



Тесла

[Text Embeddings - Serbian Language Applications. #7276](#)



Фонд за науку
Републике Србије

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 osobnosti 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 prikaz)

Statistički modeli

- distribucije po podkorporusima (faktorska analiza, klaster analiza),
- visoku ili nisku zastupljenost u potkorporusima (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

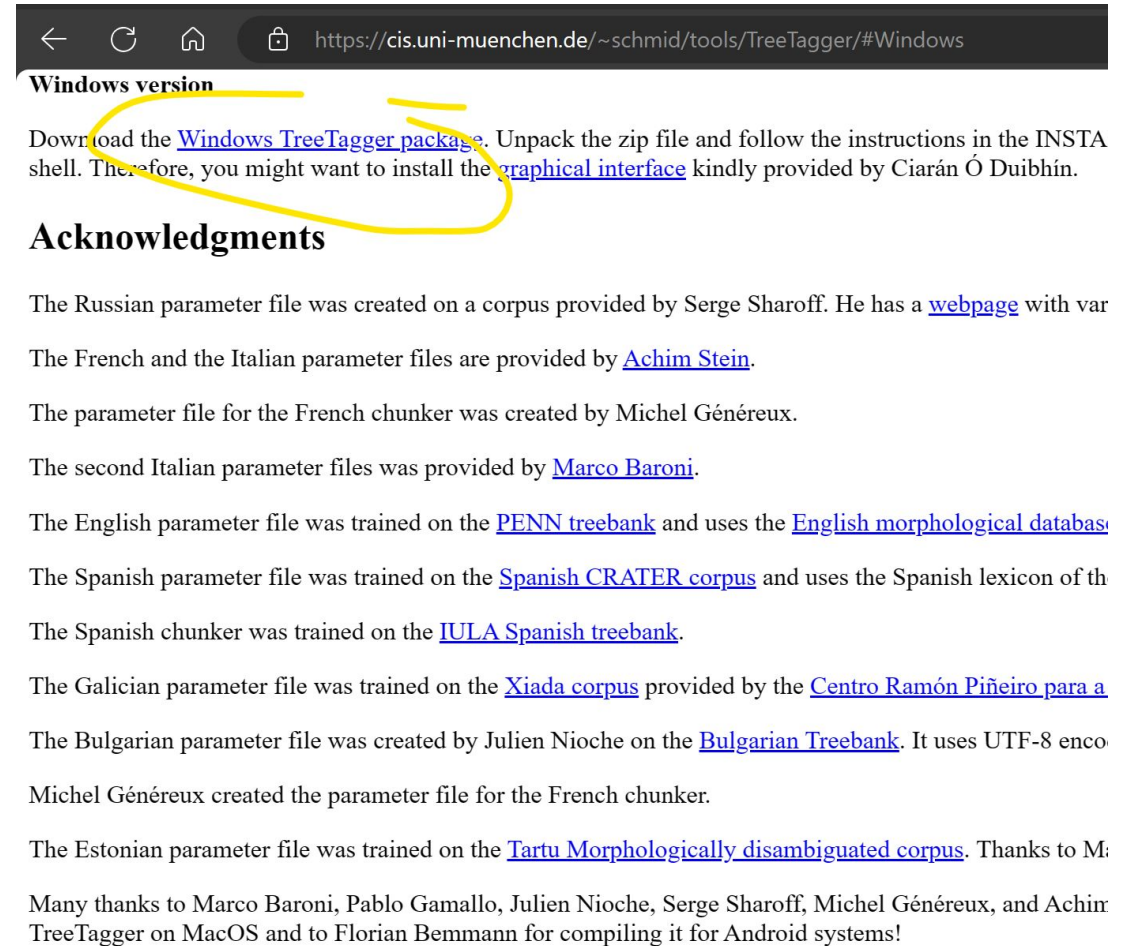
TXM 0.8.4 installation software for Windows, Mac, and Linux



Podešavanje TXM-a

- Preuzeti TreeTagger za aplikaciju za obeležavanje vrsta reči i lematiza
- Instalirati 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>



Windows version

Download the [Windows TreeTagger package](#). Unpack the zip file and follow the instructions in the INSTALL file. Therefore, you might want to install the [graphical interface](#) kindly provided by Ciarán Ó Duibhín.

Acknowledgments

The Russian parameter file was created on a corpus provided by Serge Sharoff. He has a [webpage](#) with various corpora.

