

Serbian module for NooJ

Faculty of Philology, Belgrade
Miloš Utvić, misko@matf.bg.ac.yu



Faculty of Mining and Geology, Belgrade
Ranka Stanković, ranka@rgf.bg.ac.yu
Ivan Obradović, ivano@rgf.bg.ac.yu



Faculty of Mathematics, Belgrade
Duško Vitas, vitas@matf.bg.ac.yu



1. dia

R1

RankaS; 2005.05.26.

(R+<E>)evolution of Serbian module

At the beginning it was the *Intex*,

2004: First migration to the *NooJ* (v.1.x)

2007: Problem of *XP/Vista* language support for Serbian

2008: Second migration to *Nooj* (v2): **Nooj Module**

Overview of evolution

- Overview of resources for Serbian
- Specific features of Serbian
- The approach to Intex -> Nooj migration of lexical resources (WS4LR – Convert Intex to Nooj ~ Migration software & scripts)
- Migration results – Serbian module for NooJ

Dictionaries for Serbian

- Dictionary of simple words ~ 120.000 lemmas (DELAS format)
- Dictionaries of proper names ~ 25000 lemmas (like D. Maurel: ProIntex formats)
- Dictionary of compound words (compound nouns, prepositions, conjunctions and adverbs, compound toponyms and proper names)
- Auxiliary dictionaries (special purpose filter dictionaries and auxiliary dictionaries for the processing of particular texts)

Intex transducers for Serbian

- Transducers for description of inflectional classes
- DELAS to DELAF transformation (~ 350 noun transducers, 60 adjective transducers and 440 verb transducers) → NooJ format

An example – class TACYAN

TACYAN = <E>(<E>/a+k+m+s+1+g + <E>/a+k+m+s+4+q) + <L><R>a(<E>/a+k+m+s+2+g + <E>/a+k+m+s+4+v + <E>/a+e+m+w+2+g + <E>/a+e+m+w+4+g + <E>/a+e+f+s+1+g + <E>/a+e+f+s+5+g + <E>/a+k+n+s+2+g + <E>/a+e+n+w+2+g + <E>/a+e+n+w+4+g + <E>/a+e+n+p+1+g + <E>/a+e+n+p+4+g + <E>/a+e+n+p+5+g) + <L><R>e(<E>/a+e+m+p+4+g + <E>/a+e+f+s+2+g + <E>/a+e+f+w+2+g + <E>/a+e+f+w+4+g + <E>/a+e+f+p+1+g + <E>/a+e+f+p+4+g + <E>/a+e+f+p+5+g) + <L><R>i(<E>/a+d+m+s+1+g + <E>/a+e+m+p+1+g + <E>/a+e+m+p+5+g + <E>/a+d+m+s+4+q + <E>/a+e+m+s+5+g) + <L><R>ija(<E>/b+e+m+w+2+g + <E>/b+e+m+w+4+g + <E>/b+e+f+s+1+g + <E>/b+e+f+s+5+g + <E>/b+e+n+w+2+g + <E>/b+e+n+w+4+g + <E>/b+e+n+p+1+g + <E>/b+e+n+p+4+g + <E>/b+e+n+p+5+g) + <L><R>ije(<E>/b+e+m+p+4+g + <E>/b+e+f+s+2+g + <E>/b+e+f+w+2+g + <E>/b+e+f+w+4+g + <E>/b+e+f+p+1+g + <E>/b+e+f+p+4+g + <E>/b+e+f+p+5+g + <E>/b+e+n+s+1+g + <E>/b+e+n+s+4+g + <E>/b+e+n+s+5+g) + <L><R>ijeg(<E>/b+e+m+s+2+g + <E>/b+e+m+s+4+v + <E>/b+e+n+s+2+g) + <L><R>ijega(<E>/b+e+m+s+2+g + <E>/b+e+m+s+4+v + <E>/b+e+n+s+2+g) + <L><R>ijem(<E>/b+e+m+s+3+g + <E>/b+e+m+s+7+g + <E>/b+e+n+s+3+g + <E>/b+e+n+s+7+g) + <L><R>ijemu(<E>/b+e+m+s+3+g + <E>/b+e+m+s+7+g + <E>/b+e+n+s+3+g + <E>/b+e+n+s+7+g) + <L><R>iji(<E>/b+e+m+s+1+g + <E>/b+e+m+s+4+q + <E>/b+e+m+s+5+g + <E>/b+e+m+p+1+g + <E>/b+e+m+p+5+g) + <L><R>ijim(<E>/b+e+m+s+6+g + <E>/b+e+m+p+3+g + <E>/b+e+m+p+7+g + <E>/b+e+m+p+6+g + <E>/b+e+f+p+3+g + <E>/b+e+f+p+7+g + <E>/b+e+f+p+6+g + <E>/b+e+n+s+6+g + <E>/b+e+n+p+3+g + <E>/b+e+n+p+7+g + <E>/b+e+n+p+6+g) + <L><R>ijima(<E>/b+e+m+p+3+g + <E>/b+e+m+p+7+g + <E>/b+e+m+p+6+g + <E>/b+e+f+p+3+g + <E>/b+e+f+p+7+g + <E>/b+e+f+p+6+g + <E>/b+e+n+p+3+g + <E>/b+e+n+p+7+g + <E>/b+e+n+p+6+g) + <L><R>ijih(<E>/b+e+m+p+2+g + <E>/b+e+f+p+2+g + <E>/b+e+n+p+2+g) + <L><R>ijoj(<E>/b+e+f+s+3+g + <E>/b+e+f+s+7+g) + <L><R>ijom(<E>/b+e+f+s+6+g)

