О српским корпусима и корпусима српског језика, а посебно оним који су на вебу

Душко Витас

О наслову, не баш срећно изабраном!

Једно могуће разврставање корпуса:

- српски корпуси – настали у Србији, а посебно међу члановима ГЈТ, дакле ***домаћи***

Међу њима су и корпуси без српске компоненте, нпр. савременог књижевног хрватског језика или паралелни француско-шпански корпус (из Европарл-а)

- корпуси српског језика, који садрже као компоненту материјал српског, настали овде или ***по белом свету***, као, на пример, Хенингов корпус или српски део АСПАК-а или корпуси видљиви на вебу као *SrWaC* и *SETimes*.

О наслову, не баш срећно изабраном!

С друге стране, међу овим корпусима има

- корпуса ***приватних*** чије је коришћење ограничено на мали (или врло мали) број корисника и који се не могу наћи на вебу

и

- корпуса ***доступних преко веба***, који се могу преузети или само консултовати...

Појам корпуса

У домаћим срединама се појам корпуса тумачи(о) на различите начине. На пример, било је оних који су сматрали да су листиће из грађе за РСАНУ – ***корпус***. Прецизније би било одредити овај "корпус" као конкорданце састављене над садржајем библиотеке ИСЈ где критеријум за састављање конкорданци није формално дефинисан и темељи се на префињеном разумевању прочитаних текстова.

Има и других

Шта недостаје? Пре свега, фреквенције.

Историја корпуса савременог српског језика

Први импулс – скуп ***Компјутерска обрада лингвистичких података***, одржан крајем 1977. у Сарајеву у организацији Милана Шипке.

У Хрватској – преводни Браун-корпус (Р. Филиповић), Једномилијунски (М. Могуш)

У Словенији – различити софтвер за обраду језика

у Србији – експерименти са препознавањем говора

али о обради текста ни помена у Србији!

Почеци

Први алат, уместо програма ***COCOA***, домаћи продукт ***Аурора*** (АУтоматска Рутина за Обраду РечникА):

Duško Vitas, Prikaz jednog sistema za automatsku obradu teksta, simp. ***INFORMATICA'79***, Bled, oktobar 1979, pp. 7 101

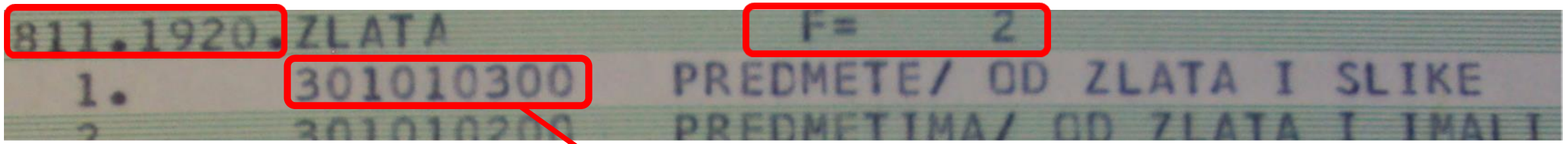
Својства: структура текста, инвертовани текст, независно од алфабета

Резултати: конкорданце, фреквенцијар, атерго...

Корпус и конкорданце


```
809.2147.ZIDARI          F=    1
  1.     302010100    +KAMENORESCI, ZIDARI I VAJARI
810. 799.ZIDOVIMA        F=    4
  1.     301010300    SLIKE PO ZIDOVIMA SVOJIH DVORACA.
  2.     301010200    NA NXENIM ZIDOVIMA, LEPO URADXENE
  3.     202010300    SLIKANI/ NA ZIDOVIMA HRAMOVA I
  4.
811.1920.ZLATA           F=    2
  1.     301010300    PREDMETE/ OD ZLATA I SLIKE
  2.
812.2031.ZLATNE          F=    1
  1.     301020200    IZRA-/ DXIVALI ZLATNE PREDMETE KA
813.1054.ZLATO           F=    1
  1.     102010100    TALE - ZLATO, OLOVO I
.3149.ZLATOM             F=    1
.      100000000    I POKRIVALE ZLATOM I/ SLONOVOM K
.2731.ZLI                F=    1
.        101 100    SHVATANXA/ DA ZLI DUHOVI ULAZE U
. 188.ZNA                F=    1
.          020100    AKO SE ZNA KADA SE
.2045.ZNAKA              F=    1
.      301020200    OD 22 ZNAKA.
.1500.ZNAKOVA            F=    1
.      202010100    +TIH JE/ ZNAKOVA BILO MNOGO
. 549.ZNALI              F=    1
.        2010300    LXUDI/ NISU ZNALI DA PISXU,
.1705.ZNALO              F=    2
.      202030100    NIJE SE ZNALO/ NI ZA
.      202030100    MATERIJAL NIJE ZNALO./
. 311.ZNAMO              F=    1
.        1020100    ZXELIMO DA ZNAMO KO-/ LIKO
.2161.ZNATNO             F=    2
.        1020200    +ON SE ZNATNO RAZLIKOVAO/ OD V
.      302010200    BIO JE ZNATNO TEZXI NEGO
. 655.ZNATNOM            F=    1
.        2010300    POSTOJI U ZNATNOM BROJU ZEMALX
824.1605.ZNACI           F=    2
  1.     202010300    PISMO./ +ZNACI SU SE
  2.     202010300    +EGIPTU SE ZNACI ZA PISANXE
```

