Тесла Text Embeddings - Serbian Language Applications, #7276

TXM: uvod, instalacija, učitavanje korpusa, svojstva, kreiranje particija

Tim projekta TESLA i Društvo za jezičke resurse i tehnologije JeRTeh

Radionica: TESLA - Tekstometrija u praksi

Beograd, RGF, 25.4.2026. 9-15h

https://tesla.rgf.bg.ac.rs/

O čemu će biti reči?

- Reč-dve o TXM-u
- Upravljanje korpusima
  - Instalacija i podešavanje alata TXM
  - Priprema tekstova i metapodataka
  - Kreiranje korpusa (import i load)
  - Kreiranje particija i podkorpusa

TEKSTOMETRIJA

Šta je tekstometrija?

- Metodologija koja omogućava nelinearno proučavanje korpusa
  - kombinujući leksikometrijska i statistička istraživanja sa korpusnim tehnologijama

Svrha tekstometrijske analize

- Opisivanje leksičkih i drugih karakteristika teksta.
- Proučavanje međuodnosa određenih elemenata teksta (kolokacije), leksičkih osobenosti particija (specifičnosti), lociranje elemenata teksta u korpusu (progresija).
- Kvantifikovanje elemenata nije nov pristup, ali je pojednostavljen primenom postojećih alata.

TXM SOFTVER

Grafičko korisničko okruženje

- CQP pretraživač i R statistički paket, bilo koji jezik koji ima model

Odlike

- Korišćenje je besplatno
- Softver otvorenog koda
- Kontinuirano ga razvija tim istraživača IHRIM laboratorije, ENS de Lyon od 2010.

Dostupne su instalacije

- desktop za Windows, Linux i Mac OS X,
- kao i veb platforma dostupna: http://portal.textometrie.org/demo/

NEKE KARAKTERISTIKE TXM SOFTVERA

Jedinice teksta

- tekstovi korpusa (knjige, članci, intervjui, …) koji mogu imati i metapodatke

Vrste korpusa

- Korpusi pisanih tekstova: TXT / XML /TEI
- Paralelni korpusi
- Korpusi transkripata (sinhronizovani sa izvornim audio ili video snimkom)

Opcione strukture teksta

- tekst može da sadrži i unutrašnje strukturne jedinice (odeljci, pasusi, upravni govor, …) koje mogu imati određena svojstva (naslov, broj, …)

Leksičke jedinice

- svaki tekst je sastavljen od niza reči koje mogu imati određena svojstva (oblik, lemu, vrstu reči, …)

NEKE KARAKTERISTIKE TXM SOFTVERA

Tekst se automatski segmentira

- prilikom uvoza u TXM okruženje, POS i lema - TreeTagger

Obrasci

- Bilo koja kombinacija ovih svojstava reči (liste frekvencija, konkordance, vizuelni prika)

Statistički modeli

- distribucije po podkorpusima (faktorska analiza, klaster analiza),
- visoku ili nisku zastupljenost u potkorpusima (analiza specifičnosti), ili analizu kolokacija.

Rezultat svake analize

- može da se izveze u tabelarnom ili grafički

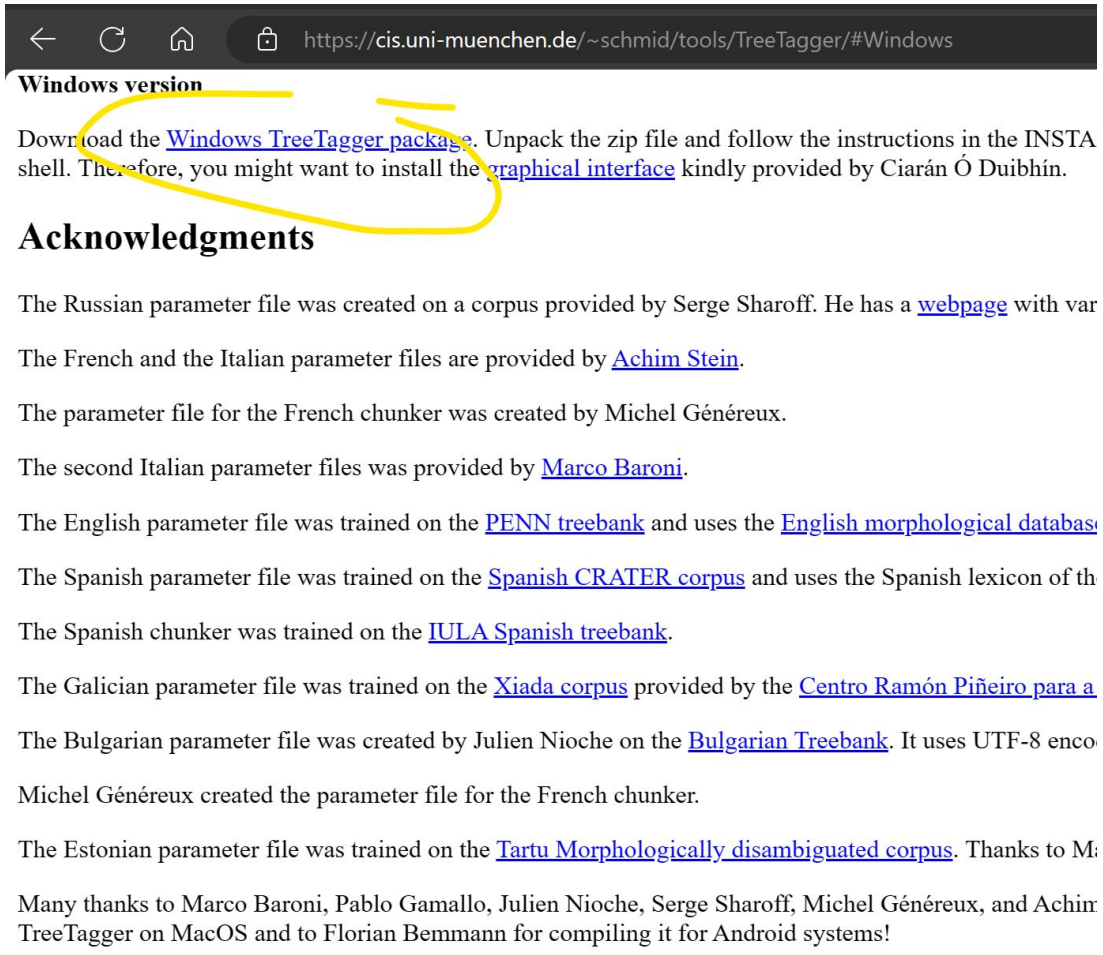
Instalacija TXM-a

- Osnovna strana sa informacijama o TXM-u:
  - FR: https://txm.gitpages.huma-num.fr/textometrie/
  - EN: https://txm.gitpages.huma-num.fr/textometrie/en/
- Aktuelna verzija (0.8.4, februar 2025) preuzeti sa:
  - https://txm.gitpages.huma-num.fr/textometrie/files/software/TXM/0.8.4/en/
  - uraditi podrazumevanu instalaciju, proveriti da ima dovoljno slobodnog mesta na disku