Serbian morphological system

glumac /actor/ ms1w

Inflectional level: *glumac*, *glumca*, *glumcu*, *glumče*...

Derivational level:

N+Dem: **glumčić** A+Poss: **glumčićev**...

N+Aug+MG: **glumčina**... A+Poss: **glumčinin**...

N+Aug+Dem+MG: **glumčinica** ...

N+GM: **glumica**... A+Poss: **glumičin**...

N+GM+Dem: **glumičica**... A+Poss: **glumičicin**...

A+Rel: **glumački**...

~ several thousand of word forms

Some Serbian FST

- Derivational transducers in Serbian
- Specific transducers (for instance, identification of acronyms or numerical expressions with appropriate inflectional and derivational properties)

dvadesetpetogodisxnxi = 25-godisxnxi
(25 years old)

Alphabets of Serbian

Use of **two** alphabets

- Official **Cyrillic** alphabet
- Serbian **Latin** alphabet (also widely used)
- Absence of a unique transliteration procedure in **any** of the standard coding schemas

š, ш → SX

đ, đ → dx

č, ч → cy

ć, ċ → cx

ž, ж → ZX

њ, њ → nX

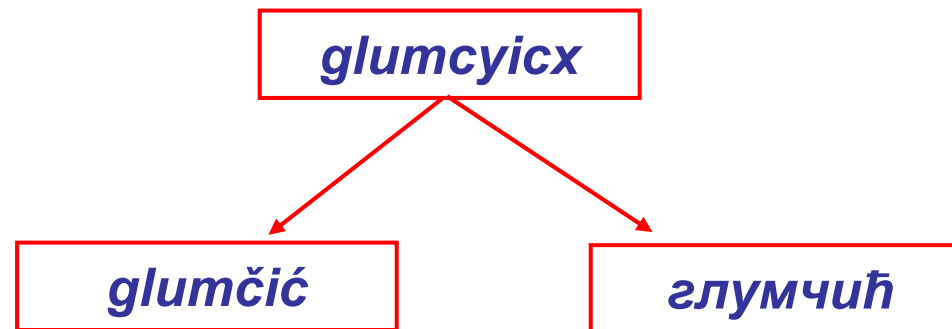
lj, љ → lX

dž, џ → dy

Eg. **Силберштајн** = **Silberštajn**

The problem in migration

Developed resources are kept in transliterated Latin alphabet
(with the adopted transliteration scheme)



