SEMINAR DRUŠTVA ZA JEZIČKE RESURSE I TEHNOLOGIJE JeRTeh

MOGUĆNOSTI UPOTREBE TXM SOFTVERA ZA ANALIZU I VIZUELNO PREDSTAVLJANJE DIGITALNIH TEKSTUALNIH KORPUSA

Jelena Jaćimović
Stomatološki fakultet Univerziteta u Beogradu

Beograd, 9. maj 2019. godine

O ČEMU ĆE BITI REČI ...

- TXM softver
- srpELTeC korpus
- Osnovni koncepti i alati TXM-a za analizu teksta

TEKSTOMETRIJA I TXM SOFTVER

ŠTA JE TEKSTOMETRIJA?

Metodologija koja omogućava nelinearno proučavanje korpusa, u kvantitativnom i kvalitativnom smislu, kombinujući leksikometrijska i statistička istraživanja sa razvijenim korpusnim tehnologijama (Unicode, XML, TEI, NLP, CQP, R).

SVRHA TEKSTOMETRIJSKE ANALIZE

Opisivanje leksičkih, morfoloških, sintaksičkih i drugih karakteristika teksta na lokalnom nivou.
Proučavanje međuodnosa određenih elemenata teksta (zajedničko pojavljivanje), leksičkih odnosa između tekstova (specifičnosti), lociranje elemenata teksta u korpusu (progresija).
Omogućava poređenje lingvističkih hipoteza ili kreiranje novih pomoću posmatranih podataka.
Omogućava pristup tekstu ne samo linearnim čitanjem, već i ciljanim posmatranjem određenih elemenata i karakteristika, izabranih u skladu sa postavljenim pitanjima.
Kvantifikovanje elemenata nije nov pristup, ali je pojednostavljen primenom postojećih alata.

TXM SOFTVER

Grafičko korisničko okruženje zasnovano na korišćenju CQP pretraživača i R statističkog paketa, koje omogućava analizu digitalnih korpusa bilo kog jezika
Korišćenje je besplatno
Softver otvorenog koda
Kontinuirano ga razvija tim istraživača IHRIM (Institut d'Histoire des Représentations et des Idées dans les Modernités) laboratorije, ENS de Lyon od 2010. godine
Dostupne su desktop instalacije za Windows, Linux i Mac OS X, kao i veb platforma

NEKE KARAKTERISTIKE TXM SOFTVERA

- Širok spektar ulaznih formata

  - Korpusi pisanih tekstova: TXT (Unicode) / XML /TEI

    - Jedinice teksta: tekstovi korpusa (knjige, članci, intervjui, …) koji mogu imati i metapodatke (autor, naslov, datum, žanr, ….)

      - Opcione strukture teksta: svaki tekst može da sadrži određene 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 (grafički oblik, lema, gramatička kategorija, …)

  - Korpusi transkripata (sinhronizovani sa izvornim audio ili video snimkom)

  - Paraelni korpusi

NEKE KARAKTERISTIKE TXM SOFTVERA

- Tekstovi su kodirani u skladu sa određenim konvencijama, poput XML kodiranja prema TEI smernicama, koje se mogu koristiti prilikom analize korpusa.

- Tekst se automatski segmentira prilikom uvoza u TXM okruženje, dok se reči označavaju svojstvima, kaošto su lema ili gramatička kategorija (TXM koristi TreeTagger).

- Bilo koja kombinacija ovih svojstava reči može se upotrebiti za navođenje njihovih obrazaca i prikaz konteksta u kojima se te reči pojavljuju (liste frekvencija, konkordance, vizuelni prikaz izdanja).

- Statistički modeli primenjeni na prebrojavanje svojstava reči omogućavaju analizu njihove distribucije po korpusima (faktorska analiza, klaster analiza), njihovu izuzetno visoku ili nisku zastupljenost u pojedinim potkorpusima (analiza specifičnosti), ili analizu privlačnosti koja postoji između pojedinih reči (analiza zajedničkog pojavljivanja).

- Rezultat svake analize može da se izveze u tabelarnom ili grafičkom obliku radi dalje analize i uređivanja pomoću nekog drugog alata.

KORPUS

srpELTeC

https://distantreading.github.io/ELTeC/srp/index.html

- Andra Gavrilović. Prve žrtve (kratka proza), 1893.
- Tadija Kostić. Gospoda seljaci (kratka proza), 1896.
- Čedomilj Mijatović. Rajko od Rasine (kratka proza), 1892.
  Ikonija, vezirova majka (pripovetka), 1891.
- Milan Milićević. Deset para (pripovetka), 1881.
- Bora Stanković. Uvela ruža (pripovetka), 1899.
- Milutin Uskoković. Potrošene reči (pripovetka), 1911.
  Došljaci (roman), 1910.
  Čedomir Ilić (roman), 1914.
- Jelena Dimitrijević. Nove (roman), 1912.
- Dragutin Ilić. Hadži Đera (roman), 1904.
- Milica Janković. Kaluđer iz Rusije (roman), 1919.
- Lazar Komarčić. Dragocena ogrlica (roman), 1880.
  Jedna ugašena zvezda (roman), 1902.
  Prosioci (roman), 1905.
- Branislav Nušić. Opštinsko dete (roman), 1902.
- Isidora Sekulić. Đakon Bogorodičine crkve (roman), 1919.

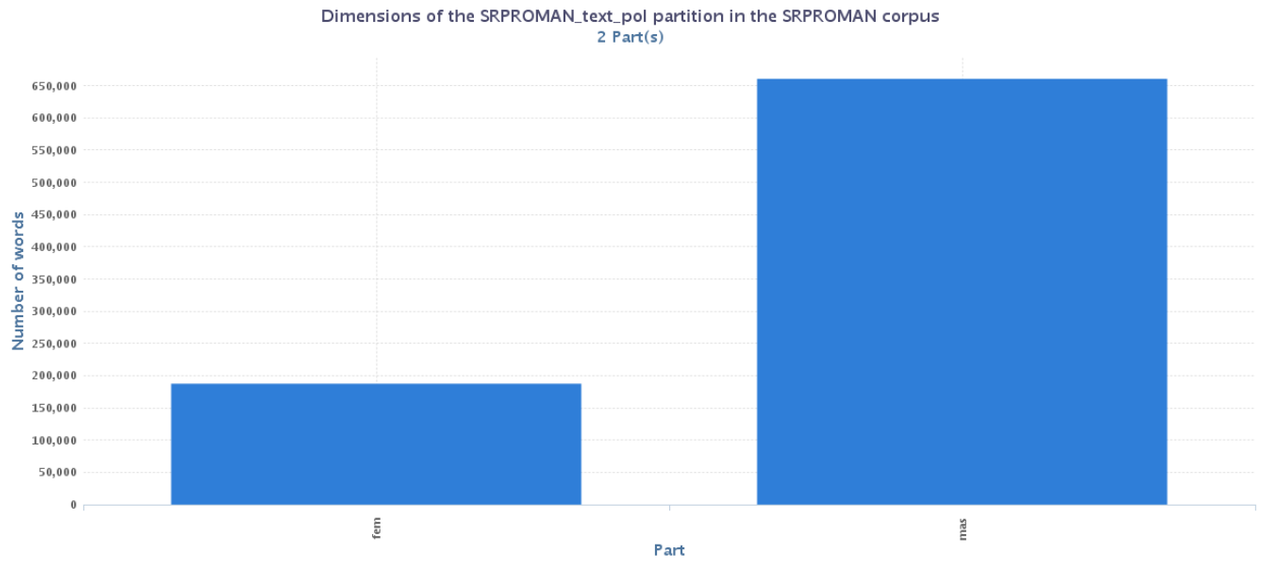
KORPUS

Modul XML - TEI Zero + CSV (uvoz u TXM)

Properties of SRPROMAN (id=SRPROMAN)

Summary Statistics

- Number of words 855412
- Number of word properties 4
- Number of structural units 15



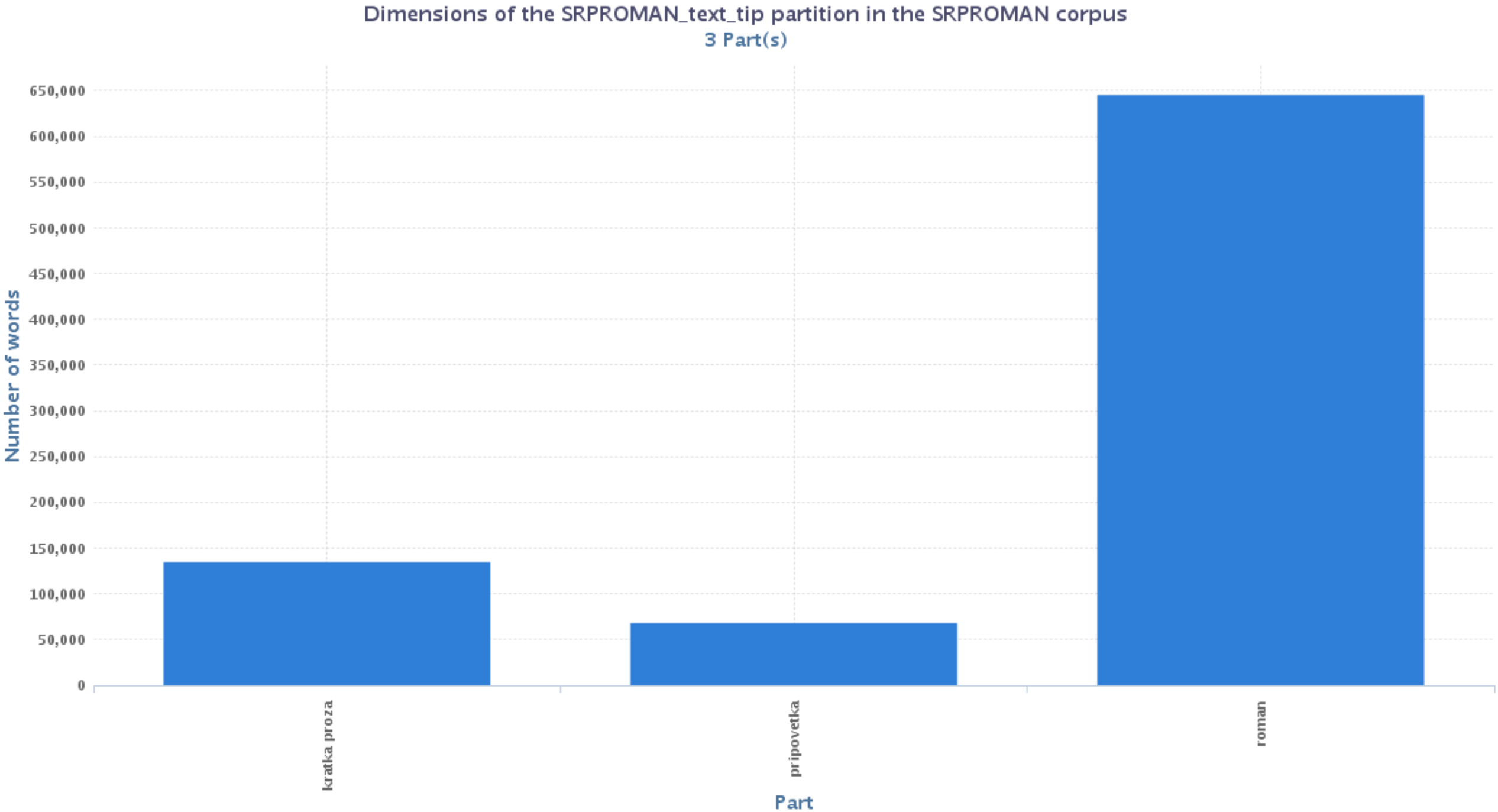
Lexical Units properties

- n : 1, 2, 3, 4, …
- srlemma : драгоцен, огрлица, :, прича, у, свој, време, …
- srpos : A, N, SENT, PREP, PRO, PUNCT, NUM, CONJ, ABB, ?, V, ADV, …
- word : Драгоцена, огрлица, :, прича, у, своје, време, …

