају се тамо амо и нечујно падају на земљу.

•J®.нег веје. Од ветра п о.тује нп трага нп гласа, а снежне пахуљпце, као белп лептирпћп, вију над замрзлпм поточпћнма, снежнпм удо- љпцама п плећатпм Руднпком, којп се п не рас- познаје од густе вејавице. Нека необпчна тп- шипа полегла на све стране, иа тп се чини да и она с паперастом белпном веје одозго. Не чује се пп жнва душа. Лескова чета украј цута повила гранчпце под спежнм теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, пзвпла кљунић и гледа како промичу крупне лептпрасте пахуљпце, по- впјају се тамо амо и нечујно падају на зем.ву.

KAKO IZGLEDA REZULTAT KORAKA SKAN I OCR *NOVE* JELENA DIMITRIJEVIĆ

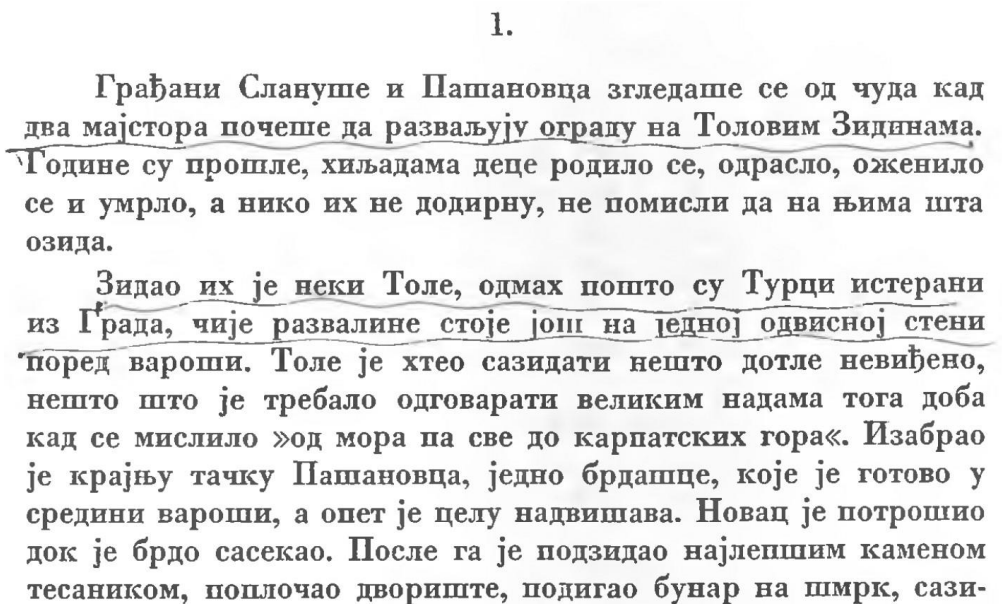


I

Пред једном лепом кућом, двоспратном, са шалонама и с три балкона, на грчком Пиргусу а турским Јалијама, где су летњиковци солунских богаташа, у једно лепо мајско јутро, зауставише се кола; и док кочијаш, Арнаутин, у арнаутском оделу, затезаше вођице танконогим ватреним вранцима с посребреном оправом, с бока тих кола скочи млад црнац, *дилавер*, у рединготу, крутој огрлици и фесу, и отворив вратанца, окрете им леђа и пружи мишицу, а из кола изиђоше три хапуме и једна робиња, све у црним *чаршафима*, то јест хаљинама за излазак, и с густим црним *печима*, веловима,

Пред једном лепом кућом, двоспратном, са шалонама и с три балкона, на грчком Пиргусу а турскнм -Јалнјама, где су летњиковци солунских богаташа, у једно лепо мајско јутро, зауставнше се кола; н Док кочијаш, Арнаутин, у арнаутском оделу, затезаше вођице танконогпм ватреним враицима с посребреном оправом, с бока тпх кола скочи млад црнац, дилавер, у редипготу, крутој огрлицп и фесу, н отворив вратанца, окрете им леђа и пружн мн- шицу, а из кола изиђоше трп хапуме и једна робиња, све у црнпм чаршафима, го јест хаљинама за излазак, и с густим црним Печима, веловима

KAKO IZGLEDA REZULTAT KORAKA SKAN I OCR *POTROŠENE REČI* MILUTIN USKOKOVIĆ



1.

Грађани Слануше и Пашановца згледаше се од чуда кад два мајстора почеше да развађују ограду на Толовим Зидинама. Године су прошле, хиљадама деце родило се, одрасло, оженило се и умрло, а нико их не додирну, не помисли да на њима шта озида.

Зидао их је неки Толе, одмах пошто су Турци истерани из Града, чије развалине стоје још на једној одвисној стени поред вароши. Толе је хтео сазидати нешто дотле невиђено, нешто што је требало одговарати великим надама тога доба кад се мислило »од мора па све до карпатских гора«. Изабрао је крајњу тачку Пашановца, једно брдашце, које је готово у средини вароши, а опет је целу надвишава. Новац је потрошио док је брдо сасекао. После га је подзидао најлепшим каменом тесаником, поплочао двориште, подигао бунар на шмрк, сази-

Грађани Слануше и Пашановца згледаше се од чуда кад два мајстора почеше да развађују ограду на Толовим Зидинама. ^Тодине су прошле, хиљадама деце родило се, одрасло, оженило се и умрло, а нико их не додирну, не помисли да на њима шта озида.