The French and the Italian parameter files are provided by [Achim Stein](#).

The parameter file for the French chunker was created by Michel Génèreux.

The second Italian parameter files was provided by [Marco Baroni](#).

The English parameter file was trained on the [PENN treebank](#) and uses the [English morphological databases](#).

The Spanish parameter file was trained on the [Spanish CRATER corpus](#) and uses the Spanish lexicon of the [Spanish treebank](#).

The Spanish chunker was trained on the [IULA Spanish treebank](#).

The Galician parameter file was trained on the [Xiada corpus](#) provided by the [Centro Ramón Piñeiro para a Galia](#).

The Bulgarian parameter file was created by Julien Nioche on the [Bulgarian Treebank](#). It uses UTF-8 encoding.

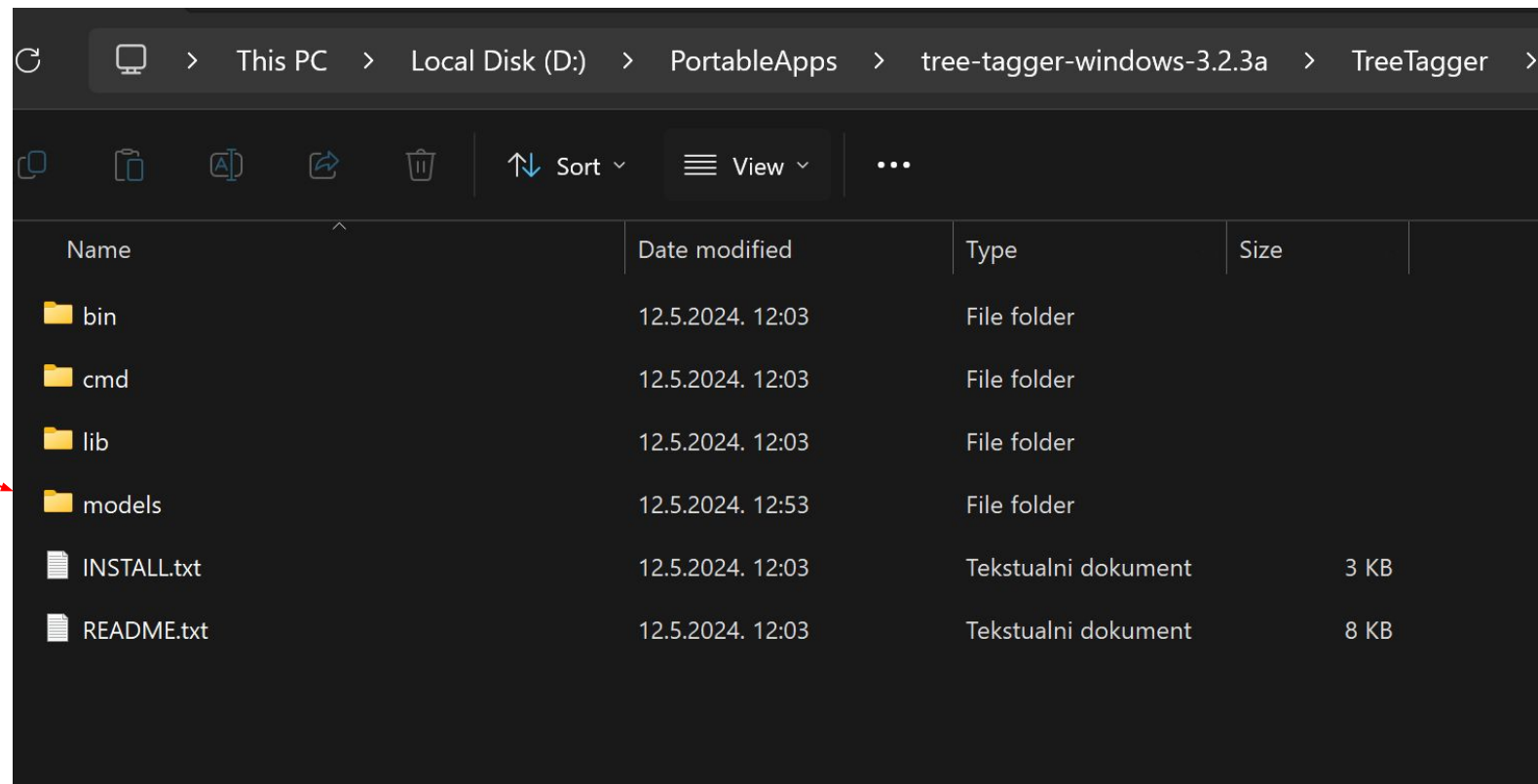
Michel Génèreux created the parameter file for the French chunker.