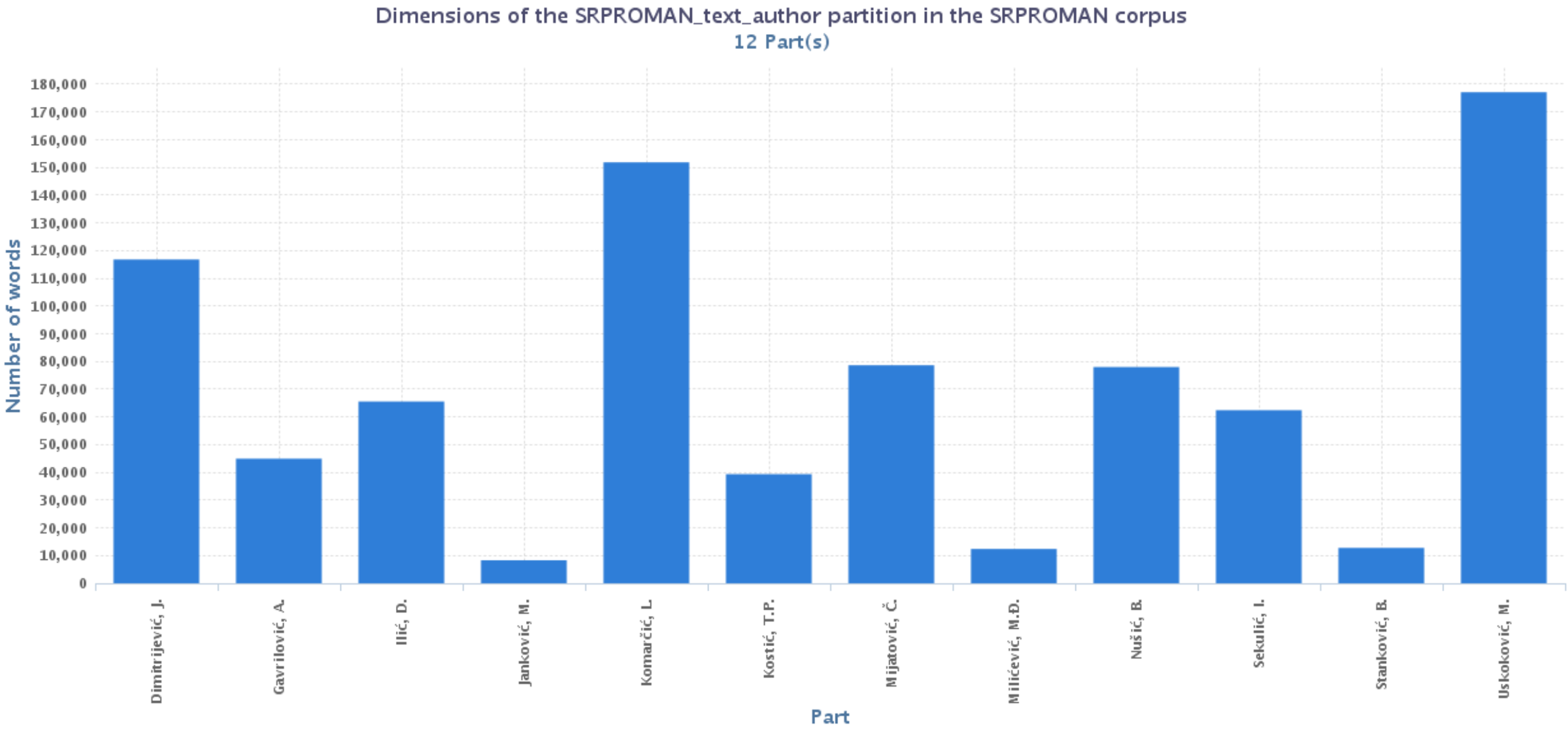
Structural Units properties

- Back, body, div, foreign, front, head, hi, l, note, p, quote, text (author, date, id, lang, pol, tip, title), title

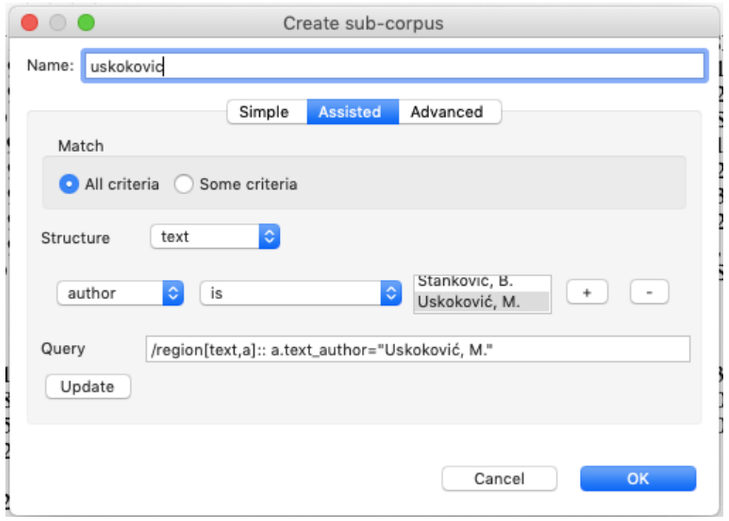
KORPUS


Dimensions of the SRPROMAN_text_tip partition in the SRPROMAN corpus
3 Part(s)
Number of words
650,000
600,000
550,000
500,000
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
kratka proza
pripovetka
roman
Part

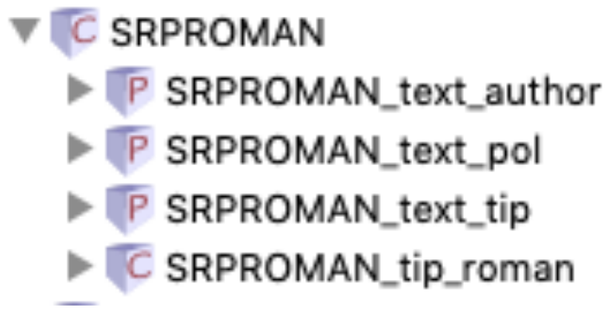
KORPUS

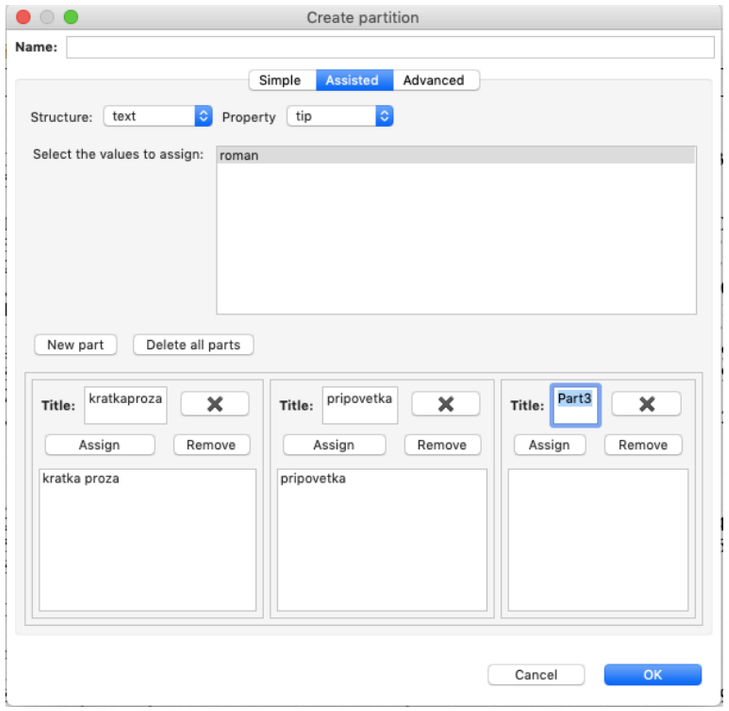

Dimensions of the SRPROMAN_text_author partition in the SRPROMAN corpus
12 Part(s)
Number of words
180,000
170,000
160,000
150,000
140,000
130,000
120,000
110,000
100,000
90,000
80,000
70,000
60,000
50,000
40,000
30,000
20,000
10,000
0
Dimitrijević, J.
Gavrilović, A.
Ilić, D.
Janković, M.
Komarčić, L.
Kostić, T.P.
Mijatović, Č.
Milićević, M.Đ.
Nušić, B.
Sekulić, I.
Stanković, B.
Uskoković, M.
Part

POTKORPUSI I PARTICIJE



POTKORPUSI





PARTICIJE

VIZUELNI PRIKAZ TEKSTOVA KORPUSA

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

LEKSIKON

Lista svojstava reči koje se pojavljuju u korpusu ili potkorpusu po rastućem ili opadajućem redosledu učestalosti

Najfrekventniji oblici su separatori, oblici koji nose malo semantičkih informacija

Srednja frekvencija: centralni rečnik korpusa

Niska frekvencija i hapaks legomenon: periferni vokabular, hapaks čini gotovo polovinu rečnika korpusa

t 855412 , v 74371 , fmin 1 , fmax 69224

word	Frequency
,	69224
.	40343
и	26068
је	23158
се	19368
да	16974
у	13821
—	13396
на	8539
!	7947
…	6266
не	5998
а	5092
од	4800
:	4698
су	4561
што	4552
као	4488
?	4262
за	4028
па	3936
то	3751
му	3154
„	3035
;	3027
га	3003
није	2948
кад	2917

1 -100 / 74371

t 855412 , v 30884 , fmin 1 , fmax 69224

srlemma	Frequency
,	69224
.	40343
јесам	35997
и	28315
се	19368
да	17750
у	15050
—	13396
тај	12453
он	11251
бити	10164
она	9188
на	9176
ја	8285
!	7947
не	6841
а	6723
…	6266
који	5681
од	5053
па	4819
хтети	4815
што	4791
:	4698
као	4691
један	4634
за	4371
овај	4277

1 -100 / 30884

srpos	Frequency
N	177889
V	157514
PUNCT	94602
PRO	88762
CONJ	81837
A	59956
PREP	57451
ADV	45545
SENT	44824
PAR	32967
NUM	8707
	2272
?	1917
INT	629
ABB	486
RN	34
PREF	20