Зидао их ^је^ неки^ Толе, одмах пошто су Турци истерани из Г^рада, чије развалине стоје још на~1едно]~~ЈДВИСНОЈ стени 'поред вароши. Толе је хтео сазидати нешто дотле невиђено, нешто што је требало одговарати великим надама тога доба кад се мислило »од мора па све до карпатских гора«. Изабрао је крајњу тачку Пашановца, једно брдашце, које је готово у средини вароши, а опет је целу надвишава. Новац је потрошио док је брдо сасекао. После га је подзидао најлепшим каменом тесаником, поплочао двориште, подигао бунар на шмрк

KAKO IZGLEDA TEKST POSLE AUTOMATSKE KOREKCIJE — *HADŽI-ĐERA*

•J®.нег веје. Од ветра п о.тује нп трага нп гласа, а снежне пахуљпце, као белп лептирпћп, вејy над замрзлпм поточпћнма, снежнпм удо- љпцама п плећатпм Руднпком, којп се п не рас- познаје од густе вејавице. Нека необпчна тп- шипа полегла на све стране, иа тп се чини да и она с паперастом белпном веје одозго. Не чује се пп жнва душа. Лескова честа украј цута повила гранчпце под спежннм теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, пзвпла кљунић и гледа како промичу крупне лептпрасте пахуљпце, по-впјају се тамо амо и нечујно падају на зем.ву.

•***J®.***нег веје. Од ветра и о.тује +++пи+++ни+++ии+++ трага +++пи+++ни+++ии+++ гласа, а снежне +++пахуљице+++, као +++бели+++ +++лептирићи+++, вејy над +++замрзлим+++ +++поточићима+++, +++снежним+++ +++удољицама+++ и +++плећатим+++ +++Рудником+++, +++који+++ се и не +++распознаје+++ од густе вејавице. Нека +++необична+++ +++тишина+++ полегла на све стране, иа +++ти+++ се чини да и она с ***паперастом +++белином+++ веје одозго. Не чује се +++пи+++ни+++ии+++ +++жива+++ душа. Лескова честа украј ***цута повила +++гранчице+++ под +++снежним+++ теретом, па се љуља с ненадна лепршања малене зебе, што је слетела на гранчицу, +++извила+++ ***кљунић и гледа како промичу крупне +++лептирасте+++ +++пахуљице+++, +++повијају+++ се тамо амо и нечујно падају на ***земву.

KAKO IZGLEDA TEKST POSLE AUTOMATSKE KOREKCIJE – *MOVE*

Пред једном лепом кућом, двоспратном, са шалонама и с три балкона, на грчком Пиргусу а турскнм -Јалнјама, где су летњиковци солунских богаташа, у једно лепо мајско јутро, зауставнше се кола; н Док кочијаш, Арнаутин, у арнаутском оделу, затезаше вођице танконогпм ватреним враицима с посребреном оправом, с бока тпх кола скочи млад црнац, дилавер, у редипготу, крутој огрлицп и фесу, н отворив вратанца, окрете им леђа и пружн мн- шицу, а из кола изиђоше трп хапуме и једна ро- биња, све у црнпм чаршафима, го јест хаљинама за излазак, и с густим црним Печима, веловима

Пред једном лепом кућом, двоспратном, са ***шалонама и с три балкона, на грчком ***Пиргусу а +++турским+++ - +++Јалијама+++, где су летњиковци солунских богаташа, у једно лепо мајско јутро, +++зауставише+++ се кола; и Док кочијаш, Арнаутин, у арнаутском оделу, затезаше вођице +++танконогим+++ ватреним +++врапцима+++вранцима+++ с посребреном ***оправом, с бока +++тих+++ кола скочи млад црнац, ***дилавер, у +++рединготу+++, крутој +++огрлици+++ и фесу, и ***отворив вратанца, окрете им леђа и +++пружи+++ +++мишицу+++, а из кола изиђоше трп +++хануме+++ и једна +++робиња+++, све у +++црпим+++црним+++ чаршафима, го јест хаљинама за излазак, и с густим црним ***Печима, веловима

KAKO FUNKCIONIŠE AUTOMATSKA KOREKCIJA

Ukratko, prvo treba ustanoviti koja su to slova koja prave problem;

- I kod *Hadži-Đere* i kod *Novih* najveći problem su ćirilična slova **п и н** koja se međusobno brkaju u raznim kombinacijama
- Zatim su česte greške **е с - љ њ - ђ ћ – га ш**
- Na osnovu toga se naprave grafovi za kritične kombinacije
- Za svaku neprepoznatu reč primenjuju se sve moguće zamene
  - (***кнша)*_*киша*_*кшиа*_*кпша*_*киша*_*кнша
  - (***крађн)*_*крађп*_*крађп*_*крађи*_*крађи*_*крађн*_*крађн
- One koje nisu reči u srpskom (prema e-rečniku) se odbacuju, pa ostaje
  - +++киша+++
  - +++крађи+++крађи+++
- Reči koje nisu mogle biti ispravljene – ili nisu u rečniku ili je „nepreviđena greška“ ostaju takve kakve su
  - ***мухалебије

KAKO FUNKCIONIŠE AUTOMATSKA KOREKCIJA

Automatska procedura uz korekciju i spaja reči koje su bile rastavljene na kraju retka, pa na takve reči primenjuje postupak korekcije (bilo da su pogrešne ili ne, jer to se nije moglo u stratu ustanoviti)