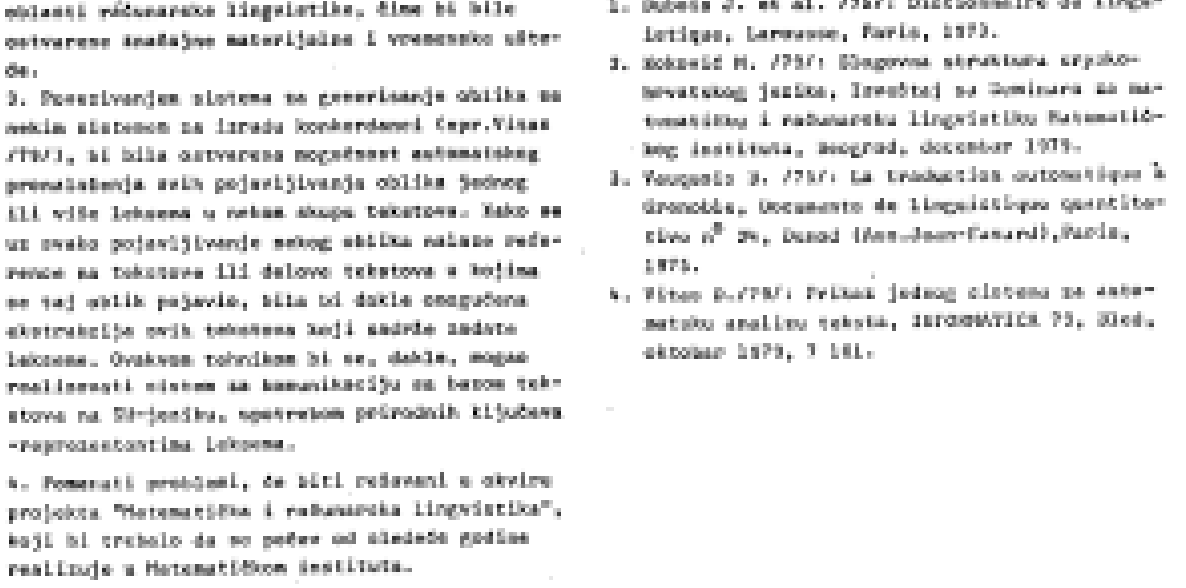
Анатомија конкорданци



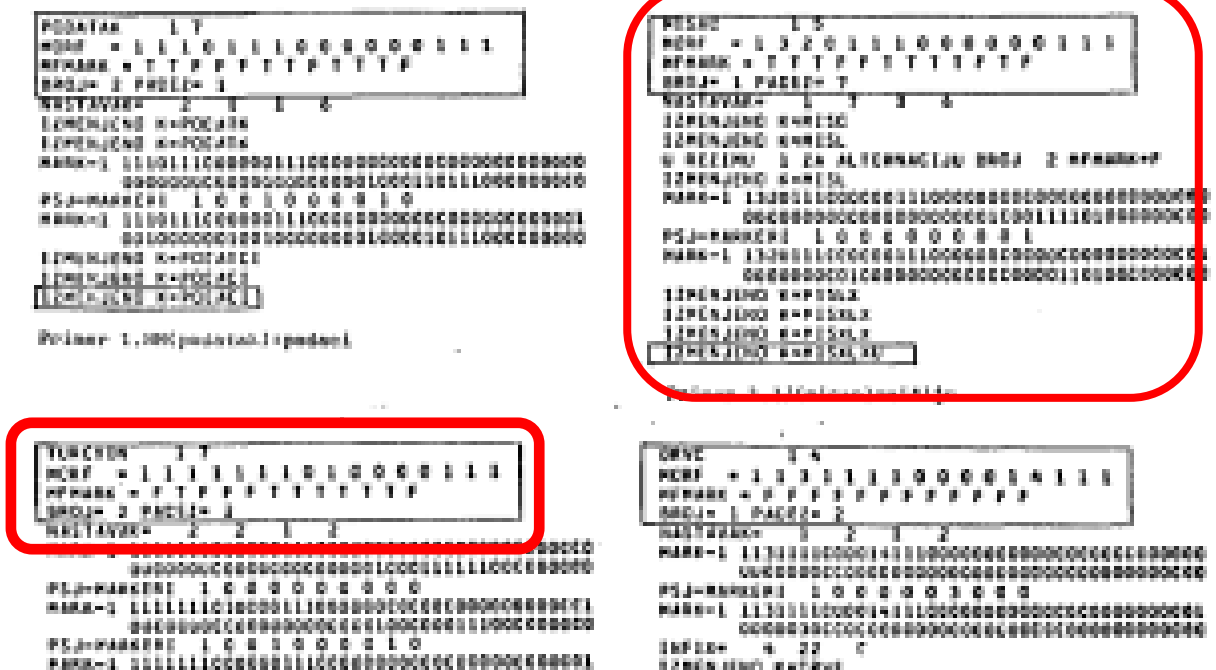
Код одговара нумерисању наслова по дубини: 3.1.1.3.

Морфологија

Duško Vitas, Generisanje imeničkih oblika u srpskohrvatskom, ***Informatica*** 80(3) izd. Slovenačko društvo za informatiku, Ljubljana, 1980, pp.49-55



мисао → мишљу



Први корпус

Почетком 80-тих је почело састављање првог корпуса који је садржавао текстове у дужини од милион речи. Корпус су чинили текстови уџбеника, новинске вести, законски и литерарни текстови.

Семинари

U OKVIRU REDOVNIH SASTANAKA SEMINARA ZA MATEMATICKU I RACUNARSKU LINGVISTIKU U NAREDNOM PERIODU CE SE ODRZATI SLEDECA PREDAVANJA.....

02.06. - DR S.VASIC - MOGUCNOSTI KVANTI ANALIZE JEZICKIH
09.06. - *** - O KORPUSU
16.06. - DR T.TOMIC - STATISTICKE KARAK SRPSKOHRVATSKOG J STANOVISTA TEORIJ
23.06. - SLOBODNA TEMA

SASTANCI CE SE ODRZAVATI U SALI 2 S.A.N.U ILI U BIBLIOTECI MATEMATICKOG INSTITUTA, KNEZ MIHAILOVA 35, SA POCETKOM U 17,00 CASOVA.

U BEOGRADU, SEKRETAR SEM