Consequences

Intex Delas entry:

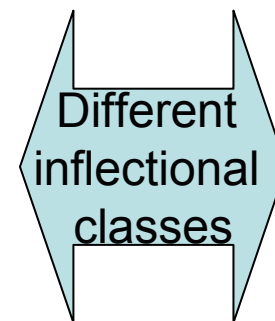
ponesxto,PRO13+Indef+ProN

Nooj:

ponesxto,ponesxto,PRO+FLX=**PRO13**+Indef+ProN

ponešto,ponesxto,PRO+FLX=**PRO13_lat**+Indef+ProN

понешто,ponesxto,PRO+FLX=**PRO13_cir**+Indef+ProN



š,ш → sx

đ,ђ → dx

č,ч → cy

ć,ћ → cx

ž,ж → zx

њ,њ → nx

љ,љ → lx

dž,џ → dy

delas-im.dic -> ascdelas-im.dic, latdelas-im.dic, cirdelas-im.dic

delas-gl.dic -> ascdelas-gl.dic, latdelas-gl.dic, cirdelas-gl.dic

....

Dilemma

Use the same lemma (transliterated option):

cyudovisxta,cyudovisxte.N+hum:ns2v:np1v:np2v:np4v:np5v

cyudovisxta,cyudovisxte,N+Hum+n+s+2+v

cyudovisxta,cyudovisxte,N+n+p+1+v

cyudovisxta,cyudovisxte,N+n+p+2+v

cyudovisxta,cyudovisxte,N+n+p+4+v

cyudovisxta,cyudovisxte,N+n+p+5+v

čudovišta,cyudovisxte.N:ns2v:np1v:np2v:np4v:np5v

чудовишта,cyudovisxte.N:ns2v:np1v:np2v:np4v:np5v

čudovišta,cyudovisxte,N+n+s+2+v

čudovišta,cyudovisxte,N+n+p+1+v