- primer 1 ноћ- ннм
  - (***ноћ-
    ннм)*_*ноћпим*_*ноћпим*_*иоћпнм*_*иоћпнм*_*иоћпим*_*иоћпим*_*ноћпнм*_*ноћпнм*_*ноћипм*_*ноћипм*_*иоћнпм*_*иоћнпм*_*иоћипм*_*иоћипм*_*ноћнпм*_*ноћнпм*_*иоћппм*_*иоћппм*_*ноћппм*_*ноћппм*_*поћинм*_*поћинм*_*поћним*_*поћним*_*поћиим*_*поћиим*_*поћннм*_*поћннм*_*поћпим*_*поћпим*_*поћпнм*_*поћпнм*_*ноћинм*_*ноћинм*_*ноћним*_*ноћним*_*ноћиим*_*ноћиим*_*иоћннм*_*иоћннм*_*иоћинм*_*иоћинм*_*ноћннм*_*ноћннм
  - Rezultat: +++ноћним+++
- Primer 2: ха- нуме
  - (***ха- нуме)*_*хапумс*_*хапуме*_*хаиумс*_*хаиуме*_*ханумс*_*хануме
  - Rezultat: +++хануме+++

AUTOMATSKA KOREKCIJA U BROJKAMA

Hadži-Đera - broj različitih nepoznatih reči (potencijalne greške)

- Posle prvih korekcija („л>“ u „љ“) – 9.062;
- Posle automatske korekcije – 1.340 (14.8%);
- Posle čitanja – 608 (akcijaši)
- Posle dopune rečnika – 177
- Novih reči u rečniku 339

Nove – broj različitih nepoznatih reči (potencijalne greške)

- Posle OCR nepoznatih reči - 11.397
- Posle automatske korekcije – 1.307 (11,5%)
- Posle čitanja – 1.089 (akcijaši)
- Posle dopune rečnika – 451
  - Zapravo još manje jer su ovde uključene i reči na francuskom i engleskom kojih ima dosta
- Novih reči u rečniku 346

ŠTA ĆE I KAKO RADITI AKCIJAŠI

Грађани Сланушe и Пашановца згледаше се од чуда кад два мајстора почеше да развaљују ограду на Толовим Зидинама. ^Тодине су прошле, хиљадама деце родило се, одрасло, оженило се и умрло, а нико их не додирну, не помисли да на њима шта озида.

ČITANJE I KORIGOVANJE

Пред једном лепом кућом, двоспратном, са ***шалонама и с три балкона, на грчком ***Пиргусу а +++турским+++ -+++Јалијама+++, где су летњиковци солунских богаташа, у једно лепо мајско јутро, +++зауставише+++ се кола; и Док кочијаш, Арнаутин, у арнаутском оделу, затезаше вођице +++танконогим+++ ватреним +++врапцима+++вранцима+++ с посребреном ***оправом, с бока +++тих+++ кола скочи млад црнац, ***дилавер, у +++рединготу+++, крутој +++огрлици+++ и фесу, и ***отворив вратанца, окрете им леђа и +++пружи+++ +++мишицу+++, а из кола изиђоше трп +++хануме+++ и једна +++робиња+++, све у +++црпим+++црним+++ чаршафима, го јест хаљинама за излазак, и с густим црним ***Печима, веловима

Svako će dobiti jedan tekst u obliku .txt dokumenta (posle automatske korekcije) i original, tj. skaniranu sliku.

- .txt dokument treba učitati i korigovati u programu po zboru – ja preporučujem i koristim Notepad++
  - Može se koristiti i Word, ali tada treba uključiti da budu vidljivi „skriveni karakteri“ i ne treba namerno ili nenamerno unositi bilo kakvo kakvo fomatiranje (italik, veći font i sl.).
- Skanirani dokument služi za korekciju – NEPOHODAN JE – možete ga i odštampati ako je tako lakše.

Kako će izgledati .txt

- Kao korigovani tekst na slajdovima 18 i 19
- Kao korigovani tekst na slajdu 17 – ovaj je posle OCR-a bio solidan pa su korigovane samo prekinute reči na kraju retka (spojene su, ako spojene daju reč iz rečnika).

KAKO ĆE AKCIJAŠI POSTUPATI (SLUČAJ 2)

U drugom, lakšem, slučaju treba samo čitati i ispravljati greške na koje se naiđe.

- Biće potrebno da se gleda u skanirani tekst samo izuzetno, kad nije baš jasno da li je u pitanju greška ili neka baš nepoznata reč (u ovim starim tekstovima će ih uvek biti baš dosta, neminovno)
- Posebno će obratiti pažnju na reči označene s *** ispred – to su spojene reči na kraju retka, može da promakne greška, tj. da ih nije trebalo spojiti.
  - Npr. posle OCR je bilo **ђу- туре**, posle automatske korekcije ***ђутуре – ostaje tako kako je
- Biće i slučajeva kad reči na kraju retka nisu spojene – to će se desiti kada spojene daju (za rečnik) nepoznatu reč. U tim slučajevima ih treba spojiti i obaviti potrebne korekcije,
  - Npr. Posle OCR je bilo **дунђе- рин**, posle automatske korekcije ***дунђе- ***рин (jer ni **дунђерин**, ni **дунђе** ni **рин** nisu u rečniku) treba spojiti u ***дунђерин.
  - Npr. Polse OCR je bilo **основ- ј ном**, posle automatske korekcije **основ-** ***ј ***ном** (jer **основ** jeste u rečniku, a **ј** i **ном** nisu); treba izbaciti **ј** viška i spojiti u **основном**.
- Kakav je značaj *** kada je neki deo već pročitan i po potrebi korigovan? Nikakav, mogu se brisati a ne moraju. Akcijaš ne mora da gubi vreme na njihovo brisanje ili dodavanje.