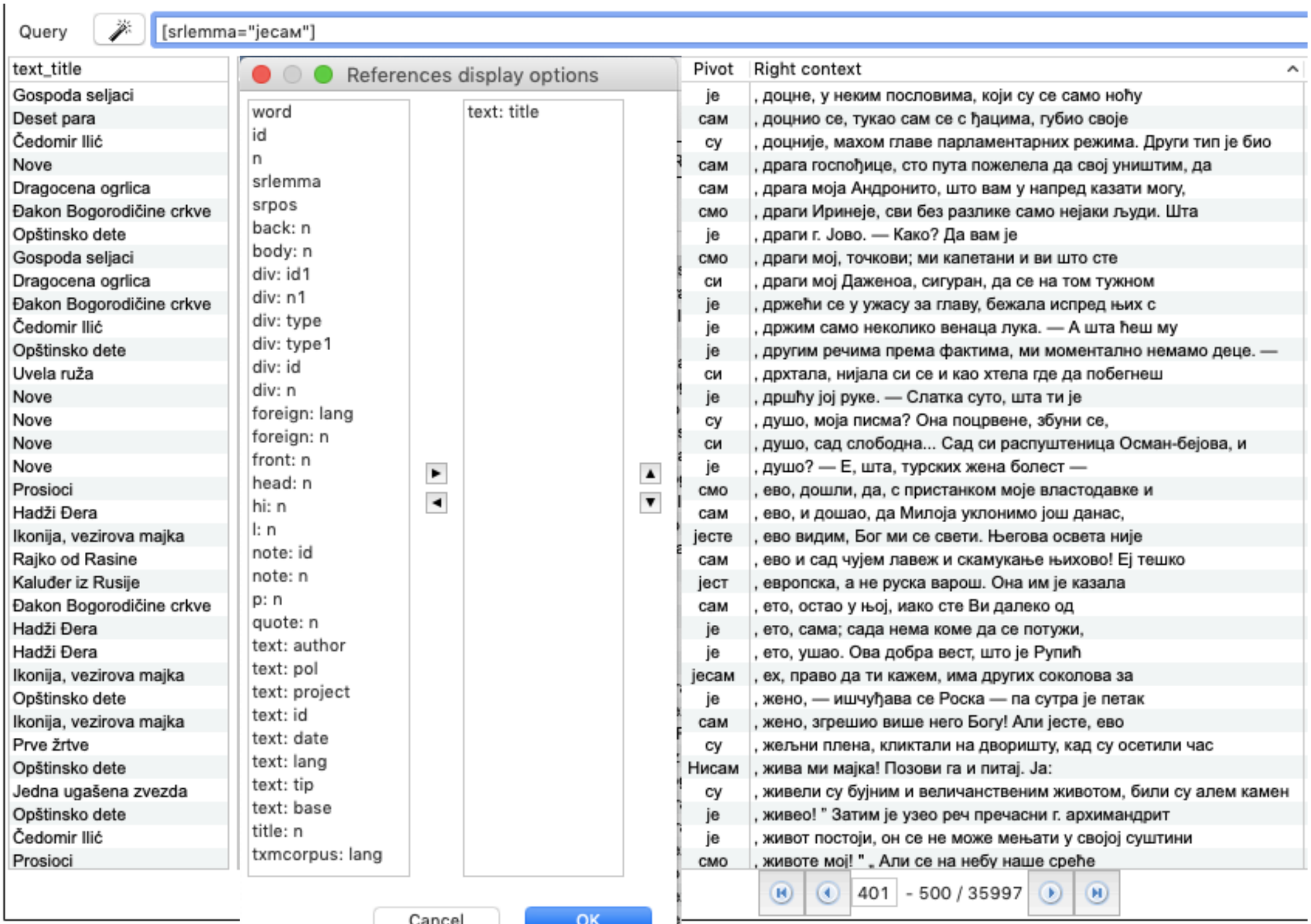
KONKORDANCE

Query [srlemma="јесам"]

text_title	Left context	Pivot	Right context
Gospoda seljaci	канцеларији и среском мутваку. А кад га	је	, доцне, у неким пословима, који су се само ноћу
Deset para	; а кад ми није воља, изостајао	сам	, доцнио се, тукао сам се с ђацима, губио своје
Čedomir Ilić	, удружљиви и безлични, ови млади људи	су	, доцније, махом главе парламентарних режима. Други тип је био
Nove	стара Јерменка воли га!... " Ја	сам	, драга госпођице, сто пута пожелела да свој уништим, да
Dragocena ogrlica	развесели и сигурним гласом дода̂де: «Сретан	сам	, драга моја Андронито, што вам у напред казати могу,
Đakon Bogorodičine crkve	нешто у себи. — У тешким часовима	смо	, драги Иринеје, сви без разлике само нејаки људи. Шта
Opštinsko dete	шнајдер — ево ову. — Дајте ми	је	, драги г. Јово. — Како? Да вам је
Gospoda seljaci	то су точкови на тој машинерији. Ми	смо	, драги мој, точкови; ми капетани и ви што сте
Dragocena ogrlica	повикао је тронуто.... «И ти	си	, драги мој Даженоа, сигуран, да се на том тужном
Đakon Bogorodičine crkve	зле, крвожедне и бестидне, и она	је	, држећи се у ужасу за главу, бежала испред њих с
Čedomir Ilić	дода̂де госпођа после неколико тренутака. — Празна	је	, држим само неколико венаца лука. — А шта ћеш му
Opštinsko dete	Сима — према стању ствари, оваквом каквом	је	, другим речима према фактима, ми моментално немамо деце. —
Uvela ruža	, и очајног. Беше устала. Зверала	си	, дрхтала, нијала си се и као хтела где да побегнеш
Nove	!... Дај да те закитим... Кити	је	, дршћу јој руке. — Слатка суто, шта ти је
Nove	па од радости заплака. — А где	су	, душо, моја писма? Она поцрвене, збуни се,
Nove	горд, јунак, мушко! Али ти	си	, душо, сад слободна... Сад си распуштеница Осман-бејова, и
Nove	истрча се са узвиком: — Шта ти	је	, душо? — Е, шта, турских жена болест —
Prosioci	само једио просто сеоско удомљење... И зато	смо	, ево, дошли, да, с пристанком моје властодавке и
Hadži Đera	и да Милоја нема у кући. Ја	сам	, ево, и дошао, да Милоја уклонимо још данас,
Ikonija, vezirova majka	жено, згрешио више него Богу! Али	јесте	, ево видим, Бог ми се свети. Његова освета није
Rajko od Rasine	— „ Није, није оче! Чула	сам	, ево и сад чујем лавеж и скамукање њихово! Ej тешко
Kaluđer iz Rusije	су говорили да је Петроград Париз, то	јест	, европска, а не руска варош. Она им је казала
Đakon Bogorodičine crkve	ми је ова црква толико омилела, да	сам	, ето, остао у њој, иако сте Ви далеко од
Hadži Đera	виђена и заклоњена, а сада? Сада	је	, ето, сама; сада нема коме да се потужи,
Hadži Đera	у траг Рамову убици. А сад му	је	, ето, ушао. Ова добра вест, што је Рупић
Ikonija, vezirova majka	, толико се изумила нисам. А да	јесам	, ех, право да ти кажем, има других соколова за
Opštinsko dete	иде у општину. — Ама што ти	је	, жено, — ишчуђава се Роска — па сутра је петак
Ikonija, vezirova majka	Иконија озбиљно и тужно..» Теби	сам	, жено, згрешио више него Богу! Али јесте, ево
Prve žrtve	за ловачки поход буду што свечаније. Соколови	су	, жељни плена, кликтали на дворишту, кад су осетили час
Opštinsko dete	, ти поклони Паји писару. Перса:	Нисам	, жива ми мајка! Позови га и питај. Ја:
Jedna ugašena zvezda	син Клева и његова прекрасна кћер Аруџа-Дара били	су	, живели су бујним и величанственим животом, били су алем камен
Opštinsko dete	и цео остали збор кликну: „ Тако	је	, живео! " Затим је узео реч пречасни г. архимандрит
Čedomir Ilić	пустити се у њега. Јер какав да	је	, живот постоји, он се не може мењати у својој суштини
Prosioci	цвете мој миљени?... " „ Срећни	смо	, животе мој! " „ Али се на небу наше среће

401 - 500 / 35997

KONKORDANCE



KONKORDANCE

Query [srlemma="јесам"]

text_title	Left context	Pivot	Right context
Gospoda seljaci	канцеларији и среском мутваку. А кад га	је	, доцне, у неким
Deset para	; а кад ми није воља, изостајао	сам	, доцнио се, тук
Čedomir Ilić	, удружљиви и безлични, ови млади људи	су	, доцније, махом
Nove	стара Јерменка воли га!... " Ја	сам	, драга госпођиц
Dragocena ogrlica	развесели и сигурним гласом додаде: «Сретан	сам	, драга моја Анд
Đakon Bogorodičine crkve	нешто у себи. — У тешким часовима	смо	, драги Иринеје,
Opštinsko dete	шнајдер — ево ову. — Дајте ми	је	, драги г. Јово. –
Gospoda seljaci	то су точкови на тој машинерији. Ми	смо	, драги мој, точк
Dragocena ogrlica	повикао је тронуто.... «И ти	си	, драги мој Даже
Đakon Bogorodičine crkve	зле, крвожедне и бестидне, и она	је	, држећи се у уж
Čedomir Ilić	додаде госпођа после неколико тренутака. — Празна	је	, држим само не
Opštinsko dete	Сима — према стању ствари, оваквом каквом	је	, другим речима према фактима, ми моментално немамо деце. —
Uvela ruža	, и очајног. Беше устала. Зверала	си	, дрхтала, нијала си се и као хтела где да побегнеш
Nove	!... Дај да те закитим... Кити	је	, дршћу јој руке. — Слатка суто, шта ти је
Nove	па од радости заплака. — А где	су	, душо, моја писма? Она поцрвене, збуни се,
Nove	горд, јунак, мушко! Али ти	си	, душо, сад слободна... Сад си распуштеница Осман-бејова, и
Nove	истрча се са узвиком: — Шта ти	је	, душо? — Е, шта, турских жена болест —
Prosioci	само једио просто сеоско удомљење... И зато	смо	, ево, дошли, да, с пристанком моје властодавке и
Hadži Đera	и да Милоја нема у кући. Ја	сам	, ево, и дошао, да Милоја уклонимо још данас,
Ikonija, vezirova majka	жено, згрешио више него Богу! Али	јесте	, ево видим, Бог ми се свети. Његова освета није
Rajko od Rasine	— „ Није, није оче! Чула	сам	, ево и сад чујем лавеж и скамукање њихово! Ej тешко
Kaluđer iz Rusije	су говорили да је Петроград Париз, то	јест	, европска, а не руска варош. Она им је казала
Đakon Bogorodičine crkve	ми је ова црква толико омилела, да	сам	, ето, остао у њој, иако сте Ви далеко од
Hadži Đera	виђена и заклоњена, а сада? Сада	је	, ето, сама; сада нема коме да се потужи,
Hadži Đera	у траг Рамову убици. А сад му	је	, ето, ушао. Ова добра вест, што је Рупић
Ikonija, vezirova majka	, толико се изумила нисам. А да	јесам	, ех, право да ти кажем, има других соколова за
Opštinsko dete	иде у општину. — Ама што ти	је	, жено, — ишчуђава се Роска — па сутра је петак
Ikonija, vezirova majka	Иконија озбиљно и тужно..» Теби	сам	, жено, згрешио више него Богу! Али јесте, ево
Prve žrtve	за ловачки поход буду што свечаније. Соколови	су	, жељни плена, кликтали на двориште, кад су осетили час
Opštinsko dete	, ти поклони Паји писару. Перса:	Нисам	, жива ми мајка! Позови га и питај. Ја:
Jedna ugašena zvezda	син Клева и његова прекрасна кћер Аруџа-Дара били	су	, живели су бујним и величанственим животом, били су алем камен
Opštinsko dete	и цео остали збор кликну: „ Тако	је	, живео! " Затим је узео реч пречасни г. архимандрит
Čedomir Ilić	пустити се у њега. Јер какав да	је	, живот постоји, он се не може мењати у својој суштини
Prosioci	цвете мој миљени?... " „ Срећни	смо	, животе мој! " „ Али се на небу наше среће