čudovišta,cyudovisxte,N+n+p+2+v

čudovišta,cyudovisxte,N+n+p+4+v

čudovišta,cyudovisxte,N+n+p+5+v

чудовишта,cyudovisxte,N+n+s+2+v

чудовишта,cyudovisxte,N+n+p+1+v

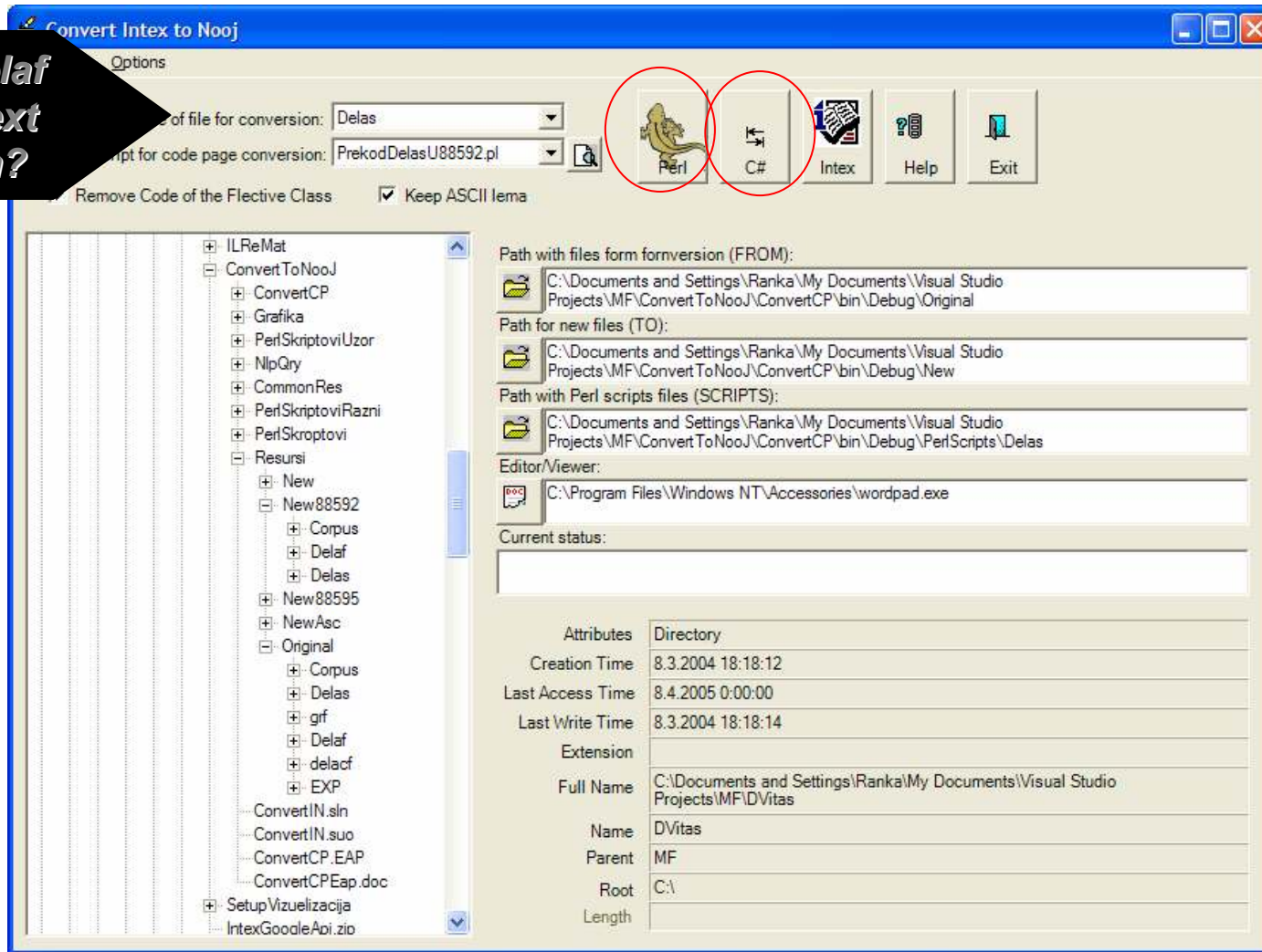
чудовишта,cyudovisxte,N+n+p+2+v

чудовишта,cyudovisxte,N+n+p+4+v

чудовишта,cyudovisxte,N+n+p+5+v

WS4LR - Overview

*Delas, Delaf
Graph, Text
Inflexion?*

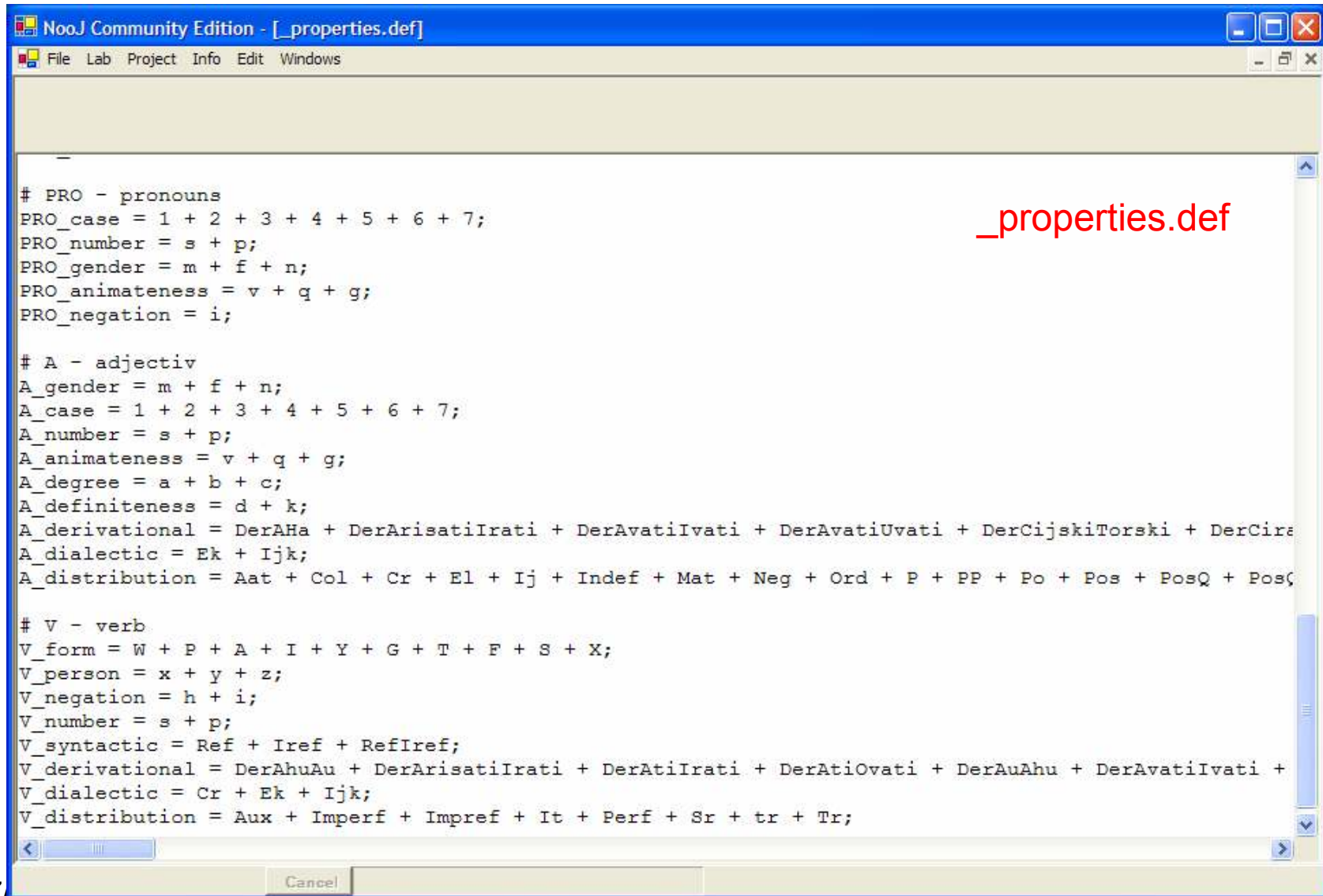


7/17/2008

The 11th NooJ Conference,
Budapest, June 8-10, 2008

13

WS4LR – properties



The screenshot shows the NooJ Community Edition software interface. The title bar reads "NooJ Community Edition - [_properties.def]". The menu bar includes "File", "Lab", "Project", "Info", "Edit", and "Windows". The main text area contains the following code:

```
# PRO - pronouns
PRO_case = 1 + 2 + 3 + 4 + 5 + 6 + 7;
PRO_number = s + p;
PRO_gender = m + f + n;
PRO_animateness = v + q + g;
PRO_negation = i;

# A - adjectiv
A_gender = m + f + n;
A_case = 1 + 2 + 3 + 4 + 5 + 6 + 7;
A_number = s + p;
A_animateness = v + q + g;
A_degree = a + b + c;
A_definiteness = d + k;
A_derivational = DerAHa + DerArisatiIrati + DerAvatiIvati + DerAvatiUvati + DerCijskiTorski + DerCira
A_dialectic = Ek + Ijk;
A_distribution = Aat + Col + Cr + El + Ij + Indef + Mat + Neg + Ord + P + PP + Po + Pos + PosQ + PosQ

# V - verb
V_form = W + P + A + I + Y + G + T + F + S + X;
V_person = x + y + z;
V_negation = h + i;
V_number = s + p;
V_syntactic = Ref + Iref + RefIref;
V_derivational = DerAhuAu + DerArisatiIrati + DerAtiIrati + DerAtiOvati + DerAuAhu + DerAvatiIvati +
V_dialectic = Cr + Ek + Ijk;
V_distribution = Aux + Imperf + Impref + It + Perf + Sr + tr + Tr;
```

The text "_properties.def" is written in red on the right side of the code area. The interface includes a scrollbar on the right and a "Cancel" button at the bottom.