KAKO ĆE AKCIJAŠI POSTUPATI (SLUČAJ 1)

U prvom, malo zahtevnijem slučaju treba čitati i ispravljati greške na koje se naiđe, a posebno obratiti pažnju:

- Na reči označene s *** - to su reči koje nisu pronađene rečnikom, pa mogu biti:
  - Ispravne reči (bar prema skanu!) ali nisu u rečniku ***абдеслуку (posebno mesto u kući gde se uzima *abdest* (Škaljić))
  - Neispravne reči (jer ima nepredviđenih grešaka) ***фе-***мшшзмом – treba popraviti u ***феминизмом; ovde je došlo do greške jer zamena шш u ини kao relativno retka nije predviđena;
  - Spojene prilikom ocerovanja (!) ***збогкорсета – treba ih rastaviti ***због корсета;
- Reči označene s +++reč+++ su korigovane reči, njih će biti najviše i one će najčešće (ali ne i obavezno!) biti ono što na tom mestu i treba da bude:
  - Posebno treba obratiti pažnju na slučajeve gde je za jednu pogrešno ocerovanu (!) reč dobijeno više mogućih ponuda +++reč1+++reč2+++ (može ih nekada biti i 4, a i više). Treba obrisati sve nepotrebne ponude, ostaviti samo pravu:
    - +++нише+++пише+++ - treba zadržati +++нише+++ ako je u pitanju imenica ***niša***, odnosno +++пише+++ ako je u pitanju glagol pisati (posle OCR-a je bilo **ипше**)
- Kao i u prethodnom slučaju, posle pročitanog i korigovanog ni *** ni +++ nemaju više značaj. Akcijaš ne mora da gubi vreme na njihovo brisanje ili dodavanje.

SITUACIJE KOJE ĆE GNJAVITI (SLUČAJ 1)

Biće nekih slučajeva koji će izgledati posebno dosadni za ispravljanje.

- Zbog čestog brkanja slova **п и н** vrlo česta reč **ni** (ćirilično) će biti pogrešno ocerovana, a onda će se kao ispravke dobiti
  - **+++пи+++ни+++ии+++** (uzvici ***pi*** i ***ii***)
- Ovde treba pristupiti funkcijama Find/Replace i zameniti svako **+++пи+++ни+++ии+++** sa **+++ни+++** (možda ipak bez globalne zamene!)
- Posle izvesnog vremena, dočeka vas neprijatno iznenađenje. Opet isto!
  - **+++ни+++ии+++пи+++**
- Vidimo da su ponude iste ali nije redosled; do toga dolazi zato što su u tekstu, u ovom konkretnom slučaju, mogle da budu razne mogućnosti **пп нн пн нп ип ин**. Treba ponovi Find/Replace.
- Neki slučajevi su posebno osetljivi (a česti)
  - **+++пије+++није+++**
- Ovde treba biti strpljiv jer zaista može da bude i jedno i drugo.

O ČEMU POSEBNO TREBA VODITI RAČUNA (SVI)

- Svi skanirani tekstovi će biti na ćirilici (s obzirom da će mahom biti originalna izdanja), pa će se i korekcije vršiti na ćirilici a to znači:

OBAVEZNO TREBA UKLJUČITI ĆIRILIČNU TASTATURU

- Ovo je naročito delikatno jer su neka slova ista pa može izgledati da je korekcija ispravna, a nije (u poslednjoj proveri će rečnik na tim mestima da signalizira, ali bolje je da toga i nema)
- Treba zadržati **TVRDI KRAJ REDA** jer će na osnovu njega automatska procedura obeležiti pasuse.
  - Prilikom ocerovanja program uglavnom zadržava kraj pasusa, ali ne uvek – recimo kod prekida stranice obično prekine pasus.
  - Kako ćemo znati gde je taj „tvrdi kraj reda“?
    - U Notepad++ će svaki „pasus“ biti zaseban red sa zasebnom numeracijom;
    - U Wordu će se to lepo videti ako se uključe „skriveni karakteri“. Kraj pasusa se tada vidi kao karakter ¶

PROBLEM KOJI ĆE SE JAVITI U NEKIM TEKSTOVIMA

- Rečeno je da će skanirani tekstovi biti na ćirilici, pa će se prilikom ocerovanja izabrati opcija „prepoznaj sve kao ćirilicu“ jer se to pokazalo kao najbolje.
- Ali, u nekim tekstovima se pojavljuju manji delovi na stranom jeziku (francuskom, engleskom...).
  - Misli se na delove koji su na stranom jeziku i napisani alfabetom i pravopisom tog jezika, a ne recimo da zvuče turski ali je prilagođeno sprskom pravopisu i pismu.
- Ti delovi s obzirom da je oabrana „samo ćirilica“ nikako ne mogu biti dobri. Izgledaće otprilike ovako:
  - **БопЧ вреак зо јооИзту! Umesto *Don't speak so foolishly!***
- Ni ne liči! Ovde je jedino rešenje da se pogleda u skanirani tekst, uključi latinična tastatura, prekuca tekst na originalnom jeziku, **VRATI ĆIRILIČNA TASTATURA**.
- Kolika je verovatnoća da vas pri korigovanju ovo snađe:
  - Može se desiti da toga uopšte nema (u *Hadži-Đeri* ne pričaju ni na kom stranom jeziku);
  - Malo i veoma retko (*Došljaci*);
  - Poprilično (*Nove* – mladoturkinje znaju i vole da pričaju i engleski i francuski).