The Estonian parameter file was trained on the [Tartu Morphologically disambiguated corpus](#). Thanks to Michel Génèreux.

Many thanks to Marco Baroni, Pablo Gamallo, Julien Nioche, Serge Sharoff, Michel Génèreux, and Achim Stein for providing the parameter files for TreeTagger on MacOS and to Florian Bemann for compiling it for Android systems!

Podešavanje TXM-a

- Na raspakovanom folderu napraviti folder “models”



Podešavanje TXM-a

- 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.

<https://live.european-language-grid.eu/catalogue/Id/9296>

The screenshot shows the European Language Grid website. The header includes the logo and navigation links: Technologies, Resources, Community, Events, Documentation, and About ELG. A 'RELEASE 2' badge is visible. A 'Go to catalogue' link is in the top right. The main content area features the title 'SrpKor4Tagging-TreeTagger' with a version of 1.0.0 (automatically assigned). Below the title are tabs for 'Overview' and 'Download'. A description states: 'TreeTagger models for tagging using Universal POS and SrpLemKor tagsets, trained using the SrpKor4Tagging annotated corpora and SrpMD4Tagging lexicons.' The page is divided into sections: 'Keyword' (part-of-speech tagging, lemmatization), 'Domain' (General), 'Intended application' (Part-of-Speech Tagging, Lemmatization), and 'Model function' (unspecified). On the right side, there are sections for 'Export' (XML) and 'All versions' (SrpKor4Tagging-TreeTagger (1.0.0 (automatically assigned))).

Edit → Preferences →

TXM → Advanced → NLP → TreeTagger

podesiti obe
putanje: za
aplikaciju i za
modele

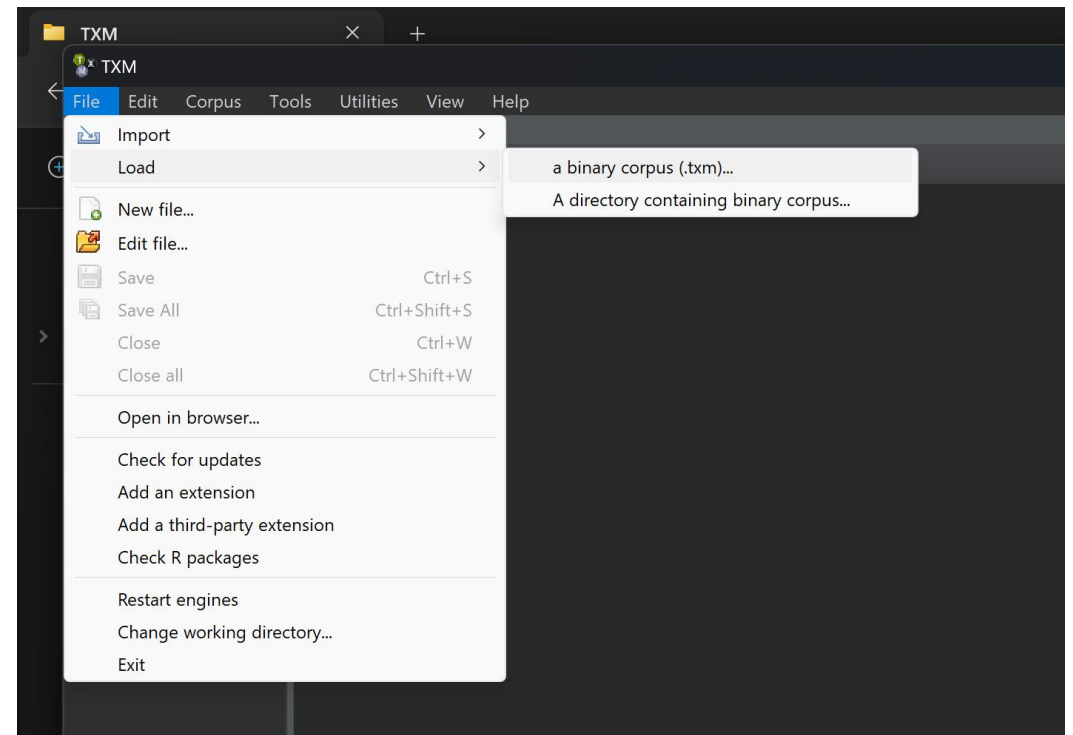
The image shows a screenshot of the TXM software interface. The main window displays a file browser on the left with a list of corpora including DANKATVIT, ELTEC-SRP-LAT-100-BEZHE, FRANERDR, FRANERFULL, GRAAL, POLI, POLINERFULL, POLINERMID, POLINERSIMP, PPTXM, SLAVIC1, SLAVICA2, SRPELTEC, SRPNER-2, TEZEKORP, VOEUX, and ZAIMPORTUTXM. The 'Preferences' dialog box is open, showing a tree view of settings. The path 'TXM → Advanced → NLP → TreeTagger' is highlighted with a yellow circle. A yellow arrow points from the 'File' menu in the top toolbar to the 'Preferences' dialog. In the background, another 'Preferences' dialog is visible, showing the 'Automatic updates' section with the checkbox 'Check for updates at startup' unchecked. The 'TreeTagger' dialog box is also open, showing fields for the install directory and linguistic models directory, both pointing to 'D:\PortableApps\tree-tagger-windows-3.2.3\TreeTagger'. It includes sections for 'TreeTagger process options' and 'TreeTagger Training process options' with various checkboxes and input fields. At the bottom, there are buttons for 'Restore Defaults', 'Apply', 'Apply and Close', and 'Cancel'.