Left context size 8

Right context size 12

Cancel OK

401 - 500 / 35997

KONKORDANCE

Query [srlemma="јесам"]

text_title	Left context	Pivot	Right context
Gospoda seljaci	канцеларији и среском мутваку. А кад га	је	, доцне, у неким пословима, који су се само ноћу
Deset para	; а кад ми није воља, изостајао	сам	, доцнио се, тукао сам се с ђацима, губио своје
Čedomir Ilić	, удружљиви и безлични, ови млади људи	су	, доцније, махом главе парламентарних режима. Други тип је био
Nove	стара Јерменка воли га!... " Ја	сам	, драга госпођице, сто пута пожелела да свој уништим, да
Dragocena ogrlica	развесели и сигурним гласом дода̂де: «Сретан	сам	, драга моја Андронито, што вам у напред казати могу,
Đakon Bogorodičine crkve	нешто у себи. — У тешким часовима	смо	, драги Иринеје, сви без разлике само нејаки људи. Шта
Opštinsko dete	шнајдер — ево ову. — Дајте ми	је	, драги г. Јово. — Како? Да вам је
Gospoda seljaci	то су точкови на тој машинерији. Ми	смо	, драги мој, точкови; ми капетани и ви што сте
Dragocena ogrlica	повикао је тронуто.... «И ти	си	, драги мој Даженоа, сигуран, да се на том тужном
Đakon Bogorodičine crkve	зле, крвожедне и бестидне, и она	је	, држећи се у ужасу за главу, бежала испред њих с
Čedomir Ilić	дода̂де госпођа после неколико тренутака. — Празна	је	, држим само неколико венаца лука. — А шта ћеш му
Opštinsko dete	Сима — према стању ствари, оваквом каквом	је	, другим речима према фактима, ми моментално немамо деце. —
Uvela ruža	, и очајног. Беше устала. Зверала	си	, дрхтала, нијала си се и као хтела где да побегнеш
Nove	!... Дај да те закитим... Кити	је	,
Nove	па од радости заплака. — А где	су	,
Nove	горд, јунак, мушко! Али ти	си	, и
Nove	истрча се са узвиком: — Шта ти	је	,
Prosioci	само једио просто сеоско удомљење... И зато	смо	,
Hadži Đera	и да Милоја нема у кући. Ја	сам	,
Ikonija, vezirova majka	жено, згрешио више него Богу! Али	јесте	,
Rajko od Rasine	— „ Није, није оче! Чула	сам	,
Kaluđer iz Rusije	су говорили да је Петроград Париз, то	јест	,
Đakon Bogorodičine crkve	ми је ова црква толико омилела, да	сам	,
Hadži Đera	виђена и заклоњена, а сада? Сада	је	,
Hadži Đera	у траг Рамову убици. А сад му	је	,
Ikonija, vezirova majka	, толико се изумила нисам. А да	јесам	,
Opštinsko dete	иде у општину. — Ама што ти	је	, жено, — ишчуђава се Роска — па сутра је петак
Ikonija, vezirova majka	Иконија озбиљно и тужно..» Теби	сам	, жено, згрешио више него Богу! Али јесте, ево
Prve žrtve	за ловачки поход буду што свечаније. Соколови	су	, жељни плена, кликтали на двориште, кад су осетили час
Opštinsko dete	, ти поклони Паји писару. Перса:	Нисам	, жива ми мајка! Позови га и питај. Ја:
Jedna ugašena zvezda	син Клева и његова прекрасна кћер Аруџа-Дара били	су	, живели су бујним и величанственим животом, били су алем камен
Opštinsko dete	и цео остали збор кликну: „ Тако	је	, живео! " Затим је узео реч пречасни г. архимандрит
Čedomir Ilić	пустити се у њега. Јер какав да	је	, живот постоји, он се не може мењати у својој суштини
Prosioci	цвете мој миљени?... " „ Срећни	смо	, животе мој! " „ Али се на небу наше среће

Sort options

id
n
srlemma
srpos

word

Cancel OK

401 - 500 / 35997

REGULARNI IZRAZI

Regularni izraz je izraz koji opisuje skup niski, u skladu sa određenim sintaksnim pravilima.

Primer

сународници, међународни, народске, народности, …

odgovara regularnom izrazu **[srlemma = ".*народ.*"]**

word	Frequency
народности	2
народске	2
Народни	1
Народну	1
међународни	1
народнога	1
народност	1
сународница	1
сународнице	1
сународници	1

SRPROMAN:[word="међународни"]

Query: [word="међународни"] Keyword: word Edit Search

sort keys: #1 None #2 None #3 None #4 None

1 - 1 / 1 Hide settings

text_id	Left context	Keyword	Right context
SRP18960…	, господин Проко, на премер, као	међународни	човек, што међу народом живит

SRP сународници 1 54

треба да искажем у овом тренутку свога живота, почаствован од овог одличног скупа да ступим у чувену и одличну партију, којом руководи на далеко чувени Вођа и господин Шеф наш, а коју у нашем срезу и округу тако мудро предводи многопоштовани дугогодишњи посланик, господин Продан Жмурић … И, заиста, ви, господин Проко, на премер, као међународни човек, што међу народом живите, знате колика потреба захтева, да у народу има изображени људи, који ћ n:10941 srpos:A srlemma:међународан и повести путем општег народног напретка и које ја, на премер,

default 54 / 190

UPITI U TXM OKRUŽENJU

- Jednostavna pretraga reči: vrednosti

- Pretraga svojstava

- Pretraga niski leksičkih jedinica

- Korišćenje informacija iz konteksta

Primeri

данас **.*народ.***

[word = “златан”] **ту[гж].***

мир|рат **.+патри.***

[srlemma = “пријатељ”]

[srpos = “A”]

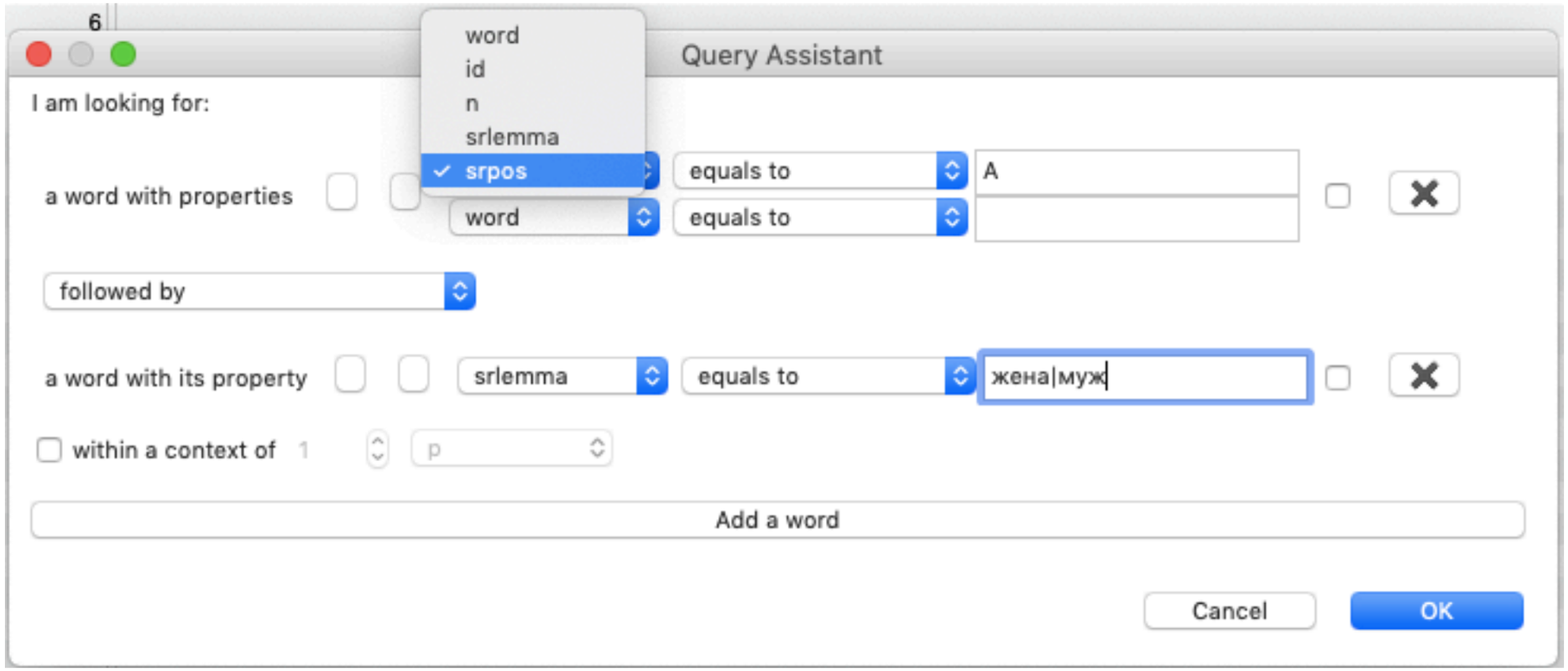
[srpos = “V” & word = игра.*]

[srlemma = “тежак”] [srpos = “N”]

[word = “срце” & _.text_pol! = “fem”]