ŠTA NE TREBA RADITI

- Ne treba normalizovati tekst. Pod tim se podrazumeva da tekst ne treba prilagođavati ni na koji način današnjem pravopisu. On treba da ostane veran originalu. Primeri:
  - **по крупној неравној чаршискоj калдрми**... Ne treba popravljati **чаршиски** u **чаршијски**;
  - **Биће... како да кажем...нечег из Официрске Задруге**? Ne preba popravljati **Официрске Задруге** u **Официрске задруге**.
- Treba ispraviti samo evidentne greške u kucanju iz same knjige (treba konsultovati, ako postoji, *Errata* na kraju knjige):
  - Npr. U *Nove* se dva puta javilo **налуне** umesto **нануле**. Ja sma zaključila da je greška i da **налуне** nije neki stari naziv za **нануле**. **Pogrešan zaključak! Zato ne treba ispravljati NIŠTA osim ako se našlo u Errata na kraju knjige.**
- Šta bi možda trebalo raditi, ali ja nisam radila:
- U nekim tekstovima se mnogo više nego što je danas običaj reči zapisuju s znakom naglašavanja
  - **и преводи преко *срата*...** Ovde bi trebalo da bude kao umlaut iznad *p*.
  - Trebalo bi to uraditi, ali prvo bi za svaki takav slučaj trebalo pronaći mesto u Unicode-u... U ovom trenutku nemamo resurse za to. Možda jednom kasnije.

AKO ŽELITE DA VIDITE BOJE U NOTEPAD++

Treba sa sajta Jerteha da skinete ELTeC12.zip i da ga otpakujete

U Notepad++ treba da otvorite meni „Language“ a onda u njemu birate opciju „Define your language“;

Otvoriće se nova kartica, u njoj u levom gornjem uglu birate dugme „Import“.

- Pitaće Vas šta da importujete: izabraćete ELTeC1.xml;
- Ponovo ćete pritsnuti dugme „Import“ i izbarati ELTeX2.xml

Izaći ćete iz Notepad++ i ponovo ga pokrenuti sa Vašim tekstom:

- Ako želite da budu obojene samo nekorigovane reči, u meniju „Language“ ćete odabrati jezik „ELTeC1“
- Ako želite da budu obojene i ispravljane reči, u meniju „Language“ ćete odabrati jetik „ELTeC2“.

KAKAV JE ZNAČAJ ČITAČICA I ČITAČA

Da li je zapravo potrebno da akcijaši baš sve čitaju? Zar nije dovoljno da samo idu od *** do *** i ispravljaju potencijalno pogrešne reči?

Značaj čitanja je veliki, jer ima slučajeva, nažalost ne malo, kada se reč oceruje pogrešno, ali u neku drugu legalnu reč. U tom slučaju, nemamo ni naznaku o grešci ni pokušaja ispravljanja.

- Mogle bi da se „ispravljaju“ sve reči, a ne samo neprepoznate, ali to bi tek dalo preterano mnogo nepotrebnih ponuda ispravki.

A to znači, da se mora čitati da bi se te greške uočile. Evo nekih čestih primera:

- **пего** umesto **него** (imenica **пега**);
- **трн** umesto **три**;
- **жепа** umesto **жена** (neka imenica **жеп**).

UNOŠENJE ETIKETA – BROJEVI STRANICA

Brojevi stranica **<pb n="12"/>**

- Ovu etiketu treba uneti na svakom mestu gde u originalu počinje nova stranica – gornja etiketa će biti stavljena na početak stranice 12 (stranice koja je tako numerisana);
- Desiće se da se pasus nastavlja na drugoj stranici – nikakav problem, ova etiketa se stavlja unutar pasusa;
- Ako se reč prekida u poslednjem redu na stranici, onda reč treba spojiti (ili je već automatski spojena), a ovu oznaku uneti iza prelomljene reči.
  - U poslednja dva slučaja ne treba zaboraviti da se spoji pasus, tj. da se ukine „tvrdi kraj reda“.
- Da li je ovo označavanje neophodno?
  - Nije neophodno, ali je vrlo korisno ako je potrebno i posle čitanja uneti još neke ispravke (a uvek je potrebno) da bi se locirale pogrešne reči u skaniranom tekstu.
  - Biće vrlo korisne i za povezivanje teksta s originalom u digitalnoj biblioteci.

UNOŠENJE ETIKETA - POGLAVLJA

Etiketa ** **

- Ovu etiketu treba staviti na početak svakog poglavlja (ako je roman podeljen na poglavlja, što obično jeste; novele možda i nemaju takvu podelu).
- Kada se završi jednog poglavlje, pre početka sledećeg se stavlja oznaka ** **

Poglavlja obično imaju jedan ili više naslova: **<head>**

```
</div>
<div type="chapter" n="4">
<head>ГЛАВА ЧЕТВРТА</head>
<head>АЛИ ЈЕДНОГ ДАНА...</head>
```

```
</div>
<div type="chapter" n="2">
<head>II</head>
```

UNOŠENJE ETIKETA - STIHOVI

Povremeno se u proznom tekstu kakav je roman pojavi deo teksta koji se po prirodi stvari deli u fiksne redove (čija dužina ne zavisi od formata stranice i veličine fonta) – to su recimo neke pesme.