Results of dictionary conversion

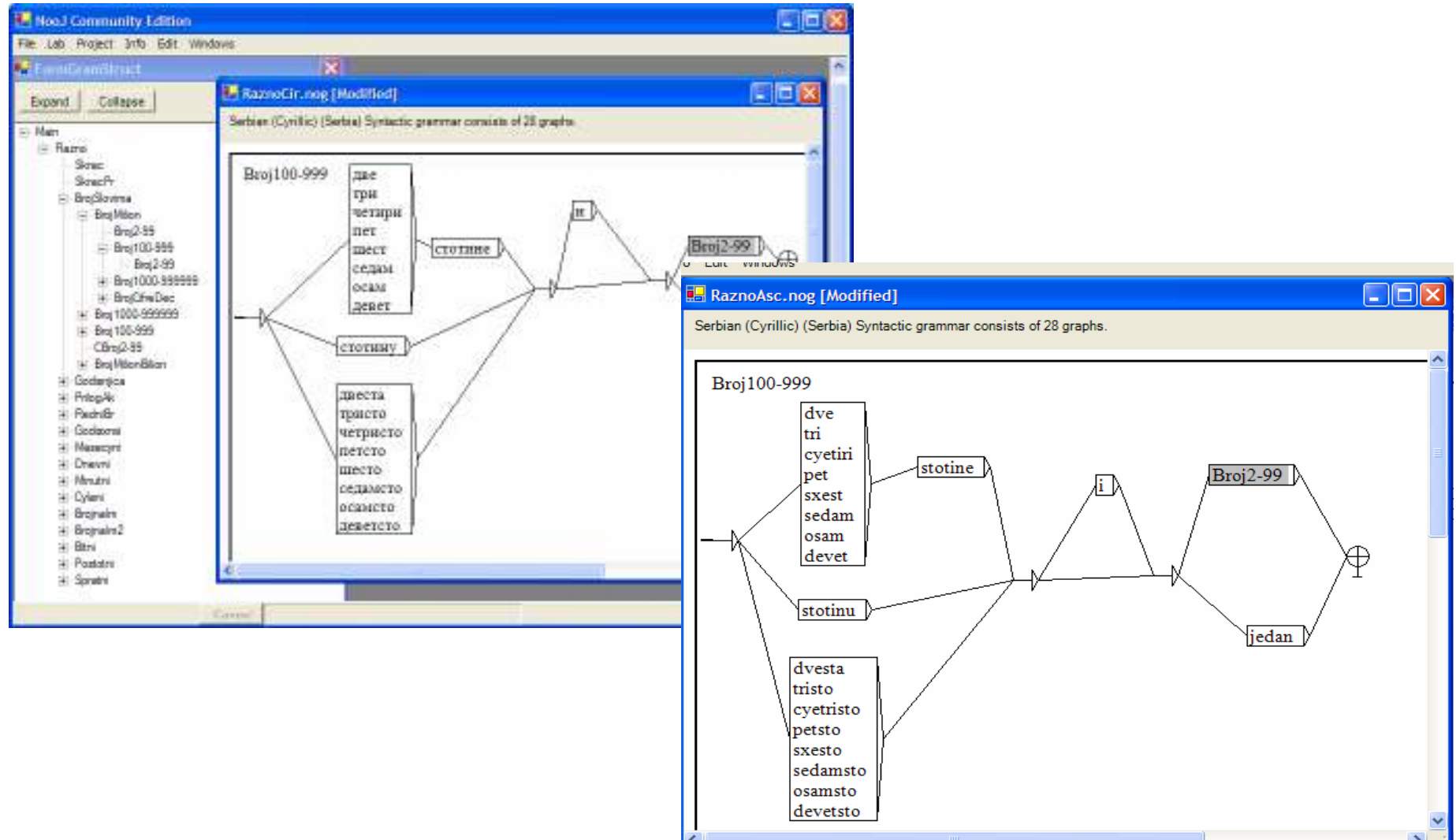
Dictionary	Intex	Nooj
delaf-ad	2.997	6.981
delaf-br	1.061	9.200
delaf-dummy+	194	1.587
delaf-ENimena	1.267	3.908
delaf-ENprez	6.504	20.460
delaf-imena	16.102	49.988
delaf-im-nove	159	989
delaf-imPoznati	74	225
delaf-int	142	308
delaf-par	92	205
delaf-pre	170	361
delaf-prefix+	76	153
delaf-prez	105.239	593.694
delaf-top-	34.531	332.588
delaf-zm	1.452	11.135
delacf-ad	98	216
delacf-con	9	19
delacf-im	57	266
Dissamb	576	2.957
FamousPeople	1.060	2.141
slskracenice-	3	6
sluzvici	10	24
spredlozi	129	291