Podešavanje TXM-a

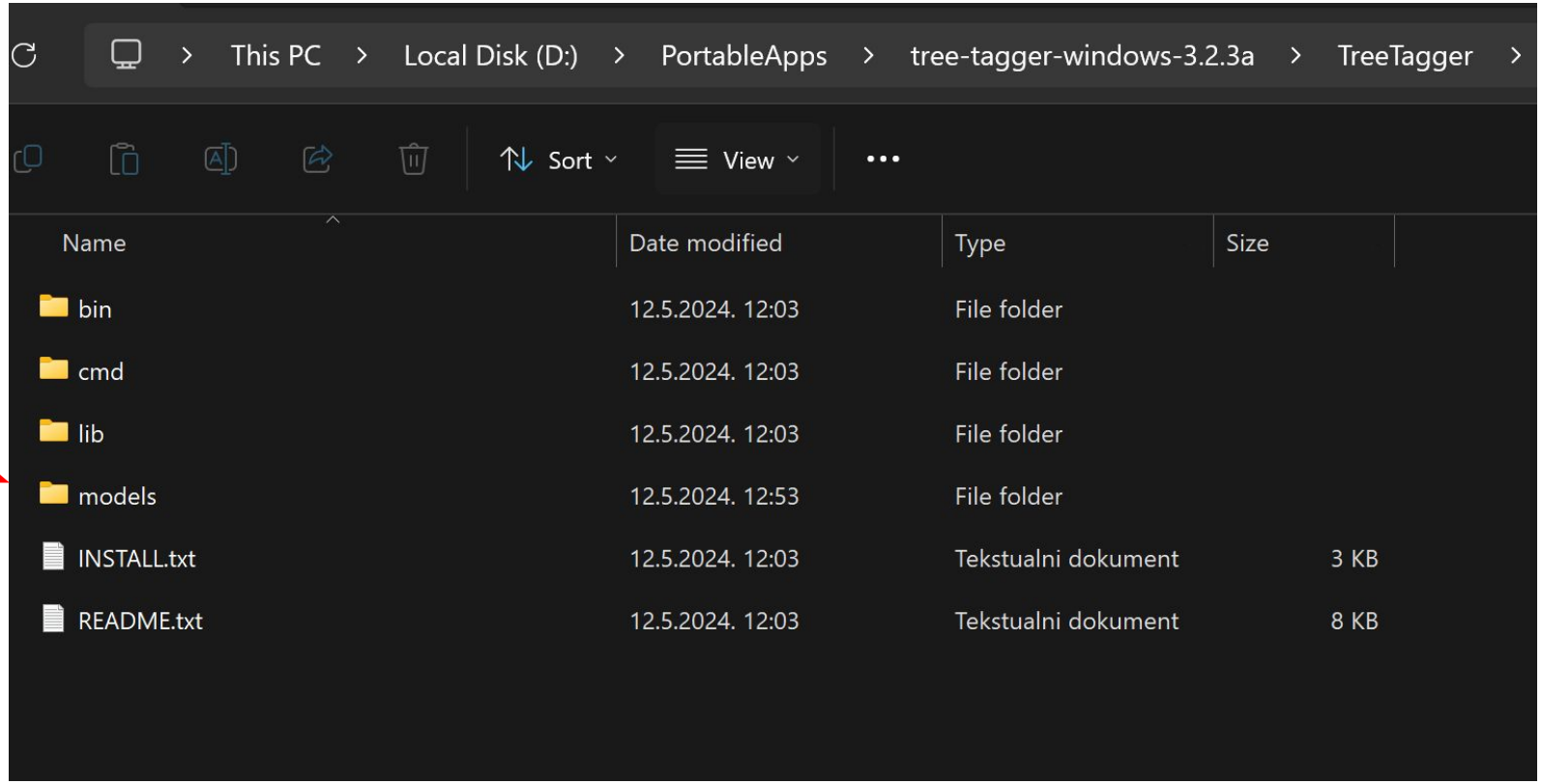
- Preuzeti TreeTagger za aplikaciju za obeležavanje vrsta reči i lematiza
- Instalitati Treetagger (dovoljno samo raspakovati zip)
- Preporuka:
  - napraviti neki folder PortableApps, ukoliko ga već nemate, na koji bi stavljali aplikacije koje ne zahtevaju proces instalacije
- Na primer:
  - D:\PortableApps\tree-tagger-windows-3.2.3a