- Za celu „pesmu“ se koristi etiketa <lg> za početak i </lg> za kraj pesme;
- Iste etikete se koriste i za strofe unutar „pesme“;
- Za pojedinačne redove se koristi etiketa <l> za početak reda i </l> za kraj reda.
- Primer

```
<lg>
<l>Црним сам се газом повезала,</l>
<l>Тужна сам ти, сестро, невесела:</l>
<l>Кроз густи сам кафез погледала,</l>
<l>Вид'ла сам га, па га заволела.</l>
</lg>
```

UNOŠENJE ETIKETA – TEKSTUALNI ELEMENTI

Već je rečeno da su to elementi za deo teksta na stranom jeziku, za naslove i sl. i uopšte za neki naglašeni deo.

Primeri:

— <foreign lang="FR">Quelle idée</foreign>!
узвикну она. За кога?

...обасјавало је уредништво <title>Препорода</title>,

Ми смо <hi>нове</hi>, а оне су Европљанке, Францускиње.

UNOŠENJE ETIKETA - FUSNOTE

Treba obeležavati samo autorske fusnote, a ne uredničke.

- Kako ćemo znati koja je koja? Hm... Osim ako ne piše *prim.ur.*

U nekim tekstovima uopšte neće biti fusnota, u drugima vrlo malo, u nekima baš dosta.

U tekstu, na mestu fusnote treba staviti etiketu **<ref target="#note11"/>**, a sam tekst fusnote pisati u **odvojenoj datoteci** za fusnote između etiketa **<note xml:id="note11">** i **</note>**.

- Svi akcijaši će uz svoj tekst dobiti i „model“ ove datoteke za beleške.

```
<hi>Сут-ана<hi>!<ref target="#note11"/> моја слатка сута како лепо име има,...
```

Fusnote treba numerisati redom od jedan pa naviše (bez obzira kako su numerisane u samom tesktu)

Primer:

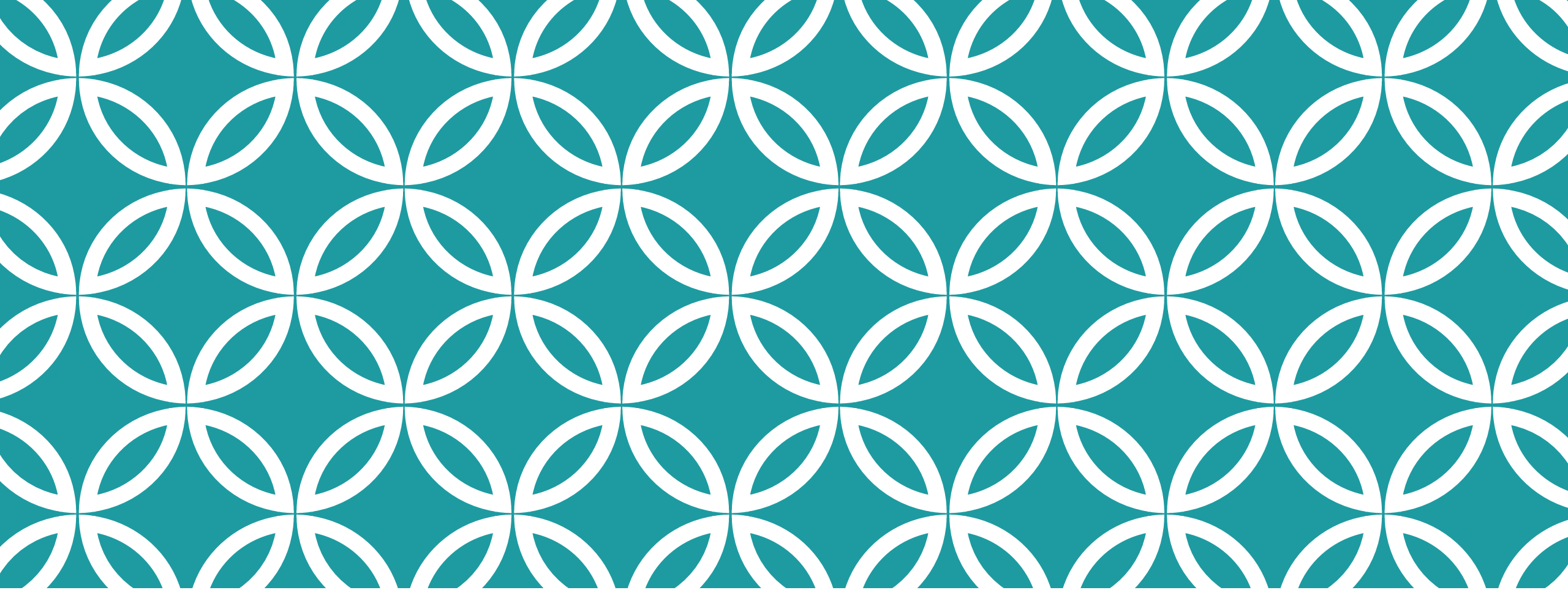
```
<note xml:id=„note11">Госпођа стрина.</note>
```

POMOĆ U RADU

Broj etiketa koje treba uneti nije veliki, ali ipak treba zapamtiti kako svaka tačno izgleda, a osim toga, one su latinične pa treba stalno prebacivati tastaturu s ćirlične na latiničnu i obrnuto.