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 (jeden 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?

The screenshot displays the TXM software interface. On the left, a tree view shows the corpus structure with folders for GRAAL, P158, and SRPNER-2022-16-108-ROMANA. The right pane shows the 'Properties of SRPNER-2022-16-108-ROMANA' dialog box. The 'General Statistics' section reports 6,244,192 words and 32 structures. The 'Word properties' section lists properties such as srlemma and srpos. The 'Structures properties' section lists structural elements like addname, back, body, and demo.

Properties of SRPNER-2022-16-108-ROMANA

General Statistics

- Number of words: 6,244,192
- Number of word properties: 4 (word, n, srlemma, srpos)
- Number of structures: 32 (addname, back, body, demo, div, div1, div2, event, foreign, front, gap, head, hi, l, loc, loc1, milestone, note, org, p, pb, pers, pers1, quote, ref, role, role1, s, text, title, trailer, work)

Word properties

[n](#) [srlemma](#) [srpos](#) [word](#)

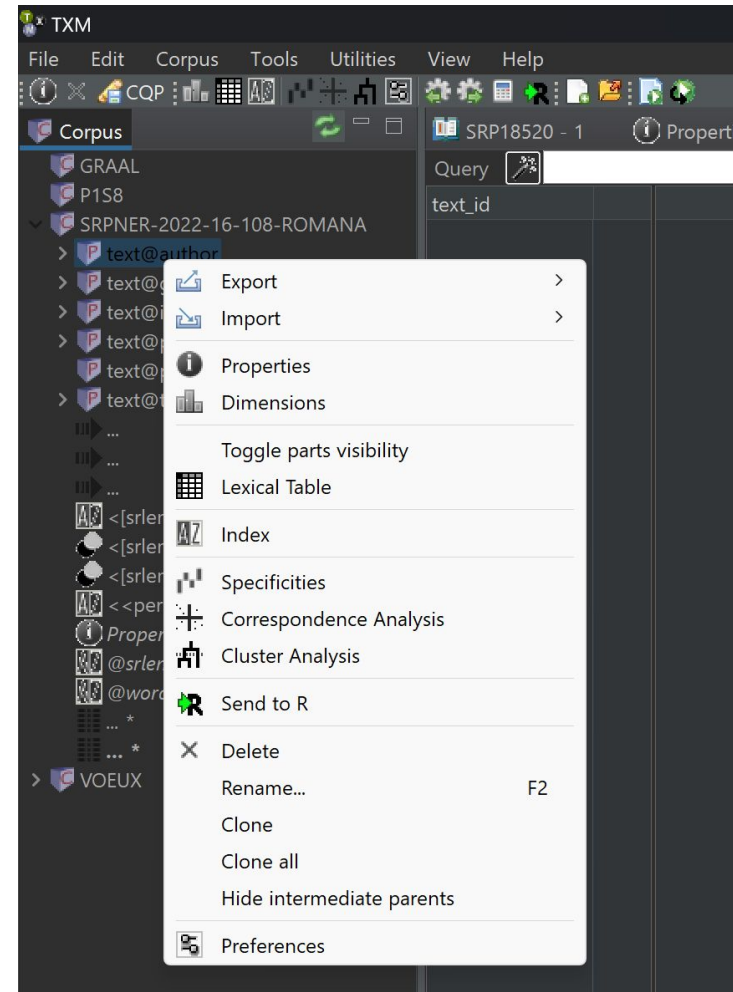
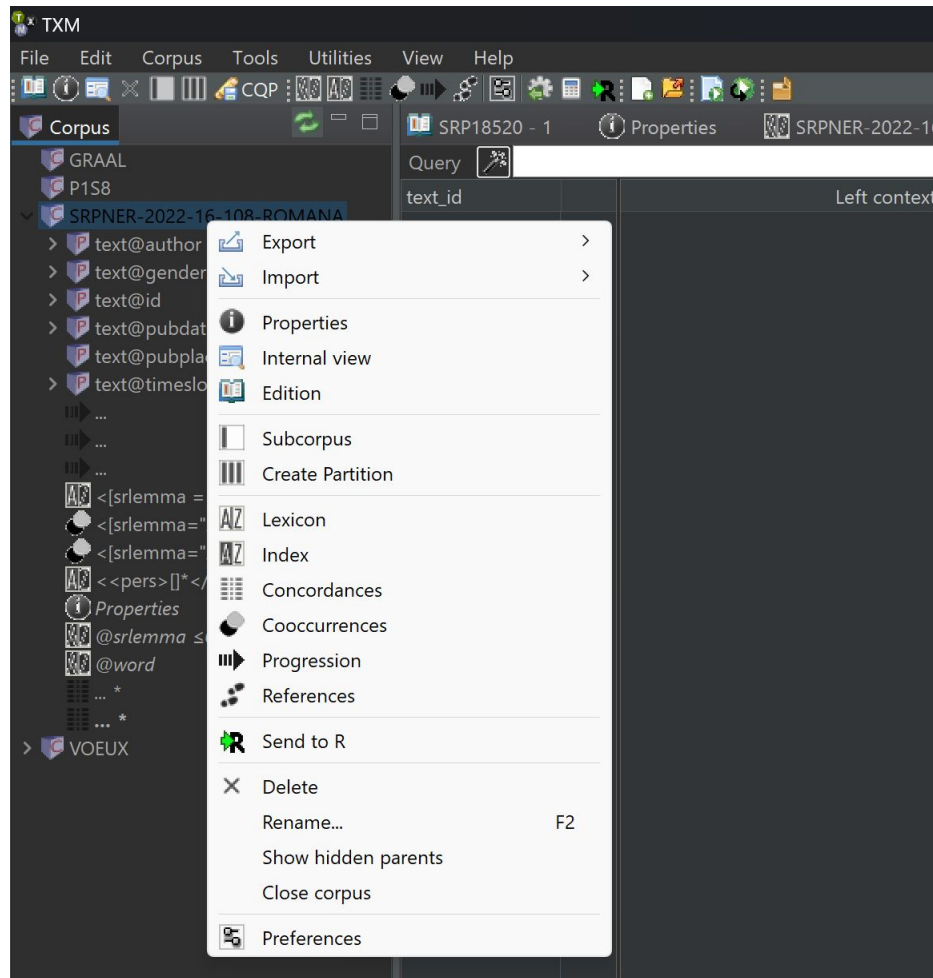
- n : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10...
- srlemma : @card@, srpski, KNJIŽEVNA, zadruga, dva, idol, napisati, Bogoboj, ATANACKOVIĆ
- srpos : NUM, ADJ, NOUN, VERB, PROP
- word : 10, SRPSKA, KNJIŽEVNA, ZADRUGA, DVA, IDOLA, NAPISAO, BOGOBOJ, ATANACKOVIĆ

Structures properties

[addname](#) [back](#) [body](#) [demo](#) [div](#) [div1](#) [div2](#) [event](#) [foreign](#) [front](#) [gap](#) [head](#) [hi](#) | [loc](#) [loc1](#) [milestone](#) [note](#) [org](#) [p](#) [pb](#) [pers](#) [pers1](#) [quote](#) [ref](#) [role](#) [role1](#) [s](#) [text](#) [title](#) [trailer](#) [work](#)