23.05.1980. DUSKO VITAS
O

O REPREZENTATIVNIM UZORCIMA PRIRODNOG JEZIKA (TEORIJA KORPUSA)

Seminar će se održati 16.,17. i 18. decembra 1981. u sali Medjunarodnog slavističkog centra,na Filološkom fakultetu, Beograd, Knez Mihailova 40
Predavači na Seminaru su: prof. H.CIMERMAN, Regenzburg, i
dr V.TOJBERT, pom.direktora IDS, Manhajm.
Predavanja će se održavati svakog dana od 16 do 19 časova, na engleskom jeziku uz simultano prevodjenje. Diskusije su predvidjene za dane 17. i 18. decembar od 11 do 13 časova. Detaljan program seminara je dostavlje u prilogu. Distribucija radnih materijala će biti obezbedjena na početk rada Seminara.

Za diskusiju je obezbedjeno i učešće jednog informatičara iz Manhajmsko Matematičkog instituta koji se stara o održavanju programske opreme na- menjene istraživanjima nad korpusom. Molimo da svoje učešće na seminaru što hitnije potvrdite, kako bi se blagovremeno mogao obezbediti smeštaj i radni materijal.

U Beogradu, 16.11.1981.

SEKRETAR
SEMINARA ZA MATEMATIČKU LINGVISTIKU
Mr D.Vitas,s.r.