https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/#Windows



Podešavanje TXM-a

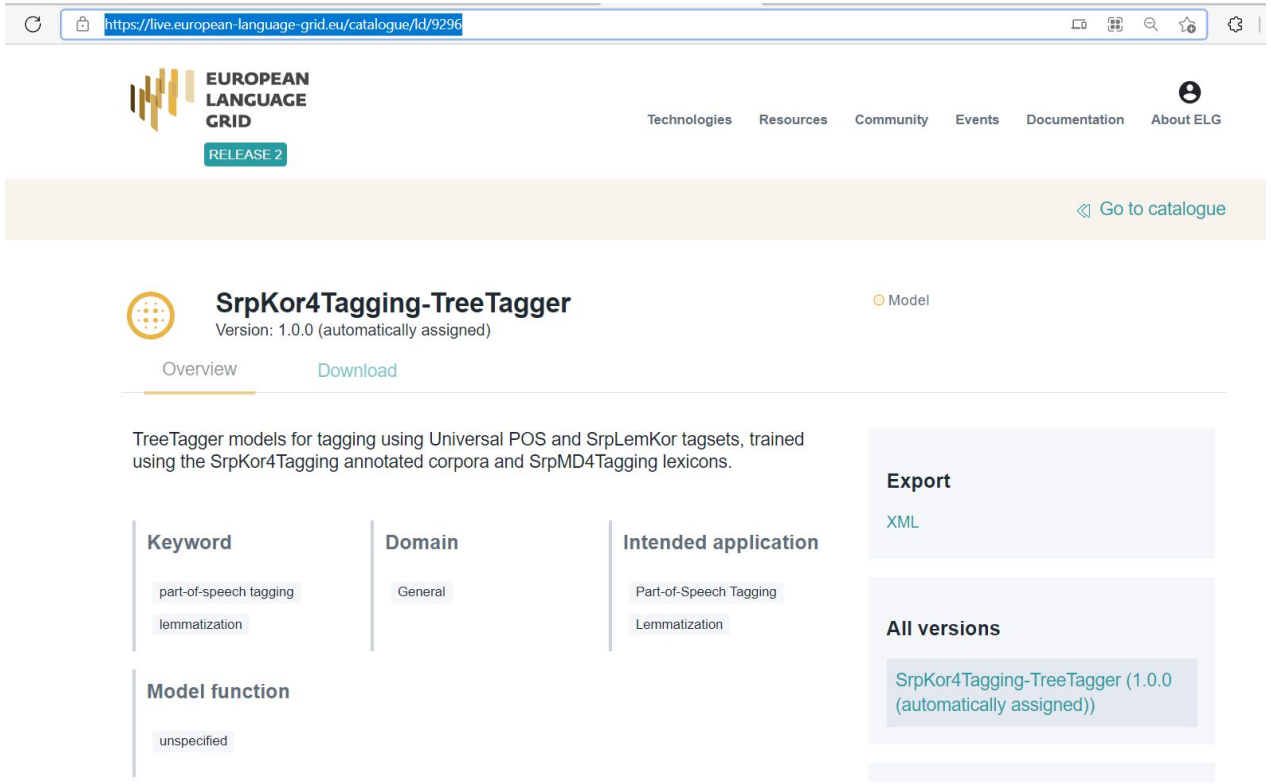
- Na raspakovanom folderu napraviti folder “models”



Podešavanje TXM-a

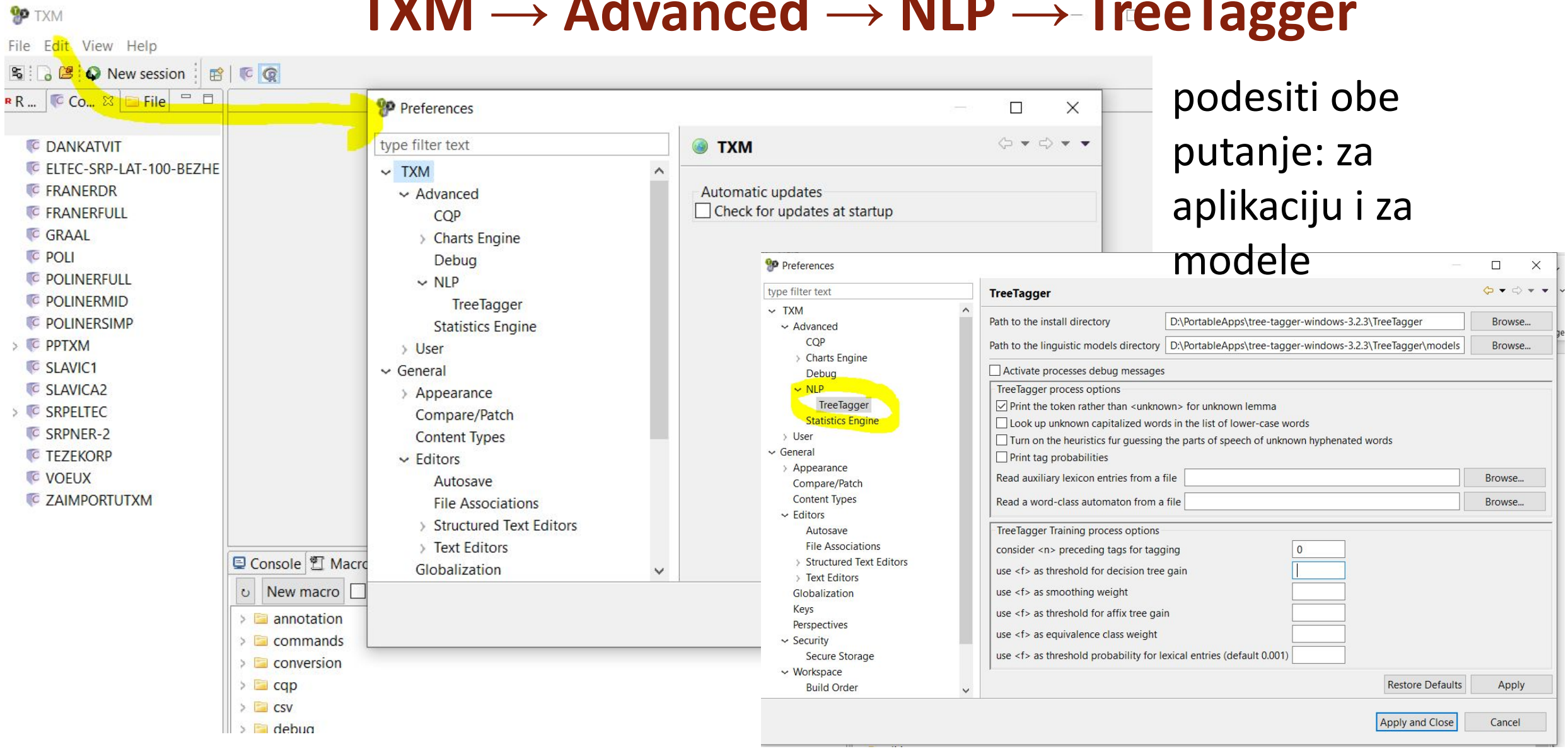
https://live.european-language-grid.eu/catalogue/ld/9296

- Preuzeti model za tagiranje TreeTagger za srpski sa ELG-a
- Raspakovati modele za srpski i staviti ih u folder ".../Treetagger/models"
- Preimenovati u "sr.par" onaj sa kojim želimo da radimo (UD)
- Pokrenuti TXM (ako već nije pokrenut)



- Ranka Stankovic, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Skoric. 2020. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3954–3962, Marseille, France. European Language Resources Association.