Ja odvojim zasebnu datoteku sa sledećim sadržajem, i odatle kopiram šta mi je potrebno:

```
</div>
<pb n="283"/>
<div n=“12“>
<head>XII</head>
<hi> </hi>
<lg> </lg>
<l> </l>
<foreign lang="FR">Quelle idée</foreign>
<foreign lang="EN">What?</foreign>
<ref target="#note55"/>
```



ZAŠTO SVE OVO RADIMO?

KOJA JE KORIST ZA SVE?

Inicijalni podstrek je ravnopravno učešće u jednoj COST akciji.

- Naše učešće tamo već nije nezapaženo, mnogi mlađi saradnici su dobili priliku za dalje usavršavanje.

Razviće se korpus kao značajan resurs za raznovsrna lingvistička, filološka i informatička istraživanja.

- Sadržaće materijal koji nije obuhvaćen sa SrpKor (romani pre 1900 godine);
- Taj korpus će biti u skladu sa „state of the art“, što znači da ćemo imati priliku da savladamo nove standarde i alate.

Mnogi akcijaši će imati priliku da nauče nešto novo, što možda do sada nisu radili (a mislili su da je ne znam koliko komplikovano).

Kao i svaka radna akcija pružiće priliku za rad u prijateljskoj atmosferi.

A osim toga...

DIGITALNA BIBLIOTEKA AURORA

Svi tekstovi korpusa ELTeC će posatti deo digitalne biblioteke Aurora koju razvija Jerteh.

Ova digitalna biblioteka već sada sadrži i mnoge druge tekstove koji nisu deo ELTeC

A kada se naša akcija uhoda, sadržaće, nadamo se, još više dela srpske književnosti 1840-1920. i kasnijih.

AURORA

KAKO SE PRIKLJUČITI AKCIJI?

Treba poslati e-mejl na adresu CvetanaJK@gmail.com sa temom „100 romana“.

Dobiće se link ka tabeli u kojoj će biti raspoloživi romani

Treba sačekati da se tekst pripremi

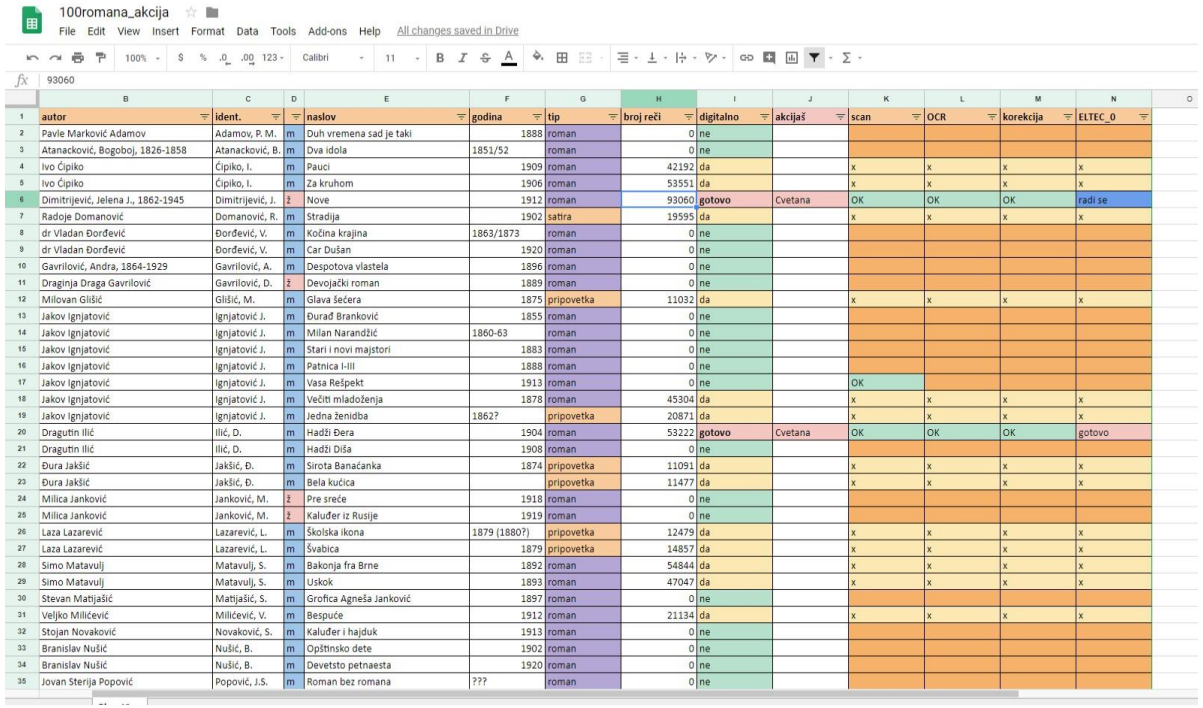
Ali, i Cvetani će biti potrebna pomoć oko pripreme tekstova

- Neko ko se dobro snalazi s Unitex-om i voli grafove
- Dobrovoljci neka se jave...