PROGRAM SEMINARA

Teme predavanja dr V.TOJBERT-a

I. Aspekti konstruisanja korpusa

1. Korpusi u IDS (istorijat nastanka, kriterijumi pri izboru teksta, obim, itd.)
   -Manhajmski korpusi pisanog jezika
   -Frajburški korpus standarnog govornog jezika
   -Korpus dijaloških struktura
   -Ponski korpus novinskih tekstova

Периоди

до 90-тих: експерименти са текстом (коректор, слогови, паралелни текстови)

90-те: ***Languages Industries*** – Синклер, Замполи, Грос, Тојберт...

90-те: почетак рада на морфолошком е-речнику српског за систем ***Intex***

Multext-East: предлог позиционог стандарда за морфолошку анотацију, анотирана Орвелова ***1984***, паралелизована ***1984***

2002.

Пројекат Министарства науке Србије "Интеракција граматике и речника" (Математички и Филолошки факултет УБ, Филозофски факултет УНС)

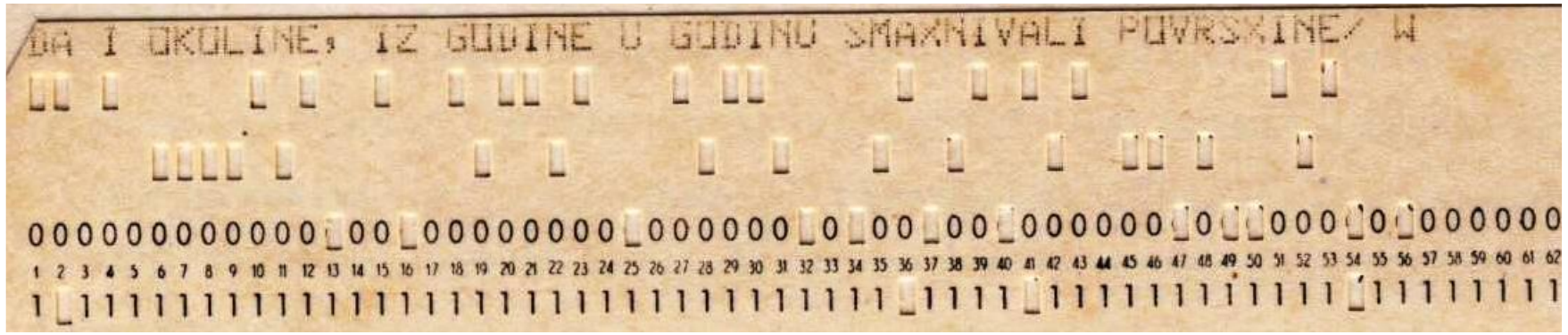
Најзначајнији резултат: први корпус прераживе преко веба са 24 милиона речи. Претрага користећи ***IMS CQP*.**

2013.

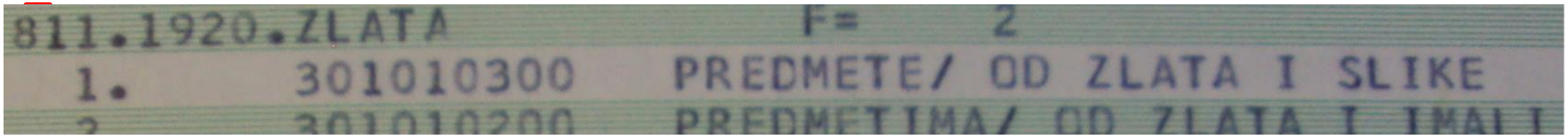
Надградња претходног корпуса на 124 милиона речи (са истим софтвером и новим сервером). Лематизиран са ознаком врсте речи, анотација извора. Описан у Утвићевој докторској тези.

Дуго путовање текстова

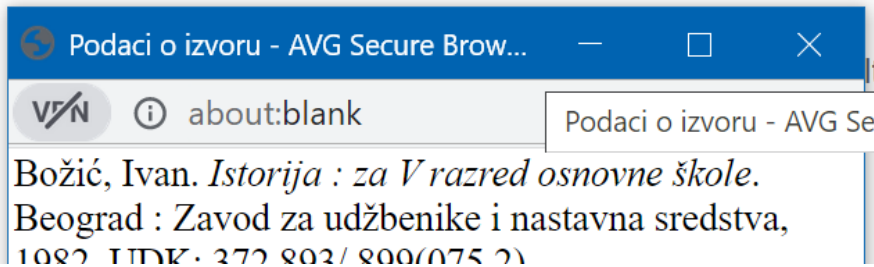
Први текстови корпуса на бушеним картицама на IBM360



а затим следе небројене миграције, прво на IBM4381, па на персоналне рачунаре и сервере...