Edit → Preferences → TXM → Advanced → NLP → TreeTagger



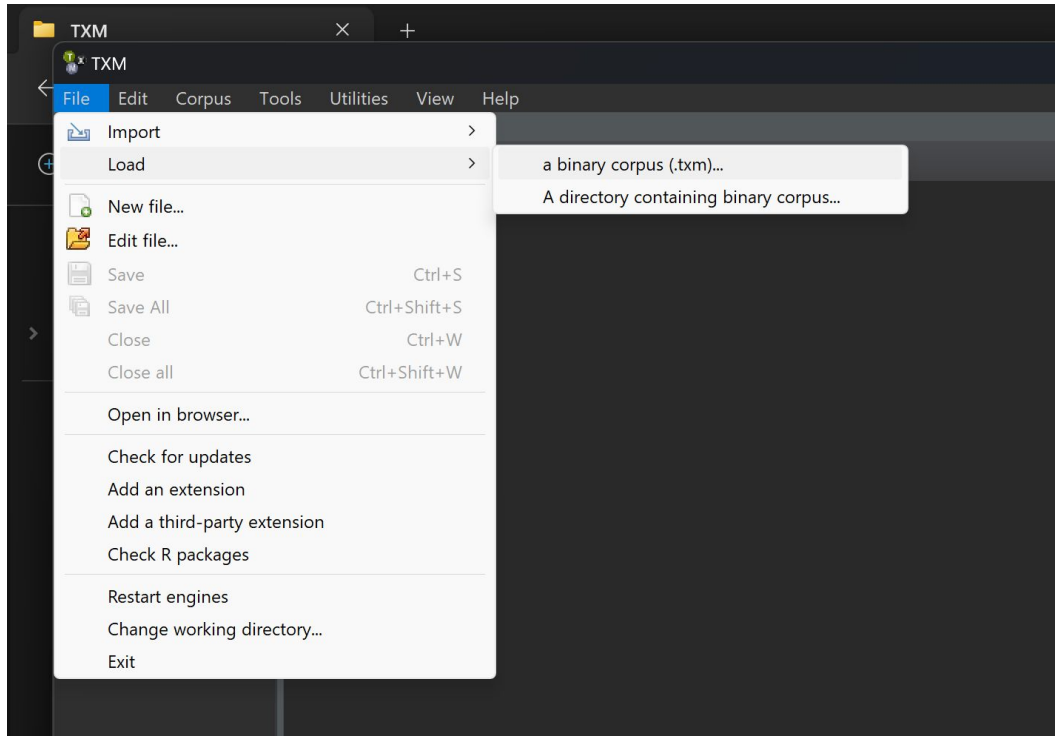
podesiti obe putanje: za aplikaciju i za modele

Uputstva za TXM

- TXM Manual
  - https://txm.gitpages.huma-num.fr/txm-manual/
  - https://txm.gitpages.huma-num.fr/textometrie/en/Documentation/
  - ili iz samog softvera
- TXM Wiki https://groupes.renater.fr/wiki/txm-users/index
- TXM demo portal http://portal.textometrie.org/demo
- Tuto@Mate#34: Flora Badin présente TXM, un logiciel de structuration d'analyse de corpus (youtube.com) (FR)
- Primeri korišćenja TXM-a:
  - Jelena Jaćimović: Tekstometrijske metode i TXM platforma za analizu i vizuelnu prezentaciju korpusa, Infoteka - Časopis za digitalnu humanistiku pdf
  -
  - i u još nekim radovima kasnije…