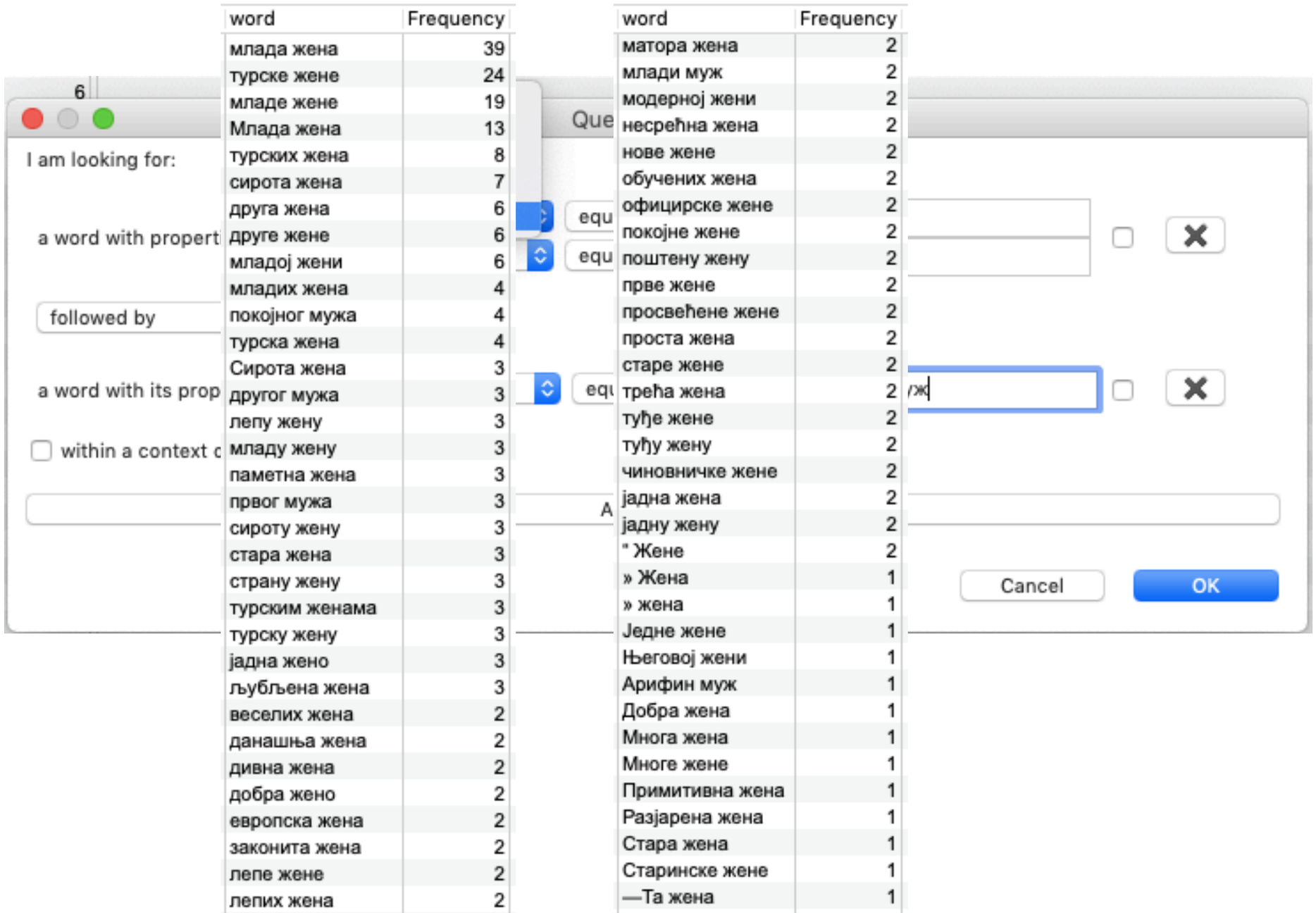
INDEKS

Lista frekvencija svojstava određenih pojavljivanja jednog CQL upita izvršenog nad korpusom, potkorpusom ili particijom



INDEKS

Lista frekvencija svojstava određenih pojavljivanja jednog CQL upita izvršenog nad korpusom, potkorpusom ili particijom

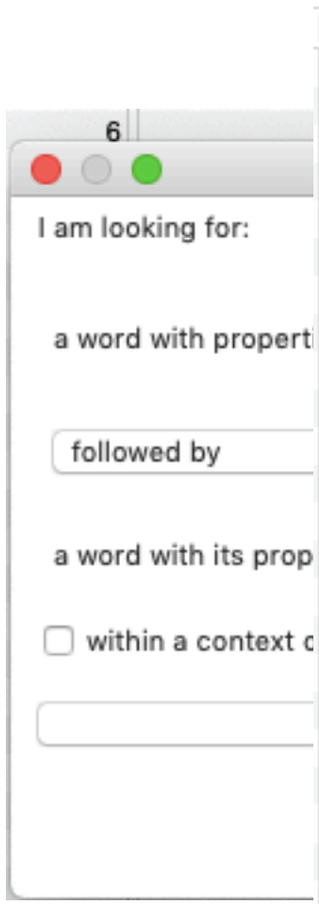


word	Frequency
млада жена	39
турске жене	24
младе жене	19
Млада жена	13
турских жена	8
сирота жена	7
друга жена	6
друге жене	6
младој жени	6
младих жена	4
покојног мужа	4
турска жена	4
Сирота жена	3
другог мужа	3
лепу жену	3
младу жену	3
паметна жена	3
првог мужа	3
сироту жену	3
стара жена	3
страну жену	3
турским женама	3
турску жену	3
јадна жено	3
љубљена жена	3
веселих жена	2
данашња жена	2
дивна жена	2
добра жено	2
европска жена	2
законита жена	2
лепе жене	2
лепих жена	2

word	Frequency
матора жена	2
млади муж	2
модерној жени	2
несрећна жена	2
нове жене	2
обучених жена	2
официрске жене	2
покојне жене	2
поштену жену	2
прве жене	2
просвећене жене	2
проста жена	2
старе жене	2
трећа жена	2
туђе жене	2
туђу жену	2
чиновничке жене	2
јадна жена	2
јадну жену	2
“ Жене	2
» Жена	1
» жена	1
Једне жене	1
Његовој жени	1
Арифин муж	1
Добра жена	1
Многа жена	1
Многе жене	1
Примитивна жена	1
Разјарена жена	1
Стара жена	1
Старинске жене	1
—Та жена	1

INDEKS

Lista frekvencija svojstava određenih pojavljivanja jednog CQL upita izvršenog nad korpusom, potkorpusom ili particijom



word	Frequency
млада жена	39
турске жене	24
младе жене	19
Млада жена	13
турских жена	8
сирота жена	7
друга жена	6
друге жене	6
младој жени	6
младих жена	4
покојног мужа	4
турска жена	4
Сирота жена	3
другог мужа	3
лепу жену	3
младу жену	3
паметна жена	3
првог мужа	3
сироту жену	3
стара жена	3
страну жену	3
турским женама	3
турску жену	3
јадна жено	3
љубљена жена	3
веселих жена	2
данашња жена	2
дивна жена	2
добра жено	2
европска жена	2
законита жена	2
лепе жене	2
лепих жена	2

word	Frequency
матора жена	2
млади муж	2
модерној жени	2
несрећна жена	2
нове жене	2
обучених жена	2
официрске жене	2
покојне жене	2
поштену жену	2
прве жене	2
просвећене жене	2
проста жена	2
старе жене	2
трећа жена	2
туђе жене	2
туђу жену	2
чиновничке жене	2
јадна жена	2
јадну жену	2
" Жене | 2 |
» Жена | 1 |
» жена | 1 |
Једне жене | 1 |
Његовој жени | 1 |
Арифин муж | 1 |
Добра жена | 1 |
Многа жена | 1 |
Многе жене | 1 |
Примитивна жена | 1 |
Разјарена жена | 1 |
Стара жена | 1 |
Старинске жене | 1 |
—Та жена | 1 |

srlemma	Frequency
млад жена	86
турски жена	43
други жена	13
сирот жена	13
леп жена	11
стар жена	11
јадан жена	8
туђ жена	6
љубљен жена	6
покојни муж	5
први муж	5
стран жена	5
добар жена	4
модеран жена	4
први жена	4
вољен жена	3
данашњи жена	3
други муж	3
европски жена	3
млад муж	3
паметан жена	3
поштен жена	3
прост жена	3
слаб жена	3
" жена	3
» жена	2
весео жена	2
диван жена	2
добар муж	2
жив жена	2
законит жена	2
заслужан муж	2
заљубљен жена	2

ZAJEDNIČKA POJAVLJIVANJA

Query [srlemma = "град.*"]

Parameters

Cooccurrents properties: srlemma Edit Thresholds: Fmin ≥ 2 Cmin ≥ 2 Score ≥ 2.0

Context: word structure back Use the left context Use the right context include the structure containing the pivot in the count

from - 9 to - 0 and from 0 to 9

Cooccurrent	Frequency	CoFrequency	Score	Mean distance
зидина	54	13	16	1.9
бедем	30	11	16	2.3
подграђе	16	9	15	3.4
у	15050	179	14	3.2
овај	4277	75	13	3.4
заповедник	21	7	10	1.6
властела	11	5	8	4.4
од	5053	69	8	3.2
платно	43	7	8	1.0
стража	22	5	6	.0
Љубоја	69	7	6	3.4
главни	139	9	6	4.3
кула	81	7	6	4.1
негда	36	5	5	3.8
царски	104	7	5	1.3
војска	147	8	5	5.0
град	196	9	5	4.7
рушевина	19	4	5	3.2
старешина	21	4	5	3.0
градски	47	5	4	5.6
деспотовина	25	4	4	5.2
царевина	28	4	4	6.5
муслиман	10	3	4	2.3
ударити	97	6	4	4.5

t pivot 280, v cooc 132, t cooc 0, T corpus 855412

ZAJEDNIČKA POJAVLJIVANJA

Query [srlemma = "град.*"]

Parameters

Query ([srlemma = "град.*"] []* [srlemma="зидина"]) | ([srlemma="зидина"] []* [srlemma = "град.*"]) within 10

Parameters

Sort keys #1 None #2 None #3 None #4 None Sort

	References	Left context	Pivot	Right context
View	text_id ...	word ...	word ...	word ...
Sort	text_id ...	word ...	word ...	word ...
Size		8		12