Small dictionaries (all in one)

Bigger dictionaries (one each)

Dictionary	Alphabets	Intex	Nooj	Compiled
delaf-gl[Asc]	Asc	439.525	556.027	yes
delaf-glLat	Lat		556.027	yes
delaf-glCir	Cyr		556.027	no
delaf-im[Asc]	Asc	230.406	512.679	yes
delaf-imLat	Lat		512.679	yes
delaf-imCir	Cyr		512.679*	no
delaf-pr[Asc]	Asc	358.038	1.710.409	yes
delaf-prLat	Lat		1.710.409	yes
delaf-prCir	Cyr		1.710.409	no

Graph conversion

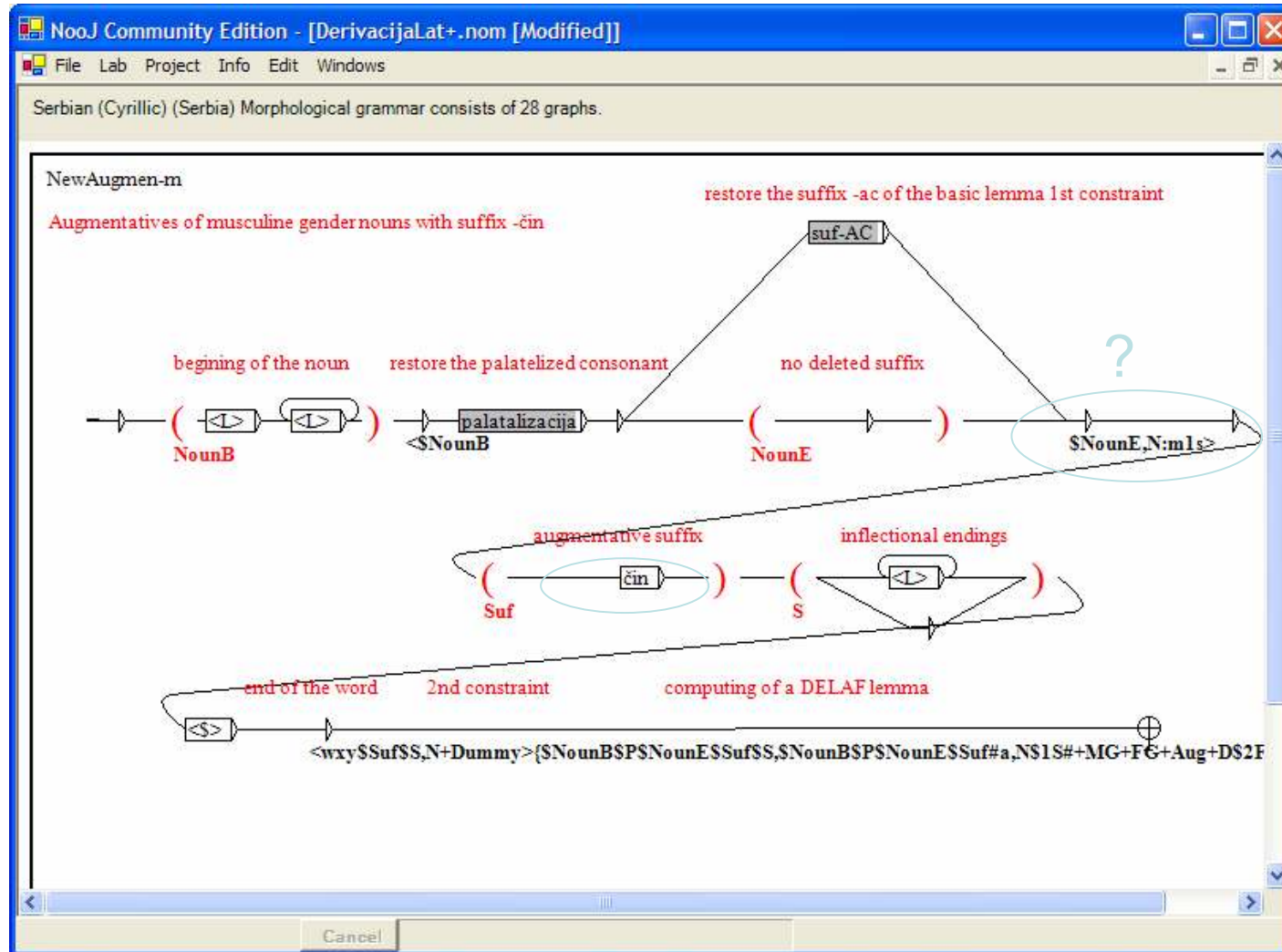


7/17/2008

The 11h NooJ Conference,
Budapest, June 8-10, 2008

16

Results of graph conversion



Corpus of Serbian

- Three short stories:
 - Filip David: *Priča o turskom časovničaru*
 - Pavle Ugrinov (V. Popović): *Vožnja tramvajem*
 - Aleksandar Tišma: *Hiljadu i druga noć*



About the corpus

Size: ~ 2.500 diff. simple words

The most frequent: *i* (177), *u* (152), *je* (124)

The most frequent bigrams (4608 diff.): *da je* (14)

<N+NProp> 79 (Gabrijel, Turčin, Švajcarska, ...)

UNKNOWN (~40): *bledunxavomusavih, ugasitozeleni,...*

or ne+Adj, naj+Adj,... aprox. by productive grammars

<A+Col> <N-Hum>

an-tisma.not lucyno spusxtenim zaliscima kao **zift crne kose** - ocyas sam
an-tisma.not su na nxima gradile lake **bele krune** , tekla je recy,