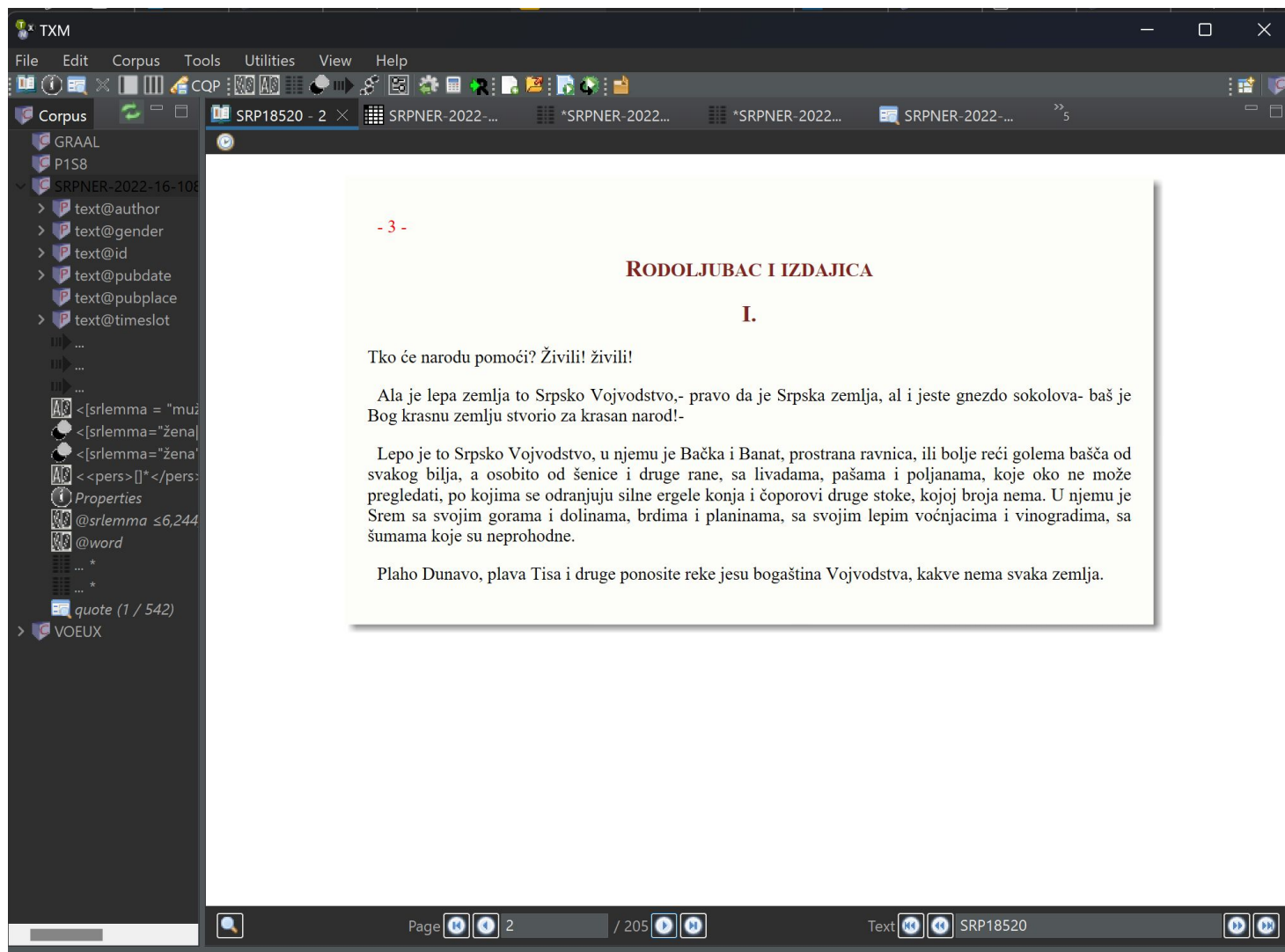
- addname
 - n (1) = 0
- back
 - n (1) = 0
- body
 - n (1) = 0
- demo
 - n (1784) = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9...
- .:iv

Pogledajmo kontekсни meni: na celom korpusu i na particijama



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.



The screenshot displays the TXM software interface. On the left, a 'Corpus' panel shows a tree view of a corpus structure with various text elements like 'text@author', 'text@gender', etc. The main window shows a rendered document with the following content:

- 3 -

RODOLJUBAC I IZDAJICA

I.

Tko će narodu pomoći? Živili! živili!