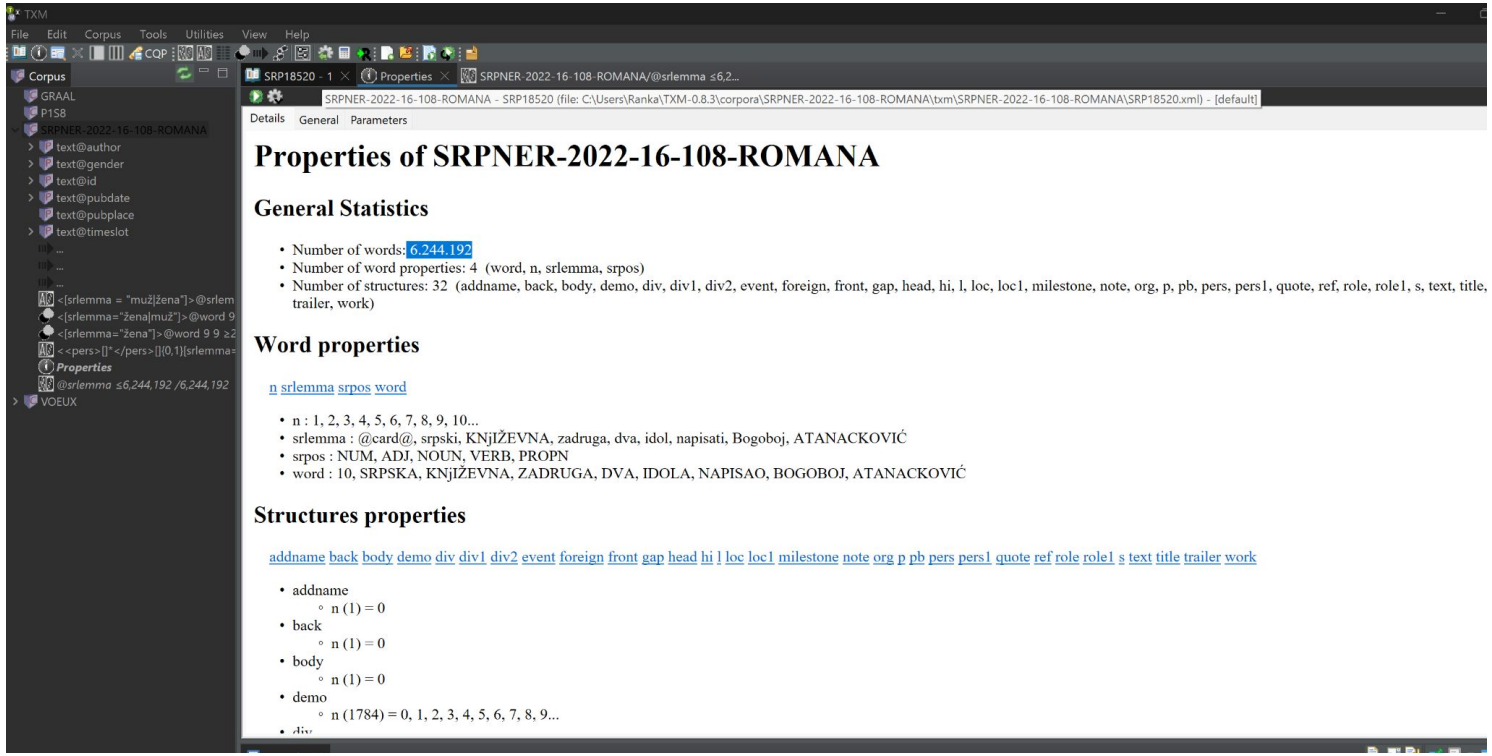
Učitavanje pripremljenog korpusa

- Korpus SrpELTeC-TXM-108-2022 (pronaći samostalno)
  - Preuzeti sa ELG-a (veličina 450MB zip fajl, ali je potrebno >3GB na disku gde su korpus TXM-a)
  - napraviti folder **korpusi** i u njega raspakovati zip (dobiće se fajl ekstenzije .txm)
- Isečak iz korpusa SrFudKo
  - preuzeti sa Mudla
- Učitati korpuse (jedn po jedan) u TXM (traje neko vreme):
  - **Load →a binary corpus (.txm)...**