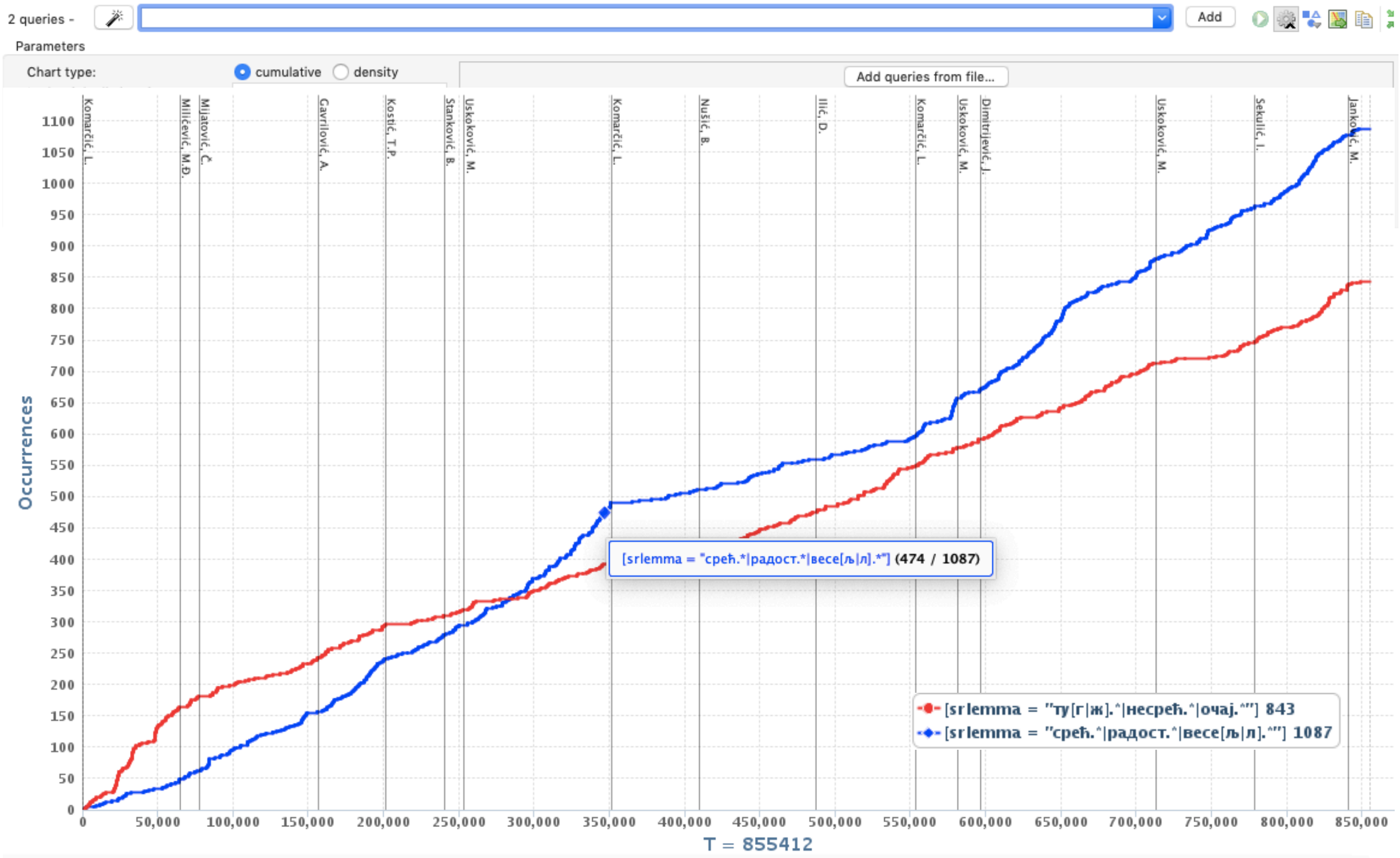
text_id	Left context	Pivot	Right context
SRP18920_	платна можете најбоље да видите, како је	град некада имао двојаких зидина	, као оно што кажу да Цариград има! Па онда баш
SRP18920_	. 2 Ваљало им је проћи поред западне	зидине од града	. Уз ту зидину, која се прилично одржала, прислоњена је
SRP18920_	Ваљало им је проћи поред западне зидине од	града. Уз ту зидину	, која се прилично одржала, прислоњена је била овде и онде
SRP18920_	опреми своју сиву ластавицу. 5 Под самим	градом кнеза Лазара, а прислоњена уз обрушену једну зидину	, која по свој прилици бејаше остатак од некадашњега спољњег зида градског
SRP18920_	џиновски стражари у панцирима од сребра стражарили по	зидинама градским	и кроз све јунаштво своје кришом погледали к прозорима дворским од биљура
SRP18920_	будио успаване људе на посао. Озго испод	зидина градских	зачу се куцкање чекића о наковањ. „ Ево је иде!
SRP18920_	један вис, на коме из далека видеше	зидине од некаквог старог порушеног града	. Али се већ мрак ухватио био кад изађоше на тај вис
SRP18920_	бих, окупана у сузама, заспала под	зидином оног старог града	у Крушевцу, ето ње к мени, на сну праћена још
SRP18920_	неколике цеви, што их видех горе у	зидинама од града	! " рече Маргита; па онда настави пуна збиље и сетности
SRP18920_	! Па онда појуре напоље, па на	зидине од града	; нањушише на цеви оловне што воде ваздух доле, па да
SRP18920_	сувог грања горела је на највишој и најширој	зидини од развалъеног града	. То је стари пустиник давао знак најближем селу под брдом,
SRP18930_	ловачки су се рогови пробе ради разлегали преко	градских зидина	. Био је већ крај месеца октобра, кад и издашна јужна
SRP19000_	црна као та ноћ, и у хладу	градских зидина	које су виделе толика столећа. Гледајући се очи у очи,

1 - 13 / 13

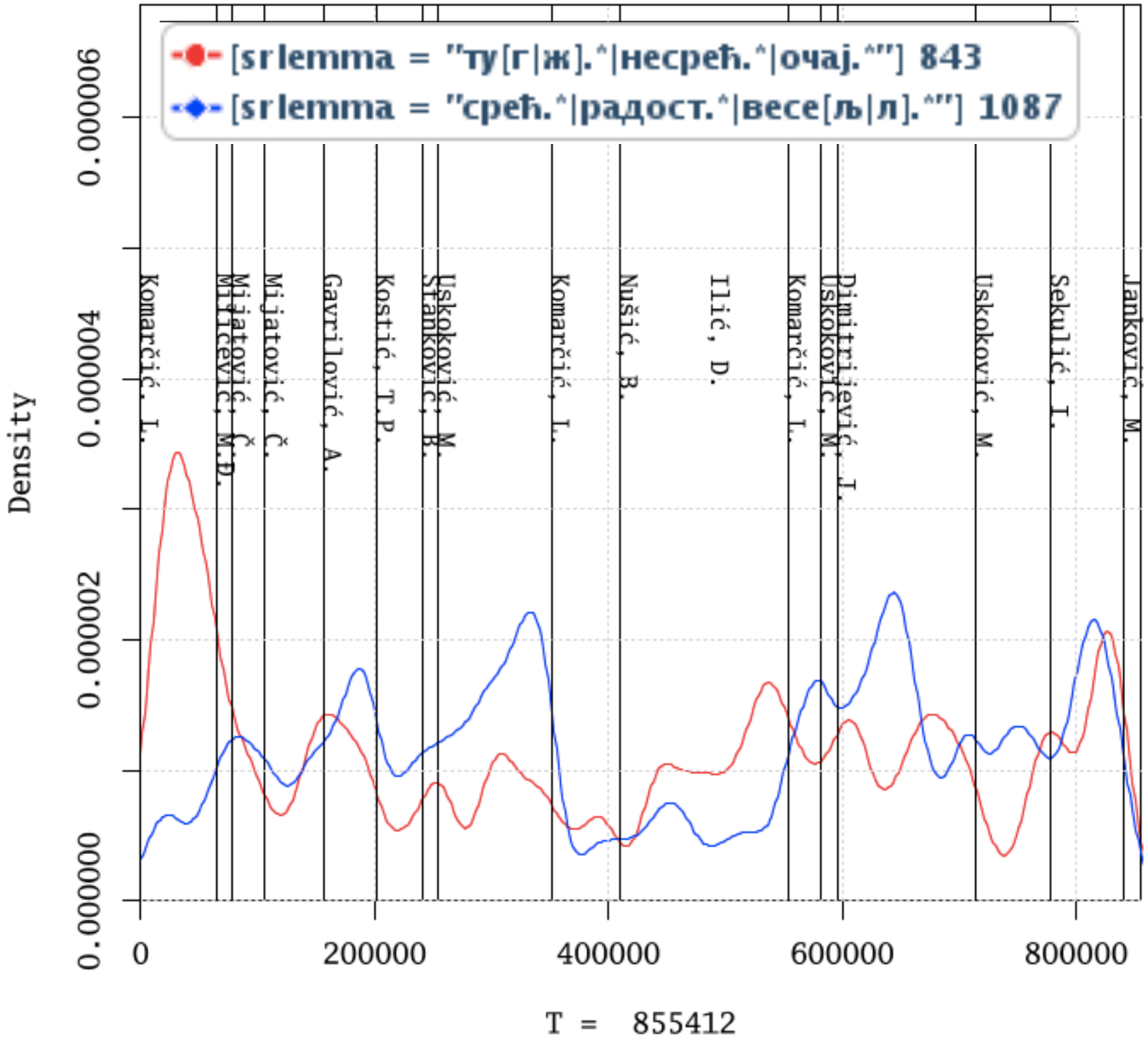
	муслиман	10	3	4	2.3
	ударити	97	6	4	4.5

t pivot 280, v cooc 132, t cooc 0, T corpus 855412

PROGRESIJA


2 queries -
Add
Parameters
Chart type:
cumulative
density
Add queries from file...
Komarčić, L.
Milićević, M.Đ.
Mijatović, Č.
Gavrilović, A.
Kostić, T.P.
Stanković, B.
Uskoković, M.
Komarčić, L.
Nušić, B.
Ilić, D.
Komarčić, L.
Uskoković, M.
Dimitrijević, J.
Uskoković, M.
Sekulić, I.
Janković, M.
Occurrences
[srlemma = "сре.*|радост.*|весе[ље].*"] (474 / 1087)
[srlemma = "ту[г|ж].^|несре.^|очај.^""] 843
[srlemma = "сре.^|радост.^|весе[ље].^""] 1087
T = 855412

PROGRESIJA


[srlemma = "ту[г|ж].^|несрећ.^|очај.^"] 843
[srlemma = "срећ.^|радост.^|весе[ље|л].^"] 1087
Density
0.000006
0.000004
0.000002
0.000000
Janković, M.
Sekulić, I.
Uskoković, M.
Dimitrijević, ...
Uskoković, M.
Komarčić, L.
Ilić, D.
Nušić, B.
Komarčić, L.
Uskoković, M.
Stanković, B.
Kostić, T.P.
Gavrilović, A.
Mijatović, Č.
Milićević, M.D.
Mijatović, Č.
Komarčić, L.
0
200000
400000
600000
800000
T = 855412

LEKSIČKE TABELE

Property srlemma

Parameters

Thresholds

Fmin 2 Number of lines 200

Merge or Delete columns Merge or Delete rows

srlemma	Frequency	Dimitrijević, J. t=80571	Gavrilović, A. t=27979	Ilić, D. t=42176	Janković, M. t=5751	Komarčić, L. t=101388	Kostić, T.P. t=25515	Mijatović, Č. t=51471
«	1104	0	0	0	51	575	0	355
»	1040	0	0	1	49	555	0	380
г	771	0	0	1	0	640	23	0
Милош	482	0	0	0	0	0	0	6
Зорка	423	0	0	0	0	0	6	1
"	469	70	5	0	2	71	4	126
;	3004	750	22	129	10	977	39	409
-	1464	14	6	5	5	22	21	19
'	695	46	85	20	0	27	214	69
госпођа	449	61	9	0	0	21	28	81
господин	528	2	0	0	3	58	116	19
одговорити	627	46	53	148	13	103	24	24
овде	490	69	34	38	2	95	13	94
седети	501	134	13	52	1	26	43	48
твој	742	46	53	27	0	120	15	137
узети	553	127	17	15	2	63	17	92
година	691	109	24	26	8	221	28	50
ти	592	74	39	35	2	54	18	110
„	2390	730	2	45	0	719	127	704
те	783	98	22	103	0	96	46	104
преко	466	63	28	38	4	85	42	40
радити	438	59	13	35	4	74	40	20
код	593	118	28	19	2	91	33	12
колико	499	59	32	47	3	100	28	26
немати	430	77	21	31	4	79	25	35
наћи	581	44	33	47	9	127	15	56
соб	478	127	15	13	1	54	35	14
”	3034	802	12	44	0	871	136	806
право	535	47	27	11	3	148	11	73
мајка	785	268	14	39	3	119	10	43
три	461	109	16	21	1	79	22	56
отићи	623	173	39	52	7	88	20	41
тек	531	33	14	64	2	61	21	81
сунце	524	28	10	8	4	274	6	27
наш	1213	163	85	11	2	429	92	83
страна	543	34	65	58	1	130	14	48