1. Istorij5_n.txt:



. Ostavili su veličanstvene grobnice , mnoge predmete od zlata i slike po zidovima svojih dvoraca . Od
ltura naziva se mikenska ku

Ukupno prikazanih rezultata: **1** od **1**

simple zlata i slike 1 (0.01 per million)

KWIC

	Details	Left context	KWIC	Right context
1	doc#4391	su veličanstvene grobnice , mnoge predmete od	**zlata i slike**	po zidovima svojih dvoraca . </s><s> Od Krićana

[pos="A" & lemma=".*ski"]{2} [pos="N"]

97. hazarski.txt:

of , kako se hrišćanski učesnik u hebrejskim i hrišćanskim (grčkim) izvorima naziva , u stvari je vizantijska univerzitetska titula i ne treba ga uzimati u uobičajenom značenju reči . Halevijev tekst u bazelskom latinskom izdanju Dž

98. ruka.txt:

i smelost ; on naglo otvori teška vrata , i postavi se na prag s isukanim mačem , kao arhanđel pred zemaljskim rajskim vratima . - Šta traži ta noćna skitnica ? Ta dronjava ispičutura što loče cevariku i razbija tuđe prozore ?

99. Sudbina_je.xml:

slio : " Ala su prepredeneeeeeee " . . . KAKO JE MAČAK ČARLI PRODAVAO MARGARETINE ŠEŠIRE Prodavnica ženskih napuljskih šešira nije isto što i brod koji plovi po plavom moru . Zato što u prodavnici šešira ima samo šešira i otm

100. kos_price.txt:

ji se vodi u povodu , na koga kiri - džija stavlja svoj prtljag jemenija - - marama za povezivanje u ženskoj muslimanskoj nošnji kavaz - - stražar , pandur , telo - hranitelj kamilavka - valjkasta kapa kaluđera , a zatim i drugi

[lemma="nov"] [lemma="godina"]

o odgovor , pa je odustao . Čarli je odlutao u mislima i poveo njega za sobom . - Ovaj je tu još od Nove godine - , začuo je dobacivku kojom je glupan animirao dva ženska tupana . Lebdeo je na stepeništu pored k

53. Romi-leš.xml:

ze - le - na traaa - va . . . Jebiga , bar je promenio ploču ! Poslednja predstava pre pauze između novih godina takođe je započela pod utiskom Lenkinog nestanka , odnosno jutrošnjeg ispitivanja policije . Nekako

54. UPotpalub.txt:

oholosti , računali , kao i zabrinuti više nego ikad za Lazarovu sudbinu , tešili smo se mislima o Novoj godini . Besparica u koju smo zapali , bila je nepodnošljiva . Nikada nam lošije nije išlo . Srećom , zaok

55. Diary_SbL.xml:

ama na glavama . U redu : za sledeću godinu reaktiviraću novogodišnje odluke , dodajući sledeće : U novoj godini ću : Prestati da budem neurotična i plašljiva . U novoj godini neću : Više nikada spavati , niti im

56. Diary_SbL.xml:

t i doživljaj sebe kao prave žene , kompletne i bez dečka , kao najbolji način da dečka steknem . U Novoj godini ću : Prestati da pušim . Piti manje od četrnaest alkoholnih jedinica nedeljno . Smanjiti obim butin

57. TismaLic_n.txt:

le lupajući dlanovima po stolu , stvorile su se odnekud stolice i prazne čaše i svi su nazdravljali Novoj godini . Gina ni ove ljude nije poznavala ; činilo joj se jedino