Ala je lepa zemlja to Srpsko Vojvodstvo,- pravo da je Srpska zemlja, al i jeste гнездо sokolova- baš je Bog krasnu zemlju stvorio za krasan narod!-

Lepo je to Srpsko Vojvodstvo, u njemu je Bačka i Banat, prostrana ravnica, ili bolje reći golem bašča od svakog bilja, a osobito od šenice i druge rane, sa livadama, pašama i poljanama, koje oko ne može pregledati, po kojima se odranjaju silne ergele konja i čoporovi druge stoke, kojoj broja nema. U njemu je Srem sa svojim gorama i dolinama, brdima i planinama, sa svojim lepim voćnjacima i vinogradima, sa šumama koje su neprohodne.

Plaho Dunavo, plava Tisa i druge ponosite reke jesu bogaština Vojvodstva, kakve nema svaka zemlja.

The interface also shows a status bar at the bottom with 'Page 2 / 205' and 'Text SRP18520'.

Vizuelni prikaz

(ukoliko su obeležene strane, naslovi, poglavlja)

SRP18800_DRAGOCENA OGRLICA

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

- I -

ДРАГОЦЕНА ОГРЛИЦА
ПРИЧА У СВОЈЕ ВРЕМЕ
написао
Л. Комарчић
БЕОГРАД
Штампарија Н. Стефановића и Друга
1880

- 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 дочепавоа 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

«Метни дрво уз дрво да боље гори!» —
Проводац. 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.