T 560648 V 200 Fmin 416 Fmax 69029

SPECIFIČNOSTI

Model verovatnoće (Lafon, 1980) zasnovan na hipergeomatrijskoj raspodeli omogućava:

- Proučavanje distribucije učestalosti reči/svojstava u (pot)korpusu podeljenom na nekoliko delova
- Poređenje delova u smislu specifične (višak/deficit) ili osnovne upotrebe reči/svojstava

Indeks specifičnosti
znak (+/-) ako je posmatrana frekvencija veća ili manja nego u “normalnoj” distribuciji (uzimajući u obzir veličinu dela u odnosu na celinu)

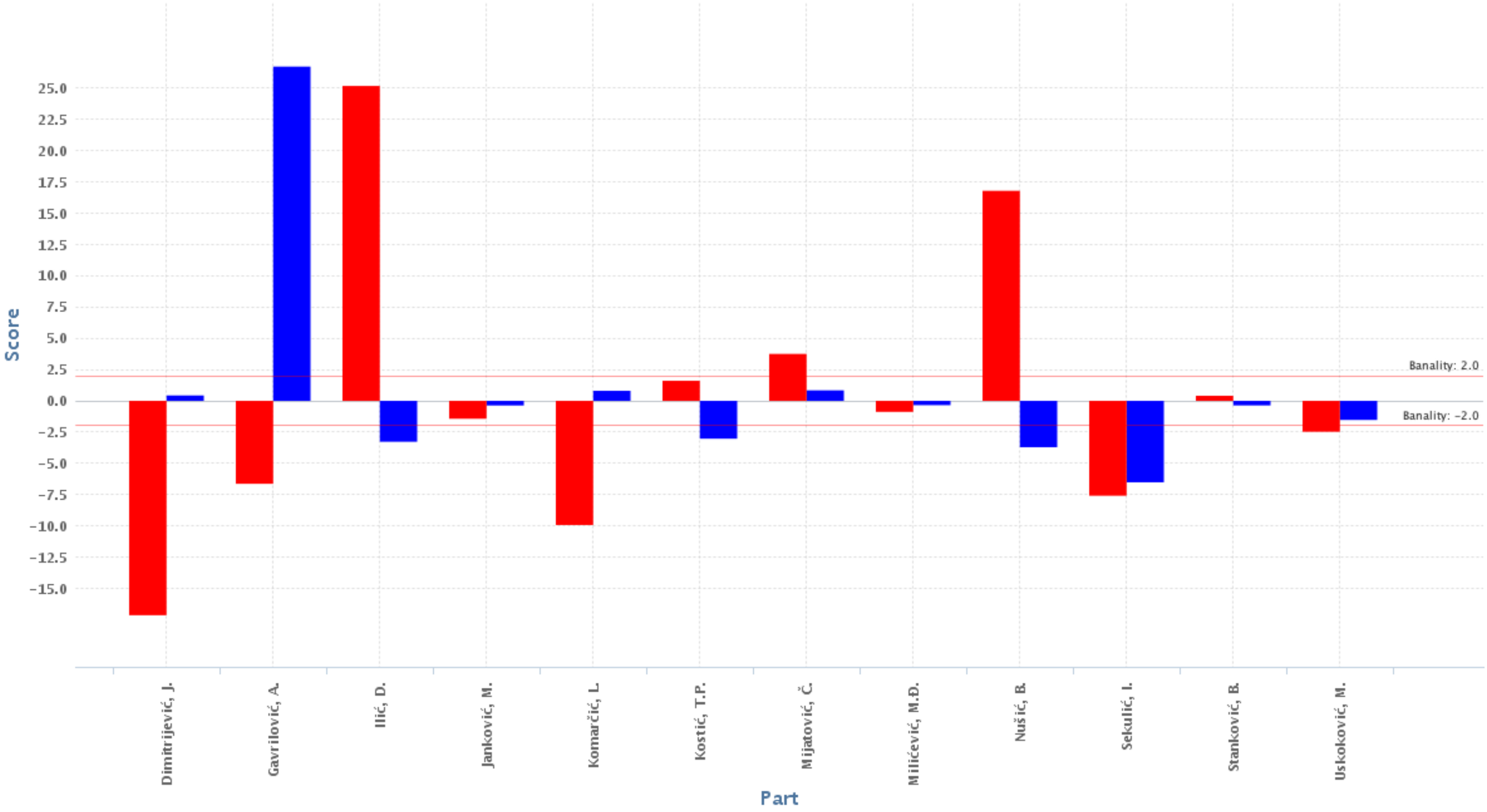
Query [srlemma = "град|село"]

srlemma	Frequency T=847992	Dimitrijević, J. t=116782	Gavrilović, A. t=44929	Ilić, D. t=65554
село	335	3	1	90
град	196	29	58	4

Property srlemma

Units	Frequency T 847992		Dimitrijević, J. t=116782	index	Gavrilović, A. t=44929	index	Ilić, D. t=65554	index
#RESTE#	847461		116750	8.1	44870	-7.0	65460	-13.3
град	196		29	0.4	58	26.7	4	-3.3
село	335		3	-17.1	1	-6.6	90	25.2