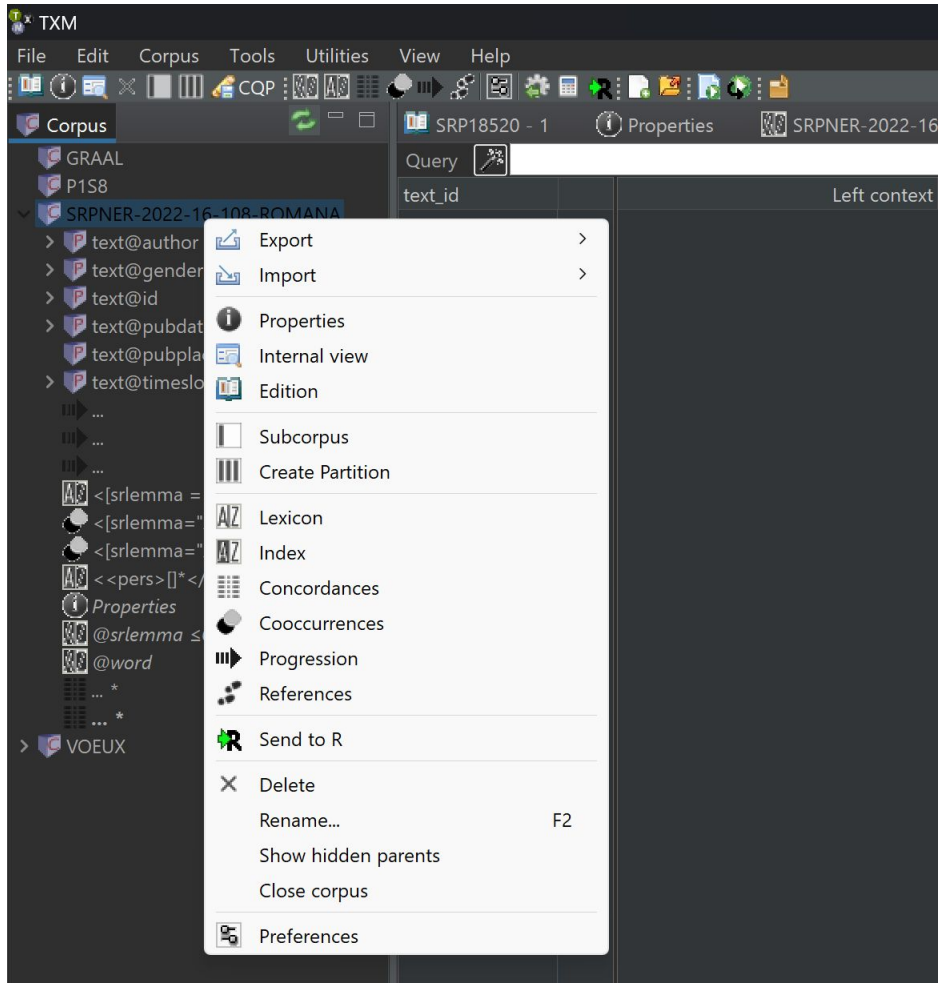


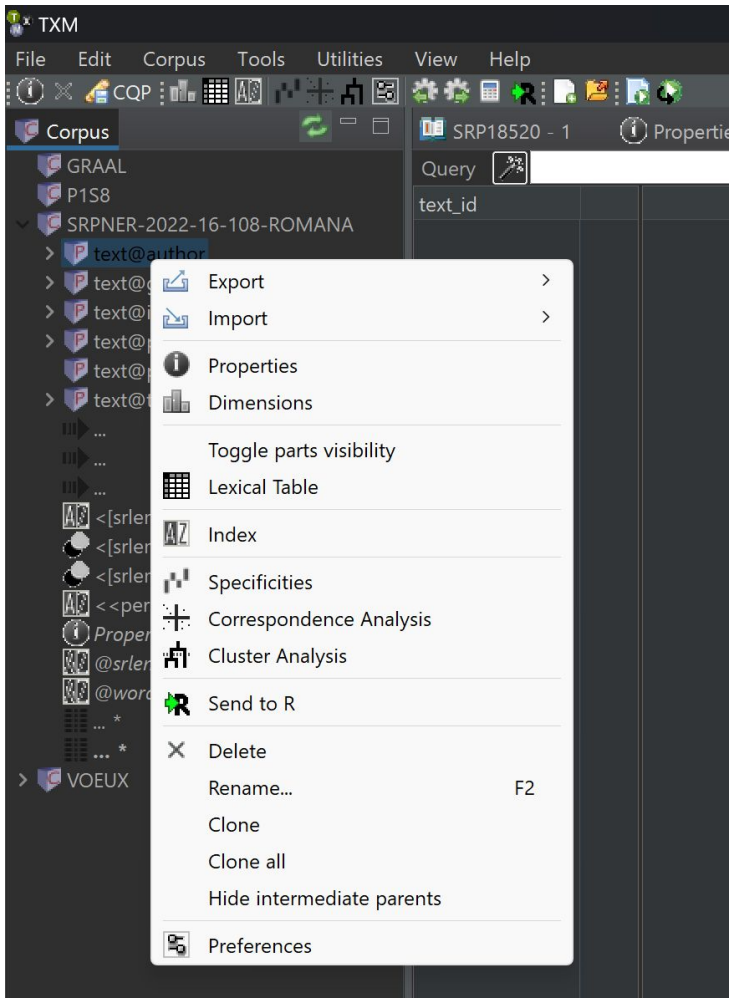
Pregledanje svojstava korpusa

- Na ime korpusa desni tester miša: **Properties**
- Šta vidimo?
  - Koliko ima reči
  - Kako je obeleženo
  - Da li imamo već pripremljene neke podele (particije)
  - Da li su i neki upiti već sačuvani i kako su se tu našli?



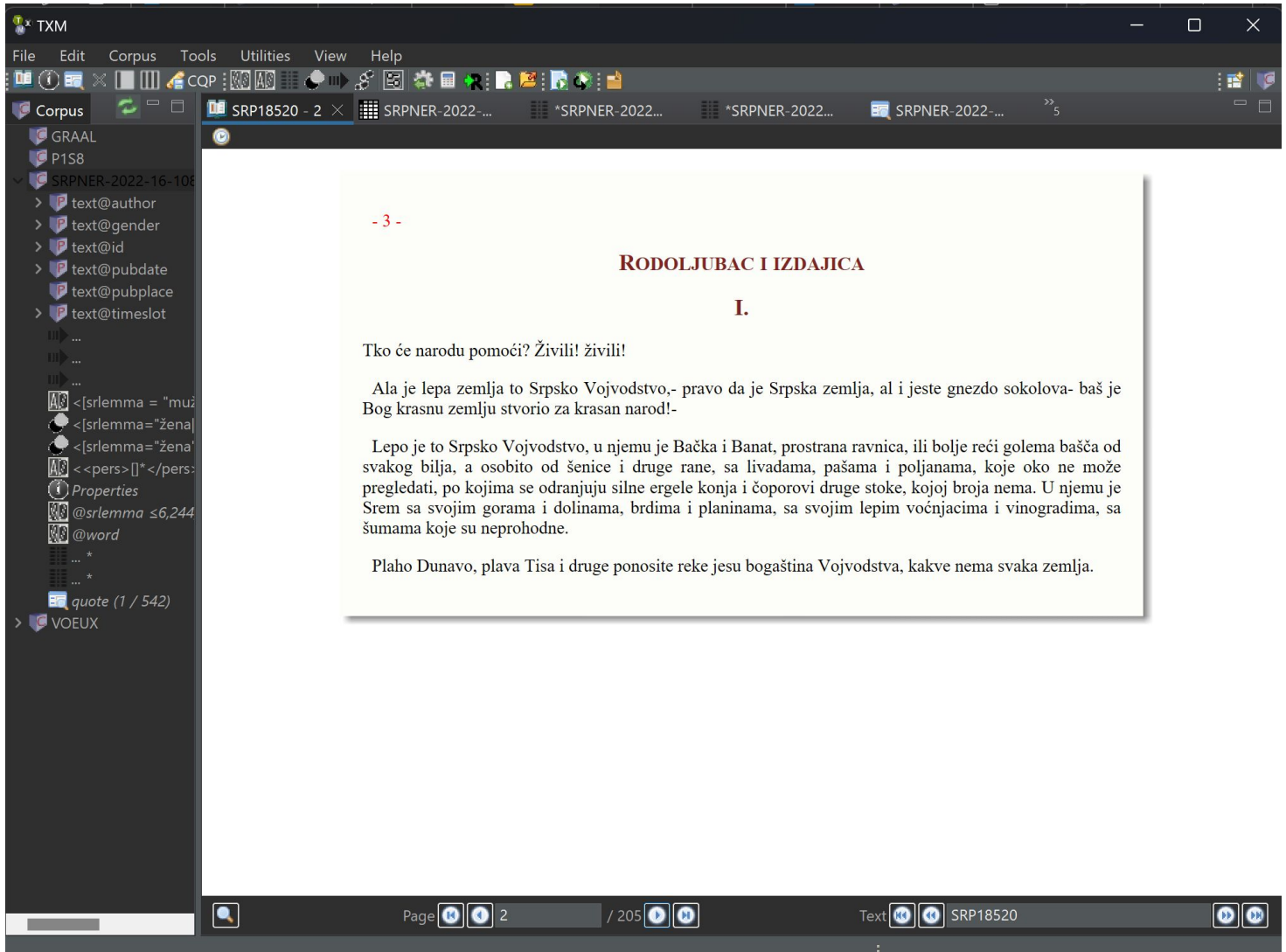
Pogledajmo konteksni meni: na celom korpusu i na participijama