1 / 242

2 / 242

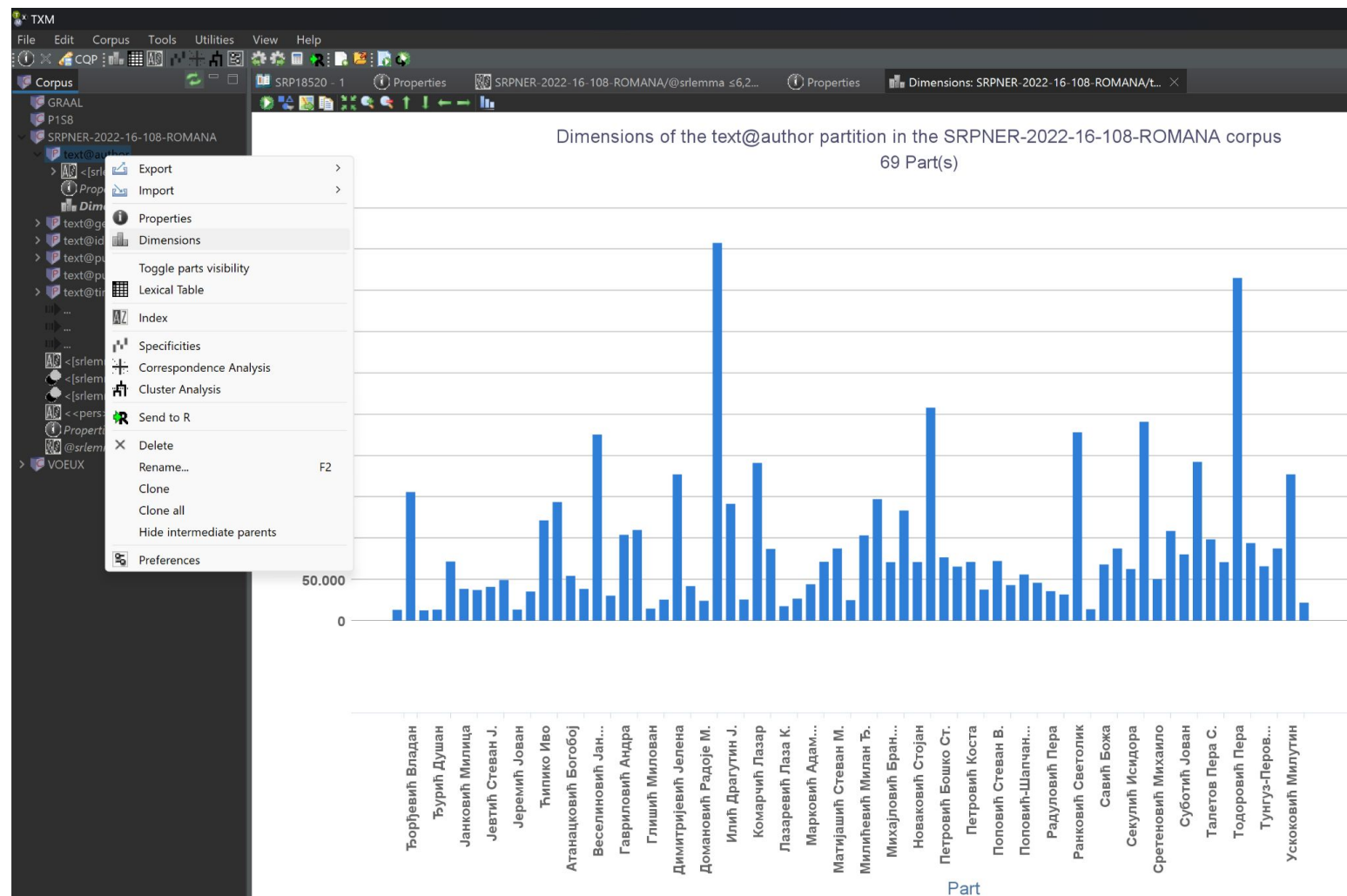
12 / 242

8 / 81

Navigation icons: back, forward, search, etc.