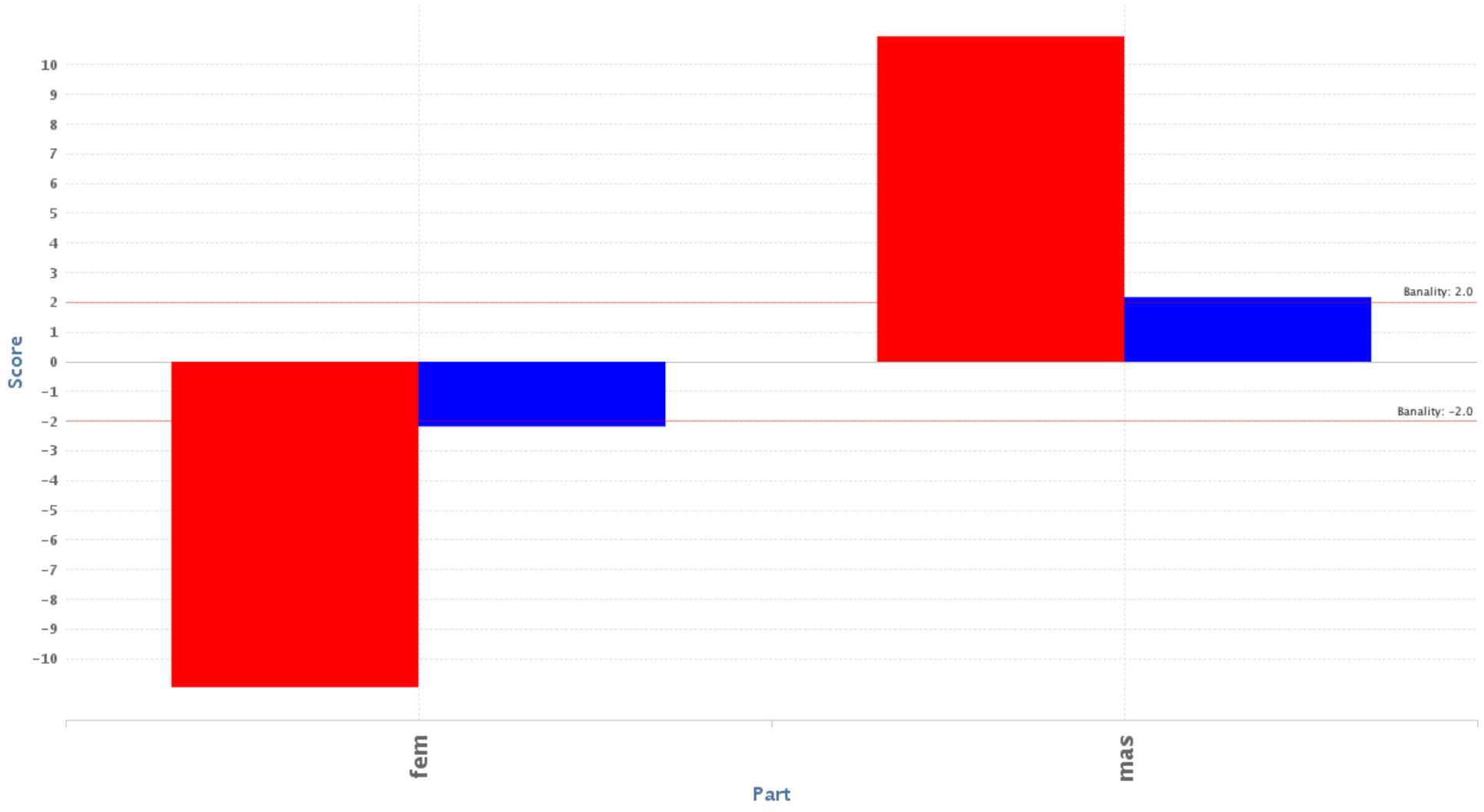
GRAFIČKI PRIKAZ SPECIFIČNOSTI



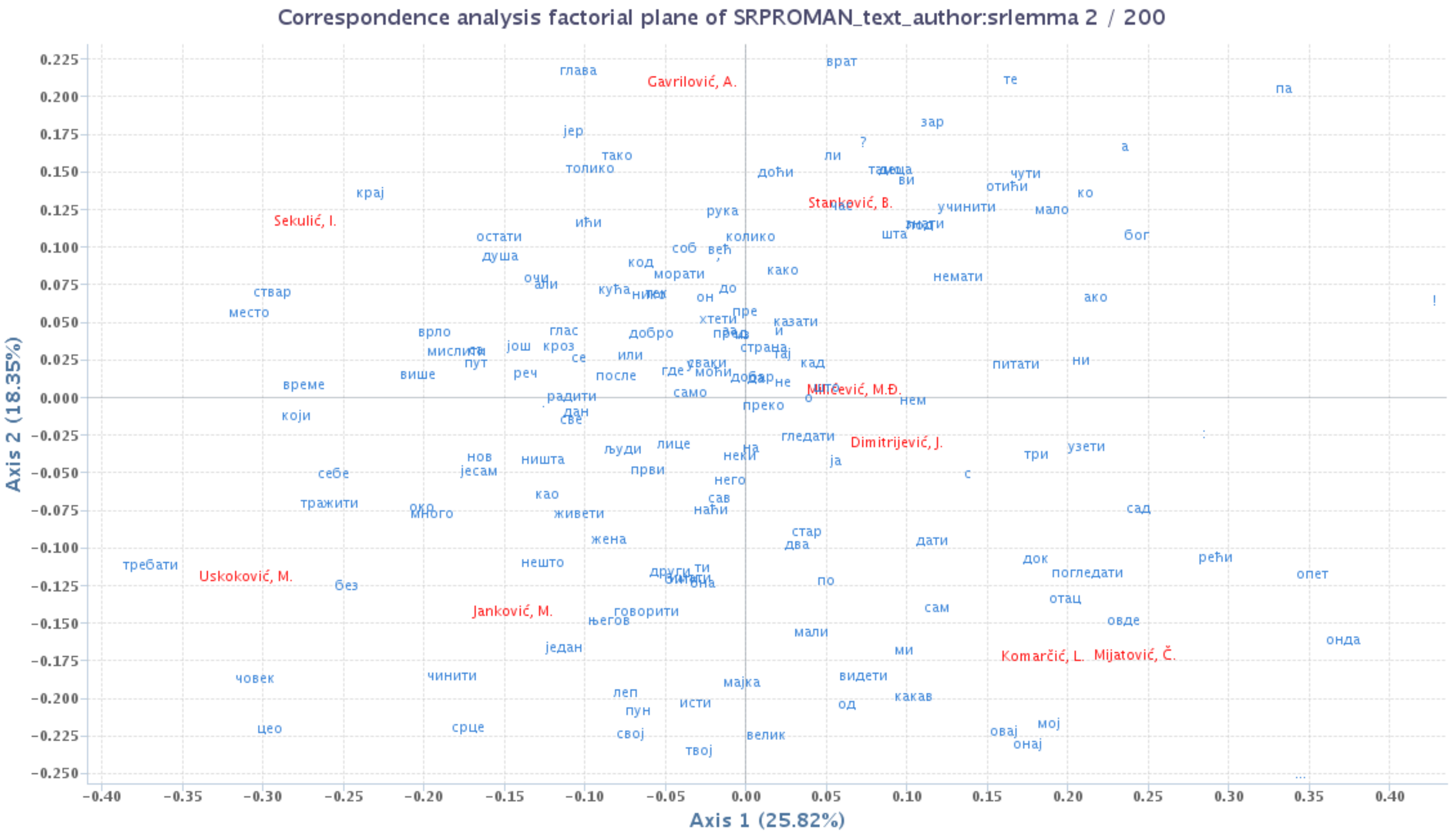
SPECIFIČNOSTI

Property word

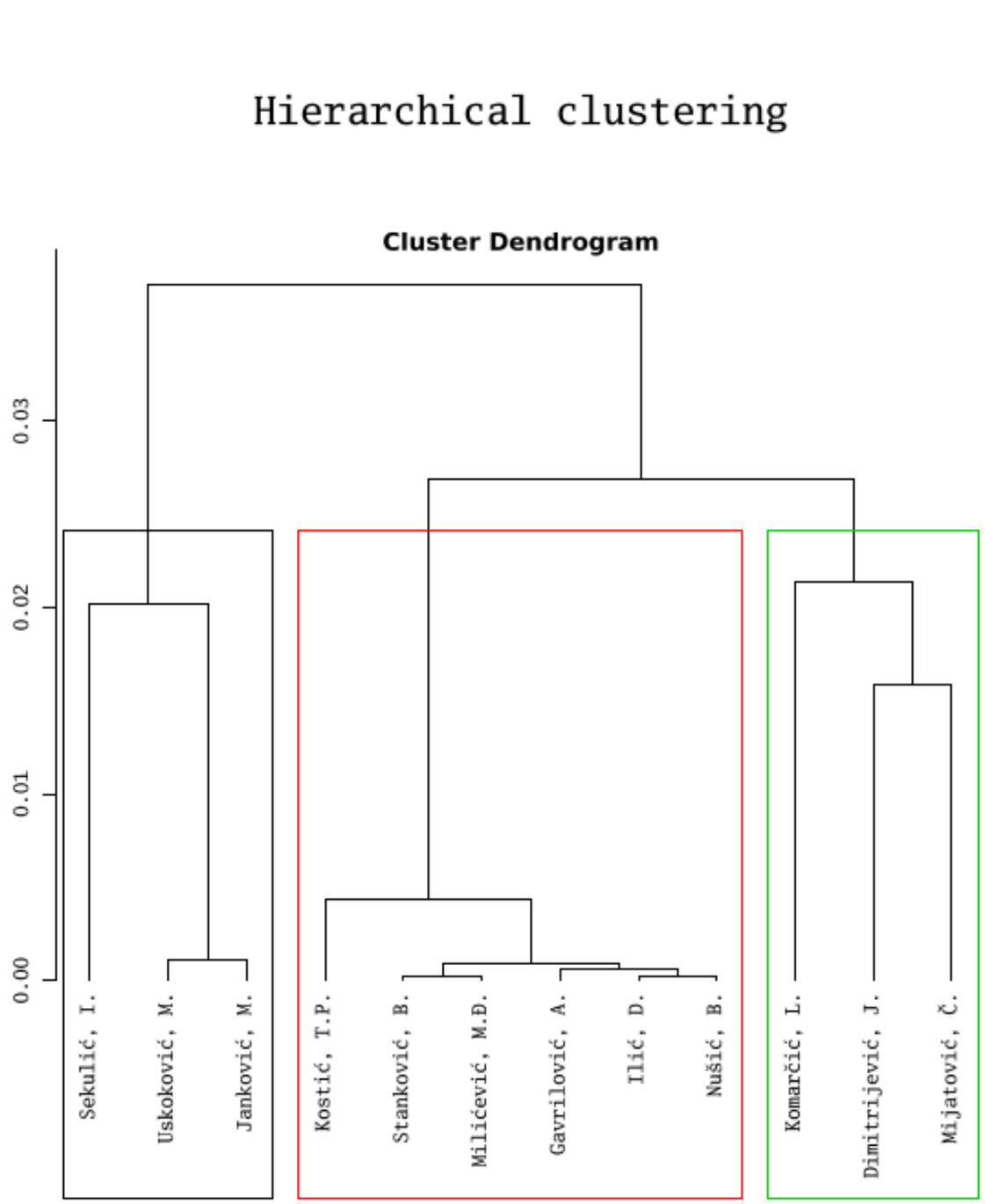
Units	Frequency T 847992		fem t=187475	index	mas t=660517	index
#RESTE#	847827		187469	10.7	660358	-10.7
град	64		6	-2.2	58	2.2
село	101		0	-11.0	101	11.0

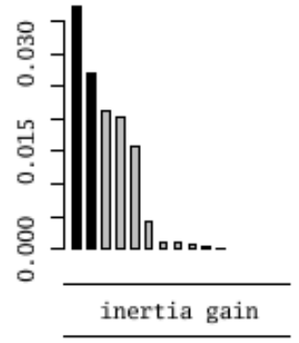


FAKTORSKA ANALIZA

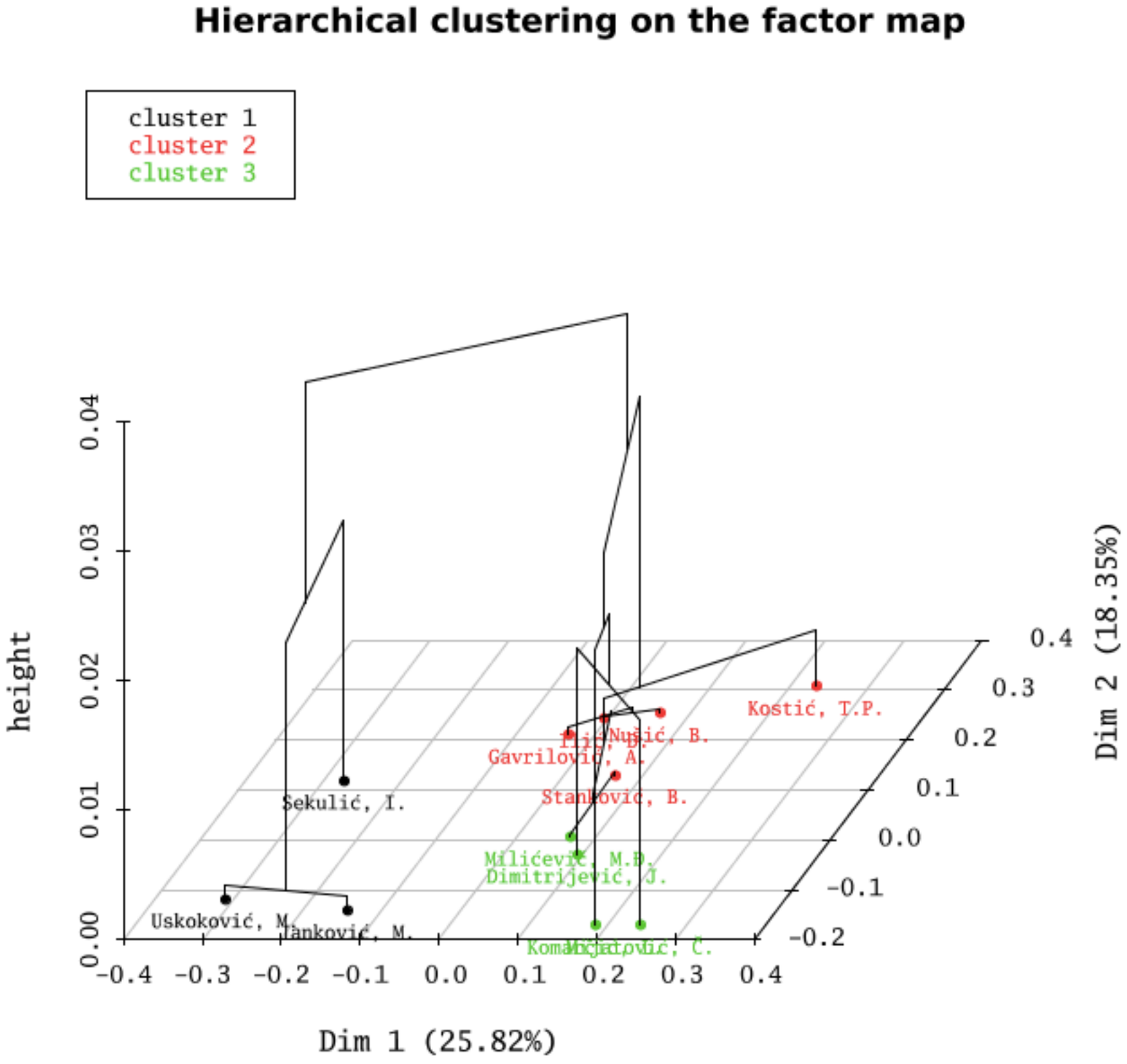


KLASTER ANALIZA





KLASTER ANALIZA



ANOTACIJA KORPUSA

Nekoliko alata za anotaciju korpusa razvijeno je u okviru TXM okruženja:

- Na osnovu CQP modela korpusa
  - Anotiranje svojstava reči, koje čine konkordance, kreiranjem novih ili ispravljanjem postojećih svojstava reči
  - Jednostavno ili napredno anotiranje reči ili niski reči, koje su predstavljene konkordancama, novim strukturama tekstova i njihovim vrednostima
- Na osnovu integrisanog URS (Unit-Relation-Scheme) modela za anotiranje reči ili niski reči

ŠTA BISMO JOŠ MOGLI

Usavršiti TreeTgger model za srpski jezik

Anotacija korpusa

Radionice za lingviste i istraživače iz oblasti humanističkih nauka Srbije

HVALA NA PAŽNJI!

jelena.jacimovic@stomf.bg.ac.rs

Acknowledgements

COST Action 16204 – Distant Reading for European Literary History

STSM-CA16204-42562

Host institution: Institut d'Histoire des Représentations et des Idées dans les Modernités UMR 5317, École Normale Supérieure de Lyon, Lyon, France