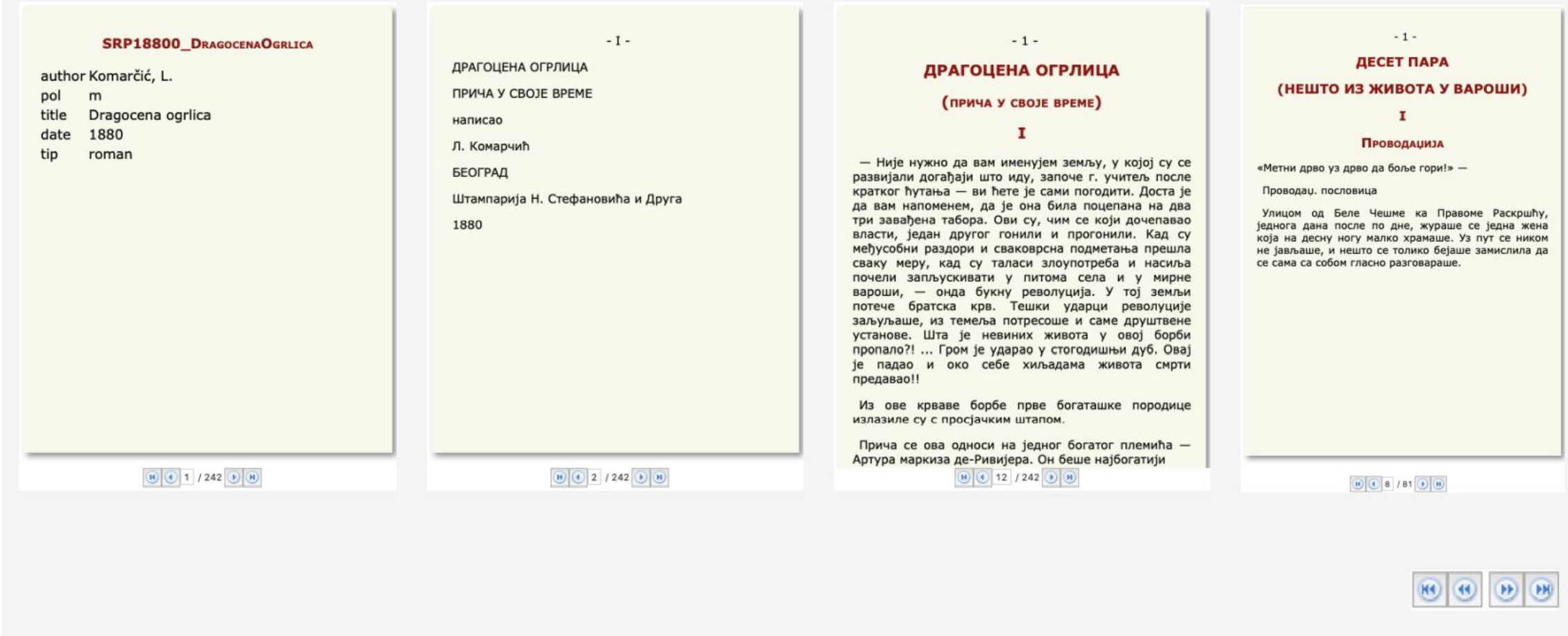
Prelistavanje

- Opcija: **Edition**
- Izgleda se može prilagođavati jer je određen css stilom
- U korpusu su strukturno obeležena poglavlja, pasusi, strane, tako da je prikaz teksta prilično veran originalu.



Vizuelni prikaz

(ukoliko su obeležene strane, naslovi, poglavlja)



SRP18800_DragocenaOgrlica

author Komarčić, L.
pol m
title Dragocena ogrlica
date 1880
tip roman

1 / 242

- I -

ДРАГОЦЕНА ОГРЛИЦА

ПРИЧА У СВОЈЕ ВРЕМЕ

написао

Л. Комарчић

БЕОГРАД

Штампарија Н. Стефановића и Друга

1880

2 / 242

- 1 -

ДРАГОЦЕНА ОГРЛИЦА

(ПРИЧА У СВОЈЕ ВРЕМЕ)

I

— Није нужно да вам именујем земљу, у којој су се развијали догађаји што иду, започе г. учитељ после кратког ћутања — ви ћете је сами погодити. Доста је да вам напоменем, да је она била поцепана на два три завађена табора. Ови су, чим се који дочепавао власти, један другог гонили и прогонили. Кад су међусобни раздори и сваковрсна подметања прешла сваку меру, кад су таласи злоупотреба и насиља почели запљускивати у питома села и у мирне вароши, — онда букну револуција. У тој земљи потече братска крв. Тешки ударци револуције заљуљаше, из темеља потресоше и саме друштвене установе. Шта је невиних живота у овој борби пропало?! ... Гром је ударао у стогодишњи дуб. Овај је падао и око себе хиљадама живота смрти предавао!!

Из ове крваве борбе прве богаташке породице излазиле су с просјачким штапом.

Прича се ова односи на једног богатог племића — Артура маркиза де-Ривијера. Он беше најбогатији

12 / 242

- 1 -

ДЕСЕТ ПАРА

(НЕШТО ИЗ ЖИВОТА У ВАРОШИ)