an-ugrinov.not prlxavom sjaju ogledao jedan niski **zxuti oblak** u zxurnom
an-ugrinov.not povijen, pogleda uprta u taj **zxuti oblak** , zxuti veliki
an-ugrinov.not jurio po povrsxini prlxave reke, **zxutu poslasticynicu**
an-ugrinov.not jedna velika suncyana pega. Velika **zxuta ladxa** iznutra
an-ugrinov.not najzad u tacyku, u mali, **beli beleg**. Ponovo se otkrivala
an-ugrinov.not prodavca sladoleda kraj svojih **plavih kola** sa dve sjajne kupe,
an-ugrinov.not iznicye nasred sxiroke rondele tog **zelenog ostrvca**, nxegovu
nesumnxivu i konacynu

<V> <N+VN>

an-david.not sebe!" Ove recyi **probudisxe**
interesovanxe visxeg savetnika. On lxubazno umoli
an-david.not vremena. Ukratko, dragi prijatelxu, on
izrazxava misxlxenxe da vreme ne mozxe biti
an-david.not su mnogi, cyuvsvxi zvuk vremena, **imali**
osecxanxe da slusxaju muziku
an-david.not oblasti vremena nisu, na zxalost, **izazvala**
interesovanxe onih kojima su bila namenxena
an-david.not svakom slucyaju, ustalxeno **je misxlxenxe**
da rastrzana i neuralgicyna Evropa
an-tisma.not potragu, znajucxi da me inacye **cyeka**
bdenxe u zxeleznicykoj cyekaonici, ukoliko me

Thank you for your attention!