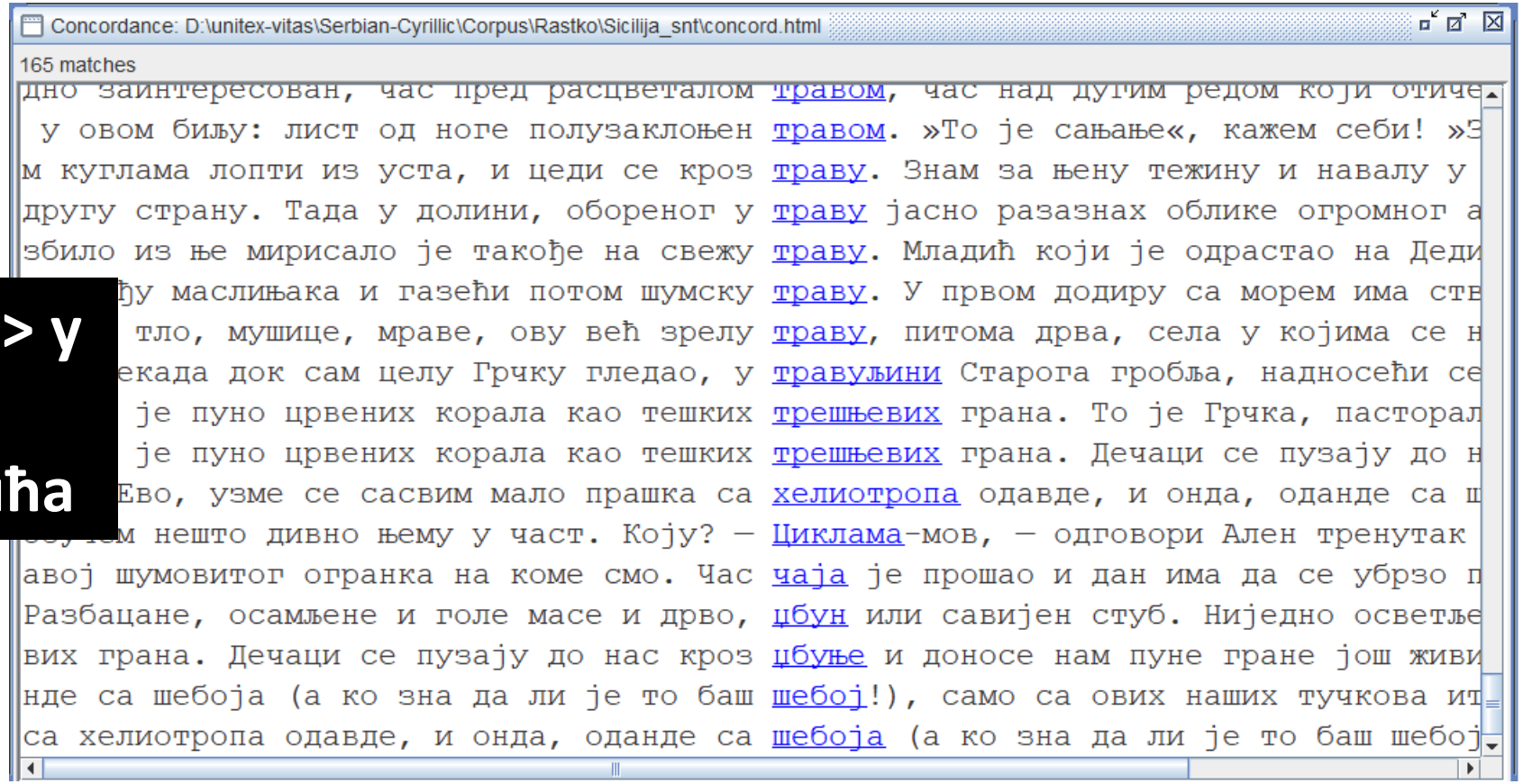
Други, махом приватни, корпуси

- Паралелни корпуси
  - енглеско-српски паралелни корпус
  - француско-српски паралелни корпус
  - немачко-српски паралелни корпус
  - италијанско-српски паралени корпус
  - хрватско-српски паралелни корпус
  - српско-српски паралелни корпус
- Књижевни корпуси и корпуси писаца (ЕЛТЕК, а затим Јаше Игњатовића, Бране Ћосића, Растка Петровића...)
- Специјализовани корпуси (терминологије, а посебно ***корпус у мрвама*** – говор о храни)

Реч о "приватним" корпусима

Обрада махом помоћу система **Unitex** и система е-речника.

Параметар <Bot> у Сицилији Растка Петровића



Колико су се ови корпуси користили?

- ~ 900 регистрованих корисника из земље и света
- ~ 100.000 сеанси (логовања корисника)
- ~ 900.000 упита
- 1 корисник ~ 1.000 упита
- ~ 250 упита дневно (од 2013.)

Захвалност

О корпусима су се старали Милош Утвић, Цветана Крстев, Ранка Станковић, Михајло Шкорић...

Прилоге за корпус су обезбедили многе драге колеге – непознати јунаци корпуса - и њима корисници дугују најискренију захвалност!

Хвала на стрпљењу и пажњи!