I

Проводаџија

«Метни дрво уз дрво да боље гори!» —

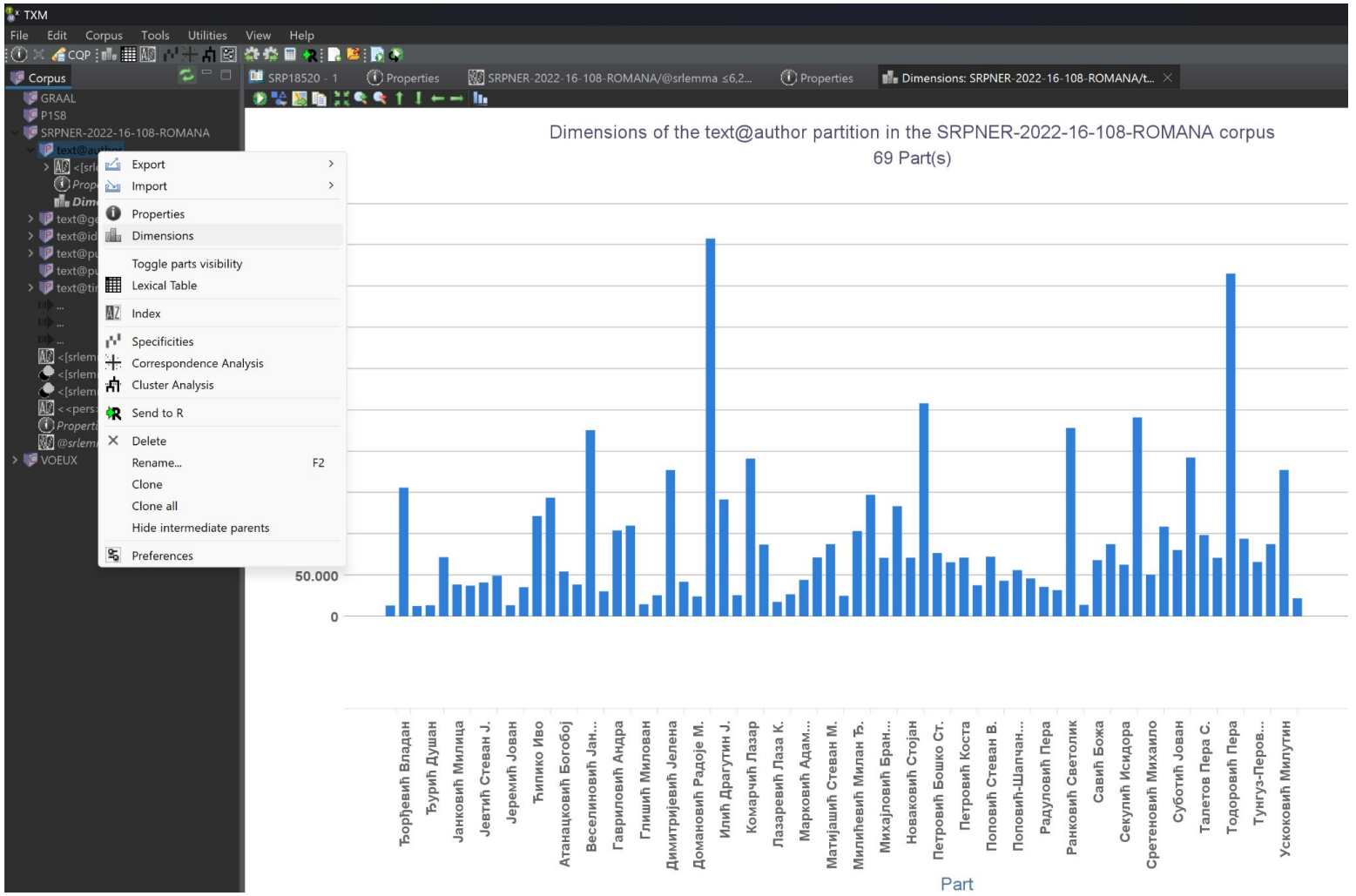
Проводаџ. пословица

Улицом од Беле Чешме ка Правоме Раскршћу, једнога дана после по дне, жураше се једна жена која на десну ногу малко храмаше. Уз пут се ником не јављаше, и нешто се толико бејаше замислила да се сама са собом гласно разговараше.

8 / 81

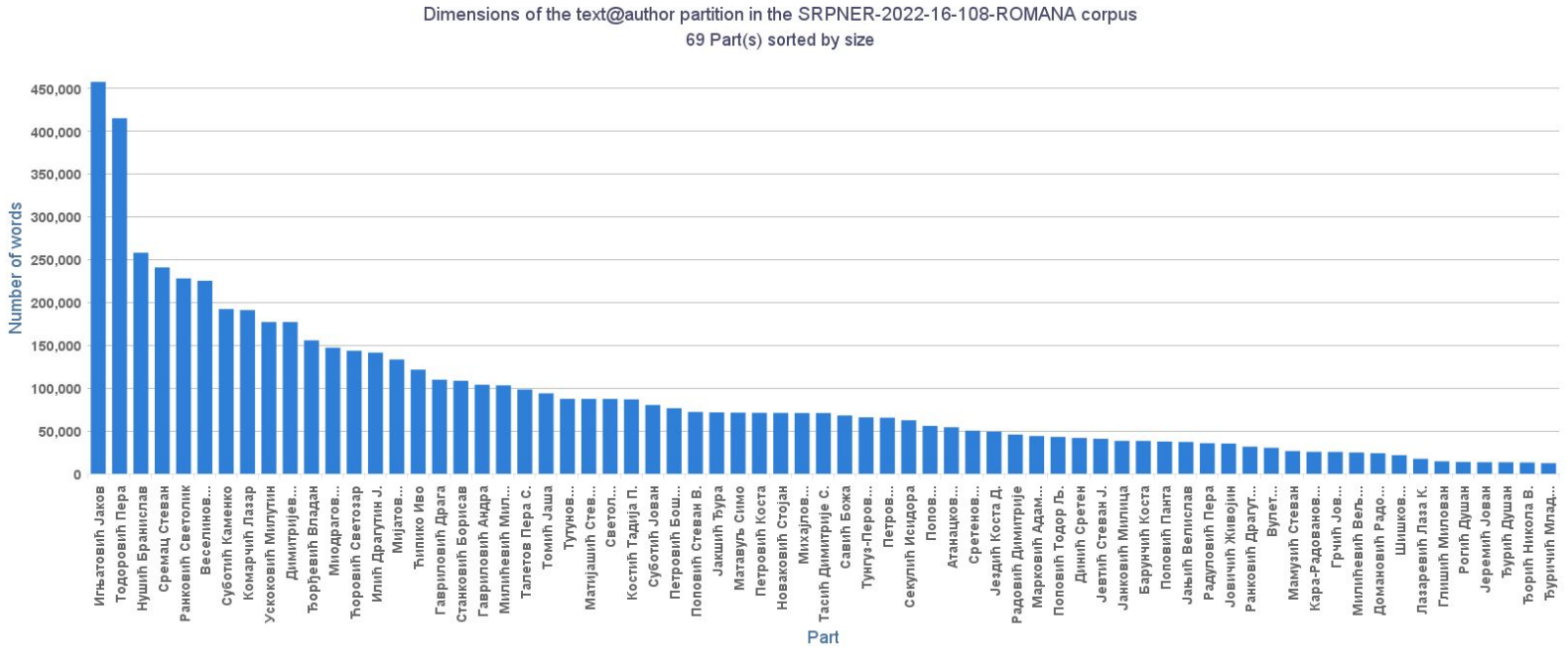
Pregledanje svojstava korpusa

- Pogledajmo particije po dimenzijama (iz kontekstnog menija **Dimensions**)
  - Po autorima (*author*)
  - Po polu (*gender*)
  - Po periodima (*timeslot*)



Pregledanje svojstava korpusa

- Sortirati u opadajućem redosledu histogram
- Od particija po autorima, najduže su *Jakova Ignjatovića*, *Pere Todorovića* i *Branislava Nušića*



Napomene: Ako se ne vidi legenda grafikona, treba uključiti prvo dugme sa leve strane - toolbar.