ŠTA SE NUDI JE U TABELI

Da bi se izabralo delo, treba:

- Filtrirati kolonu „digitalno“ da u njoj stoji samo vrednost „ne“
- Filtrirati kolonu „akcijaš“ da sadržaj bude prazan
- Ako ne možete da čekate, onda filtrirajte kolonu „scan“ da u njoj stoji „OK“
- Izaberite Vaš roman
- Nemojte sami menjati ništa u tabeli, javite šta je Vaš izbor an adresu CvetanaJK@gmail.com



	autor	ident.		naslov	godina	tip	broj reči	digitalno	akcijaš	scan	OCR	korekcija	ELTEC_0
2	Pavle Marković Adamov	Adamov, P. M.	m	Duh vremena sad je taki	1888	roman	0	ne					
3	Atanacković, Bogoboj, 1826-1858	Atanacković, B.	m	Dva idola	1851/52	roman	0	ne					
4	Ivo Ćipiko	Ćipiko, I.	m	Pauci	1909	roman	42192	da		x	x	x	x
5	Ivo Ćipiko	Ćipiko, I.	m	Za kruhom	1906	roman	53551	da		x	x	x	x
6	Dimitrijević, Jelena J., 1862-1945	Dimitrijević, J.	ž	Nove	1912	roman	93060	gotovo	Cvetana	OK	OK	OK	radi se
7	Radoje Domanović	Domanović, R.	m	Stradija	1902	satira	19595	da		x	x	x	x
8	dr Vladan Đorđević	Đorđević, V.	m	Kočina krajina	1863/1873	roman	0	ne					
9	dr Vladan Đorđević	Đorđević, V.	m	Car Dušan	1920	roman	0	ne					
10	Gavrilović, Andra, 1864-1929	Gavrilović, A.	m	Despotova vlastela	1896	roman	0	ne					
11	Draginja Draga Gavrilović	Gavrilović, D.	ž	Devojački roman	1889	roman	0	ne					
12	Milovan Glišić	Glišić, M.	m	Glava šećera	1875	pripovetka	11032	da		x	x	x	x
13	Jakov Ignjatović	Ignjatović J.	m	Đurađ Branković	1855	roman	0	ne					
14	Jakov Ignjatović	Ignjatović J.	m	Milan Narandžić	1860-63	roman	0	ne					
15	Jakov Ignjatović	Ignjatović J.	m	Stari i novi majstori	1883	roman	0	ne					
16	Jakov Ignjatović	Ignjatović J.	m	Patnica I-III	1888	roman	0	ne					
17	Jakov Ignjatović	Ignjatović J.	m	Vasa Rešpekt	1913	roman	0	ne		OK			
18	Jakov Ignjatović	Ignjatović J.	m	Večiti mladoženja	1878	roman	45304	da		x	x	x	x
19	Jakov Ignjatović	Ignjatović J.	m	Jedna ženidba	1862?	pripovetka	20871	da		x	x	x	x
20	Dragutin Ilić	Ilić, D.	m	Hadži Đera	1904	roman	53222	gotovo	Cvetana	OK	OK	OK	gotovo
21	Dragutin Ilić	Ilić, D.	m	Hadži Diša	1908	roman	0	ne					
22	Đura Jakšić	Jakšić, Đ.	m	Sirota Banaćanka	1874	pripovetka	11091	da		x	x	x	x
23	Đura Jakšić	Jakšić, Đ.	m	Bela kućica		pripovetka	11477	da		x	x	x	x
24	Milica Janković	Janković, M.	ž	Pre sreće	1918	roman	0	ne					
25	Milica Janković	Janković, M.	ž	Kaluđer iz Rusije	1919	roman	0	ne					
26	Laza Lazarević	Lazarević, L.	m	Školska ikona	1879 (1880?)	pripovetka	12479	da		x	x	x	x
27	Laza Lazarević	Lazarević, L.	m	Švabica	1879	pripovetka	14857	da		x	x	x	x
28	Simo Matavulj	Matavulj, S.	m	Bakonja fra Brne	1892	roman	54844	da		x	x	x	x
29	Simo Matavulj	Matavulj, S.	m	Uskok	1893	roman	47047	da		x	x	x	x
30	Stevan Matijašić	Matijašić, S.	m	Grofica Agneša Janković	1897	roman	0	ne					
31	Veljko Milićević	Milićević, V.	m	Bespuće	1912	roman	21134	da		x	x	x	x
32	Stojan Novaković	Novaković, S.	m	Kaluđer i hajduk	1913	roman	0	ne					
33	Branislav Nušić	Nušić, B.	m	Opštinsko dete	1902	roman	0	ne					
34	Branislav Nušić	Nušić, B.	m	Devetsto petnaesta	1920	roman	0	ne					
35	Jovan Sterija Popović	Popović, J.S.	m	Roman bez romana	???	roman	0	ne					

DRUGARSKA SE PESMA ORI
PESMA KOJA SLAVI RAD
SRCE GROMKO NEK NAM ZBORI
DA NAM ŽIVI, ŽIVI RAD!!!
PODIGNIMO U VIS ČELA MI - JUNACI RADA SVOG,
NAŠA BIĆE ZEMLJA CELA,
DA NAM ŽIVI, ŽIVI RAD!
U DIVLJAKA LUK I STRELA,
ŽELJEZNICA, SELO I GRAD,
TO SU NAŠIH RUKU DELA,
DA NAM ŽIVI, ŽIVI RAD!!!

WWW.SECANJA.COM