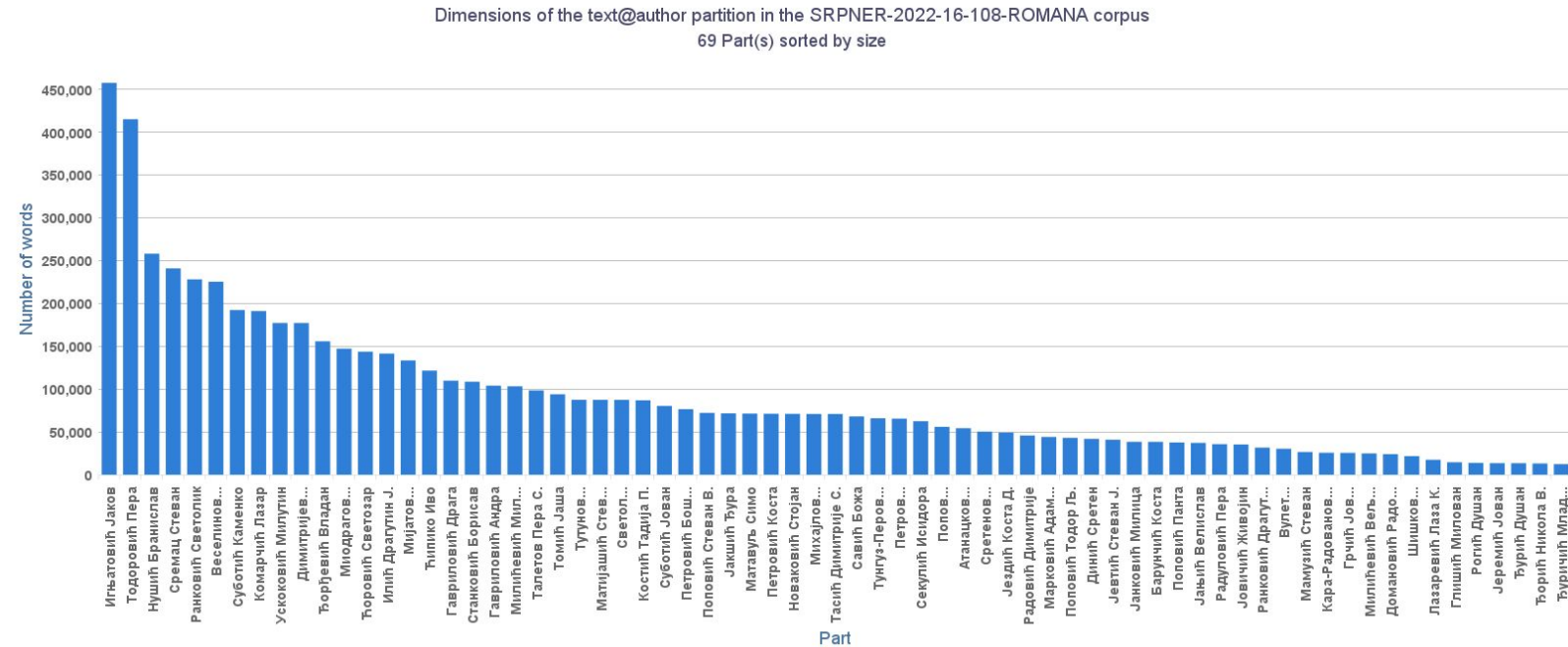
Pregledanje svojstava korpusa

- 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.